Contribution à l'étude des facteurs évolutifs déterminant les profils moléculaires des gènes HLA des populations humaines

Goeury, Thomas

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Section de biologie
Département de génétique et évolution
Professeure Alicia Sanchez-Mazas

Contribution à l'étude des facteurs évolutifs déterminant les profils moléculaires des gènes HLA des populations humaines

THÈSE

présentée aux Facultés de médecine et des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences en sciences de la vie,
mention Écologie et Évolution

par

Thomas GOEURY

de

Annecy (France)

Thèse N° 68

GENÈVE

ReproMail

2020

LIFE SCIENCES PhD SCHOOL GENEVA

DOCTORAT ÈS SCIENCES EN SCIENCES DE LA VIE DES FACULTÉS DE MÉDECINE ET DES SCIENCES

MENTION ECOLOGIE ET EVOLUTION

Thèse de Monsieur Thomas GOEURY

intitulée :

«Contribution à l'étude des facteurs évolutifs déterminant les profils moléculaires des gènes HLA des populations humaines»

Les Facultés de médecine et des sciences, sur le préavis de Madame A. SANCHEZ-MAZAS, professeure ordinaire et directrice de thèse (Département de génétique et évolution, Unité d'anthropologie), Monsieur B. CHOPARD, professeur ordinaire (Département d'informatique), Monsieur J. M. NUNES, docteur (Département de génétique et évolution, Unité d'anthropologie), Monsieur J. VILLARD, professeur (Faculté de médecine, Département de médecine interne des spécialités) et Madame P. GERBAULT, docteure (Department of Life Sciences, University of Westminster, London, United Kingdom), autorisent l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 3 septembre 2020

Thèse - 68 -

Le Doyen
Faculté de médecine

Le Doyen
Faculté des sciences

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Remerciements

Ce travail a été possible grâce au soutien du Fonds National Suisse pour la Recherche Scientifique (subsides FNRS n° 31003A_144180 et 310030_188820, attribués à A. Sanchez-Mazas).

Au terme de ces cinq années de recherches, j'aimerais remercier l'ensemble des personnes qui m'ont permis, grâce à leur aide, de terminer ce doctorat.

Je remercie tout d'abord la Professeure Alicia Sanchez-Mazas, qui a initié et dirigé ces recherches et m'a donné l'opportunité de réaliser mon doctorat au sein de son laboratoire. Son expertise et ses connaissances m'ont été d'une aide précieuse tout au long de ce travail. Je la remercie également pour l'excellent travail de supervision qu'elle a pu réaliser et pour le temps qu'elle a passé à apporter des corrections à ce document.

Je remercie également les Docteurs Pascale Gerbault, Jean Villard, Bastien Chopard et José Manuel Nunes qui ont aimablement accepté d'être membres de mon jury de thèse, ainsi que pour leurs commentaires éclairés à propos de ce manuscrit.

Je remercie aussi spécifiquement le Docteur José Manuel Nunes pour avoir toujours répondu à mes nombreuses questions, pour ses enseignements et pour son aide précieuse dans le développement de MADaM.

Je remercie les Docteurs Da Di et Stéphane Buhler, pour toutes les informations qu'ils m'ont transmises sur le HLA et la génétique des populations.
Les Docteurs Estella Poloni et Mathias Currat pour avoir toujours pris le temps de répondre à mes questions, quelles qu'elles soient.
Stephan Weber et David Roessli pour leur aide, et surtout leur grande disponibilité, sur tous mes problèmes de nature informatique.
La Docteur Anne Mayor, pour ses conseils et informations précieuses sur l'histoire des peuples d'Afrique de l'ouest ainsi que pour m'avoir donné accès à sa thèse de doctorat sur les peuples de la boucle du Niger.

Mes remerciements vont aussi à l'ensemble de l'équipe de l'Unité d'Anthropologie pour le cadre de travail remarquable admirable qu'ils offrent : Jérémy, pour nos longues discussions autour de nos sujets de recherche respectifs, David Glauser pour nos discussions du midi, Maria, Miriam, Nuno, Claudio et Juliano pour tous les moments de détente après le travail, Luis pour toutes ces fondues organisées au bord du lac.

Je remercie aussi Louis et Médéric pour les moments de détente et de fou rire en fin de semaine, Anys, Ange, mais aussi "Plop" pour sa solidarité sans faille, Florent pour nos longues discussions au travers de l'Atlantique durant les nuits d'insomnie, et Thomas pour nos discussions autour de la littérature science-fiction.

Je n'oublie bien sûr pas ma famille, sans qui rien n'aurait été possible et pour avoir cru en moi tout au long de mes études. Je les remercie pour leur soutien et leur aide tout au long de ces années.
Je remercie tout particulièrement Caroline, pour m'avoir soutenu, aidé (et supporté...) sans failles durant ces cinq années de doctorat et je suis fier de pouvoir lui dédier.

Résumé

Ce travail de thèse porte sur l'étude des facteurs évolutifs contribuant à la distribution du polymorphisme du complexe des antigènes leucocytaires humains (HLA), aussi connu sous le nom de complexe majeur d'histocompatibilité (MHC) humain. Utilisant des données moléculaires HLA issues de séquençages ADN, il vise à apporter des éléments de réponse sur l'importance de tels facteurs, selon trois niveaux d'analyse. Le premier niveau concerne les mécanismes susceptibles de générer de la diversité HLA au sein des populations, conduisant, par exemple, à l'apparition de nouveaux allèles. Le second niveau se concentre sur la façon dont s'est distribuée cette diversité entre les différentes populations humaines, au travers de l'histoire du peuplement et de ses mécanismes, à savoir les migrations et la dérive génétique mais aussi la sélection naturelle selon les adaptations des populations à leurs environnements. Le troisième niveau s'intéresse aux mécanismes qui ont pu faire varier ce polymorphisme au sein de diverses espèces (en l'occurence, ici, l'humain et le chimpanzé) sur une échelle de temps plus longue, conduisant, par exemple, à l'apparition de nouveaux gènes. Ce travail se compose ainsi de trois études principales, qui ont étudié la diversité HLA sur trois échelles : d'abord à l'échelle intra-populationnelle, puis à l'échelle inter-populationnelle et, pour finir, à l'échelle inter-spécifique.

La première étude porte sur l'analyse parallèle de deux populations d'origines très différentes, les Mandenka du Sénégal et les Cham du Vietnam, dont la variabilité moléculaire de 8 gènes HLA a été analysée grâce au séquençage complet de ces gènes. Cette analyse a révélé une importante diversité des codons codant pour le site de reconnaissance de l'antigène sur les exons 2 (et aussi les exons 3 pour les gènes de classe I) au sein des deux populations. Cette variabilité confirme, pour la première fois à l'échelle populationnelle, l'hypothèse précédemment énoncée d'une forte sélection agissant sur ces codons, en lien avec la fonction de présentation des antigènes des molécules HLA. En revanche, nos analyses ont révélé des différences entre les deux populations au niveau des haplotypes et de leurs déséquilibres de liaison. Chez les Mandenka, un seul haplotype HLA (de classe II) présente à la fois une fréquence élevée et un fort déséquilibre de liaison, résultat probable d'une sélection positive en réponse à un ou plusieurs pathogènes. Cet haplotype inclut l'allèle HLA-DRB1*13:04, dont l'analyse attribue de manière convaincante l'origine à une conversion génique. Chez les Cham, cette analyse a révélé deux groupes d'allèles en déséquilibre de liaison, signal probable non pas d'une sélection, mais d'une histoire démographique particulière, rejoignant l'hypothèse (linguistique) d'une origine duelle de cette population résultant d'un mélange entre des habitants locaux d'Asie continentale du sud-est et des migrants austronésiens venant de Taïwan. En complément, cette étude a permis de comparer, de manière détaillée, trois techniques de typage différentes appliquées aux gènes HLA (PCR-SSO, séquençage des exons 2 et séquençage des gènes complets) sur les données des Mandenka, grâce au fait que les mêmes individus de cette population ont été typés par ces trois techniques en l'espace de 25 ans.

La deuxième étude est une analyse de plus de 2'000 individus provenant de 36 populations de la bande du Sahel et d'Afrique du nord principalement, mais aussi de Syrie et de Slovaquie, comme références extérieures. Tous les individus avaient été séquencés par une méthode de séquençage haut-débit, dite « NGS-454 », aux exons 2 des quatre gènes de classe II HLA-DRB1, -DQA1, -DQB1 et -DPB1. Cette technique engendrant de nombreuses difficultés, il nous a d'abord fallu développer une nouvelle procédure bioinformatique (appelée `MADaM`) pour trier spécifiquement, avec le maximum de précision possible, le volume de données produit par ces séquençages. L'analyse de ces données

a ensuite permis de comparer entre elles toutes ces populations sur le plan génétique, et de mettre en évidence des signaux relatifs à l'histoire des migrations (signaux démographiques) et aux adaptations des populations à leurs environnements (signaux de sélection), notamment un possible rôle protecteur de plusieurs allèles des quatre loci étudiés vis-à-vis de la malaria en Afrique.
La troisième étude comporte deux volets. Dans le premier, l'ensemble des séquences de la base de données d'IPD/IMGT-HLA ont été utilisées pour décrire le contenu en information (selon la théorie de l'information de C. Shannon) des différents introns et exons des gènes HLA. Cette étude a mis en évidence une importante information dans les exons 3 des gènes HLA-DQA1, -DQB1 et -DPB1, information peu redondante avec celle des exons 2 de ces gènes. Ce résultat a d'abord permis d'expliquer les différences observées dans les résultats obtenus, dans notre première étude, sur les mêmes données testées par les trois techniques de typage appliquées au Mandenka. L'analyse de la banque de données IPD/IMGT-HLA a aussi révélé que les régions codantes (exons) contenaient en général 10 fois plus d'information que les régions non-codantes (introns et UTR). Le second volet de cette étude a consisté à décomposer en chaînes de Markov les séquences des différents exons et introns de 11 gènes HLA ainsi que de quatre gènes Patr (le MHC du chimpanzé) de classe I. Cette analyse a permis de soutenir l'hypothèse d'une origine commune des 3 gènes HLA de classe I et des 4 gènes Patr de classe I, sur la base des similarités observées entre les différentes régions géniques. Pour les gènes HLA de classe II, elle a révélé une grande similarité des régions des gènes $II\beta$, cohérente avec l'hypothèse de leur origine commune, mais aussi une différence marquée entre les exons et introns des gènes de la région HLA-DRB, compatible avec l'hypothèse d'une origine distincte de ces deux régions.
En conclusion, ce travail montre l'intérêt d'utiliser des données de séquences des gènes HLA pour des études de génétique des populations et de génétique évolutive, par rapport à l'utilisation de données classiques décrivant la variation par des allèles HLA nominaux. L'analyse de la diversité au niveau nucléotidique de ces gènes permet de déterminer plus précisément les effets de différentes forces évolutives agissant sur chacun des loci et de mieux interpréter la variation par rapport à l'histoire des populations et des espèces.

Abstract

This thesis work investigates the evolutionary factors determining the distribution of the human leukocyte antigens (HLA) polymorphism, also known as the human major histocompatibility complex (MHC). Using molecular data from DNA sequencing, it aims at providing clues about the extent of such factors, according to three levels of analysis. The first level is about the mechanisms able to generate the HLA diversity within populations, leading, for example, to the emergence of new alleles. The second level focuses on how this diversity has been distributed among the different human populations, through the peopling history and its mechanisms, namely migrations and genetic drift but also natural selection, according to the adapation of populations to their environments. The third level focuses on the mechanisms that could have driven the differentiation of different species (here, humans and chimpanzees) on a longer time scale, leading, for example, to the birth of new genes. This work thus consists of three main studies, that have investigated the HLA diversity at three scales : first, at the intra-population level, then at the inter-population level and, finally, at the inter-specific level.

The first study is a parallel analysis of two populations of very different origins, the Mandenka from Senegal and the Cham from Vietnam, for which the variability of 8 HLA genes has been analyzed by whole-gene sequencing. This analysis revealed a great variability of the codons coding for the antigen recognition sites on exons 2 (and exons 3 for class I genes), in the two populations. This variability confirms, for the first time at the population level, the previously emitted hypothesis of a strong selection acting on these codons, in relation to the antigen presentation function of the HLA molecules. On the other hand, our analysis revealed differences between the two populations at the level of their haplotypes and linkage disequilibria. In the Mandenka, a single class II haplotype shows both a high frequency and a strong linkage disequilibrium, probably stemming from a positive selection by one or several pathogens. This haplotype includes the HLA-DRB1*13:04 allele, for which the analysis convincingly attributes an origin through gene conversion. Concerning the Cham, this analysis revealed two clusters of alleles in linkage disequilibrium, as a probable signal, not of selection, but of their particular demographic history, sustaining the (linguistic) hypothesis of its dual origin, resulting from an admixture between local inhabitants of mainland south-east Asia and Austronesians migrating from Taiwan. In addition, this study allowed us to compare, in detail, three different typing techniques applied to the HLA genes (PCR-SSO, exon 2 sequencing and whole-gene sequencing) of the Mandenka dataset, thanks to the typing of the same individuals of this population by these three techniques in the space of 25 years.

The second study is an analysis of more than 2'000 individuals belonging to 36 populations of, mainly, the Sahel Belt and North Africa, but also Syria and Slovakia, as external references. All individuals had been sequenced by a high-throughput sequencing method, called « NGS-454 », at exons 2 of the four genes HLA-DRB1, -DQA1, -DQB1 and -DPB1. This technique giving rise to several difficulties, we first had to develop a new bioinformatic procedure (called `MADaM`) to specifically sort, with as much accuracy as possible, the amount of data generated by the sequencing. This data analysis then allowed us to compare genetically all these populations between them, and to highlight signals related to the migration history (demographic signals) and adaptations of these populations to their environment (selection signals), especially a putative protective role of several alleles of the four studied loci against malaria in Africa.

The last study has two parts. In the first one, all the sequences from the IPD/IMGT-HLA

database were used to describe the information content (according to C. Shannon's information theory) of the different introns and exons of HLA genes. This study disclosed important information in exons 3 of the HLA-DQA1, -DQB1 and -DPB1, which is not redundant with that of exons 2 of these genes. This result allowed us to explain the observed discrepancies found in the results obtained within the frame of our first study, on the same data tested by the three typing techniques applied to the Mandenka. This study also revealed that coding regions (exons) generally contain 10 times more information than non-coding regions (introns and UTRs). The second part of this study is a Markov chains decomposition of the sequences of the different exons and introns of 11 HLA genes as well as four Patr genes (chimpanzee's MHC) of class I. This analysis supports the hypothesis of a common origin of the 3 HLA class I genes and the 4 Patr genes of class I, based on similarities between their different gene regions. For HLA class II genes, this study reveals a greater similarity of the regions of the $II\beta$ genes, consistent with the hypothesis of their common origin, but also a marked difference between the exons and introns of the different HLA-DRB genes, consistent with the hypothesis of a distinct origin of these two regions.

In conclusion, this work shows the interest of using sequencing data of HLA genes for population genetic and evolutionary studies, compared to classical data describing the variability with nominal HLA alleles. The analysis of the diversity, at the nucleotidic scale, allows to determine more precisely the different evolutionary forces acting on each locus and better interpret the variation in relation to the history of populations and species.

Table des matières

Chapitre 1

Introduction générale

1 Historique de la découverte du HLA

Le complexe HLA (Antigènes Leucocytaires Humains [1]) fait référence à un ensemble de gènes extrêmement polymorphiques localisés sur le chromosome 6 et impliqués dans les processus de défense immunitaire de l'organisme.

Il a été mis en évidence par Jean Dausset [Dausset, 1958] une vingtaine d'années après la première mise en évidence d'un système similaire chez une autre espèce, le système H-2 de la souris [Gorer, 1937]. J. Dausset a observé que la mise en contact de sérum sanguin d'un patient leucopénique avec un échantillon de moelle osseuse d'un autre patient provoquait une réaction rapide d'agglutination des leucocytes. Plus tard, il fera l'observation d'une absence d'agglutination avec des leucocytes provenant de trois donneurs de sang et en déduira un polymorphisme des antigènes leucocytaires qu'il baptisera MAC, acronyme formé des initiales de ces trois patients [Dausset, 1984]. J. Dausset a dès le début pointé le fait que « dans un avenir lointain », l'étude de ces antigènes leucocytaires pourra avoir une importance dans les greffes d'organes et de moelle osseuse [Dausset, 1958].
Très rapidement, d'autres équipes de recherche découvriront des systèmes similaires : à Leiden (Pays-Bas) l'antigène 4a4b par J.J. Van Rood [Van Rood et al., 1958, van Rood and Van Leeuwen, 1963] et à San Francisco (USA) les trois antigènes LA1, LA2 et LA3 par Rose Payne [Payne and Rolfs, 1958, Payne et al., 1964].
Ces découvertes similaires ont mené au premier atelier de travail collaboratif en histocompatibilité en 1964 [Amos et al., 1965] (ateliers à présent connus sous le nom de « *International HLA and Immunogenetics workshops (IHIW)* ») réunissant à Durham (Caroline du nord, USA), entre autres, J. Dausset et son système MAC (renommé alors Hu-1 [2]), J.J. Van Rood et R. Payne. C'est après les ateliers de 1965 à Leiden (Pays-Bas) [IHWG, 1965] et de 1967 à Turin (Italie) [Curtoni et al., 1967] que ces systèmes antigéniques seront regroupés sous l'appellation HLA.
Etant donné l'implication majeure de ces gènes dans l'histocompatibilité, on désigne généralement ce système sous le nom de « Complexe Majeur d'Histocompatibilité », ou CMH (MHC en anglais pour « *Major Histocompatibility Complex* »). MHC (ou CMH) est aussi, par extension, le terme utilisé pour désigner ce complexe chez d'autres organismes (voir plus loin).

1. En anglais : *Human Leukocyte Antigene*.
2. Nommé de manière similaire au complexe analogue de la souris, H-2 [Gorer, 1937].

C'est en 1977 que le complexe HLA a été localisé sur le bras court du chromosome 6, dans la région 6p21 [Francke and Pellegrino, 1977] et l'intérêt porté à ce complexe génétique a conduit au séquençage de la région en 1999 [The MHC sequencing consortium, 1999], bien avant le séquençage du chromosome 6 en entier en 2003 [Mungall et al., 2003]. C'est en 2009 que la région HLA a été totalement cartographiée et annotée [Shiina et al., 2009].

2 Histoire évolutive du MHC

Le développement depuis plus de 20 ans des techniques de séquençage ADN et leur généralisation à d'autres espèces permettent, à l'aide de la génomique comparative, d'étudier l'apparition du MHC et son évolution. Des gènes de classe I et II ont été retrouvés dans l'ensemble du clade monophylétique des Gnathostomes (vertébrés à mâchoire) [Kasahara et al., 1995, Flajnik et al., 1999] tandis que les Agnathes (vertébrés sans mâchoire) tels que les lamproies en sont dépourvus. Ainsi il est possible de déterminer que l'apparition d'un système immunitaire adaptatif s'est fait après la séparation des Agnathes et Gnathostomes, mais avant la diversification de ces derniers.

La comparaison des différents MHC des Gnathostomes, depuis les poissons cartilagineux jusqu'aux mammifères, a permis d'inférer la structure du MHC originel, visible sur la Figure 1.1.

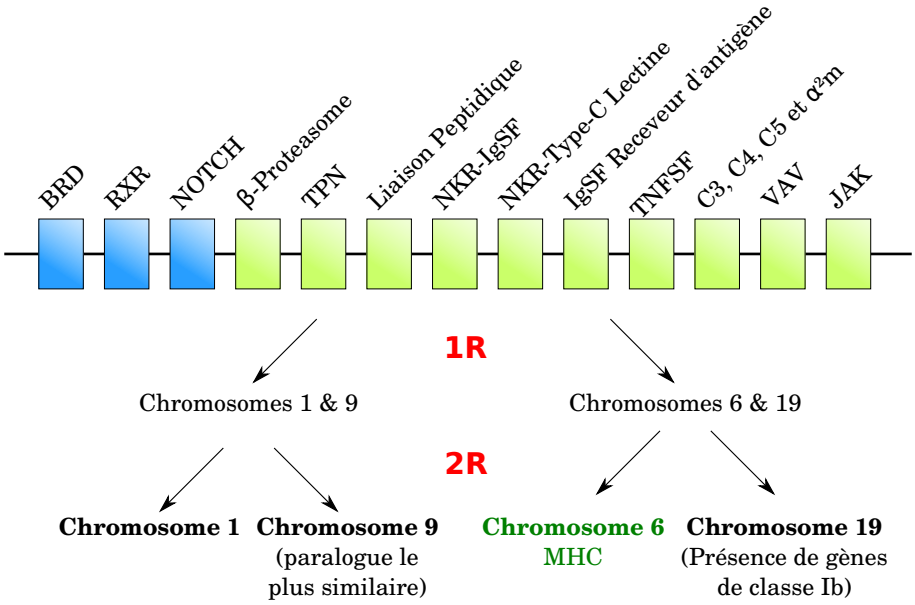


FIGURE 1.1 – Structure théorique du proto-MHC, adaptée de [Flajnik, 2018], ainsi que les deux événements de duplication génomique (1R et 2R) ayant mené à l'apparition du MHC et de ses trois complexes paralogues. Les numéros des chromosomes correspondent aux chromosomes humains, mais ces paralogues sont retrouvés dans l'ensemble des vertébrés. Les noms abrégés des gènes sont les suivants : BRD : Protéine à bromodomaine ; RXR : Récepteur X des rétinoïdes ; NOTCH : Protéine homologue du locus neurogénique Notch ; TPN : tapasine ; NKR-IgSF : Récepteur au IgSF des cellules tueuses naturelles ; NKR-C-type lectin : Récepteur aux lectines de type C des cellules tueuses naturelles ; IgSF : Superfamille des immunoglobulines ; TNFSF : Superfamille des facteurs de nécrose tumorale ; C3, C4, C5 : Composant du complément C3, C4, C5 ; α2m : α2 macroglobuline ; JAK : Janus kinase.

Ce proto-MHC aurait ensuite subi deux événements d'allopolyploïdisation (duplication génomique suite au croisement de deux espèces proches), le premier dans un ancêtre commun à tous les vertébrés et le second dans un ancêtre commun à tous les Gnathostomes après leur divergence avec les Agnathes [Ohno, 1999, Wolfe, 2001]. Ces deux événements de duplication génomique ont donné lieu à ce qui est appelé le « Big Bang évolutif » de l'immunité, qui est caractérisé par l'apparition dans une échelle de temps

géologique courte (∼50 millions d'années [Bernstein et al., 1996, Laird et al., 2000]) d'un grand nombre de caractéristiques immunitaires présentes chez l'humain, incluant (parmi d'autres) les immunoglobulines M et D, les cellules à récepteurs T $\alpha\beta$ et $\gamma\delta$, la rate, les cellules dendritiques et le MHC [Flajnik, 2018].

Il est toutefois difficile de déterminer si ce sont les molécules de classe I ou II qui sont apparues en premier et cette question rejoint celle de l'apparition de la région de liaison au peptide (PBR[3]), c'est-à-dire de savoir si l'état ancestral de cette PBR est plus proche de ce qui est actuellement observé pour les loci de classe I (formée par les deux premières sous-unités de la chaîne α) ou de classe II (formée par les premières sous-unités des chaînes α et β). Hashimoto *et al.* [Hashimoto et al., 1990] avancent que la PBR aurait évolué à partir de domaines V (Variables) des immunoglobulines tandis que Flajnik *et al.* [Flajnik et al., 1991] avancent que la PBR serait dérivée des protéines chaperonnes hsp70 (*heat shock protein*). Cette dernière hypothèse soutient que les molécules de classe I seraient apparues en premier, issues d'un événement de recombinaison entre une immunoglobuline et une hsp70. Toutefois, cette hypothèse est peu probable puisque l'analyse des domaines de liaison peptidique des hsp70 bactériennes a montré peu de similarités avec les domaines de liaison peptidique des molécules de classe I [Zhu et al., 1996].
L'hypothèse la plus parcimonieuse est de considérer que les gènes de classe I sont issus des gènes de classe II [Klein and Sato, 1998]. La molécule ancestrale du MHC serait alors un homodimère de deux molécules composées d'un seul domaine de liaison peptidique inclus dans une simple immunoglobuline membranaire.
Une étude menée en 2010 par Günther *et al.* [Gunther et al., 2010] a montré que les molécules HLA de classe II étaient capables de lier des peptides présentés dans les deux directions, tandis que les molécules de classe I ne peuvent lier que des peptides présentés avec une orientation N-C[4]. Pour les auteurs, cette observation soutient l'hypothèse d'une molécule homodimère originelle.
La duplication du gène codant pour cette molécule aurait alors permis l'apparition des premiers gènes de classe II α et β codant pour un hétérodimère. Les gènes de classe I seraient issus d'un remaniement des exons, par exemple, par recombinaison entre deux gènes, conduisant au transfert d'un deuxième exon codant pour le domaine de liaison peptidique dans un autre gène [Kaufman, 1988].

Afin de comprendre les mécanismes évolutifs expliquant l'organisation actuelle du HLA, il nous faut étudier son évolution depuis son apparition chez un ancêtre commun à tous les vertébrés.
Le texte qui suit présente une description (non exhaustive) des principales caractéristiques du MHC au sein des grands ordres de Gnathostomes : les poissons cartilagineux, les poissons osseux, les amphibiens et les archausoriens. Le MHC des mammifères sera présenté par la description du système HLA (MHC humain), modèle d'étude génétique de ce travail de doctorat.

3. En anglais : *Peptide Binding Region*, voir page 15.
4. Pour une protéine, l'extrémité N-terminale réfère à l'extrémité se terminant par un acide aminé avec une fonction amine libre (-NH2), tandis que l'extrémité C-terminale se caractérise par la présence d'une fonction carboxyle (-COOH).

2.1 Poissons cartilagineux

Chez les poissons cartilagineux tels que les requins, les gènes de classe IIα et IIβ sont liés sur le chromosome, dans une même région, de même qu'avec le gène de classe I (Ia), suggérant que la duplication génique ayant généré les gènes de classe II α et β, ainsi que les gènes de classe I, est survenue très tôt dans l'évolution (ces gènes étant retrouvés dans des taxons évolutivement très distants tels que les humains et les requins) [Ohta et al., 2000]. De la même façon, les gènes du complément (C4) ont aussi été identifiés, définissant une région de classe III liée à ces régions de classe I et II, suggérant une origine ancienne de cette association [Terado et al., 2003].

2.2 Poissons osseux

Les poissons osseux (Téléostéens) présentent la caractéristique d'avoir leurs classes I et II non liées et réparties sur plusieurs chromosomes (Figure 1.3)[Klein and Sato, 1998]. Cette observation soulève la question de savoir si l'association entre les régions du MHC est un caractère ancestral ou dérivé. Dans le cas d'un caractère ancestral, les gènes de classe I et II seraient issus d'un même gène primordial, tandis que dans le cas d'un caractère dérivé, les gènes de classe I et II seraient apparus sur des chromosomes différents puis associés dans un ancêtre des tétrapodes [Ohta et al., 2000]. Toutefois, seuls les Téléostéens présentent ce caractère, toutes les autres espèces de Gnathostomes présentant une liaison (même si certaines fois partielle) entre les régions du MHC, ce qui amène à conclure, par principe de parcimonie, au caractère dérivé du MHC des Téléostéens.

Les gènes de classe I présentent la particularité d'être conservés entre les espèces de Téléostéens tandis que les gènes de classe II montrent plus de plasticité [Bingulac-Popovic et al., 1997, McConnell et al., 1998, Sato et al., 2000] (certains ordres tels que les Gadiformes étant même dépourvus de gènes de classe II [Star et al., 2011, Star and Jentoft, 2012, Malmstrøm et al., 2016]). L'association étroite entre les gènes de classe I, TAP [5] et du protéasome chez les Téléostéens expliquerait cette relative stabilité des gènes de classe I [Ohta et al., 2000].

2.3 Amphibiens

Chez les amphibiens, les gènes de classe I, II et III sont liés, à l'instar de tous les autres clades à l'exception des Téléostéens. Des différences ont été observées dans l'expression des gènes des différentes classes, ceux de classe I étant exprimés seulement au stade adulte, mais de manière ubiquitaire, tandis que les gènes de classe II sont exprimés au stade larvaire sur les lymphocytes B et cellules présentatrices d'antigènes auxquelles se rajoutent les cellules T lors du stade adulte [Salter-Cid et al., 1998, Flajnik et al., 1987]. Il est à noter aussi la présence de gènes non-classiques de classe I hors du MHC [Flajnik et al., 1993] mais sur le même chromosome [Courtet et al., 2001], ainsi que l'apparition, pour ce taxon, des gènes de classe II DM[Flajnik, 2018].

5. En anglais : *Transporter associated with Antigen Processing*, transporteur associé au traitement des antigènes.

2.4 Archosauriens

Les Archausoriens sont un clade regroupant les oiseaux et les reptiles [St John et al., 2012] qui a été inégalement étudié du point de vue du MHC, puisque peu de données sont disponibles pour les reptiles.

Une étude, de Jaratlerdsiri *et al.*, parue en 2014 et menée sur le MHC du crocodile marin *Crocodylus porosus* a permis d'identifier quelques caractéristiques du MHC des reptiles [Jaratlerdsiri et al., 2014]. Il apparaît que la structure du MHC du crocodile marin est composée de deux régions de classe I indépendantes et de deux régions de classe II elles aussi indépendantes.
Au sein des régions de classe I, trois gènes classiques ont été retrouvés (Crpo-UA, -UB and -UC) ainsi que six pseudogènes. À l'instar du MHC du poulet, une association étroite a été observée entre les gènes classiques de classe I et les gènes TAP (codant pour un transporteur de peptides vers le réticulum endoplasmique [Bouvier, 2003]).
Pour les régions de classe II, six loci ont été identifiés dont trois sont des gènes fonctionnels. Ces trois gènes codent pour deux chaînes β (DAB1 et DAB2) et une chaîne α (DAA).

L'étude du MHC aviaire a surtout été menée sur le poulet (*Gallus gallus*), qui possède un MHC réparti en deux régions principales, MHC-B et MHC-Y, toutes deux localisées sur le même chromosome (GGA16) mais non associées car séparées par un point chaud de recombinaison méïotique [Miller et al., 1996]. La région du MHC-B est caractérisée par une extrême compaction, puisque regroupant 19 gènes sur 92kb [Kaufman et al., 1999] (en comparaison, le HLA comprend 253 gènes sur 3.8mb, voir aussi Figure 1.2). Toutefois, en 2007, une étude menée par Shiina *et al.*[Shiina et al., 2007] a mis en évidence l'existence de 25 gènes supplémentaires au-delà de cette région (étendant la région à 46 gènes sur 242kb), la plupart ayant des fonctions (encore mal définies) relatives à l'immunité.
Au sein du MHC-B, la région de classe I est intégrée entre la région de classe II et celle de classe III [Shiina et al., 2007] et les gènes TAP1 et TAP2 sont localisés dans la région de classe I [Kaufman et al., 1999] (au contraire du HLA où ils sont localisés dans la région de classe II).
L'extrême compaction du MHC du poulet est aussi observée chez le tétras lyre (*Tetrao tetrix*) [Wang et al., 2012] et le faisan doré (*Chrysolophus pictus*) [Ye et al., 2012], avec respectivement 19 et 20 gènes pour des régions de 88 et 97kb. D'autres espèces d'oiseaux ont aussi été étudiées, telles que la caille du Japon (*Coturnix japonica*) [Shiina et al., 2004] et la dinde sauvage (*Meleagris gallopavo*) [Chaves et al., 2009] montrant des régions du MHC-B plus grandes (respectivement 41 et 34 gènes sur 180 et 197kb) mais de taille toutefois largement inférieure (de plusieurs ordres de grandeur) à celle de la région HLA chez l'humain.

2.5 Mammifères

La comparaison du MHC humain (HLA) et murin par Amadou en 1999 [Amadou, 1999] a mis en évidence le caractère paralogue des gènes classiques de classe I des mammifères (issus de duplications d'un gène ancestral dans l'espèce étudiée), associé à la présence de gènes orthologues à l'humain et la souris (issus de duplications d'un gène ancestral dans un ancêtre commun aux deux). L'hypothèse privilégiée est que ces gènes orthologues sont des régions très conservées, représentant des points d'ancrage, dont les modifications seraient délétères, et qui délimitent des positions au sein desquelles les loci (les gènes paralogues) sont libres d'évoluer de manière indépendante, par duplications et délétions.

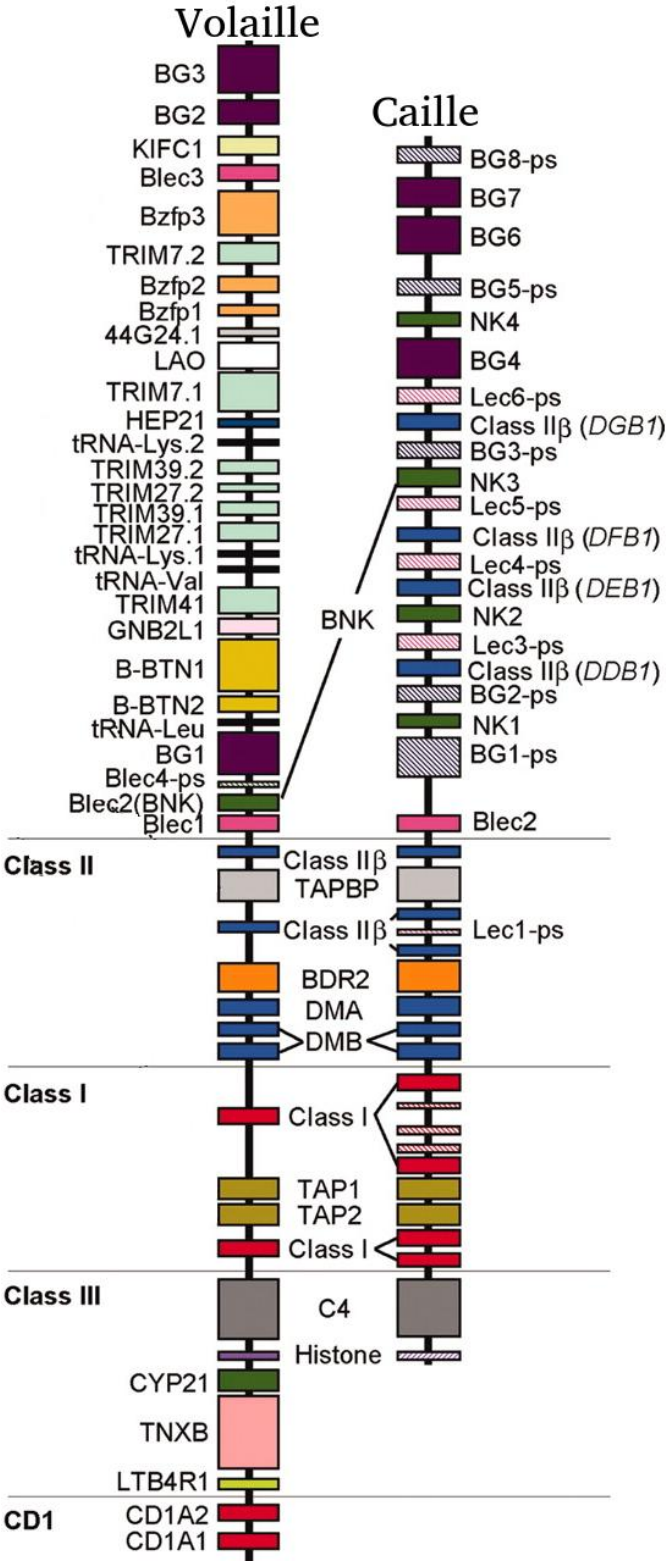


FIGURE 1.2 – Représentation schématique du MHC-B du poulet (gauche) et de la caille (droite), illustrant l'extrême compaction du MHC aviaire. Figure adaptée de [Shiina et al., 2007].

Ces blocs présentent différents niveaux de polymorphisme, dérivant tous d'un même bloc originel appelé duplicon et dont les analyses de génomique comparative ont déterminé qu'il devait être composé de plusieurs gènes d'hémochromatose (HCG), d'un gène MIC, d'un rétrovirus HERV-16 et d'un gène de classe I [Kulski et al., 1999, Kulski et al., 2000].

Cinq régions ont été décrites chez les mammifères, qui portent les noms de régions α, β, κ, ζ et olfr. Les trois premiers blocs sont retrouvés chez la plupart des mammifères, à l'exception du cochon pour lequel le bloc α est absent [Velten et al., 2008]. Les blocs ζ et olfr ne sont retrouvés que chez les rongeurs (rat et souris), le bloc ζ étant retrouvé en position centromérique, dans la région de classe II étendue (après les gènes de classe II) tandis que le bloc olfr est retrouvé de manière télomérique au bloc α [Amadou, 1999, Günther and Walter, 2001, Ioannidu et al., 2001, Kumánovics et al., 2002].
De plus, les différentes espèces de mammifères montrent des différences dans le nombre de gènes de chacun de ces blocs, les primates (humains et chimpanzés) possédant le plus grand bloc α (10 à 11 gènes), tandis que les rongeurs possèdent les plus grands blocs κ (plus de 50 gènes) et β (de 15 à 16 gènes) [Buhler, 2007].
Cette structure en blocs, propre aux mammifères, associée à la grande variation inter-spécifique du nombre de gènes dans ces blocs illustre une fois de plus la grande variabilité et plasticité de ce système génétique. Cette variabilité inter-spécifique s'observe aussi pour les gènes de classe II. Par exemple chez la vache et le mouton, les gènes HLA-DP sont remplacés par les gènes HLA-DI/DY, tandis que chez le chat, la région HLA-DQ est absente mais la région HLA-DR montre de nombreuses duplications de gènes [Kulski et al., 2002, Kelley et al., 2005].

La Figure 1.3 présente de manière synthétique la diversité du MHC, partant des taxons les plus éloignés évolutivement (poissons cartilagineux tels que le requin) et parcourant l'arbre des Gnathostomes jusqu'aux mammifères. Il est alors possible de dégager trois tendances : 1) tout d'abord une duplication en bloc du MHC [Kulski et al., 2002] ayant conduit de la dizaine de gènes du proto-MHC (théorique) de la Figure 1.1 à la centaine du système HLA (Figure 1.5), 2) un regroupement des loci dans une même région génomique, ce qui aurait facilité la co-évolution entre ces derniers [Trowsdale, 2002], et 3) l'important dynamisme de la région du MHC, caractérisée par des changements (gains ou pertes) de liaisons entre les régions, des duplications ou des pertes de loci, voire de régions entières.

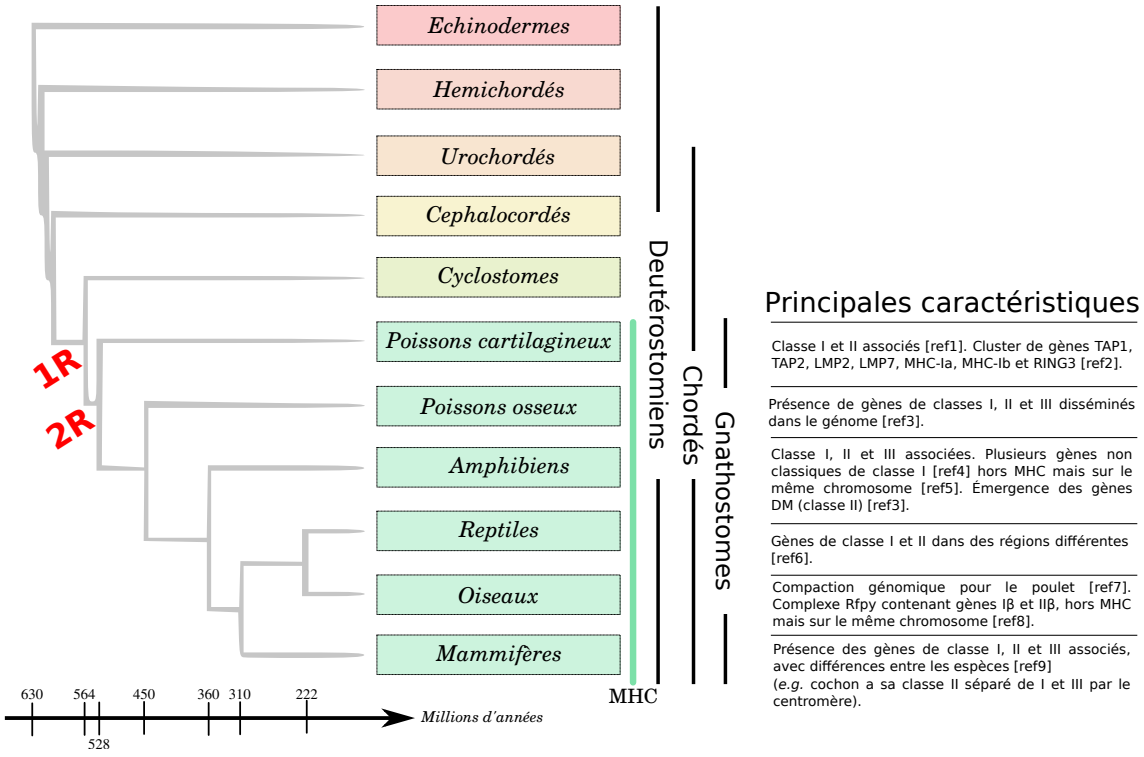


FIGURE 1.3 – Résumé des analyses de comparaison évolutive sur le MHC chez les Gnathostomes (adapté de [Flajnik, 2018]). La colonne « principales caractéristiques » n'a pas pour vocation de décrire *in extenso* l'ensemble des spécificités du MHC de chaque espèce de chaque clade, mais de dresser un tableau des traits principaux des MHC au sein de ces clades. 1R et 2R signalent les événements d'allopolyploïdisation des génomes ayant mené au « Big Bang évolutif » de l'immunité adaptative (voir [Ohno, 1999, Wolfe, 2001, Flajnik, 2018]). Les références sont : ref1 : [Ohta et al., 2000], ref2 : [Flajnik and Kasahara, 2001], ref3 : [Flajnik, 2018], ref4 : [Flajnik et al., 1993], ref5 : [Courtet et al., 2001], ref6 : [Jaratlerdsiri et al., 2014], ref7 : [Kaufman et al., 1999], ref8 : [Miller et al., 1996], ref9 : [Kulski et al., 2002].

3 Présentation du système HLA

La région du complexe HLA s'étend sur 3.7Mb, sur le bras court du chromosome 6 (position 6p21). La dernière cartographie en date du HLA, réalisée en 2009 par Shiina *et al.* [Shiina et al., 2009], a identifié 253 gènes (133 gènes codant pour des protéines, 22 pour des ARN non codants [6], 79 pseudogènes et 19 gènes de fonction indéterminée) et est représentée sur la Figure 1.5. À part les régions contenant des loci géniques, des éléments répétés tels que des SINE [7], LINE [8] et LTR [9] représentent respectivement 17.7, 16.7 et 10.7% de la région du HLA [Shiina et al., 2009].

3.1 Fonction des gènes « classiques »

Les gènes HLA sont répartis dans trois grandes régions qui ont été appelées classes I, II et III, auxquelles s'ajoute une région dite « de classe II étendue ». Les loci dis « classiques » correspondent aux loci codant pour les molécules HLA responsables de la détection des peptides antigéniques et leur présentation aux lymphocytes afin d'initier la réponse immunitaire [Ploegh and Watts, 1998].
La Figure 1.4 est une représentation schématique des domaines extra-cellulaires des molécules classiques de classes I et II, représentées en train de lier un peptide antigénique.

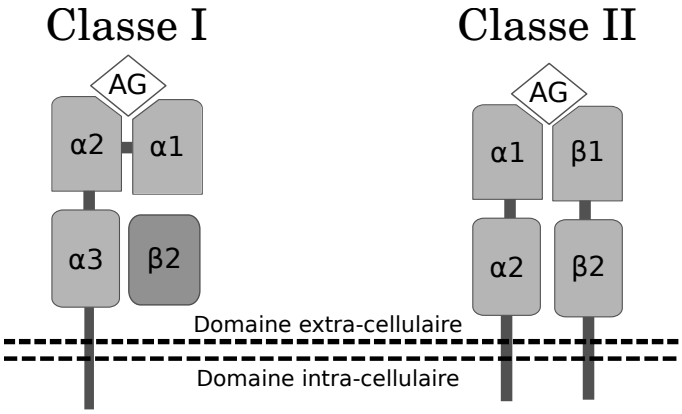


FIGURE 1.4 – Représentation schématique des domaines extra-cellulaires des molécules HLA « classiques » de classes I (à gauche) et II (à droite). Le losange AG représente l'antigène lié dans le site de reconnaissance de l'antigène, la double ligne en pointillés représente la membrane cellulaire et les régions $\alpha 1$ à $\alpha 3$ ainsi que $\beta 1$ et $\beta 2$ représentent les différents domaines des molécules (le domaine $\beta 2$ des molécules de classe I est codé par le gène de la $\beta 2$ micro-globuline, localisé hors de la région HLA).

Les loci classiques de classe I codent pour des molécules exprimées à la surface de l'ensemble des cellules nucléées, à l'exception des cellules spermatiques et certains neurones [Piertney and Oliver, 2006]. Elles sont composées de trois domaines extracellulaires α. Les deux premiers, $\alpha 1$ et $\alpha 2$, forment la région de liaison au peptide tandis

6. Aussi connus sous l'acronyme ARNnm, pour ARN non messager, ce sont des ARN non traduits en protéines par les ribosomes.
7. *Short Interspersed Nuclear Elements* : Petits Éléments Nucléaires Intercalés.
8. *Long Interspersed Nuclear Elements* : Longs Éléments Nucléaires Intercalés.
9. *Long Terminal Repeat Sequence* : Séquence terminale Longue Répétée.

que le domaine $\alpha 3$ est responsable de l'interaction avec les cellules T. La chaîne β est codée par le gène de la $\beta 2$ microglobuline, localisé hors du HLA, sur le chromosome 15 [Goodfellow et al., 1975]. Ces molécules présentent de petits peptides, généralement de 8 à 10 acides-aminés [Steven et al., 2000], dérivés d'antigènes endogènes (principalement d'origine virale) aux cellules T cytotoxiques CD8+ [Meyer et al., 2006].
Les gènes classiques de classe II codent pour des hétérodimères exprimés uniquement par les cellules présentatrices d'antigènes (telles que les cellules dendritiques ou les lymphocytes B). Les gènes fonctionnent par paires, par exemple HLA-DQA1 et HLA-DQB1 codent respectivement pour la chaîne α et β de la molécule HLA-DQ. La région de liaison au peptide est formée par les deux domaines $\alpha 1$ et $\beta 1$ codés par les exons 2, tandis que les exons 3 codent pour le domaine impliqué dans l'interaction avec les cellules T. Les autres exons codent pour des régions trans-membranaires ou cytosoliques. Ces molécules présentent de plus grands peptides, généralement entre 15 et 20 acides-aminés (en lien avec la conformation de la poche de liaison au peptide puisque seuls neufs acides-aminés sont liés dans le site de reconnaissance de l'antigène [Stern and Wiley, 1994, Madden, 1995]), dérivés d'antigènes exogènes (d'origine bactérienne ou parasitaire) aux cellules T auxiliaires CD4+ [Meyer et al., 2006].

3.2 Le site de reconnaissance de l'antigène

Les domaines $\alpha 1$ et $\alpha 2$ des molécules de classe I et les domaines $\alpha 1$ et $\beta 1$ des molécules de classe II forment la région de liaison au peptide (PBR [10]). Au cœur de cette région se situe le site de reconnaissance de l'antigène (ARS [11]), une poche composée d'une quarantaine d'acides-aminés et directement responsable de la détection et la liaison des peptides antigéniques [Reche and Reinherz, 2003]. Cette poche est caractérisée par une importante variabilité moléculaire des codons qui codent pour ces acides-aminés. Cette variabilité se manifeste par un taux de mutations non-synonymes [12] de ces codons bien supérieur au taux de mutations synonymes [13] (d'un facteur 3.2 à 4 pour les gènes HLA de classe I) suggérant une sélection de type « avantage de l'hétérozygote » [Hughes and Nei, 1988, Hughes and Nei, 1989c, Bitarello et al., 2016].
La Figure 1.6 illustre les positions des codons ARS au sein de la PBR telles qu'utilisées dans ce travail.

3.3 Nomenclature

La nomenclature HLA a été établie en 1987 suite au 10ème Atelier *IHIW* pour faire face au nombre croissant d'allèles HLA identifiés. La Figure 1.7 illustre la façon d'écrire un allèle HLA, selon la dernière mise-à-jour de 2010 [Marsh et al., 2010] :

10. En anglais : *Peptide Binding Region.*

11. En anglais : *Antigen Recognition Site.* Le nom exact de cette région n'est pas standardisé et il est possible d'observer d'autres noms dans la littérature, tels que (en anglais) *Antigen Bonding Site, Antigen Binding Site, Peptide Binding Groove...*. Par commodité, pour l'ensemble de ce travail, nous appellerons ARS la poche dans laquelle les peptides sont liés et PBR (*Peptide Binding Region*) la région de la molécule formée par les domaines $\alpha 1$ et $\alpha 2$ (classe I) et $\alpha 1$ et $\beta 1$ (classe II) et abritant l'ARS.

12. Mutation ponctuelle de l'ADN induisant un changement d'acide aminé codé par le codon muté.

13. Mutation ponctuelle de l'ADN n'induisant pas de changement d'acide aminé codé par le codon muté (mutation dite « silencieuse »).

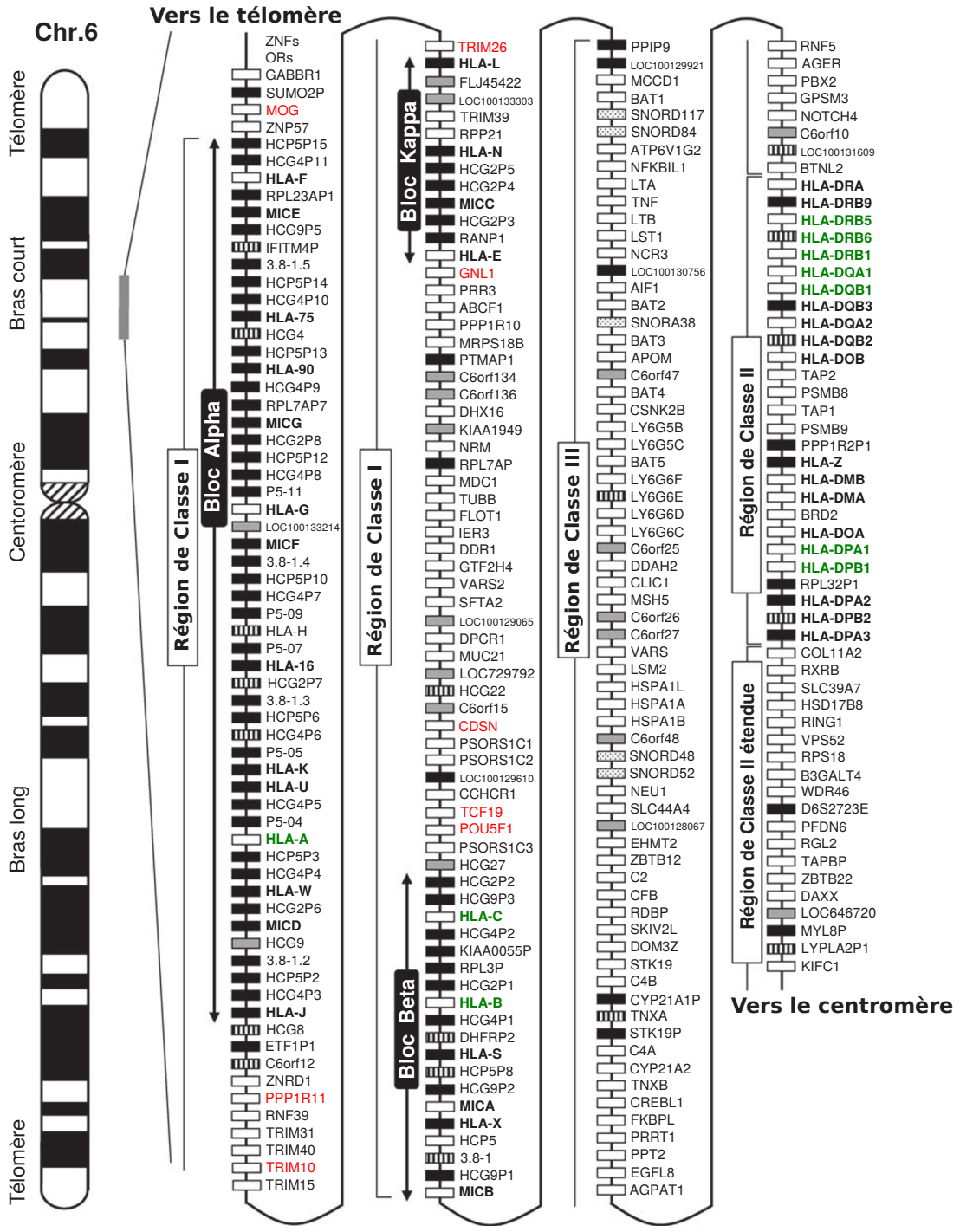


FIGURE 1.5 – Carte de la région du complexe HLA, adaptée de [Shiina et al., 2009], qui s'étend des coordonnées génomiques (chromosome 6) 29 677 984 ($GABBR1$) à 33 485 635 ($KIFC1$) sur l'assemblage 36.3 du génome humain du NCBI (National Center for Biotechnology Information). Les boîtes noires verticales encadrées de flèches représentent les trois blocs α, β et κ de la région de classe I. Les gènes dont le nom est en rouge correspondent à des gènes identifiés par[Amadou, 1999] comme étant des gènes d'ancrage (voir page 18). Les boîtes blanches donnent l'étendue des quatre régions HLA : régions de classe I, II, II étendue et III. Les boîtes horizontales respectivement blanches, grises, à bandes et noires représentent respectivement des gènes exprimés, des loci candidats à une fonction de gène, des gènes non-codants et des pseudogènes. Les gènes dont le nom apparaît en vert représentent les gènes étudiés dans ce travail de doctorat.

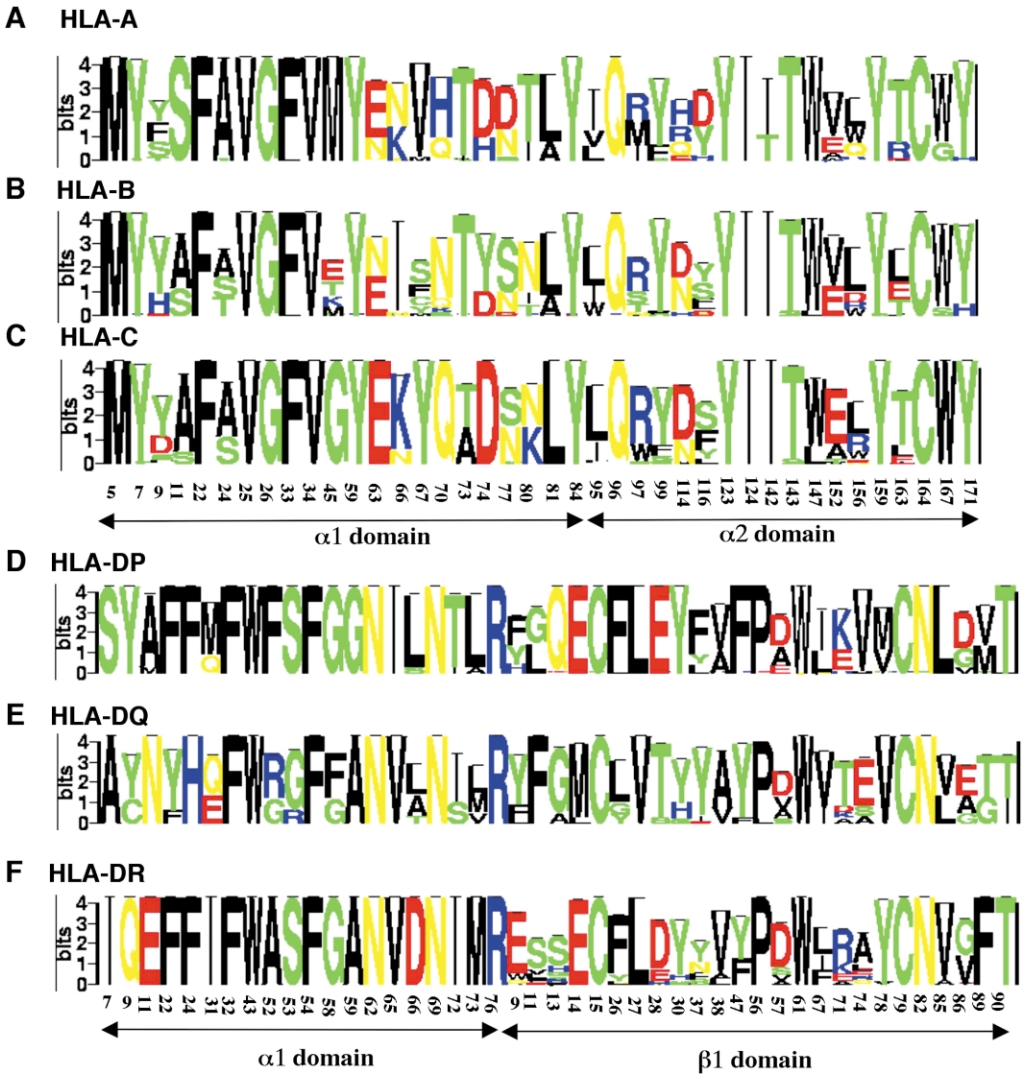


FIGURE 1.6 – Représentation logo des acides aminés observés sur les sites ARS, donnant la fréquence relative des différents acides-aminés observés à chaque site. La fréquence relative (en ordonnées) a été calculée à l'aide de la formule de l'entropie de Shannon [Shannon, 1948] (pour le détail de cette formule, voir le Chapitre 5). Le système utilisant 20 symboles, l'information maximale à une position est de $log2(20) = 4.3$ bits (s'il n'y a qu'un seul acide-aminé). Les numéros sur l'axe des abscisses représentent les positions des acides-aminés dans la séquence protéique. Les couleurs des acides-aminés sont arbitraires. Figure issue de [Reche and Reinherz, 2003] avec l'autorisation de l'auteur.

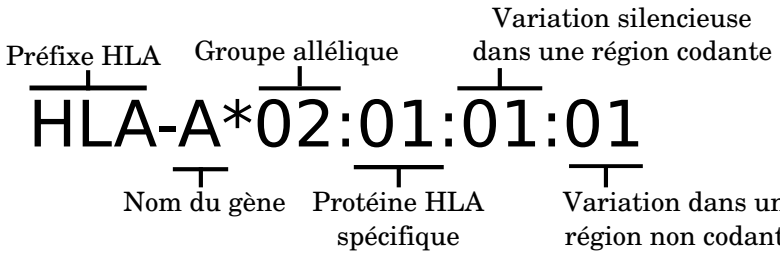


FIGURE 1.7 – Illustration de la nomenclature officielle des allèles HLA. Adaptée de `https://hla.alleles.org/nomenclature/naming.html`

Le nom complet d'un allèle HLA est d'abord composé du terme HLA, suivi d'un tiret et du nom du gène (*e.g.* A, B, C, DRB1 ...). Après l'astérisque viennent les différents champs indiquant le niveau de résolution de l'information disponible pour cet allèle. Le premier champ définit le groupe allélique, c'est-à-dire la spécificité sérologique (à l'exception de HLA-DPB1 qui a été défini par titrage cellulaire *in-vitro* [Buhler, 2007]). Le deuxième champ définit une variation non-synonyme au sein de la protéine, tandis que le troisième champ définit une variation silencieuse dans une des régions codantes du gène (exons). Le quatrième champ définit une variation dans une région non codante (introns ou UTR).
À la suite de ce nom peuvent apparaître les suffixes **L**, **S**, **N** ou **Q**, pour indiquer, respectivement, une expression faible (*low*), une protéine uniquement excrétée (*secreted*), un allèle « *nul* » non exprimé ou une expression discutable (*question*). Deux autres suffixes ont été définis, **C** et **A**, soit pour une expression cytoplasmique (et non à la surface de la cellule), soit pour une expression aberrante, mais n'ont pas encore été utilisés (du moins en mars 2017).

Ainsi, quand il est précisé qu'un allèle est défini au premier, second, troisième ou quatrième champ, cela donne une indication sur le niveau de résolution que l'on a sur cet allèle. Habituellement, le séquençage ADN permet d'obtenir au moins une résolution au troisième champ (séquençage des exons) et au mieux au quatrième champ quand les introns et UTR sont eux aussi séquencés.

3.4 Région HLA de classe I

La région de classe I regroupe, sur les 1.8Mb les plus télomériques du HLA, 128 loci incluant 19 gènes HLA de classe I, dont trois sont appelés « gènes classiques » (HLA-A, -B et -C) et trois sont appelés « non classiques » (HLA-E, -F et -G). Cette région inclut, aussi sept gènes MIC [14], dont seuls deux sont exprimés (MICA et MICB).

La région de classe I est organisée en trois blocs évolutifs, appelés α, β et κ, délimités par des gènes « d'ancrage » [15] [Amadou, 1999].
Le bloc α, le plus télomérique et le plus étendu des trois, comprend le gène classique HLA-A ainsi que quatre gènes MIC (MIC-E, -G, -F et -D, tous les quatre pseudogènes). Le bloc β, le plus centromérique des trois, comprend quant à lui les gènes classiques HLA-B et -C ainsi que deux gènes MIC (MIC-A et -B). Entre les deux se situe le bloc

14. En anglais : *MHC class I polypeptide-related sequence*, Séquence liée au polypeptide du MHC de classe I
15. En anglais : *anchor genes* ou *framework genes*.

κ, le plus petit en termes de taille bien que comprenant tout les éléments du duplicon originel : le gène MIC-C (pseudogène), les pseudogènes HCG2P-5, -4 et -3 (similaires au gène d'hémocromatose) et les deux gènes HLA-N (pseudogène) et HLA-E.

Ces régions présentent un important polymorphisme en nombre d'allèles, principalement dû aux trois gènes classiques de classe I, ces derniers comptabilisant [16] 18'742 allèles différents (voir la Table 1.1). Ces régions présentent aussi des polymorphismes de taille, par exemple au bloc β les individus porteurs de l'allèle HLA-B*48:01 montrent une délétion complète d'une portion d'une centaine de kilobases incluant le gène MIC-A [Komatsu-Wakui et al., 1999].

3.5 Région HLA de classe II

La région HLA de classe II est localisée sur la partie centromérique du HLA, d'une taille de 0.7Mb et contient 27 loci (dont cinq pseudogènes) [Shiina et al., 2009], dont plusieurs gènes hautement polymorphiques (HLA-DRBx [17], -DQA1, -DQB1, -DPA1 et -DPB1). Cette région abrite aussi d'autres gènes et pseudogènes moins polymorphiques tels que HLA-DQA2 et -DQB2, HLA-DOA et -DOB, ou les gènes PSMB8 et PSMB9 codant pour le protéasome (réalisant la protéolyse des antigènes dont les peptides seront présentés aux molécules de classe I) ainsi que les gènes TAP1 et TAP2, impliqués dans le transport des peptides du cytoplasme vers le réticulum endoplasmique (où ils seront présentés par les molécules de classe II) [Kulski et al., 2002].

L'analyse du déséquilibre de liaison des loci de classe II a mis en évidence la présence d'un point chaud de recombinaison méïotique près du gène TAP2 (2.3% de recombinaison par génération pour cette région, contre 0.31% pour la région de classe I) [Martin et al., 1995, Cullen et al., 1997, Jeffreys et al., 2001], ce qui a permis de définir deux blocs (indépendants des blocs de classe I) appelés δ et ϵ.

Le bloc δ comprend les gènes HLA-DR et -DQ (les plus polymorphiques des loci de classe II, voir Table 1.1). Les gènes HLA-DR présentent une organisation sous forme de blocs haplotypiques, toujours composés du gène HLA-DRB1 (très polymorphe, voir Table 1.1), du gène HLA-DRA (longtemps considéré comme monomorphe car présentant très peu de polymorphisme, voir Table 1.1) et du gène HLA-DRB9, auxquels s'associent jusqu'à trois autres loci. Cinq configurations DR ont été identifiées chez différents individus et sont représentées sur la Figure 1.8, les * indiquant les pseudo-gènes.

Le bloc ϵ, quant à lui, regroupe les gènes HLA-DPA1, -DPB1, -DPA2 et -DPB2 et constitue un bloc évolutivement stable puisque la structure (ordre des gènes, orientations et éléments *Alu*) est identique chez les humains (*Homo sapiens*), chimpanzés (*Pan sp.*) et orang-outans (*Pongo sp.*) (appartenant au micro-ordre des catarrhiniens, ou « singes de l'ancien monde ») ainsi que chez le tamarin à crête blanche (*Saguinus oedipus*, représentant du micro-ordre des platyrrhiniens, ou « singes du nouveau monde ») [Grahovac et al., 1993]. La divergence entre les catarrhiniens et les platyrrhiniens remontant à 38 millions d'années [Perelman et al., 2011] le bloc ϵ est donc une structure plus ancienne et très conservée tout au long de l'évolution des simiiformes (infra-ordre des

16. Recensement en janvier 2020.

17. Les gènes HLA-DRB présentent un important polymorphisme, caractérisé par des présences et absences de 10 gènes différents (voir plus loin).

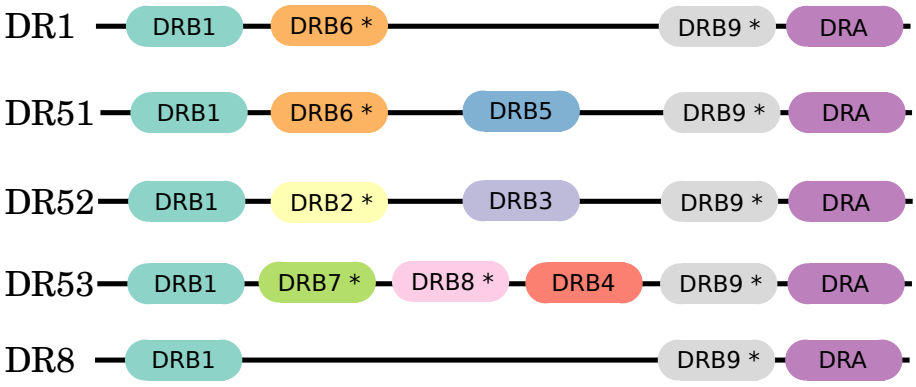


FIGURE 1.8 – Représentation schématique des configurations génomiques de la région DR selon les cinq haplotypes identifiés. Les * signalent les pseudogènes. Sources : [Hohjoh et al., 2001, Doxiadis et al., 2008b].

primates regroupant les catarrhiniens et platyrrhiniens).
Ce bloc inclut aussi les gènes HLA-DMA et -DMB, codant pour des protéines dont la fonction est de faciliter l'assemblage des molécules de classe II [Ceman et al., 1995]. Ces gènes sont retrouvés à partir de l'évolution des amphibiens [Flajnik, 2018], suggérant une origine ancienne (au moins 360 millions d'années, voir Figure 1.3).

Au-delà de la région de classe II, dans la partie la plus centromérique du HLA, se trouve la région dite « de classe II étendue ». Cette région ne contient aucun gène HLA classique, mais regroupe plusieurs gènes impliqués dans les processus immunitaires, tels que RXRB (facteur de transcription régulateur de l'expression des gènes de classe I [Dey et al., 1992]) ou TAPBP (une immunoglobuline impliquée dans le chargement des peptides facilitant la liaison entre les molécules HLA de classe I et la molécule TAP [Paulsson et al., 2002]). D'autres gènes tels que KIFC1, DAXX, ZBTB22 sont retrouvés dans des espèces évolutivement éloignées de l'humain telles que le poisson globe japonais (*Fugu rubripes*) [Clark et al., 2001] ce qui suggère qu'ils sont conservés depuis au moins 450 millions d'années (voir Figure 1.3).

	Gène	**Nb. Allèles**	**Nb. Protéines**	**Taille (nt.)**
Classe I	A	5'907	3'720	4'625
	B	7'126	4'604	3'341
	C	5'709	3'470	3'382
Classe II	DRA	29	2	5'205
	DRB1	3'331	2'357	11'080
	DRB3/4/5	479	458	13'068*/14'972*/12'945
	DQA1	229	98	18'884
	DQB1	1'795	1'194	8'917
	DPA1	168	65	16'210
	DPB1	1'537	1'006	13'771

TABLE 1.1 – Nombre d'allèles nominaux HLA connus (pour les gènes classiques) ainsi que nombre de protéines différentes codées par ces gènes (source : IMGT/HLA, v3.39, Janvier 2020) et la taille (en nucléotides) de ces gènes (source : `https://www.genecards.org/`, GRCh37/hg19 sauf pour "*" : GRCh38/hg38).

3.6 Région HLA de classe III

La région HLA de classe III, localisée entre la région de classe I et la région de classe II, regroupe 75 loci sur un segment de 0.9Mb [Shiina et al., 2009]. Il s'agit de la région la plus dense en gènes du génome humain [Milner, 2001, Kulski et al., 2002]. Cette région ne possède pas de gènes classiques HLA, mais regroupe un grand nombre de gènes liés au processus inflammatoire tels que des gènes de cytokines (LTA, LTB) ou le facteur de nécrose tumorale (TFN) dans sa partie télomérique (proche de la région de classe I). Dans sa portion plus centromérique (proche de la région de classe II) sont retrouvés des gènes du complément (C2, C4A, C4B), si bien que la grande différence entre la partie télomérique et la partie centromérique de la région de classe III a conduit certains à proposer de renommer la première partie « région de classe IV » [Gruen and Weissman, 1997].

4 Génération et variation du polymorphisme HLA

Maintenant que nous avons présenté l'apparition et les processus évolutifs à long terme ayant permis la diversification du MHC au sein des vertébrés, ainsi que la structure du HLA, nous allons nous intéresser aux processus évolutifs à plus court terme qui ont permis la génération et le maintien du polymorphisme du MHC/HLA au sein des espèces. Dans un premier temps nous nous intéresserons au modèle appelé « naissance et mort », processus ayant mené aux loci HLA actuels, puis nous présenterons les méthodes évolutives générant le polymorphisme au sein de ces loci. Finalement, nous présenterons les processus qui ont permis le maintien et la variation de ce polymorphisme, c'est-à-dire, d'une part, la sélection naturelle (les différents types, l'origine de la sélection et sa cible) et, d'autre part, l'effet de la démographie.

4.1 Génération du polymorphisme

Théorie de la « naissance et mort »

Le modèle de « naissance et mort » est proposé pour décrire les mécanismes évolutifs observés dans certaines familles multigéniques telles que HLA [Hughes and Nei, 1989b, Nei and Hughes, 1992, Klein et al., 1993b, Nei et al., 1997]. Ce modèle postule que de nouveaux gènes sont créés par des événements de duplication génique, tandis que certains deviennent des pseudogènes suite à des altérations génétiques (telles que des décalages du cadre de lecture induisant la formation de codons stop ou la perte de régions promotrices) ou sont tout simplement supprimés du génome.

Il a été mis en évidence que des espèces évolutivement distantes (telles que l'humain, la souris et le xénope) possèdent des gènes de classe I différents, illustrant ce processus de duplication et perte de gènes [Klein and Figueroa, 1986, Hughes and Nei, 1989b]. Par exemple chez l'humain, la région de classe I comprend 18 gènes HLA, dont 12 sont des pseudogènes [Robinson et al., 2015]. Les loci classiques A, B et C ne sont partagés que par les hominoïdes [18], tandis que les singes du nouveau monde (platirrhiniens) possèdent d'autres gènes classiques dans la région de classe I, bien que certaines espèces telles que le tamarin (*Saguinus oedipus*) possèdent des loci non classiques orthologues avec les humains [Chen et al., 1992].
Chez les félins, le MHC (appelé FLA [19]) possède 12 gènes de classe I fonctionnels [Yuhki et al., 2007] dont trois (FLAI-E, -H et -K) sont des gènes classiques [Yuhki et al., 2008, Holmes et al., 2013] tandis que chez les humains, HLA-H et -K sont des pseudogènes [Robinson et al., 2015]. De la même façon, les gènes de classe I du HLA ne sont pas orthologues avec les gènes de classe I du MHC de la souris [Rogers, 1985, Hughes and Nei, 1989a], ce dernier possédant lui aussi des pseudogènes [Trowsdale, 1995].
Pour résumer, les gènes de classe I ne sont pas orthologues chez les mammifères, indiquant une origine différente de ces gènes et cohérente avec la théorie de « naissance et mort » [Hughes and Nei, 1989b].

Les gènes de classe II montrent une orthologie plus importante entre les différentes

18. L'une des deux grandes familles des catarrhiniens, comprennant les gibbons, les orang-outans, les gorilles, les chimpanzés et les humains.
19. Pour *Feline Leukocyte Antigen*.

espèces de mammifères : les régions DR, DQ, DP et DO sont partagées entre les humains et les rongeurs [Nei et al., 1997], mais les bovins ont perdu la région DP et gagné une région DY/DI [Band et al., 1998], de même que la région DP est absente des chiens, les chats domestiques ne possèdent pas de région DQ [Yuhki et al., 2007] mais montrent une expansion de la région DR [Yuhki, 2003]. Au contraire, si les gènes de classe II des mammifères sont orthologues, les gènes de classe II de taxons évolutivement plus éloignés (par exemple les oiseaux et amphibiens) ne montrent pas d'orthologie avec ceux des mammifères [Hughes and Nei, 1990].

On observe aussi que le modèle évolutif des gènes de classe II est similaire à celui des gènes de classe I [Hughes and Nei, 1990], les premiers évoluant aussi selon un modèle de « naissance et mort » bien que leur renouvellement soit plus long [Nei et al., 1997, Takahashi, 2000, Piontkivska, 2003].

Mutations ponctuelles et recombinaisons

L'un des facteurs principaux de l'évolution moléculaire de génération en génération est la mutation ponctuelle de l'ADN. Une étude de Satta *et al.* de 1996 a estimé le taux de mutation dans la région HLA-DR (introns et 3'UTR) entre $0.79 \cdot 10^{-9}$ et $1.27 \cdot 10^{-9}$ par site et par an [Satta et al., 1996a]. Ces résultats sont similaires à ceux obtenus précédemment par Satta *et al.* en 1993, qui avait estimé des taux de mutation de l'ordre de 10^{-9} par site et par an (HLA-A : $1.37 \pm 0.61 \cdot 10^{-9}$, HLA-B : $1.84 \pm 0.40 \cdot 10^{-9}$, HLA-C : $3.87 \pm 1.05 \cdot 10^{-9}$, HLA-DRB1 : $1.18 \pm 0.36 \cdot 10^{-9}$), utilisant des simulations informatiques basées sur l'ensemble des exons de ces gènes pour sept espèces de primates différentes (humain, chimpanzee, gorille, orang-outan, gibbon, macaque et tamarin) [Satta et al., 1993]. Le taux de mutation des loci HLA semblent alors similaire à ce qui est observé dans le reste du génome humain [Nachman and Crowell, 2000].

Toutefois, le mécanisme de conversion allélique (ou recombinaison intragénique) est aussi proposé comme une force évolutive importante pour HLA [Gyllensten et al., 1991, Andersson and Mikko, 1995, Bergström et al., 1998, Richman et al., 2003b, Richman et al., 2003a]. La conversion allélique consiste en un échange de segments courts (habituellement moins de 35 nucléotides pour HLA [Parham et al., 1995]), de manière non réciproque, entre deux allèles d'un même locus.
Par rapport à la mutation ponctuelle, ce mécanisme présente l'avantage de générer de nouveaux variants de manière plus rapide, puisque les segments échangés proviennent d'allèles ayant déjà passé le filtre de la sélection naturelle et de la dérive génétique [Hedrick, 1998]. De plus, il est avancé que la répartition des codons ARS tout le long de l'exon 2 (et exon 3 pour les classe I) permet de créer plus efficacement de nouveaux variants, fonctionnellement différents, par recombinaison intragénique [Mona et al., 2008].

Les goulots d'étranglements [20], consécutifs à une diminution de taille des populations, sont des événements causant une réduction de la diversité à l'échelle du génome, donc aussi du MHC, et dont le signal a été observé dans de nombreuses populations animales telles que l'oryx d'arabie (*Oryx leucoryx*) [Hedrick et al., 2000], l'élan (*Alces alces*) [Mikko and Andersson, 1995], le chamois des Pyrénées ou des Alpes [Schaschl et al., 2005] ou d'autres espèces de ruminants sauvages [Smulders et al., 2003]. Dans ce cas, il sem-

20. *Bottleneck*

blerait que la diversité génétique puisse être rétablie rapidement par conversion allélique [Andersson and Mikko, 1995].
Plusieurs exemples de conversion allélique ont aussi été rapportés chez l'humain. L'allèle HLA-B*53, identifié par plusieurs études comme étant protecteur vis-à-vis de la malaria [Hill et al., 1991, Adamek et al., 2015, Sanchez-Mazas et al., 2017] résulterait apparament d'une conversion allélique impliquant HLA-B*35 [Allsopp et al., 1991]. Des études menées sur des populations natives d'Amérique du Sud ont aussi mis en évidence de nombreux allèles qui résulteraient de conversions alléliques [Belich et al., 1992, Watkins et al., 1992, Parham et al., 1997]. Certains de ces allèles n'étant pas retrouvés dans des populations asiatiques, on suppose que ces allèles sont récents (d'un point de vue évolutif) [Hedrick, 1998], puisque le peuplement des Amériques, depuis l'Asie par le détroit de Béring, daterait, selon la plupart des spécialistes, de 15'000 à 25'000 ans [Goebel et al., 2008, Skoglund and Reich, 2016, Bourgeon et al., 2017].
Finalement, une étude de 1995 étudiant la fréquence de la conversion allélique dans l'exon 2 de HLA-DPB1 dans des spermatozoïdes de personnes hétérozygotes [Zangenberg et al., 1995] a mis en évidence un taux de conversion allélique de 1 pour 10'000 gamètes, illustrant le rôle important de la recombinaison à ce locus.

L'importance de la conversion allélique dans la génération du polymorphisme fait toutefois débat. Une étude menée en 2006 sur 18 chamois des Pyrénées [Schaschl et al., 2005] a mis en évidence une contribution de la conversion allélique dix fois plus importante que celle de la mutation ponctuelle. Toutefois, en 2008, une nouvelle étude [Mona et al., 2008] a remis en cause cette mesure[21], avançant une contribution similaire de la conversion allélique et de la mutation ponctuelle, cette dernière ayant été identifiée comme une source de polymorphisme importante pour HLA [Takahata et al., 1992, Klein et al., 1993c, Takahata and Satta, 1998].

4.2 Variations du polymorphisme par sélection naturelle

Le rôle premier du système HLA est celui de l'immunité adaptative, à savoir de détecter la présence de pathogènes dans l'organisme via les peptides antigéniques liés, puis de déclencher la réponse immunitaire. Ainsi son polymorphisme est principalement maintenu par pression pathogénique, exercée sur les populations [Doherty and Zinkernagel, 1975, Prugnolle et al., 2005b] bien que les processus démographiques aient aussi joué un rôle important [Di and Sanchez-Mazas, 2014, Pechouskova et al., 2015].
Le polymorphisme HLA est alors un équilibre entre une grande diversité permettant de détecter un large spectre de peptides antigéniques (et donc une large gamme de pathogènes) et une fréquence suffisante d'allèles protecteurs spécifiques en cas de prévalence élevée d'un pathogène particulier [Meyer and Thomson, 2001, Sommer, 2005, Piertney and Oliver, 2006]. Les deux mécanismes principalement rencontrés avec le HLA, et présentés ci-après, sont la sélection balancée et la sélection directionnelle.

21. [Mona et al., 2008] expliquent que [Schaschl et al., 2005] ont réalisé leurs calculs sur les séquences uniques et non en considérant l'ensemble des séquences de chaque individu, comme exigé par les méthodes basées sur la coalescence.

Cible de la sélection

Avant de présenter les différents mécanismes de sélection maintenant le polymorphisme HLA, il est nécessaire d'identifier la cible de cette sélection.
La fonction des molécules HLA est de détecter, lier et présenter des peptides antigéniques aux lymphocytes T (voir page 14). L'analyse de la structure tri-dimensionnelle des molécules de classe I [Bjorkman et al., 1987b] et II [Brown et al., 1993] a permis d'identifier les résidus en contact avec les peptides antigéniques. Ces résidus forment une poche (voir Figure 1.4) appelée site de reconnaissance de l'antigène (ARS [22]) et sont donc directement impliqués dans la liaison des peptides antigéniques [Reche and Reinherz, 2003].

Plusieurs études ont pu déterminer que les codons ARS étaient plus largement ciblés par la sélection que le reste de la région de liaison au peptide (PBR). Chez le chamois des Alpes (*Rupicapra rupicapra*), des études ont montré que 64 à 82% [Schaschl et al., 2004, Mona et al., 2008] des sites polymorphiques observés sur l'exon 2 de rupi-DRB étaient localisés sur les codons ARS (ces derniers ne représentant, pour rupi-DRB, que 22% des sites de l'exon 2). Ces codons ARS montrent aussi un plus grand nombre de mutations non-synonymes que de mutations synonymes [Schaschl et al., 2004, Schaschl et al., 2005, Mona et al., 2008], jusqu'à un facteur 7 pour le chamois des Pyrénées (*Rupicapra pyrenaica*) [Alvarez-Busto et al., 2007].

Sélection balancée

La sélection balancée est un processus sélectif qui regroupe plusieurs phénomènes évolutifs : l'avantage des hétérozygotes, la sélection fréquence-dépendante et la variation dans le temps et l'espace de la sélection [Doherty and Zinkernagel, 1975, Slade and McCallum, 1992, Hedrick, 2002]. C'est un processus naturel par lequel plusieurs allèles sont maintenus dans les populations à des fréquences plus importantes que celles attendues par simple dérive génétique.

Un des effets de la sélection balancée est la rétention du polymorphisme trans-spécifique. Le polymorphisme trans-spécifique est la présence d'allèles identiques (par ascendance) au sein de plusieurs espèces. La théorie de la coalescence postule que la proportion d'allèles partagés entre deux espèces diminue avec le temps depuis la divergence mais que la sélection balancée permet de maintenir plus longtemps ce nombre d'allèles partagés [Takahata, 1990, Takahata and Nei, 1990].
Un grand nombre d'études ont mis en évidence du polymorphisme trans-spécifique au sein du MHC, notamment chez des salmonidés [Garrigan and Hedrick, 2001, Miller and Withler, 2004], des ongulés [Van Den Bussche et al., 2002], des pinnipèdes [Hoelzel et al., 1999], des rongeurs [Musolf et al., 2004], les geckos [Radtkey et al., 1996], les passereaux [Richardson and Westerdahl, 2003] et pour finir, les primates [Klein et al., 1993d, Suarez et al., 2003].

Un exemple notable de polymorphisme trans-spécifique est observable chez les chamois des Alpes (*Rupicapra rupicapra*) et des Pyrénées (*Rupicapra pyrenaica*). Ces deux espèces ont divergé il y a 1.6 million d'années [Lalueza-Fox et al., 2005] et ont ensuite subi un déclin démographique (dû à la déglaciation et à la fragmentation de leur habitat) induisant une dérive génétique rapide. Cette différenciation génétique

22. *Antigen Recognition Site.*

s'observe particulièrement au niveau des séquences mitochondriales (cytochrome B) puisque le temps de coalescence estimé est 2.5 fois supérieur à la date de la divergence, les deux espèces montrant une monophylie stricte pour ce locus. Malgré cette importante différenciation entre les deux espèces, celles-ci possèdent encore plusieurs allèles similaires dans leurs gènes MHC respectifs, illustrant l'importance de la sélection balancée dans la conservation du polymorphisme trans-spécifique [Mona et al., 2008].
Finalement, en 2003, la comparaison des résultats de 48 études sur plusieurs espèce, n'a rapporté qu'une seule étude ne mettant pas en avant un effet de la sélection balancée sur le MHC [Bernatchez and Landry, 2003].

Nous allons maintenant voir en détail les différents phénomènes évolutifs qui composent la sélection balancée, à savoir l'avantage de l'hétérozygote, la sélection fréquence-dépendante et la variation dans le temps et l'espace de la sélection.

Avantage de l'hétérozygote

L'hypothèse de l'avantage de l'hétérozygote postule que les individus hétérozygotes à un locus HLA possèdent un avantage sélectif en étant capables de présenter davantage de peptides antigéniques que les homozygotes et donc d'être mieux protégés dans un environnement riche en pathogènes [Doherty and Zinkernagel, 1975, Hedrick, 1998]. L'avantage de l'hétérozygote peut être symétrique (tout les hétérozygotes ont la même valeur sélective [Doherty and Zinkernagel, 1975]) ou asymétrique avec une valeur sélective plus importante pour les hétérozygotes présentant des allèles fortement divergents [Wakeland et al., 1990]. Cet avantage de l'hétérozygote peut survenir via une simple dominance, où un seul des deux allèles confère une protection contre un pathogène [Penn et al., 2002], ou via une sur-dominance, où les deux allèles participent à la défense immunitaire contre 1) soit un seul pathogène présentant plusieurs peptides antigéniques liés par plusieurs molécules HLA [Kurtz et al., 2004], 2) soit plusieurs pathogènes [Penn et al., 2002, McClelland et al., 2003, Wegner et al., 2003, Wegner et al., 2004], 3) soit plusieurs souches co-infectantes d'un même pathogène [Hughes and Nei, 1992].
Que ce soit dans le cas d'une simple dominance ou d'une sur-dominance, les hétérozygotes ont une valeur sélective au minimum égale (même dans le cas de la présence d'un allèle causant une susceptibilité à un pathogène [Worley et al., 2010]) et souvent supérieure aux homozygotes [Hughes and Nei, 1988]. La distinction est importante puisque seule la sur-dominance permet de maintenir la diversité génétique : en cas de simple dominance on tendrait vers une homozygotie puisque l'allèle qui n'est pas impliqué serait éliminé par dérive génétique [Takahata and Nei, 1990, McClelland et al., 2003].
Et finalement, à l'échelle individuelle, la valeur sélective n'est pas optimale lorsque que l'hétérozygotie est maximale [Nowak et al., 1992]. En effet, les molécules HLA liant aussi bien des peptides d'origine antigénique que endogéniques (les peptides du *soi*), une plus grande hétérozygotie implique plus de peptides du *soi* liés. Or, plus de peptides du *soi* présentés implique une sélection négative plus drastique sur les cellules T lors de leur maturation dans le thymus[23], menant à une baisse de leur diversité [Nowak et al., 1992, Messaoudi, 2002, Radwan et al., 2020].

23. La maturation des lymphocytes T dans le thymus est composée de deux étapes. La première, appelée sélection positive, ne va retenir que les lymphocytes T capables de se lier aux molécules HLA et la seconde, la sélection négative, va éliminer tous les lymphocytes qui se lient à des molécules HLA présentant des peptides du *soi*, ou auto-antigènes [Reece et al., 2007].

Plusieurs études ont déjà mis en évidence l'existence de l'avantage de l'hétérozygote pour des locis du MHC/HLA. Chez les chamois des Alpes (*Rupicapra rupicapra*) [Schaschl et al., 2012] et le babouin (*Papio ursinus*) [Huchard et al., 2010], l'hétérozygotie est liée à la longévité des individus mâles, les mâles hétérozygotes montrant une durée de vie plus élevée que les homozygotes. Une étude menée dans des populations de saumon royal (*Oncorhynchus tshawytscha*) en Colombie-Britannique [Evans and Neff, 2009] a montré une plus faible proportion d'individus infectés (par un panel de 80 bactéries identifiées) chez les hétérozygotes que chez les homozygotes au locus de classe II B1. Chez la souris, une étude basée sur la co-infection par *Salmonella enterica* et le virus de l'encéphalomyélite murine de Theiler a mis en évidence une charge pathogénique plus faible chez les hétérozygotes [McClelland et al., 2003]. Chez les humains, une étude menée sur une cohorte gambienne a mis en évidence une diminution des cas symptomatiques de l'hépatite B chez les individus hétérozygotes aux loci HLA-DQ et -DR [Thursz et al., 1997] et une autre étude menée en 2005 [Prugnolle et al., 2005b] a montré que l'hétérozygotie aux loci HLA-A et -B était corrélée avec la richesse en pathogènes des milieux dans lesquels vivent les populations. En 2012, Sanchez-Mazas *et al.* ont obtenu des résultats similaires pour les loci HLA-A et -B pour 80 populations, avec une corrélation entre l'hétérozygotie de ces loci et la richesse en pathogènes [Sanchez-Mazas et al., 2012]. Les auteurs ont toutefois souligné que 1) cette corrélation n'était plus significative lorsqu'étaient retirées les populations soumises à une dérive génétique rapide et 2) une corrélation négative était observée au locus HLA-DQB1, suggérant des mécanismes évolutifs différents. En 2017, dans une étude sur des populations africaines, Sanchez-Mazas *et al.* n'a toutefois pas montré de corrélation entre la diversité des allèles HLA-A et -B, effet peut-être dû à une aire géographique plus restreinte [Sanchez-Mazas et al., 2017]. Une corrélation entre la diversité des pathogènes et du HLA a aussi été mise en évidence pour les deux gènes de classe II du rat taupe (*Spalax ehrenbergi*), où l'hétérozygotie est corrélée avec la charge pathogénique (en acariens gamasides et vers parasites) du milieu [Nevo and Beiles, 1992].

Sélection fréquence-dépendante

Un autre type de sélection balancée s'appelle la sélection fréquence-dépendante, aussi connue sous le nom de l'avantage de l'allèle rare. Deux mécanismes sont avancés pour expliquer cette sélection.
Dans un premier cas, le pathogène peut développer des adaptations qui lui permettent d'échapper aux allèles HLA les plus communs dans une population (co-évolution hôte-parasite). Dans ce cas il est possible que ce soient les allèles les plus rares, pour lesquels le pathogène n'a pas évolué pour lui permettre d'y échapper, qui apportent une résistance. Ces allèles verront leurs fréquences augmenter par sélection directionnelle positive [Takahata and Nei, 1990], exerçant à leur tour une pression de sélection sur le pathogène qui évoluera éventuellement en de nouvelles souches capables d'échapper à ce nouvel allèle, et ainsi de suite. Ce mécanisme engendre un processus dynamique de co-évolution hôte-pathogène, où la valeur sélective marginale de chaque allèle fluctue de manière cyclique, permettant de maintenir un polymorphisme génétique au niveau du HLA si ces fluctuations sont plus rapides que la perte des allèles rares par dérive génétique [Slade and McCallum, 1992].

La sélection fréquence-dépendante n'est pas non plus incompatible avec l'avantage de l'hétérozygote, puisque les individus hétérozygotes peuvent avoir une meilleure valeur sélective juste parce qu'ils portent un allèle rare et non parce qu'ils sont simplement

hétérozygotes [Penn, 2002]. De plus, il peut y avoir une composante fréquence-dépendante à l'avantage de l'hétérozygote : puisqu'il y a de faibles probabilités qu'un hétérozygote hérite de deux allèles rares, ces derniers seront sur-représentés chez les hétérozygotes [24]. Ces deux mécanismes sélectifs, avantage de l'hétérozygote et avantage des allèles rares, peuvent aussi varier dans le temps et l'espace puisque la pression pathogénique varie [Spurgin and Richardson, 2010], tout en étant difficilement identifiables séparément par des tests classiques de sélection [Takahata and Nei, 1990, Ejsmond et al., 2010].

Sélection fluctuant dans le temps et l'espace

Finalement, la dernière forme de sélection balancée, la sélection fluctuant dans le temps et l'espace, postule que la variation spatiale et temporelle des différents pathogènes et de leurs abondances peut aussi maintenir la diversité observée aux gènes HLA. La fluctuation des différents pathogènes va engendrer une fluctuation dans les fréquences des allèles sous sélection, soit dans le temps, soit dans l'espace entre différentes populations et mesurable à l'aide de F_{ST} (voir page 46). Deux critères sont néanmoins nécessaires pour expliquer ce modèle : premièrement, que la sélection d'un allèle par la pression d'un pathogène soit directionnelle et, deuxièmement, que ces fluctuations de pressions pathogéniques soient déterminées par un facteur externe et non interne, telle que la co-évolution hôte-pathogène puisque, dans ce cas là, des évolutions du pathogène pourraient lui permettre d'échapper au système immunitaire de l'organisme (tel qu'observé dans le cas de sélection fréquence-dépendante) [Spurgin and Richardson, 2010]. Des simulations mathématiques ont montré que, en théorie, la sélection fluctuant dans le temps et l'espace pouvait à elle seule expliquer la diversité observée pour HLA [Hedrick, 2002].

Une étude menée en 2008 portant sur la comparaison des allèles MHC-DRB et de marqueurs micro-satellites de trois méta-populations [25] de grand campagnol (*Arvicola amphibius*) en Écosse (Royaume-Uni), sur la période 1998-2005, a mis en évidence la présence de plusieurs forces évolutives, dont une sélection fluctuant dans le temps et l'espace [Oliver et al., 2008].
Entre les méta-populations, sur de grandes distances géographiques (les trois méta-populations étant distantes de plus de 100km), la variabilité du MHC semble principalement influencée par la dérive génétique.
À l'intérieur des méta-populations par contre, la diversité du MHC apporte des preuves d'une sélection agissant pour maintenir une diversité fonctionnelle. Un signal de sélection balancée a été observé dans deux méta-populations sur une échelle de temps courte (de l'ordre de l'année), ces deux méta-populations montrant alors un excès d'hétérozygotes, indiquant une sélection sous la forme d'un avantage de l'hétérozygote. Pour une de ces mêmes méta-populations, un signal de sélection directionnelle a aussi été observé pour une autre année. De plus, au sein d'une des méta-populations, l'étude n'a identifié de signaux de sélection naturelle que pour trois ($\sim$6 générations) des six années durant lesquelles cette méta-population a été étudiée. Les mesures pour les six années d'étude suggèrent même que les facteurs neutres, tels que la migration et la dérive génétique, soient les premiers facteurs de variabilité génétique à long terme.
Cette étude illustre alors les variations rapides de régimes de sélection qui peuvent être

24. Puisque la probabilité de posséder deux allèles rares est bien plus petite que celle de posséder un allèle rare et un allèle plus commun.

25. Puisqu'étant formées de plusieurs colonies indépendantes, échantillonnées en trois localités différentes.

observés au MHC, puisque ces méta-populations montrent aussi bien des signaux de sélection balancée que directionelle selon les années, tout en soulignant le rôle prépondérant de la dérive génétique. Les auteurs concluent aussi sur l'intérêt à porter à la conception des expériences, car la sélection varie temporellement, mais aussi géographiquement puisque différents signaux de sélection peuvent être observés selon l'échelle géographique considérée.

Sélection directionnelle

Plusieurs études ont mis en évidence que la sélection balancée n'était pas la seule force évolutive agissant sur les gènes HLA.
Une étude menée dans le cadre du projet UK10K [26] [The UK10K Consortium, 2015] en 2016 et étudiant des traces de sélection récente dans des génomes britanniques a mis en évidence, pour trois loci HLA [27], des signaux de sélection positive survenue il y a moins de 3'000 ans [Field et al., 2016]. Une autre étude de 2016 [Lindo et al., 2016], basée sur 25 génomes de populations natives actuelles d'Amérique et 25 génomes de populations natives d'Amérique ayant vécu avant la colonisation européenne a mis en évidence un changement de régime de sélection pour le locus HLA-DQA1, qui serait passé d'une sélection directionnelle positive à une sélection directionnelle négative après la colonisation. Dans une récente étude, Barquera *et al.* ont analysé les prédictions *in-silico* de liaisons peptidiques de 438 molécules HLA de classes I et II pour plusieurs virus [28] pour 158 à 374 populations (selon le locus) réparties mondialement [Barquera et al., 2020]. Les auteurs ont rapporté une fréquence élevée des molécules identifiées comme des « ligands forts » (c'est-à-dire liant plus de 1% des peptides présentés) chez les populations natives américaines, associée à une faible fréquence des molécules identifiées comme « ligands faibles » (liant moins de 1% des peptides). Ces populations n'étant pas nécéssairement proches géographiquement, le partage de ces molécules « ligands forts » est proposé par les auteurs comme le résultat d'une sélection positive sur ces molécules, probablement sous la forme d'un balayage sélectif *doux* [29].
Une étude menée sur le chamois (*Rupicapra rupicapra*) en Autriche [Schaschl et al., 2012] a mis en évidence une sélection positive de l'allèle rupi-DRB*169, qui est deux fois plus fréquent dans les populations ayant été exposées à la gale (*Sarcoptes scabiei*) que dans celles n'ayant pas été exposées, allant jusqu'à causer une diminution importante de l'hétérozygotie à ce locus chez les populations exposées.
Des signaux de sélection positive au locus MHC-DRB ont aussi été observés chez trois autres espèces d'ongulés, le bouquetin des Alpes (*Capra ibex*), le bouquetin d'Espagne (*Capra pyrenaica*) et le Jahral (*Hemitragus jemlahicus*) [Schaschl et al., 2006]. Cette étude a déterminé que les codons ARS étaient la cible de cette sélection directionnelle.
Finalement, un exemple notable de sélection directionnelle chez l'humain concerne l'allèle HLA-B*53 et son association avec la malaria. En 1991, une étude menée par Hill *et al.* sur une cohorte de 2'000 Gambiens a montré que cet allèle était beaucoup plus fréquent (25%) chez les individus sains ou présentant des formes modérées de malaria, alors qu'il était moins fréquent (16%) chez les individus présentant une forme sévère de malaria [Hill et al., 1991]. L'année d'après, en 1992, Hill *et al.* a mis en évidence le mécanisme moléculaire expliquant ce résultat comme étant la reconnaissance, par la molécule

26. *United Kingdom 10'000 genomes.*
27. SNP
28. Trois virus de type influenza, un lentivirus et trois coronavirus, incluant SARS-CoV2.
29. C'est-à-dire un balayage sélectif touchant plusieurs allèles ou molécules en même temps.

HLA-B*53, d'un peptide nonamérique dérivé d'un antigène du parasite *Plasmodium falciparum* (responsable de la malaria en Afrique de l'ouest) et exprimé durant l'étape hépatique de l'infection (LSA-1) [Hill et al., 1992b]. Finalement, une étude menée en 2017 a mis en évidence une corrélation directe entre la prévalence de *P. falciparum* et la fréquence de l'allèle HLA-B*53:01:01 en Afrique, suggérant une sélection positive de cet allèle en réponse à une protection vis-à-vis de la malaria [Sanchez-Mazas et al., 2017].

Ces études montrent que la sélection balancée n'est pas la seule force évolutive agissant sur les loci HLA, mais que des signaux de sélection directionnelle (positive ou négative) peuvent être observés.

4.3 Effets de la géographie et de la démographie

Bien que la sélection due aux pathogènes soit un facteur important de la diversité génétique HLA, la démographie et la géographie vont elles aussi influencer les profils alléliques HLA des populations.

Plusieurs études ont révélé une composante géographique dans la structure génétique des populations aux loci HLA. En 2001, une étude menée par Sanchez-Mazas [Sanchez-Mazas, 2001] a montré que, pour plusieurs loci impliqués dans des réactions immunitaires (système Rhésus, systèmes GM et HLA-DPB1), il existait une corrélation importante entre les distances génétiques et géographiques des populations, soutenant un modèle d'isolation par la distance [Slatkin, 1993]. HLA-DPB1 (comme le système GM) montrait la plus forte corrélation avec la géographie ($r = 0.61$) ainsi qu'une importante différenciation entre groupes de populations ($F_{CT} = 0.107$) lorsque ces derniers étaient définis par des régions géographiques (Afrique sub-saharienne, Europe et Asie de l'est). En 2014, une étude de Di *et al.* sur le peuplement de l'Asie avait mis en évidence que pour ces populations, les effets démographiques prévalaient sur la sélection naturelle pour expliquer la répartition du polymorphisme HLA [Di and Sanchez-Mazas, 2014].

En 2005, Prugnolle *et al.* ont mis en évidence que, pour les trois loci de classe I HLA-A, -B et -C, les distances géographiques [30] expliquaient, à elles seules, 17 à 39% de la variabilité de l'hétérozygotie estimée, la part restante étant due à la pression pathogènique (estimée par la richesse en pathogènes) [Prugnolle et al., 2005b]. Ces résultats sont similaires à ceux de Qutob *et al.* 2012 qui ont montré que la géographie (distance terrestre) expliquait 36 à 41% de la perte d'hétérozygotie aux trois loci de classe I quand la distance des populations vis-à-vis d'une origine géographique supposée en Afrique sub-saharienne [31] augmentait [Qutob et al., 2012].

En 2014, une étude de Di et Sanchez-Mazas a étudié l'histoire du peupleument asiatique au travers de la diversité HLA [Di and Sanchez-Mazas, 2014]. Cette étude, basée sur 90 populations d'Asie de l'est a révélé une différenciation génétique nord-sud de ces populations. Ce résultat illustre le rôle prépondérant de la géographie dans la distribution de la variabilité HLA, par rapport à la sélection naturelle.
Une observation similaire a été faite pour les populations des îles d'Asie et du Pacifique (Austronésie, Micronésie, Mélanésie et Polynésie). En 2000, Mack *et al.* ont analysé la

30. Itinéraires les plus courts à travers les continents et, entre les continents, par les ponts terrestres.
31. Très exactement depuis le point de latitude -12° et longitude 25°, avancé dans cette étude comme étant le point d'origine des humains anatomiquement modernes [Betti et al., 2009, Manica et al., 2007].

diversité des gènes HLA de classe II (HLA-DRB1, -DQB1 et -DPB1) chez 11 populations de ces régions [Mack et al., 2000]. Les relations phylogénétiques entre ces populations, estimées à l'aide de la méthode de *Neighbor-Joining* [Saitou and Nei, 1987] sur les fréquences des haplotypes HLA-DRB1∼DQB1, montrent que les populations sont regroupées par régions géographiques, illustrant ici aussi l'effet de la géographie sur la distribution des polymorphismes HLA.

Une étude menée en 2011 par Buhler et Sanchez-Mazas [Buhler and Sanchez-Mazas, 2011] sur la diversité nucléotidique des exons codant pour la région de liaison au peptide (exons 2 et 3) de sept loci HLA (HLA-A, -B, -C, -DRB1, -DQA1, -DQB1 et -DPB1) au sein d'un ensemble de 56 à 106 (selon le locus considéré) échantillons de populations à l'échelle mondiale a mis en évidence une corrélation entre la distance géographique et la distance génétique (mesurée par le coefficient de Reynolds). Cette corrélation variait entre les loci, allant de $r = 0.167$ pour HLA-B à $r = 0.468$ pour HLA-DQB1, mais était généralement plus réduite lorsqu'estimée à l'intérieur des régions géographiques, confirmant le rôle de la géographie dans la variabilité moléculaire des loci HLA à l'échelle mondiale.
Cette étude intégrait aussi des populations de Taiwan, d'Océanie et des populations natives d'Amérique et la corrélation entre les distances génétiques et les distances géographiques augmentait (à l'exception de HLA-DQA1 et -DQB1) lorsque ces populations étaient retirées de l'étude. Ce résultat peut s'expliquer par des effets démographiques très forts (dérive génétique) agissant sur ces populations. En effet, celles-ci montrent une importante baisse de diversité génétique, notamment jusqu'à deux fois plus d'homozygotes que dans les autres populations, et des valeurs de diversité nucléotidique plus réduites. Ces résultats sont dus à des distributions de fréquences alléliques particulières, montrant un ou deux allèles à hautes fréquences et un petit nombre d'allèles moins fréquents, dont la cause est probablement une dérive génétique rapide due à de petites tailles de populations.

En effet, bien que la sélection joue un rôle important, la diversité génétique est aussi influencée par les effets démographiques tels que les goulots d'étranglements et une dérive génétique rapide. En revanche, l'effet de la démographie (et de la géographie) n'est pas limité au HLA mais est visible sur l'ensemble du génome [Bamshad and Wooding, 2003]. Par exemple, pour les populations pastoralistes Peuls étudiées par Černý *et al.* en 2011, l'analyse de la diversité des génomes mitochondriaux a révélée une diversité limitée (peu d'haplotypes différents), qui a été interprétée comme le signal d'une contraction démographique du côté des lignées maternelles [Černý et al., 2011]. Ces effets démographiques s'observent majoritairement dans de petites populations, telles que les populations natives d'Amérique, de Taïwan ou des îles du Pacifique qui ont subi une dérive génétique rapide liée à leur histoire migratoire [Buhler et al., 2016], mais aussi chez des populations isolées d'autres régions, telles que les chasseurs-cueilleurs Aka d'Afrique tropicale chez qui cette dérive génétique rapide a induit une forte différenciation génétique vis-à-vis des autres populations d'Afrique sub-saharienne [Renquin et al., 2001].
Une étude menée en 2001 par Chu *et al.* [Chu et al., 2001] sur les populations natives de Taïwan a mis en évidence une très haute fréquence de l'allèle HLA-A*24:01 (jusqu'à 86% chez les Paiwan) associée à une hétérozygotie réduite (comparée aux autres populations du Pacifique et d'Asie [Bugawan et al., 1999]), probable résultat d'un effet fondateur et de plusieurs goulots d'étranglement, mais aussi peut-être d'une sélection positive pour cet allèle (d'autres marqueurs tels que HLA-DRB1 ou microsatellites [32] montrant en revanche

32. Séquences d'ADN formée d'une répétition de courts motifs nucléotidiques.

davantage de diversité).
Cette même étude a aussi identifié une diversité génétique réduite au locus HLA-DPB1 pour des populations mélanésiennes, certaines d'entre elles, comme les habitants des îles Trobriand, ne montrant que 5% d'hétérozygotie à ce locus, suggérant un fort impact d'un isolement géographique (et probablement culturel) sur la diversité des loci HLA.
Des travaux menés par Sanchez-Mazas *et al.* en 2017 sur les populations (semi-) nomades d'Afrique tels que les Bédouins Rashaida ou les Beja Hadendoa ont mis en évidence des fréquences particulièrement élevées de certains allèles (par exemple HLA-A*02 ou -B*50) associées à une diversité génétique réduite (plus petit nombre d'allèles et hétérozygotie réduite à ces loci) [Sanchez-Mazas et al., 2017]. Ces résultats ont été interprétés comme le résultat d'une dérive génétique rapide de ces populations, conséquence de leur petites tailles de populations et leurs flux géniques réduits avec les populations voisines.

Toutefois, une réduction de la diversité aux loci HLA n'implique pas forcément une diminution de la fonctionnalité de ces loci. Dans une étude parue en 2016 et comparant les répertoires de peptides présentés par les trois loci de classe I, Buhler *et al.* ont montré que les populations ayant une perte de diversité génétique (suite à de la dérive génétique rapide) à l'un ou l'autre des loci HLA ne montraient pas de diminution dans le répertoire de peptide présentés par leurs allèles si l'on considérait les trois loci de classe I conjointement [Buhler et al., 2016]. Ce résultat s'explique par un modèle de « sélection asymétrique conjointe divergente » [33] qui agirait sur l'ensemble des trois loci de classe I et compenserait une perte de diversité à un locus en particulier.

33. En anglais : *Joint Divergent Asymmetric Selection.*

5 Techniques de typage HLA

Une partie de ce travail de doctorat utilise et compare des résultats de typages HLA provenant de plusieurs techniques différentes (Chapitre 4). Nous présentons ici une description des principales méthodes habituellement utilisées afin d'expliquer les différents problèmes qui sont rencontrés lors de l'étude de loci hautement polymorphiques.
Ces typages se répartissent en deux grandes catégories : les typages basés sur la sérologie et les typages basés sur l'analyse de l'ADN.

5.1 Typages sérologiques

Les typages sérologiques reposent sur la détection des molécules HLA exprimées à la surface des cellules lymphocytaires.
Les lymphocytes de l'individu à tester sont incubés avec un ensemble d'anticorps anti-HLA dans une plaque de Terasaki contenant entre 60 et 72 puits. Chaque puits est rempli d'un anticorps spécifique à une sérologie HLA qui va se lier aux molécules HLA des lymphocytes contre lesquelles les anticorps sont dirigés. Des protéines du complément sont ajoutées (sérum de lapin) afin de déclencher une cascade d'activation amenant les lymphocytes liés à des anticorps à subir une lyse membranaire par le complexe d'attaque membranaire. Cette perméabilisation de la membrane cytoplasmique va permettre à un fluorochrome, le bromure d'éthidium, de pénétrer dans la cellule. Les cellules ainsi marquées deviennent visibles en microscopie, permettant d'identifier la sérologie HLA de ces dernières.

Cette méthode est principalement utilisée pour détecter les allèles HLA-A, -B et dans une plus petite mesure HLA-DR [Buhler, 2007] puisque ce sont les molécules HLA les plus abondamment exprimées à la surface des cellules. Les molécules HLA-C étant moins exprimées (d'un facteur 10 [Zemmour and Parham, 1992] et variable selon les allèles [Apps et al., 2013, Kaur et al., 2017]), les typages sérologiques conduisent régulièrement à des non-détections d'allèles [Turner et al., 1998]. Une étude menée en 1997 par Mytilineos *et al.* et comparant des typages sérologiques et des typages PCR-SSP (méthode décrite plus loin) pour HLA-C chez les mêmes individus a mis en évidence une différence de typages de l'ordre de 33%, principalement due à des allèles non détectés par sérologie [Mytilineos et al., 1997].
De plus, certaines spécificités sérologiques différentes appartiennent à des groupes dits « de réaction croisée », possédant des épitopes[34] similaires et donc détectés par les mêmes anticorps anti-HLA. Par exemple le groupe de réaction croisée 10C regroupe les spécificités *HLA-A*10, *11, *28, *32, *33, *43 et *74* [Wade et al., 2007].
D'autres molécules HLA telles que HLA-DQ2 (codées par les gènes HLA-DQA2 et -DQB2) sont spécifiquement exprimées par certains types de cellules (les cellules de Langerhans [Lenormand et al., 2012]) et ne sont donc pas détectées par cette méthode.

5.2 Typages nucléotidiques

La deuxième catégorie de typages HLA a été nommée ici typages nucléotidiques puisque reposant sur l'identification des séquences nucléotidiques codant pour les molé-

34. Partie d'un antigène qui peut être reconnue par la partie variable d'un anticorps.

cules HLA.

PCR-RFLP : PCR et polymorphisme de tailles de fragments de restriction

Cette méthode se base sur la technique de RFLP [35], où les produits de la PCR (les premières applications n'utilisaient pas de PCR mais uniquement de l'ADN génomique issu de cultures cellulaires [Bidwell et al., 1988]) sont digérés par des enzymes de restriction puis mis à migrer sur un gel d'électrophorèse. Cette méthode a été la première méthode non sérologique à être appliquée sur les gènes HLA, plus spécifiquement sur les gènes de classe II [Bidwell et al., 1988].

PCR-SSO : PCR et oligonucléotides séquences-spécifiques

Cette méthode repose sur une amplification des régions d'intérêt (exons 2, ainsi que 3 pour les gènes HLA de classe I) par PCR, suivie d'une immobilisation des produits de la PCR sur un support solide (typiquement une membrane de nylon) et une hybridation directe avec un ensemble de sondes moléculaires séquences-spécifiques marquées par un fluorochrome. Seules les sondes liées à leurs séquences spécifiques subsistent après lavage du support, le fluorochrome permettant de les détecter [Roosnek et al., 2015].

Cette technique a tout d'abord été appliquée aux loci de classe II [Bunce et al., 1997], avec les premiers typages de HLA-DQA1 en 1986 [Saiki et al., 1986], HLA-DRB1 en 1988 [Tiercy et al., 1988] et HLA-DQB1 en 1989 [Eliaou et al., 1989]. L'application aux gènes de classe I a été plus tardive, en raison d'un manque de séquences ADN complètes (afin de concevoir les amorces PCR et sondes oligonucléotidiques) disponibles pour ces gènes, avec un premier typage de HLA-B en 1991 [Allsopp et al., 1991], suivi de HLA-A en 1992 [Fernandez-Viña et al., 1992]. Il a fallu attendre deux ans pour que la technique se démocratise pour les gènes de classe I [Yoshida et al., 1992, Oh et al., 1993, Levine and Yang, 1994].

À l'exception des problèmes rencontrés au début des typages PCR-SSO des gènes de classe I (manque de séquences complètes), les principaux problèmes concernent les amorces PCR (leur robustesse et spécificité), le développement des sondes oligonucléotidiques (nécessitant de connaître la séquence ciblée par la sonde) et le temps de génotypage pour l'ensemble des loci. En effet, ce dernier est particulièrement long (>24h) pour obtenir le génotype HLA complet d'un individu (HLA-A, -B, -C, DRB1/3/4, -DQA1, -DQB1, -DPA1 et -DPB1) [Bunce et al., 1997], bien que des méthodes plus rapides telles que la « PCR-SSO inverse » (où les sondes sont fixées sur la membrane et les produits de PCR libres) aient été développées, cette dernière restant plus compliquée à mettre en œuvre de par le grand nombre de sondes requises pour un typage en parallèle des gènes de classes I et II.

PCR-SSP : PCR par amorces séquences-spécifiques

Cette méthode repose sur le principe de la spécificité des amorces PCR pour réaliser l'amplification d'un locus. Chaque région polymorphique est mise en présence d'un couple d'amorces spécifiques à un allèle. Plusieurs combinaisons de couples

35. *Restriction Fragment Length Polymorphism* : polymorphisme de tailles de fragments de restriction.

d'amorces sont testées et l'amplification de l'ADN (vérifiée par une migration sur gel d'électrophorèse et un marquage au bromure d'éthidium) permet l'identification de l'allèle.

L'un des avantages de cette méthode est la résolution de la phase, puisque cette méthode est capable de détecter les polymorphismes portés par un seul chromosome (*cis*) alors que la PCR-SSO détecte les polymorphismes sur les deux chromosomes (*cis* et *trans*) [Bunce et al., 1997].
La principale limite de cette méthode réside dans la résolution et le nombre d'amorces nécessaires, puisqu'un typage basse résolution de HLA-A, -B et -DR nécessite 72 amorces, tandis que ce nombre peut monter jusqu'à 400 pour un typage complet de HLA-A, -B, -C, -DRB1/3/4/5 et -DQB1 [Roosnek et al., 2015].

PCR-SSCP : PCR et polymorphisme conformationnel d'ADN simple brin

Cette méthode utilise le repliement spécifique des molécules d'ADN simple brin non dénaturées [Hayashi, 1992]. En effet, une seule mutation suffit à changer la conformation tri-dimensionnelle des molécules d'ADN et donc ses propriétés de migration dans un gel d'électrophorèse en conditions non dénaturantes. Cette méthode est toutefois plus compliquée à interpréter, puisque l'effet des mutations étant difficile à prédire, certaines mutations ne sont pas détectées.

PCR-SBT : PCR et typages basés sur les séquences

La dernière méthode présentée est une méthode qui a énormément évolué ces dernières années. Le principe de base de la PCR-SBT (*Sequence Based Typing*, Typages Basés sur les Séquences) repose sur une amplification des régions-cibles par PCR, un séquençage ADN des produits de PCR et une comparaison avec les séquences connues des bases de données.
Les premiers typages SBT utilisaient la technique de séquençage Sanger [Santamaria et al., 1993], mais cette méthode a rapidement rencontré plusieurs problèmes. Le premier concerne la conception des amorces PCR, qui doivent allier robustesse (c'est à dire amplifier seulement le locus ciblé et non des loci similaires, notamment sur des gènes paralogues) et spécificité (amplifier de manière similaire l'ensemble des allèles du locus), ces deux caractéristiques étant difficiles à assurer par l'important polymorphisme des loci HLA (en particulier les exons 2 et 3). L'autre problème concerne la réalisation de la phase : dans le cas du séquençage de plusieurs exons, tel que réalisé sur les gènes de classe I pour des individus hétérozygotes, il s'agit de déterminer si les séquences appartiennent au même chromosome (*cis*) ou sont portées par des chromosomes différents (*trans*). La résolution de ce problème implique alors d'utiliser des lignées cellulaires clonales (rendues artificiellement homozygotes) ou d'utiliser un traitement statistique basé sur des séquences déjà connues.

Avec l'arrivée des séquenceurs de nouvelle génération, permettant un séquençage à haut débit, mais aussi en parallèle d'un grand nombre de séquences (de plusieurs loci pour un grand nombre d'individus), cette technique est devenue la méthode de référence, notamment dans le domaine de la génétique des populations.
On a vu alors l'application de deux stratégies différentes de typage, la première reposant sur des séquençages d'amplicons des régions haute-

ment polymorphiques (exons 2 et 3) pour un grand nombre d'échantillons [Bentley et al., 2009, Gabriel et al., 2009, Holcomb et al., 2011, Moonsamy et al., 2013] et la deuxième reposant sur une PCR à longue portée amplifiant, pour quelques individus, l'ensemble d'un gène de classe I (cette stratégie restant difficilement applicable aux gènes de classe II à cause de leur plus grande taille et de la présence de longues régions composées d'éléments répétés), suivie d'une fragmentation, d'un séquençage *shotgun* et d'un assemblage [Lind et al., 2010].
L'évolution des capacités techniques et technologiques de ces séquenceurs permet de séquencer des segments de gènes plus longs (en dehors des exons 2 et 3), avec un taux d'erreur réduit (notamment en augmentant la profondeur de séquençage). Les séquenceurs commercialisés par PacificBioscience® ont comme particularité de pouvoir séquencer des fragments de gènes très longs (3'000 pb en moyenne et pouvant aller jusqu'à 14'000 pb [De Santis et al., 2013]), permettant de résoudre presque intégralement le problème de la phase [Mayor et al., 2014]. Ces séquenceurs ont toutefois comme limite leur taux d'erreurs très élevés en comparaison des autres technologies disponibles (14.1% d'erreur pour le PacBio-RS-C2, contre 0.32% pour le Illumina MiSeq [De Santis et al., 2013]), ce qui implique soit un très grand volume de séquençage (et l'hypothèse que les erreurs de séquençage sont vraiment aléatoires), soit de coupler des séquençages basés sur des lectures longues avec des séquençages de lectures courtes, les premiers servant de référence afin de réaliser l'alignement et la phase des seconds.

Une dernière méthode est à signaler, celle dite de l'imputation. Le développement des puces de séquençage (puces à SNP [36]) permettant d'obtenir un aperçu du génome complet s'est généralisé ces dernières années dans les études de génétique des populations. Ces puces à SNP ne couvrent toutefois pas les gènes HLA à cause de limitations techniques (haut taux de polymorphisme combiné avec la présence de SNP tri- ou tétra-nucléotidiques). L'imputation repose sur le principe de déséquilibre de liaison entre les SNP hors de la région HLA (mais proches) et les SNP à l'intérieur des gènes HLA (définissant les allèles). Ainsi, il est possible de prédire, dans une certaine mesure, les allèles d'un individu sur la base de SNP voisins (si l'on dispose d'un ensemble de séquences de référence, à haute résolution). L'essor des puces à SNP a conduit à un développement des recherches sur l'imputation HLA et plusieurs logiciels sont disponibles : MAGPrediction [Li et al., 2011], HLA*IMP:02 [Dilthey et al., 2013], HIBAG [Zheng et al., 2014] et GRIMM [Maiers et al., 2019]. En plus d'identifier les allèles HLA d'un individu, cette méthode permet de détecter la variabilité hors des gènes traditionnellement typés, mais aussi de comparer la contribution des gènes HLA à celle des gènes non-HLA quant à la variabilité phénotypique [Sanchez-Mazas and Meyer, 2014].
Cette méthode présente toutefois des limites, 1) dans la résolution qu'il est possible d'avoir, puisqu'elle fonctionne très bien pour définir des allèles au premier champ (>98% de concordance) mais beaucoup moins bien pour définir des allèles au second champ [Zheng et al., 2014] et 2) étant donné que les motifs de déséquilibre de liaison diffèrent selon les populations, il faut un ensemble de séquences de référence pour chaque population.

36. *Single Nucleotide Polymorphism* : Polymorphismes Mono-Nucléotidiques.

6 Méthodes utilisées dans ce travail

6.1 Hardy-Weinberg

Le test d'équilibre de Hardy-Weinberg est plus compliqué à mettre en œuvre pour HLA que pour d'autres loci, notamment à cause des ambiguïtés (lorsque la méthode de typage ne permet pas de différencier entre plusieurs allèles sur un même locus d'un même chromosome) ou, selon la méthode de typage, la présence d'allèles « blancs » (correspondant à des allèles potentiellement présents mais non observés).
Pour cela, nous avons adopté l'outil Gene[Rate] [Nunes, 2016]. Il s'agit d'une suite d'outils en ligne (`https://hla-net.eu/`) permettant notamment un test de l'équilibre de hardy-Weinberg ne reposant que sur un seul degré de liberté.
La technique implémentée dans Gene[Rate] consiste en un test du rapport de vraisemblances (LRT [37]) [Nunes, 2016]. Ce test compare la vraisemblance des fréquences sous l'hypothèse de l'équilibre de Hardy-Weinberg avec la vraisemblance des fréquences sous un modèle alternatif, incluant un coefficient de consanguinité. Ce dernier modèle peut s'écrire sous la forme :

$$L_0 = f(p_i, F) \tag{1.1}$$

ÉQUATION 1.1 – Équation de la vraisemblance L_0 des fréquences alléliques p_i sous l'hypothèse d'un modèle tenant compte d'un coefficient de consanguinité F.

Il s'agit donc d'un modèle de vraisemblance basé sur les distributions de fréquences alléliques et sur un paramètre F (allant de 0 à 1) mesurant la déviation par rapport au modèle d'équilibre de Hardy-Weinberg. Dans ce cas, le modèle assumant l'équilibre de Hardy-Weinberg (HWE) est un cas particulier du précédent, où $F = 0$:

$$L_{HWE} = f(p_i, F = 0) = f(p_i) \tag{1.2}$$

ÉQUATION 1.2 – Équation de la vraisemblance L_{HWE} des fréquences alléliques p_i sous l'hypothèse de l'équilibre de Hardy-Weinberg.

Il est ensuite possible de comparer ces deux valeurs de vraisemblance puisque le double de la différence entre les deux vraisemblances suit une distribution du χ^2 à un degré de liberté (puisqu'il n'y a qu'un seul paramètre de différence entre les deux modèles).

6.2 Estimation des fréquences alléliques ou haplotypiques

Deux méthodes existent pour estimer les fréquences alléliques d'une population à partir de la liste des génotypes individuels d'un échantillon de la population. La méthode du comptage (ou dénombrement) permet d'obtenir les fréquences alléliques de chacun des allèles dans le cas d'allèles codominants, en l'absence d'allèles blancs et sans ambiguïtés (il s'agit d'un cas particulier de l'algorithme EM décrit après). Dans le cas contraire, si un des allèles n'a pas pu être compté (ambiguïté de typage, allèle récessif ou impossible à typer), l'algorithme d'*Expectation Maximization* (EM) permet d'estimer les fréquences alléliques.

37. En anglais : *Likelihood Ratio Test.*

Méthode du comptage

La méthode du comptage est simple et est décrite dans l'équation 1.3 :

$$\begin{aligned} p_i &= \frac{2N_{ii} + \sum_{i<j}^{k} N_{ij}}{2N} \\ \sigma^2(p_i) &= \frac{p_i(1-p_i)}{2N} \end{aligned} \tag{1.3}$$

ÉQUATION 1.3 – Calcul de la fréquence p_i de l'allèle i et de la variance de cette fréquence, dans un système multi-allélique diploïde avec k allèles codominants. N_{ii} : nombre d'homozygotes pour l'allèle i ; N_{ij} : nombre d'hétérozygotes pour les allèles i et j ; N : nombre d'individus.

Algorithme *Expectation Maximization*

Dans le cas où le génotypage serait ambigu, il est impossible d'estimer la fréquence d'un allèle non recensé à l'aide de la méthode de comptage. Pour cela on peut appliquer un cas particulier de l'algorithme EM (dont le nom provient des deux étapes d'*expectation* et de *maximization* répétées de manière itérative [Dempster et al., 1977]). La version actuelle [Nunes, 2005] implémentée dans Gene[Rate] est une extension de la méthode de comptage proposée en 1955 par Ceppellini *et al.* [Ceppellini et al., 1955], par la suite étendue aux haplotypes en 1995 [Excoffier and Slatkin, 1995, Hawley and Kidd, 1995, Long et al., 1995].

Cet algorithme consiste à utiliser des valeurs provisoires des fréquences alléliques pour estimer les valeurs attendues des génotypes non observables directement (plusieurs génotypes étant regroupés sous un même phénotype), permettant alors l'estimation des fréquences alléliques par comptage des génotypes. Ces étapes sont répétées jusqu'à la convergence, c'est-à-dire jusqu'à ce que la variation de la vraisemblance soit plus faible qu'un seuil prédéfini.

L'algorithme commence avec un jeu arbitraire (ou aléatoire) de fréquences alléliques $\theta_0 = (p_1^0, \ldots, p_k^0)$.

L'**étape de prévision** (*expectation*) consiste à estimer les fréquences haplotypiques, à partir de ces fréquences alléliques sur la base des proportions de Hardy-Weinberg (l'équilibre de Hardy-Weinberg est donc une précondition de cet algorithme), pour calculer la probabilité d'assigner chaque phénotype observé (les *priors*) à un des génotypes possibles.

L'**étape de maximisation** (*maximization*) utilise ensuite ces fréquences génotypiques relatives pour ré-estimer les fréquences alléliques par comptage direct. Ces fréquences alléliques sont alors ré-utilisées dans l'étape de prévision jusqu'à la convergence.

Selon l'importance des ambiguïtés du jeu de données initial, plusieurs distributions de fréquences alléliques peuvent être obtenues selon le θ_0 initial. Il faut donc s'assurer qu'il n'y ait qu'une seule solution possible (c'est-à-dire que les mêmes résultats soient obtenus peu importe θ_0). Pour cela, l'algorithme est initialisé avec plusieurs θ_0 différents et le nombre de solutions trouvées est rapporté à la fin. Si une seule solution est rapportée, alors les estimations de fréquences alléliques sont fiables ; dans le cas contraire, il faut essayer d'utiliser un jeu de données plus grand ou de réduire la résolution des allèles HLA (afin de faire disparaître une partie des ambiguïtés) [Nunes, 2016].

6.3 Déséquilibres de liaison

Les tests de déséquilibre de liaison implémentés dans Gene[Rate] consistent en trois tests, un test de déséquilibre de liaison haplotypique et deux tests de déséquilibre de liaison global [Nunes, 2016].

Déséquilibre de liaison haplotypique

Pour un haplotype (formé de deux allèles), le déséquilibre de liaison est le fait que la fréquence de l'haplotype soit différente du simple produit des fréquences des deux allèles pris séparément (l'appariement entre les deux allèles n'est pas aléatoire).
Le test du déséquilibre de liaison haplotypique repose sur les résidus du test du χ^2, selon la formule suivante :

$$res = \frac{N_{obs} - N_{att}}{\sqrt{N_{att}}} \tag{1.4}$$

Équation 1.4 – Calcul des résidus (res) du test du χ^2 appliqué au test du déséquilibre de liaison haplotypique. N_{obs} : nombre de fois que l'haplotype est observé ; N_{att} : nombre de fois que l'haplotype est attendu (selon le produit des fréquences alléliques individuelles).

Ces résidus sont ensuite standardisés, en les divisant par leur variance :

$$stdres = \frac{res}{\sigma(RES)} \tag{1.5}$$

Équation 1.5 – Calcul des résidus standardisés pour le test de déséquilibre de liaison haplotypique. res est la valeur du résidus du χ^2 calculé pour un haplotype, RES est l'ensemble des résidus calculés pour chaque haplotype (utilisés pour estimer la variance de ces résidus) et $stdres$ représente la valeur du résidu standardisé.

La statistique $stdres$ obtenue s'interprète alors en la comparant à une distribution de loi normale centrée réduite, l'hypothèse nulle étantl'absence de déséquilibre de liaison (exemple donné au seuil $\alpha = 0.05$) :

- $stdres \geq 1.96$: déséquilibre de liaison **positif** et significatif, la fréquence observée de l'haplotype est supérieure à celle attendue par le simple produit des fréquences alléliques ;
- $-1.96 < stdres < 1.96$: la fréquence observée de l'haplotype est similaire à celle attendue par le produit des fréquences alléliques, il n'y a pas de déséquilibre de liaison ;
- $stdres \leq -1.96$: déséquilibre de liaison **négatif** et significatif, la fréquence observée de l'haplotype est inférieure à celle attendue par le simple produit des fréquences alléliques.

Déséquilibre de liaison global

Le déséquilibre de liaison global survient lorsque les deux loci ne sont pas indépendants au niveau de leur transmission d'une génération à une autre.
Dans le cas simple d'un système de deux loci bialléliques, le déséquilibre de liaison global est directement estimé par le test de déséquilibre haplotypique. Mais, lorsqu'il y a plus d'haplotypes à tester, l'utilisation de ce test de déséquilibre haplotypique nécessite alors des corrections pour tests multiples. Toutefois, ces corrections pour tests multiples reposent sur l'hypothèse d'une indépendance des données, ce qui n'est pas le cas dans le cadre d'un test de déséquilibre de liaison.
Deux tests ont alors été développés pour tester le déséquilibre de liaison global, sans avoir recours à la comparaison de l'ensemble des haplotypes (et donc aux corrections pour tests multiples). Le premier, le test paramétrique, repose sur l'utilisation d'un rapport de vraisemblance et le deuxième, le test non-paramétrique, consiste en une procédure de *bootstrap* appliquée au résultat du test paramétrique. Dans les deux cas l'hypothèse nulle qui est testée est l'absence de déséquilibre de liaison global ($H_0 : DL_g = 0$).

Test paramétrique de déséquilibre de liaison global

Le test paramétrique de déséquilibre de liaison global se base sur un test du rapport de vraisemblances (LRT). Ce test compare la vraisemblance d'un modèle basé sur les fréquences haplotypiques avec la vraisemblance d'un modèle basé sur le produit des fréquences alléliques (posant comme précondition l'équilibre de Hardy-Weinberg) :

$$LRT = 2 \cdot ln \frac{L(FH)}{L(PFA)} \tag{1.6}$$

ÉQUATION 1.6 – Calcul du rapport de vraisemblance LRT pour le test paramétrique de déséquilibre de liaison global. FH : Fréquences Haplotypiques, PFA : Produit des Fréquences Alléliques.

La valeur de LRT obtenue est alors comparée à une distribution de χ^2 avec autant de degrés de liberté que d'haplotypes. Ainsi, en cas de grand nombre d'haplotypes testés (ce qui est généralement le cas avec HLA), le test perd en puissance (principalement des erreurs de type II).

Test non-paramétrique de déséquilibre de liaison global

Le test non-paramétrique de déséquilibre de liaison global est une extension du premier (le test paramétrique) visant à résoudre le problème de la perte de puissance statistique en cas de grand nombre d'haplotypes.

Il s'agit d'un ré-échantillonnage aléatoire d'un grand nombre d'échantillons, basé sur les fréquences alléliques observées (l'équilibre de Hardy-Weinberg est donc une précondition de ce test), pour lesquels la statistique LRT sera aussi calculée. Ensuite, la statistique obtenue à l'aide de la méthode paramétrique est située par rapport aux statistiques obtenues par ré-échantillonnage et le test étant unilatéral, l'hypothèse nulle de l'absence de déséquilibre de liaison global est rejeté si la LRT est supérieure à 95% des LRT simulées.

6.4 Indices classiques

Les indices de diversité qui suivent sont des indices communément utilisés dans les analyses de génétique HLA des populations [Currat et al., 2010, Di and Sanchez-Mazas, 2014, Inotai et al., 2015, Souza et al., 2020]. Ils ont été choisis afin d'obtenir des mesures statistiques représentatives de la diversité allélique (hétérozygoties et richesse allélique) ou moléculaire (diversité nucléotidique et nombre de sites polymorphiques).

Hétérozygoties

L'**hétérozygotie observée** (H_o pour *Heterozygosity observed*) est la proportion d'individus hétérozygotes à un locus dans un échantillon de population.
Pour des individus diploïdes, l'**hétérozygotie attendue** est équivalente à la diversité génique [Nei, 1987]. L'hétérozygotie attendue (H_e pour *Heterozygosity expected*) se calcule à partir de la somme des carrés des fréquences alléliques (indiquant donc comme précondition l'équilibre de Hardy-Weinberg) :

$$H_e = \frac{n}{n-1}(1 - \sum_{i=1}^{k} p_i^2) \tag{1.7}$$

ÉQUATION 1.7 – Équation du calcul de l'hétérozygotie attendue H_e d'un échantillon. k : nombre d'allèles ; n : taille d'échantillon donnée par le nombre de copies du gène (2 copies par individu diploïde) ; p_i : fréquence de l'allèle i. Le coefficient $\frac{n}{n-1}$ permet de centrer l'estimateur H_e, qui est biaisé.

Nombre moyen de différences par paires de séquences et diversité nucléotidique

Le nombre moyen de différences par paires de séquences π d'un échantillon de population correspond à la distance moléculaire moyenne entre chacune des paires de séquences dans l'échantillon. On l'estime par :

$$\begin{aligned} \pi &= \frac{n}{n-1} \sum_{i=1}^{k} \sum_{j<i} p_i p_j d_{ij} \\ \sigma^2(\pi) &= \frac{3n(n+1)\pi + 2(n^2+n+3)\pi^2}{11(n^2-7n+6)} \end{aligned} \tag{1.8}$$

ÉQUATION 1.8 – Calcul du nombre moyen de différences par paires de séquences et de sa variance [Tajima, 1993]. n : taille d'échantillon donnée par le nombre de copies du gène (2 copies par individu diploïde) ; k : nombre d'allèles ; d_{ij} : distance entre les séquences i et j (en l'absence de correction particulière il s'agit de la distance de Hamming, c'est-à-dire le nombre de bases différentes entre les deux séquences) ; p_i, p_j : fréquences des allèles i et j.

La diversité nucléotidique correspond à la probabilité que deux sites nucléotidiques homologues soient différents. C'est l'équivalent de la diversité génique pour des données moléculaires.

$$\begin{aligned} \pi_n &= \frac{\pi}{L} \\ \sigma^2(\pi_n) &= \frac{n+1}{3(n-1)L}\pi_n + \frac{2(n^2+n+3)}{9n(n-1)}\pi_n^2 \end{aligned} \tag{1.9}$$

ÉQUATION 1.9 – Calcul de la diversité nucléotidique et de sa variance [Tajima, 1993]. π : nombre moyen de différences par paires de séquences (voir Équation (1.8)) ; L : nombre de sites nucléotidiques comparés ; n : taille d'échantillon donnée par le nombre de copies du gène (2 copies par individu diploïde).

Nombre de sites polymorphiques S

Le nombre de sites polymorphiques S est le nombre de sites présentant un polymorphisme entre deux (ou plus de deux) séquences alignées.

Richesse allélique

Le nombre d'allèles (maximum) observable à un locus donné dans un échantillon de n individus diploïdes est de $2n$. Ainsi, plus la taille d'échantillon est importante, plus la probabilité d'observer un allèle rare est importante. Au contraire, une petite taille d'échantillon diminue cette probabilité. Afin de pouvoir comparer le nombre d'allèles entre deux (ou plus de deux) échantillons de tailles différentes, la richesse allélique fournit un estimateur comparable entre des échantillons de différentes tailles.
Il s'agit pour cela d'estimer le nombre d'allèles qui seraient observés dans un échantillon de taille plus petite (habituellement la plus petite taille d'échantillon observée parmi tous les échantillons à comparer). Pour cela, en 1996, El Mousadik et Petit [El Mousadik and Petit, 1996] ont proposé une méthode reposant sur le principe de raréfaction défini par Hurlbert en 1971 [Hurlbert, 1971]. Cette méthode utilise les formules suivantes :

$$\begin{aligned} r_i(g) &= \sum_{k=1}^{K}(1 - P_{ik}) \\ P_{ik} &= \frac{\binom{N_i - N_k}{g}}{\binom{N_i}{g}} \end{aligned} \tag{1.10}$$

ÉQUATION 1.10 – Calcul de la richesse allélique r_i dans un échantillon de taille g. N_i : taille d'échantillon initiale, N_k le nombre de copies de l'allèle k.

La formule permettant de calculer cette richesse allélique est donnée par le nombre de copies du gène dans la population (2 par individu diploïde) N_i, le nombre de copies de l'allèle K N_k selon la formule (équation 1.10) intégrant g, où $g < N_i$ est la taille d'échantillon pour lequel on désire calculer la richesse allélique r_i. P_{ik} représente la probabilité (estimée grâce à la loi binomiale) que l'allèle k ne soit pas observé dans un échantillon de taille g.

6.5 Tests de neutralité sélective

De la même façon que pour les indices de diversité, les tests présentés ici sont communément utilisés dans les analyses de génétique HLA des populations. Ils permettent de détecter les événements dits « micro-évolutifs » [Vitti et al., 2013], c'est-à-dire datant d'avant la spéciation. Ils reposent sur une analyse de la diversité génétique (une réduction, par exemple, en cas de balayage sélectif) et permettent de détecter des évènements datant de moins de 250'000 ans [Sabeti, 2006].

Test d'Ewen-Watterson-Slatkin

Le test de neutralité sélective implémenté dans Gene[Rate] repose sur une version du test d'Ewen-Watterson-Slatkin (EWS) [Slatkin, 1994, Slatkin, 1996]. Toutefois, le test de EWS ne permet pas de s'accommoder des ambiguïtés de typage. Pour tenir compte de ce problème, le test de BEWS [38] implémenté dans Gene[Rate] introduit une génération aléatoire d'échantillons de génotypes, basée sur la distribution de fréquences alléliques estimée (l'équilibre de Hardy-Weinberg est donc une précondition de ce test) [Nunes, 2016]. Le test de EWS est alors appliqué à chacun de ces échantillons générés aléatoirement, générant un ensemble de pValeurs qui sont corrigées par la méthode *False Discovery Rate* [Benjamini and Hochberg, 1995].

$$pVal^i_{adj} = pVal^i \times (k/i) \tag{1.11}$$

ÉQUATION 1.11 – Équation du calcul d'une pValeur ajustée ($pVal_{adj}$) selon la méthode de [Benjamini and Hochberg, 1995]. i : rang de la pValeur (triées de manière croissante), k : nombre total de tests réalisés.

Le résultat de ce test est alors un intervalle de pValeurs. Le test étant bilatéral, au seuil $\alpha = 0.05$, l'hypothèse nulle de neutralité sélective est rejetée si : 1) la borne la plus basse de l'intervalle est inférieure à 0.025, dans ce cas il y a un rejet de H_0 en faveur d'un excès d'hétérozygotes, ou 2) la borne la plus haute de l'intervalle est supérieure à 0.975, dans ce cas il y a un rejet de H_0 en faveur d'un excès d'homozygotes.

Test du D de Tajima

Le test du D de Tajima, reposant sur le modèle des sites infinis, permet de tester si les séquences ADN d'une population évoluent de manière neutre [Tajima, 1989b]. L'hypothèse nulle H_0 de ce test est donc l'absence de forces évolutives et l'équilibre démographique pour la population.

Le calcul du D de Tajima repose sur le paramètre de mutation θ ($\theta = 2M\mu$, où μ est le taux de mutation et M un paramètre de taille de population [39]), estimé de deux façons différentes, se basant soit sur le nombre de sites polymorphiques S (θ_S), soit sur le nombre moyen de différences par paires de séquences θ_π :

38. En anglais : *Bootstrapped Ewen-Watterson-Slatkin.*
39. 2N pour les populations d'individus diploïdes, N pour les populations d'individus haploïdes.

$$\begin{aligned} \theta_S &= \frac{S}{\sum_{i=0}^{n-1} 1/i} \\ \theta_\pi &= \pi \end{aligned} \tag{1.12}$$

ÉQUATION 1.12 – Calculs des deux estimateurs du paramètre de mutation θ utilisés dans le calcul du D de Tajima. S : nombre de sites polymorphiques ; n : taille d'échantillon ; π : nombre moyen de différences par paires de séquences.

La valeur du D est calculée avec la formule :

$$D = \frac{\theta_\pi - \theta_S}{\sqrt{\sigma^2(\theta_\pi - \theta_S)}} \tag{1.13}$$

ÉQUATION 1.13 – Équation permettant de calculer le D de Tajima selon les deux estimateurs du paramètre de mutation calculés avec les équations (1.12).

Dans l'implémentation de Arlequin v3.5 [Excoffier and Lischer, 2010], la significativité du D_{obs} de Tajima obtenu est calculée en générant un grand nombre (>1'000) d'échantillons aléatoires (à partir du jeu de données initial), sans sélection et à l'équilibre démographique, en calculant à chaque fois une valeur du D_{sim} (valeur du D simulée). La pValeur du test correspond alors à la proportion de $D_{sim} < D_{obs}$. Le test étant bilatéral, au seuil $\alpha = 0.05$ l'hypothèse nulle de neutralité/équilibre démographique est rejetée si la pValeur est inférieure à 0.025 (D_{obs} significativement inférieur à 0) ou supérieure à 0.975 (D_{obs} significativement supérieur à 0).
L'interprétation de la valeur du D de Tajima dépend ensuite de la significativité et de la valeur de ce dernier :

- **D>0** : $\theta_\pi > \theta_S$, plus d'haplotypes que de sites polymorphes. Déficit en allèles rares, s'explique par une contraction démographique OU de la sélection balancée ;
- **D=0** : $\theta_\pi \simeq \theta_S$, chaque site polymorphique correspond à un nouveau variant. Pas de traces de sélection, la population évolue de manière neutre (par mutations et dérive) ;
- **D<0** : $\theta_\pi < \theta_S$, moins d'haplotypes que de sites polymorphes. Excès d'allèles rares. Balayage sélectif (ou déséquilibre de liaison avec un gène affecté par le balayage sélectif) OU expansion démographique.

Le test du D de Tajima peut ainsi présenter des difficultés d'interprétation, puisque les effets de la démographie et de la sélection peuvent être confondus. De plus, Aris-Brosou et Excoffier, en 1996, ont mis en évidence, par simulations informatiques, que l'observation d'un $D < 0$, lié à une expansion démographique, n'était visible qu'à partir d'une de variation d'au moins 100 fois de la taille de la population [Aris-Brosou and Excoffier, 1996]. Cette même étude a aussi mis en évidence qu'un taux de mutation hétérogène au sein des séquences (certains sites accumulant plus rapidement les mutations que d'autres) peut augmenter à lui seul la valeur du D.

Test du rapport dN/dS

Le dernier test de sélection utilisé dans ce travail est celui du rapport dN/dS, tel qu'implémenté dans MEGA7 [Kumar et al., 2016].
Le test du $\omega = dN/dS$ (aussi appelé test du Ka/Ks) est un test de neutralité sélective reposant sur le nombre de mutations synonymes dS (c'est-à-dire mutations ponctuelles n'induisant pas un changement d'acide aminé dans la protéine [40]) et le nombre de mutations non-synonymes (mutations ponctuelles induisant un changement d'acide-aminé). L'hypothèse nulle H_0 de ce test est la neutralité sélective, c'est-à-dire que le nombre de mutations synonymes et non-synonymes, dans les séquences testées, sont identiques ($dN = dS$, $\omega = 1$). La significativité est testée à l'aide du paramètre Z, donné par la formule suivante :

$$Z = \frac{(dN - dS)}{\sqrt{(\sigma^2(dN) + \sigma^2(dS))}} \tag{1.14}$$

Équation 1.14 – Équation permettant le calcul de la statistique Z estimant la significativité du rapport dN/dS. dN et dS sont, respectivement, les nombres de mutations non-synonymes et synonymes observées dans les séquences ; $Var(dN)$ et $Var(dS)$ sont les variances correspondantes.

L'équation 1.14 permet de calculer la statistique Z pour deux séquences. Dans le cas de plus de deux séquences, dN et dS sont les moyennes observées du nombre de mutations non-synonymes et synonymes, la variance étant quant à elle estimée par ré-échantillonnage aléatoire des codons (décrit à la page 55 de [Nei and Kumar, 2000]).

La pValeur de ω s'obtient en comparant la statistique Z à une loi normale centrée réduite $\mathcal{N}(0, 1)$ et s'interprète comme un test bilatéral (exemple donné pour un seuil $\alpha = 0.05$) :

- si la pValeur est inférieure à 0.025 : $\omega < 1$ et $dN < dS$, il y a moins de mutations non-synonymes que de mutations synonymes, c'est un signal de sélection purificatrice (les nouvelles formes de protéines sont éliminées) ;
- si la pValeur est supérieure à 0.975 : $\omega > 1$ et $dN > dS$, il y a plus de mutations non-synonymes que de mutations synonymes, c'est un signal de sélection positive (les nouvelles formes de protéines sont favorisées).

Toutefois, le test du rapport dN/dS a été initialement développé pour l'analyse de séquences provenant d'espèces différentes [Kimura, 1977, Muse and Gaut, 1994, Goldman and Yang, 1994] puisque ses fondements théoriques ont été développés pour l'analyse d'espèces indépendantes et divergentes [Nielsen and Yang, 2003]. Ainsi, si les séquences proviennent d'une même espèce, le rapport dN/dS n'est pas une g†fonction monotone du coefficient de sélection et des rapports $dN/dS < 1$ peuvent être observés aussi bien sous un régime de sélection négative que positive [Kryazhimskiy and Plotkin, 2008]. En conclusion, le test de rapport dN/dS est présenté ici car il a été employé dans la publication associée au Chapitre 4 [Goeury et al., 2018a], mais ce travail étant réalisé avec des séquences provenant d'une même espèce, les résultats de ce test sont sujets à caution.

40. Grâce à la redondance du code génétique.

6.6 Distances génétiques et analyses d'échelonnement multidimensionnel

Distance de Reynolds

La distance de Reynolds (ou coefficientsec :c1-methodes de co-ancestralité) est une distance évolutive entre deux groupes (par exemple des populations) [Reynolds et al., 1983]. Elle prend en compte la possibilité de l'identité par ascendance, c'est-à-dire que deux allèles d'un même gène proviennent d'un même allèle ancestral.
Il s'agit d'une mesure basée sur l'indice de fixation F_{ST}. Il a été montré que ce dernier pouvait être estimé d'après la taille de la population (supposée constante et suffisamment grande) et le nombre de générations depuis la divergence entre les deux populations, sous conditions que la seule force évolutive soit la dérive génétique et que la population soit à l'équilibre de Hardy-Weinberg [Reynolds et al., 1983]. Cela permet alors de calculer le coefficient de coancestralité entre deux populations de taille N ayant divergé il y a t générations :

$$\begin{aligned} F_{ST} &= 1 - (1 - \frac{1}{2N})^t = 1 - \exp \frac{t}{2N} \\ \Theta_w &= -ln(1 - F_{ST}) \simeq \frac{t}{2N} \end{aligned} \tag{1.15}$$

ÉQUATION 1.15 – Équations permettant : le calcul de l'indice de fixation F_{ST} entre deux populations de taille N ayant divergé il y a t générations et le calcul de la distance de Reynolds (coefficient de co-ancestralité) Θ_w entre ces deux populations.

La différence entre les F_{ST} (ou Φ_{ST} dans le cadre de données moléculaires) et la distance de Reynolds (Θ_w), réside dans le fait que la distance de Reynolds repose sur des hypothèses biologiques (divergence des populations due à la mutation et dérive génétique) et non une simple distance géométrique (telle que les F_{ST} et Φ_{ST}).

Analyses d'échelonnement multidimensionnel

L'analyse d'échelonnement multidimensionnel[41] (MDS) [Kruskal, 1964a, Kruskal, 1964b] est une méthode visant à projeter dans un espace de dimensions réduites (typiquement deux ou trois) un ensemble de points défini dans un espace de plus grandes dimensions, tout en conservant les proximités (les paires de points proches dans les observations doivent le rester dans la MDS).

En génétique des populations, les MDS sont utilisées pour représenter les relations entre les populations à partir d'une matrice triangulaire de distances métriques entre ces populations. Une distance métrique (d^*_{ij}) correspond à une distance géométrique entre deux points (i et j) dans un espace euclidien et doit respecter trois conditions :

1. Elle ne peut pas être négative ($d_{ij} > 0$ et $d_{ii} = 0$) ;
2. Elle doit être symétrique ($d_{ij} = d_{ji}$) ;
3. Elle doit respecter l'inégalité triangulaire ($d_{ij} + d_{jk} \geq d_{ik}$).

41. En anglais : *Multidimensional Scaling Analysis*, MDS.

La MDS fonctionne en cherchant à maximiser une fonction de stress, qui est une mesure de la différence des distances entre les points dans l'espace initial de hautes dimensions et des distances dans l'espace de plus faibles dimensions :

$$stress = \sqrt{\frac{\sum_{i,j=1}^{i=k,j=l}(d^*_{ij} - d_{ij})^2}{\sum_{i,j=1}^{i=k,j=l}(d^*_{ij})^2}} \tag{1.16}$$

Équation 1.16 – Équation du calcul du stress dans une MDS. d^*_{ij} représente les distances entre les points i et j dans l'espace initial (en hautes dimensions) et d_{ij} les distances entre i et j dans l'espace en plus petites dimensions.

La plus petite valeur de stress atteignable (minimum 0) détermine la meilleure projection que peut prendre cette MDS et sert ensuite à évaluer la qualité de la projection selon une règle empirique [Kruskal, 1964b] :

— < 0.05 : excellente ;
— < 0.10 : satisfaisante ;
— < 0.20 : faible (configuration de la MDS à considérer avec prudence) ;
— > 0.20 : mauvaise (configuration de la MDS ne reflétant pas la distribution initiale).

Les analyses d'échelonnement multidimensionnel ont été réalisées à l'aide de la librairie `VEGAN` [Oksanen et al., 2019] pour R [R Core Team, 2020].

Analyse de Variance Moléculaire - AMOVA

L'analyse de variance moléculaire AMOVA [42] est une adaptation de l'ANOVA (analyse de variance) [Cockerham, 1969, Cockerham, 1973, Weir and Cockerham, 1984] tenant compte de la distance moléculaire entre les haplotypes [Excoffier et al., 1992]. C'est une analyse qui estime les différentes composantes de la variance moléculaire selon une variable catégorielle particulière (par exemple, le mode de vie ou l'appartenance à un groupe linguistique).
Se basant sur différents groupes de populations, l'AMOVA est une analyse hiérarchique du partitionnement de la variance totale entre les différentes composantes de covariance [Rousset, 2008]. Ces covariances sont ensuite utilisées pour calculer les indices de fixation définis par Wright [Wright, 1949, Wright, 1965] et par Slatkin [Slatkin, 1991] :

— F_{ST} : la part de variabilité due à des différences entre les populations ;
— F_{SC} : la part de variabilité due à des différences entre les populations à l'intérieur des groupes ;
— F_{CT} : la part de variabilité due à des différences entre les différents groupes.

Dans l'implémentation de Arlequin v3.5 [Excoffier and Lischer, 2010] la significativité des différents indices est estimée par des permutations des haplotypes, individus ou populations entre (respectivement) les individus, les populations ou les groupes de populations.

42. En anglais : *Analysis of molecular variance.*

6.7 Test de Mantel

Le test de Mantel [Mantel, 1967] teste la corrélation entre deux matrices de mêmes dimensions, habituellement (dans un cadre de génétique des populations) une matrice de distances génétiques et une matrice de distances géographiques. Dans le cadre de ce travail, c'est le test de Mantel tel qu'implémenté dans `R` par le package [R Core Team, 2020] `ADE4` [Dray and Dufour, 2007, Bougeard and Dray, 2018, Chessel et al., 2004, Dray et al., 2007] qui a été utilisé.
Le test de Mantel implémenté dans `ade4` consiste en un calcul de la corrélation (r de Pearson) entre deux matrices carrées[43]. La significativité est testée en localisant le r_{obs} parmi un grand nombre de r_{sim} obtenus par permutations des lignes et colonnes d'une des matrices et le calcul d'un coefficient de corrélation.

6.8 Autres analyses statistiques

Analyse en Composantes Principales & Analyse Factorielle des Correspondances

L'Analyse en Composantes Principales (ACP) dérive des travaux de Pearson [Pearson, 1901] et a été formalisée en tant que telle par Hotelling en 1933 [Hotelling, 1933]. C'est une méthode qui permet de transformer des variables corrélées en nouvelles variables décorrélées les unes des autres : les composantes principales (ou axes). Il s'agit d'une transformation à la fois géométrique (projection dans un nouvel espace) et statistique (répartition inégale de la variance sur les différentes composantes principales).
Au contraire de la MDS, il n'y a pas de perte d'information. Toute l'information contenue dans les données initiales est conservée. Elle est juste redistribuée sur plusieurs composantes principales et concentrée sur les premiers axes, ce qui permet aussi de débruiter les données en retirant les derniers axes les moins informatifs.
L'ACP étant très sensible à la variance des variables utilisées, il est possible de réduire les variables (toutes les variables ont alors une variance de 1), permettant alors d'éviter qu'une variable à forte variance influence énormément la projection (par exemple si les ordres de grandeur sont différents). L'effet contraire peut aussi survenir lors de la réduction des variables en augmentant la variance d'une variable qui présentait une faible variance, augmentant ainsi de manière artificielle l'importance de cette variable. De la même manière, il est possible de centrer (toutes les variables ont une moyenne de 0) afin de pouvoir comparer des variables indépendamment de leurs moyennes.

L'autre analyse présentée est l'Analyse Factorielle des Correspondances (AFC) [Benzécri, 1973, Greenacre, 1984]. De manière similaire à l'ACP, il s'agit de transformer un jeu de données pour décorréler les différentes variables et concentrer la variance sur les premières composantes principales. La principale différence est que les lignes et les colonnes du tableau de données initial jouent un même rôle (alors que la distinction est nécessaire pour l'ACP), chacune des variables étant elle aussi représentée par un point projeté dans le nouveau référentiel, permettant d'analyser conjointement les observations et les variables.

Ces deux analyses sont utilisées dans ce travail via l'implémentation `ADE-4` [Dray and Dufour, 2007, Bougeard and Dray, 2018, Chessel et al., 2004,

43. Puisqu'étant, pour N populations, des matrices de distances de taille $N \cdot N$.

Dray et al., 2007] pour `R` [R Core Team, 2020].

t-SNE

La t-SNE (*t-Distributed Stochastic Neighbors Embedding*) [van der Maaten and Hinton, 2008] est une méthode de réduction de dimensionnalité non-linéaire, dans laquelle les observations similaires (proches) dans l'espace à hautes dimensions, restent proches lors de la projection sur un espace de plus faibles dimensions. Elle a été développée par van der Maaten et Hinton en 2008 en se basant sur la SNE (*Stochastic Neighbors Embedding*) [Hinton and Roweis, 2002].

La première étape de l'algorithme t-SNE consiste à construire, à partir des distances euclidiennes des points en hautes dimensions (espace X), une distribution de probabilités entre chaque paire d'observations. C'est-à-dire la probabilité conditionnelle qu'un point x_i sélectionne comme point voisin x_j si les voisins sont choisis en proportion de leur densité de probabilité dans une distribution Gaussienne centrée sur x_i.
Les points sont ensuite initialisés aléatoirement sur l'espace de plus faibles dimensions (espace Y) et leurs densités de probabilité sont aussi calculées.

$$\begin{aligned} p_{j|i} &= \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \\ q_{i|j} &= \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k\neq i}(1+\|y_i - y_k\|^2)^{-1}} \end{aligned} \tag{1.17}$$

Équation 1.17 – Équations permettant le calcul de la densité de probabilité conditionnelle de deux points x_i et x_j en hautes dimensions ($p_{j|i}$) et de leurs projections y_i et y_j en faibles dimensions ($q_{i|j}$).

La différence entre la SNE et la t-SNE résulte de l'utilisation d'une distribution de Student à un degré de liberté (au lieu d'une distribution Gaussienne) dans le calcul de $q_{i|j}$ permettant d'accentuer, dans la projection en faibles dimensions, les différences entre les points modérément différents en hautes dimensions.

Si la projection y_i et y_j reflète correctement la similarité entre x_i et x_j alors $p_{j|i}$ et $q_{j|i}$ sont égaux. Le but de la t-SNE est alors de trouver une représentation qui minimisera la différence entre $p_{j|i}$ et $q_{j|i}$. Pour cela, l'algorithme t-SNE utilise la divergence de Kullback-Leibler (mesure de différenciation usuelle pour les distributions de probabilités, similaire à l'entropie, voir équation 1.18) et cherche à minimiser cette divergence pour tous les points en utilisant la méthode du *Gradient Descent* sur une fonction de coût :

$$
\begin{aligned}
C &= \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}} \\
\frac{\delta C}{\delta y_i} &= 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}
\end{aligned}
\tag{1.18}
$$

ÉQUATION 1.18 – Équation de la fonction de coût (divergence de Kullback-Leibler) à optimiser par *Gradient Descent*. P_i et Q_i représentent la distribution conditionnelle de x_i et y_i par rapport à l'ensemble des autres points.

Le dernier paramètre, et parmi les plus importants à définir pour une t-SNE, concerne le facteur de perplexité. Ce facteur de perplexité va définir la variance σ_i de la Gaussienne définie par l'équation (1.17). Deux valeurs différentes pour σ_i vont définir deux distributions de probabilités P_i différentes et donc deux projections différentes. La perplexité est définie comme :

$$
Perp(Pi) = 2^{H(Pi)} = 2^{-\sum_j p_{j|i} log_2 p_{j|i}} \tag{1.19}
$$

ÉQUATION 1.19 – Équation de la relation entre le facteur de perplexité $Perp$ et l'entropie de Shannon H de la distribution de probabilités Pi.

La perplexité est donc un paramètre de la t-SNE, à spécifier avant d'exécuter l'algorithme, qui va définir la variance σ_i de la Gaussienne dans l'équation (1.17) et donc en d'autres termes le nombre de points qui seront considérés comme voisins de x_i. Typiquement définie entre 5 et 50 [van der Maaten and Hinton, 2008], une faible valeur de perplexité va faire apparaître des structures très locales du jeu de données (chaque point ayant « peu » de voisins) tandis qu'une perplexité plus élevée va faire apparaître des structures plus globales (chaque point ayant « beaucoup » de voisins).

Le dernier problème qui se pose avec l'utilisation de l'algorithme t-SNE est le problème d'un minimum local dans la fonction d'optimisation, causant une projection qui, bien qu'étant dans une position stable, ne représente pas la projection idéale. Pour palier à ce problème, l'algorithme est initialisé plusieurs fois avec des états initiaux différents (la projection aléatoire initiale des points dans l'espace de faibles dimensions) et seule la meilleure projection, sur la base de la plus petite valeur de divergence de Kullback-Leibler est conservée.

7 Buts de ce travail

Les questions de ce travail de doctorat concernent les facteurs évolutifs qui déterminent les profils génétiques HLA, à l'échelle moléculaire, des populations humaines au travers de trois aspects. Le premier aspect concerne la génération de la diversité HLA au sein des populations, conduisant, par exemple, à l'apparition de nouveaux allèles par recombinaison intragénique ou mutation ponctuelle, ou à la création de déséquilibre de liaison entre les différents gènes. Le second aspect porte sur la façon dont cette diversité a été distribuée et a évolué au sein des différentes populations humaines. Au cours de l'histoire du peuplement et de ses mécanismes (flux génique et dérive génétique dus, respectivement, aux migrations et à l'isolement des populations), mais aussi par sélection naturelle, reflétant les adaptations des populations à leurs environnements. Le troisième aspect s'intéresse à la génomique comparative entre l'humain et le chimpanzé, plus spécifiquement aux mécanismes qui ont pu différencier génétiquement ces deux espèces au niveau de leurs gènes MHC sur une échelle de temps plus longue, conduisant, par exemple, à l'apparition de nouveaux gènes.

Ce travail est divisé en sept chapitres.
Le Chapitre 1 (présent chapitre d'introduction) présente l'état des connaissances actuelles sur l'utilisation du système HLA en génétique des populations, ainsi que les différentes méthodes existantes pour étudier ce polymorphisme.
Le Chapitre 2 présente les résultats d'une étude menée sur deux populations d'origines très différentes, les Mandenkalu [44] du Sénégal et les Cham du Vietnam, toutes deux séquencées pour les loci de classe I et II à haute résolution. Cette étude permet d'apporter des éléments de réponse quant aux mécanismes de génération de nouveaux variants HLA ainsi qu'à l'impact d'événements démographiques et de sélection sur les profils HLA de ces populations. Les Mandenkalu du Sénégal peuvent être considérés comme une population de référence pour l'Afrique sub-saharienne, car elle est étudiée depuis plusieurs décennies d'abord au êdépartement puis à l'Unité d'anthropologie de l'Université de Genève. Plusieurs études de génétique des populations ont aussi été menées sur les gènes HLA de cette population, par l'utilisation de diverses méthodes de typages appliquées aux mêmes échantillons, et dont les résultats seront comparés entre eux afin d'estimer quels auront été les apports de 25 ans d'évolution de techniques de typage HLA. Cette comparaison permettra d'examiner les ambiguïtés de typage générées par les méthodes précédemment utilisées et de déterminer les bénéfices apportés par les nouvelles technologies de typage, basées sur les séquençages des gènes HLA complets. De plus, cette population a été étudiée pour d'autres systèmes génétiques et la comparaison de nos résultats avec ceux des autres études nous aidera à mieux identifier quelles sont les forces évolutives qui agissent spécifiquement sur les différents gènes HLA.
Le Chapitre 3 est la présentation d'un algorithme (appelé `MADaM`) de traitement de données de séquençage haut débit, développé spécifiquement pour traiter des résultats de séquençage de loci hyper-variables comme les gènes HLA, en l'occurence les résultats des typages des populations analysées au Chapitre 4.
Le Chapitre 4 présente la première étude basée sur plus de 2'000 individus provenant de 36 populations d'Afrique (Afrique du nord, de l'ouest, centrale et de l'est), mais aussi de quelques populations d'Asie de l'ouest et d'Europe centrale utilisées comme références, chacune séquencée aux exons 2 de quatre loci HLA de classe II (HLA-DRB1, -DQA1, -DQB1 et -DPB1). Les résultats des séquençages de ces populations ont été

44. Pluriel de Mandenka.

traités à l'aide de `MADaM` et intègrent aussi ceux de la population Mandenka du Sénégal du Chapitre 2. Cette étude est particulièrement intéressante car elle considère un grand nombre d'échantillons de populations composés d'individus génotypés simultanément sur les mêmes quatre loci. Elle apporte de nouvelles données pour un continent qui a peu été étudié du point de vue HLA, ces populations étant ici caractérisées pour la première fois en ce qui concerne la variabilité HLA au niveau moléculaire. L'utilisation de données moléculaires permettra d'apporter des éléments de réponse aux questions portant sur les distributions des allèles et haplotypes HLA de classe II en Afrique, notamment les relations entre ces populations (d'un point de vue génétique), ainsi que les effets de la démographie, du mode de vie (sédentaire ou nomade) et de la pression en pathogènes exercée, notamment, par la malaria.
Le Chapitre 5 présente une analyse, non plus populationnelle, mais statistique des allèles HLA, basée sur les informations (de volume considérable) des bases de données IPD/IMGT-HLA et GenBank (NCBI). Il est composé de deux études résultant de certaines questions soulevées dans les précédents chapitres. La première de ces études concerne la répartition de l'information au sein des gènes HLA d'après la théorie de l'information de C. Shannon [Shannon, 1948] et permet de mieux comprendre les différences entre les méthodes de typage des Mandenka du Sénégal étudiées au Chapitre 2. La deuxième étude utilise une décomposition en chaînes de Markov des séquences des allèles HLA, découlant d'une méthode pour trier des séquences et implémentée dans l'algorithme `MADaM` du Chapitre 2. Cette étude est une approche nouvelle permettant d'étudier l'histoire évolutive des gènes HLA chez l'humain et son plus proche cousin, le chimpanzé.
Le Chapitre 6 est une discussion générale, où les différents résultats des analyses réalisées dans ce travail de doctorat seront mise en parallèle avec les connaissances actuelles, pour proposer des réponses aux principales interrogations ayant motivé ce travail.
Finalement, le Chapitre 7 donne des conclusions générales.

Chapitre 2

Étude comparée des Mandenka du Sénégal et des Cham du Vietnam

1 Introduction

Le développement des nouvelles technologies de séquençage (NGS - *Next Generation Sequencing*) depuis une dizaine d'années a permis d'approfondir les connaissances sur l'important polymorphisme génétique de la région HLA [Shiina et al., 2009], amenant notamment à une explosion du nombre d'allèles connus (plus de 23'000 nouveaux allèles ont été rapportés ces 10 dernières années [1]). Ces nouvelles technologies s'accompagnent de nouveaux défis techniques et scientifiques à relever, notamment la comparaison entre les différentes technologies (Roche 454 [Bentley et al., 2009, Shiina et al., 2012], Illumina [Cereb et al., 2015], IonTorrent [Shiina et al., 2012, Barone et al., 2015], PacBio [Cereb et al., 2015]). Ces séquenceurs à haut débit permettent depuis plusieurs années d'améliorer la précision et le débit des typages de routine dans les laboratoires d'histocompatibilité [Allen et al., 2018] et leurs avantages bénéficient aussi aux études de génétique des populations, notamment par une meilleure caractérisation des haplotypes à plus de deux loci (appelés haplotypes étendus) mais aussi en permettant une analyse directe de la variation nucléotidique, même hors des exons 2 et 3 traditionnellement typés [Sanchez-Mazas and Meyer, 2014, Bitarello et al., 2016, Meyer et al., 2018].

L'étude présentée dans ce chapitre porte en premier lieu sur la comparaison de trois techniques de typage (PCR-SSO, NGS-454 et NGS-MiSeq) sur huit gènes HLA dans une population bien documentée : les Mandenkalu (pluriel de Mandenka) d'Afrique de l'ouest. Cette population Mandenka a été échantillonnée en 1990 lors d'une campagne de terrain au Sénégal oriental et a, depuis, été typée (pour HLA) par trois méthodes différentes. La première, appelée ici PCR-SSO [2], consiste en une amplification par PCR de certaines régions cibles (exons 2 et 3 pour les gènes de classe I, exon 2 pour les gènes de classe II) puis d'une hybridation avec des sondes oligonucléotidiques spécifiques et marquées pour détecter un signal en cas d'hybridation.
Les deux autres méthodes sont des techniques qui peuvent être regroupées sous l'appellation «Typages basés sur les séquences (nucléotidiques)», puisque reposant sur le séquençage de la totalité [3] (NGS-MiSeq) ou une partie (exons 2 des gènes de classe II par NGS-454) des loci puis une comparaison avec les bases de données pour identifier les

1. `www.ebi.ac.uk/ipd/imgt/hla/stats.html`
2. *Polymerase Chain Reaction-Sequence Specific Oligonucleotide.*
3. Ou presque, puisque certaines régions telles que l'exon 1 de HLA-DRB1 restent un défi à séquencer.

allèles correspondants.
Les résultats de cette étude ont fait l'objet d'une publication dans la revue HLA [4] (Wiley) [Goeury et al., 2018a], ainsi que d'un *Population Report* dans la même revue [Goeury et al., 2018b]. Ces deux communications sont disponibles à la fin de ce chapitre et les résultats de cette étude sont ici présentés conjointement avec ceux de la population Cham du Vietnam (résultats encore non publiés).

Les Madenkalu constituent une population bien documentée puisqu'ayant été étudiée avant même la campagne de terrain de 1990 (voir [Blanc et al., 1990] pour une revue anthropologique). Ils sont représentatifs d'un plus large groupe de populations d'Afrique de l'ouest et sont depuis plusieurs années une population de référence du HGDP-CEPH [5].

Suite à la comparaison des trois techniques de typage, les typages NGS-MiSeq seront ensuite utilisés pour clarifier la relation entre la diversité nucléotidique et les différentes forces sélectives et/ou démographiques agissant sur les loci HLA étudiés. Plusieurs autres polymorphismes génétiques ont déjà été étudiés pour cette population : HLA de classe I [Dard et al., 1992], immunoglobulines [Blanc et al., 1990, Dard et al., 1996, Dard et al., 1997], ADN mitochondrial [Graven et al., 1995], RFLP [6] pour 80 marqueurs génomiques [Poloni et al., 1995], N-acetyltransferase [Sabbagh et al., 2008], α- [Martinson et al., 1995] et β-globulines [Currat et al., 2002]), ayant mis en évidence un important niveau de diversité génétique probablement lié à une expansion démographique importante [Excoffier and Schneider, 1999].
Cette hypothèse démographique sur les Mandenkalu excluant de fait la dérive génétique comme force évolutive majeure dans cette population, cette population peut alors être un modèle d'étude des effets de la sélection naturelle sur les gènes HLA, démographie et sélection naturelle étant les deux forces évolutives majeures du HLA, dont les effets sont difficiles à séparer [Meyer et al., 2006].

Les résultats obtenus pour les Mandenkalu seront ensuite comparés à ceux de la seconde population de l'étude, les Cham du Vietnam. Les Cham, descendants des royaumes Champa fondés au deuxième siècle de notre ère, représentent une population d'à peu près 500'000 personnes parlant une langue de la famille austronésienne [Eberhard et al., 2019]. L'origine des Cham est un sujet d'intérêt en histoire des populations puisqu'ils sont l'une des deux seules populations (avec les Moken) locutrices austronésiennes en Asie continentale du sud-est (MSEA [7]), la famille austronésienne (composée de 1'257 langues différentes [Eberhard et al., 2019]) étant majoritairement retrouvée sur les îles océaniennes, de Taïwan à l'île de Pâques, ainsi qu'à Madagascar [Bellwood et al., 2006].
Deux hypothèses ont été émises quant à l'origine des Cham [Peng et al., 2010]. La première, basée sur le modèle dit de la sortie de Taïwan (« *Out of Taïwan hypothesis* ») des locuteurs de la branche malayo-polynésienne de l'austronésien il y a 4'500 ans, postule que les ancêtres des Cham seraient des Austronésiens provenant des îles d'Asie du sud-est (ISEA [8]) ayant pris pied sur le continent asiatique il y a 2'500 ans

4. Anciennement *Tissue Antigens*.
5. Human Genome Diversity Project – Centre d'Étude du Polymorphisme Humain (`http://www.cephb.fr/hgdp/main.php`
6. *Restriction Fragment Length Polymorphisms* : Polymorphismes des tailles des fragments digérés par enzymes de restrictions.
7. Anglais : Mainland South-East Asia.
8. Anglais : Island South-East Asia.

[Thurgood, 1999, Higham, 2002, Southworth et al., 2004, Bellwood, 2007] ; c'est une hypothèse démique impliquant des mouvements de populations et des flux géniques entre elles. La seconde hypothèse, connue sous le nom du « réseau maritime Nusantao de commerce et de communication », est une hypothèse postulant que l'origine des Cham en tant que locuteurs austronésiens proviendrait d'une diffusion culturelle (sans déplacement de populations) le long des routes commerciales maritimes en Asie du sud-est [Solheim et al., 2007].
En 2010, une étude menée par Peng *et al.* sur l'ADN mitochondrial (lignées maternelles) a suggéré que les Cham seraient plus proches génétiquement des Mon-Khmer (une population MSEA) que des populations ISEA [Peng et al., 2010]. Deux ans plus tard, en 2012, une étude menée cette fois-ci sur le chromosome Y (lignées paternelles) par He *et al.* a mis en évidence un flux génique récent en provenance d'ISEA [He et al., 2012] .
L'hypothèse démique serait alors favorisée actuellement pour expliquer l'origine des Cham, ces derniers étant les descendant d'un mélange populationnel entre des locuteurs Mon-Khmer continentaux et une population austronésienne migrante ayant quitté Taïwan il y a 4'500 ans.
Après les études portant sur les lignées maternelles et paternelles, cette étude analyse la diversité génétique de 11 loci autosomaux HLA dans un échantillon de 62 Cham non-apparentés du Vietnam.

Ce chapitre portera donc tout d'abord sur la comparaison de trois techniques de typage HLA sur 25 ans d'étude des gènes HLA de la population Mandenka, afin de déterminer les avantages des nouvelles méthodes, mais aussi les nouveaux défis technologiques et scientifiques qu'elles soulèvent. Puis, afin de répondre en partie à la question « quels sont les avantages de ces nouvelles méthodes ? » les gènes HLA de deux populations, les Mandenkalu du Sénégal et les Cham du Vietnam, seront analysés avec des méthodes de typage récentes (séquençages Illumina-MiSeq) afin de répondre aux questions suivantes : 1) quels sont les effets de la sélection et/ou de la démographie sur la diversité génétique HLA des Mandenkalu ? et 2) qu'est-ce que l'étude des gènes HLA peut apporter comme information sur l'origine de la population Cham ?
Pour aborder la première problématique, les résultats des typages NGS-MiSeq seront comparés à ceux des méthodes traditionnelles (PCR-SSO sur les exons 2/3) et des typages basés sur des séquençages de courtes régions géniques (NGS-454 sur les exons 2), tous les typages ayant été réalisés sur les mêmes individus de la même population Mandenka. Les deux autres questions seront traitées en estimant et comparant plusieurs indices de diversité moléculaire et en réalisant des tests de neutralité sélective pour les deux populations Mandenka et Cham.

2 Matériels et Méthodes

2.1 Échantillonnages

La population Mandenka a été échantillonnée durant une campagne de terrain entre janvier et février 1990. 20mL de sang périphérique ont été prélevés à 205 donneurs et donneuses, après avoir recueilli leurs consentements libres et éclairés. Les individus proviennent de cinq villages (Batranke, Bantata, Baraboye, Lakanta et Soucouta) dans le district de Bandafassi près de Kédougou (Figure 2.1). Les relations d'apparentement ont aussi été documentées et 101 individus ont pu être considérés comme non apparentés, selon la méthode développée par Le Than en 2007 [Le Than, 2007].
L'ADN a été extrait en suivant le protocole décrit par [Tiercy et al., 1989].

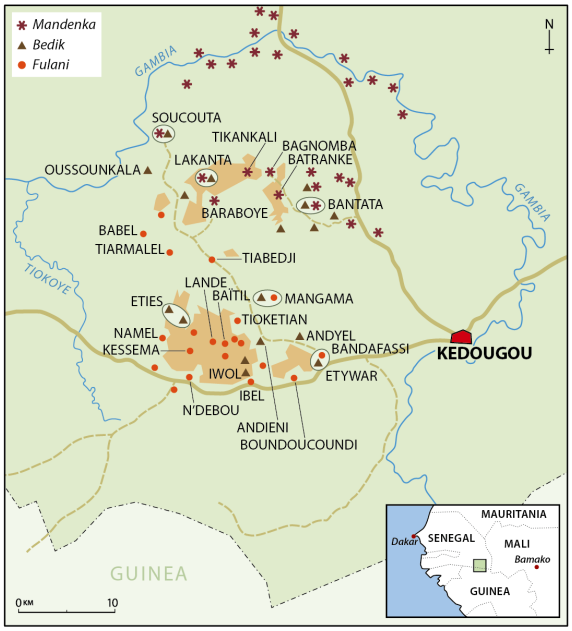


FIGURE 2.1 – Zone d'échantillonnage des Mandenkalu, près de la ville de Kédougou dans le Sénégal oriental (Région du Niokholo, district de Bandafassi). La carte illustre la proximité géographique entre les Mandenka, les Bedik et les Fulani. Crédits : David Glauser, UNIGE.

La population Cham a été échantillonnée dans la province de Binh Thuan, dans le sud du Vietnam (Figure 2.2). Un total 166 échantillons ont été obtenus dans le cadre d'une collaboration avec le Professeur An-Vu-Trieu (Laboratoire d'Immunogénétique du Département d'Immunologie et de Physiopathologie, Université de Médecine, Hanoi, Vietnam).

2.2 Typages

Les Mandenkalu ont été typés en utilisant trois méthodes différentes (PCR-SSO, NGS-454 et NGS-MiSeq), les Cham ont été typés en utilisant uniquement la méthode NGS-MiSeq.

PCR-SSO

Les gènes de classe II (DRB1, DQB1 et DPB1) ont été amplifiés par la méthode PCR-SSO, consistant en une amplification par PCR des exons 2 des gènes

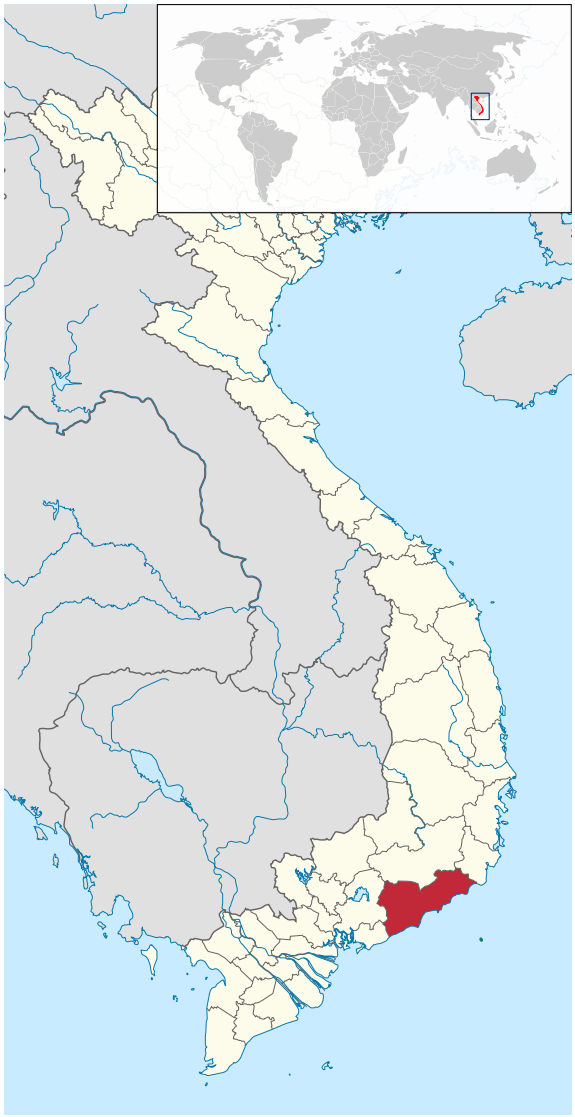


FIGURE 2.2 – Zone d'échantillonnage des Cham, dans la province de Binh Thuan, côte sud du Vietnam. Adapté de wikimedia-commons, licence CC-BY-SA 3.0.

ciblés, suivie d'une fixation sur des filtres en nylon et d'une hybridation avec des sondes oligonucléotidiques spécifiques (SSO : *Sequence-Specific Oligonucleotide*). Les sondes pour typer les gènes HLA-DRB1 et HLA-DQB1 ont été développées pour discriminer les allèles identifiés par le comité de la nomenclature HLA de 1991 [Tiercy et al., 1992, Morel et al., 1990]. Les gènes des classe I (A, B et C) ont été typés par PCR-SSO en 1996, en utilisant des sondes spécifiquement développées [Andrien et al., 1993, Tiercy et al., 1994, Grundschober et al., 1997]. En 2001 les loci HLA-A et -B ont été typés à nouveau par PCR-SSO en utilisant les protocoles du 13ème atelier international d'histocompatibilité et d'immunogénétique (basé sur 139 sondes identifiant 341 allèles différents) [Hansen, 2006]. Tous les typages ont été réalisés au laboratoire de J.M. Tiercy aux Hôpitaux Universitaires de Genève (HUG).

NGS-454

Les exons 2 de quatre gènes de classe II (DRB1, DQA1, DQB1, DPB1) ont été amplifiés et séquencés par la méthode NGS-454. Pour chaque gène, 245 individus (dont 46 réplicats) et 12 contrôles (uniquement H_2O) ont été amplifiés.
Les amplicons ont été marqués et multiplexés en suivant la méthode décrite par [Galan et al., 2010]. Les amorces PCR utilisées (Table 2.1) ont été modifiées par l'ajout d'un adaptateur *Lib-L Titanium* de 30pb en 5' (CCATCTCATCCCTGCGTGTCTCC-GACTCAG, nécessaire pour l'emPCR [9]) et des tags oligonucléotidiques en 5' et 3'. 32 tags *forward* et 21 tags *reverse* ont été conçus, afin de générer 182 combinaisons uniques pour identifier de manière non-ambigüe les amplicons. Les séquençages ont été réalisés en trois séries (trois *runs*).

Locus	**Direction**	**Amorce**
DRB1	Forward	Adaptateur + tag + CCGGATCCTTCGTGTCCCCACAGCACG
	Reverse	Adaptateur + tag + CCGAATTCCGCTGCACTGTGAAGCTCTC
DQA1	Forward	Adaptateur + tag + GTTTCTTYCATCATTTTGTGTATTAAGGT
	Reverse	Adaptateur + tag + CGGTAGAGTTGTAGCGTTTA
DQB1	Forward	Adaptateur + tag + AGGATCCCCGCAGAGGATTTCGTGTACCA
	Reverse	Adaptateur + tag + TCCTGCAGGACGCTCACCTCTCCGCTGCA
DPB1	Forward	Adaptateur + tag + GCTGCAGGAGAGTGGCGCCTCCGCTCAT
	Reverse	Adaptateur + tag + CGGATCCGGCCCAAAGCCCTCACTC

TABLE 2.1 – Amorces PCR *forward* et *reverse* utilisées pour le pyroséquencage des quatre exons 2 des gènes HLA de classe II.

Les amorces PCR utilisées ont permis le séquençage de la position 5 à 245 de l'exon 2 de DRB1 et 17 à 258 pour l'exon 2 de DQB1. Les séquences de DQA1 comprennent les 50 dernières paires de bases de l'intron 1 et vont jusqu'à la position 217 de l'exon 2 (les 32 dernières paires de bases sont manquantes). Pour DPB1, les séquences couvrent les 19 dernières paires de bases de l'intron 1, l'ensemble de l'exon 2 et les deux premières paires de bases de l'intron 2.

La préparation de la librairie a été réalisée par Lydie Brunet (Laboratoire d'anthropologie, génétique et peuplements, Université de Genève). Le séquençage a été réalisé

9. Emulsion PCR.

par Beckman Coulter Genomics (Genomic Services, Danvers, Massachuetts). Les lectures ont été filtrées avec Mothur [Schloss et al., 2009] et un PhredScore minimal de 30, puis explorées et assignées manuellement avec SESAME Barcode [Meglécz et al., 2011] par Lydie Brunet, Maeva Pasqier et Thomas Goeury (Laboratoire d'anthropologie, génétique et peuplements, Université de Genève).

NGS-MiSeq

Deux techniques complémentaires ont été utilisées pour séquencer entièrement les huit loci de classe I et II chez les Mandenka :

1. Hopitaux Universitaires de Genève (HUG) : 54 individus ont été séquencés en utilisant le kit Holotype HLA X2 (Omixon Biocomputing Ltd, Budapest, Hongrie), ciblant sept loci HLA (A, B, C, DRB1, DQA1, DQB1 et DPB1). Après séquençage sur la plateforme Illumina MiSeq, les lectures ont été traitées par HLA Twin v1.1.1 (Omixon Biocomputing Ltd, Budapest, Hongrie) ;

2. Université de Stanford : 65 individus (incluant 25 séquencés aux HUG afin de confirmer certains typages) ont été typés pour huit loci HLA (A, B, C, DRB1, DQA1, DQB1, DPA1 et DPB1) en utilisant la méthode MIA FORA NGS développée par Sirona Genomics (Immucor Inc, Norcross, Georgie, USA). Les allèles ont été assignés à l'aide du logiciel d'alignement NGS de Sirona Genomics.

Des génotypes non-ambigus (deux séquences par locus et par individu) ont été obtenus pour 51 à 86 individus (selon le locus, voir Table 2.2 pour le détail). Des ambigüités ont été observées dans 1.28% des cas et dans ce cas l'ensemble des allèles possibles ont été conservés.

Les individus Cham ont été séquencés à Stanford par la même méthode que les Mandenkalu.

2.3 Alignement des séquences MiSeq

Les séquences consensus obtenues pour chaque individu par la méthode NGS-MiSeq ont été alignées à une séquence de référence (issue de la base de données IPD-IMGT/HLA v3.25.00 [Robinson et al., 2015]) à l'aide de MAFFT [Katoh, 2002] et les exons de référence ont été alignés à l'aide de l'option «`--add`». La Table 2.2 liste le nombre de séquences disponibles à chaque locus et les séquences de référence utilisées. Une vérification manuelle a permis de retirer les séquences trop fortement divergentes (considérées comme artefacts). Les régions géniques (exons, introns et UTR [Mack, 2015]) ont été extraites de ces alignements.

Locus	Allèles de réf.	Exons de réf.	N. exons	Ns. Mandenka	Ns. Cham
A	A*01:01:01	*idem*	8	174	124
B	B*07:02:01	*idem*	7	166	124
C	C*01:02:01	*idem*	8	166	124
DRB1	DRB1*01:01:01	*idem*	6	160	122
DQA1	DQA1*01:01:02	DQA1*01:01:01	4	158	98
DQB1	DQB1*02:01:01	DQB1*05:03:01	6	154	124
DPA1	DPA1*01:03:01	*idem*	4	102	124
DPB1	DPB1*02:01:02	DPB1*01:01:01	5	166	124

TABLE 2.2 – Tableau récapitulant les séquences de référence utilisées pour aligner et nettoyer les résultats de séquençage NGS-MiSeq et identifier les différentes régions géniques (exons, introns, UTR), *idem* indique que les exons de référence proviennent de la même séquence que l'allèle de référence. La table fournit aussi le nombre de séquences (individus diploïdes) disponibles à chaque locus pour les Mandenkalu (Ns. Mandenka) et les Cham (Ns. Cham) ainsi que le nombre d'exons disponibles pour chaque locus (N. exons).

2.4 Analyses de génétique des populations

Comparaison des techniques de typage

Pour les Mandenkalu, les individus typés par au moins deux techniques différentes ont été comparés à chaque locus. Seule une correspondance parfaite sur les deux allèles (entre les deux techniques) a été considéré comme correcte (si au moins un des deux allèles était différent, la comparaison a été considérée comme fausse).
Pour le nombre d'individus typés à chaque locus et par chaque technique, une méthode de ré-échantillonnage a été appliquée pour tenir compte de cette variabilité. Pour chaque comparaison d'un locus entre deux techniques, 1'000 tirages au sort aléatoires (sans remise) ont été effectués, en prenant une taille d'échantillon de 66 (correspondant au minimum de typages comparables pour un locus, HLA-C, typé par NGS-MiSeq).

Allèles et haplotypes

Seuls les individus non apparentés pour chacune des deux populations ont été inclus dans l'analyse.
La plateforme **hla-net.eu** a été utilisée pour réaliser les tests de Hardy-Weinberg et d'Ewen-Watterson-Slatkins et pour estimer les distributions de fréquences alléliques, le nombre d'allèles, l'hétérozygotie, les distributions de fréquences des haplotypes bi-alléliques et les déséquilibres globaux et haplotypiques [Nunes et al., 2014, Nunes, 2016].
La richesse allélique a été estimée par la méthode de raréfaction (voir Chapitre 1, page 42), estimant le nombre d'allèles qui auraient été détectés à chaque locus si toutes les tailles d'échantillons avaient été aussi petites que la plus petite taille d'échantillon utilisée dans l'étude [El Mousadik and Petit, 1996].

Tests de neutralité

Pour chaque exon pris séparément (à chaque locus pour chaque population), l'indice D de Tajima [Tajima, 1989b, Tajima, 1989a], le ratio du taux de mutations non-synonymes / synonymes dN/dS [Li, 1993], l'indice π de diversité nucléotidique et le nombre de sites

polymorphiques S ont été calculés. Les D de Tajima, π et S ont été calculés avec le logiciel Arlequin v3.5 [Excoffier and Lischer, 2010] et le ratio dN/dS à l'aide du logiciel MEGA [Kumar et al., 2016] en appliquant la méthode de Nei-Gojobori [Nei and Gojobori, 1986], ne tenant pas compte des codons dégénéres (pris en compse par les méthodes de Li-Wu-Luo et Pamilo-Bianchi-Li [Li et al., 1985, Pamilo and Bianchi, 1993]).

Les codons codant pour le site de reconnaissance de l'antigène (ARS) [Reche and Reinherz, 2003] ont été analysés séparément des autres codons (non-ARS) pour les exons 2 (classe I et II) et les exons 3 (classe I). Les positions de ces codons correspondent à ceux indiqués dans la Figure 1.6, rappelés dans le Tableau 2.3.

	Région	**Positions des codons ARS**
Classe I	Exon 2	5, 7, 9, 11, 22, 24, 25, 26, 33, 34, 45, 59 63, 66, 67, 70, 73, 74, 77, 80, 81, 84
	Exon 3	95, 96, 97, 99, 114, 116, 123, 124, 142, 143 147, 152, 156, 159, 163, 164, 167, 171
Classe II	Exon 2α	7, 9, 11, 22, 24, 31, 32, 43, 52, 53, 54 58, 59, 62, 65, 66, 69, 72, 73, 76
	Exon 2β	9, 11, 13, 14, 15, 26, 27, 28,30, 37, 38, 47, 56 57, 61, 67, 71, 74, 78, 79, 82, 85, 86, 89, 90

TABLE 2.3 – Positions des codons ARS pour les exons 2 et 3 des loci de classe I et les exons 2 des loci de classe II. Les positions proviennent de [Reche and Reinherz, 2003] (voir la Figure 1.6).

Puisqu'il était impossible de distinguer des *indels*[10] structuraux de régions non séquencées, les régions géniques comportant des *indels* ont été retirées de l'analyse, à l'exception du codon 56 de l'exon 2 de HLA-DQA1 pour lequel un *indel* tri-nucléotidique est reconnu en tant que polymorphisme putatif (séquences nucléotidiques disponibles dans la base de données IPD-IMG/HLA). Ainsi, un seuil de 5% maximum de données manquantes a été utilisé pour Arlequin, MEGA étant paramétré pour considérer tout codon avec des *indels* comme une information manquante.

Le test de neutralité sélective du D de Tajima étant bilatéral (voir page 43), les pValeurs obtenues ont été ajustées de la façon suivante :

— Pour les pValeurs supérieures ou égales à 0.5 : $p_{Adj} = 2 \cdot (1 - pValeur)$;

— Pour les pValeurs inférieures à 0.5 : $p_{Adj} = 2 \cdot pValeur$.

Les pValeurs ont ensuite été corrigées pour tenir compte des tests multiples en utilisant la correction *fdr* (*False Discovery Rate*, [Benjamini and Hochberg, 1995]) telle qu'implémentée dans R [R Core Team, 2020].

La significativité du ratio dN/dS (voir page 45) a été évaluée à l'aide du test Z , où

$$Z = \frac{(dN - dS)}{\sqrt{(\sigma^2(dN) + \sigma^2(dS))}}$$

suit une distribution normale sous l'hypothèse nulle H0.

10. Insertions et délétions.

Analyses en composantes principales

Puisqu'un grand nombre de statistiques corrélées ont été calculées aux exons 2, 3, 4 et 5 (ce dernier uniquement pour les gènes de classe I) de chacun des loci (D de Tajima, nombre de sites polymorphiques S, diversité nucléotidique π, dN et dS) pour chacune des populations, ces statistiques ont été combinées dans une analyse en composantes principales (ACP). Cette analyse, réalisée à l'aide de la librairie `ade4` [Chessel et al., 2004, Dray et al., 2007, Dray and Dufour, 2007, Bougeard and Dray, 2018] pour R [R Core Team, 2020], permet de projeter les observations sur un plus petit nombre d'axes décorrélés (les composantes principales).
Trois analyses en composantes principales ont été réalisées sur ces deux populations, la première en ne considérant que les données pour les Mandenkalu, la seconde en ne considérant que les données pour les Cham et la dernière en utilisant les données des deux populations à des fins de comparaison.
Il est à noter que la première ACP, réalisée sur les données de la population Mandenka est issue de la publication associée à ce chapitre [Goeury et al., 2018a] et présente la particularité que les différentes statistiques (D, π, S, dN et dS) ont aussi été calculées pour le premier, deuxième et troisième nucléotide de chaque codon. Cette particularité est propre à l'ACP menée sur cette population et n'a été appliquée ni aux données de la population Cham ni aux données de l'ACP pour les deux populations.

3 Résultats

3.1 Résultats des typages et équilibre de Hardy-Weinberg

La Table 2.4 donne le nombre d'individus typés par chacune des trois méthodes (les Cham n'ayant été typés que par NGS-MiSeq), avec le détail du nombre total d'individus typés (entre parenthèses) et du nombre d'individus typés non apparentés inclus pour la suite des analyses. Les génotypes et identificateurs sont disponibles dans l'annexe S-28.

	Mandenka			**Cham**
Locus	**PCR-SSO**	**NGS-454**	**NGS-MiSeq**	**NGS-MiSeq**
A	72 (196)	–	72 (85)	62 (62)
B	67 (82)	–	67 (82)	62 (62)
C	54 (66)	–	54 (66)	62 (62)
DRB1	96 (188)	96 (189)	65 (78)	61 (61)
DQA1	–	66 (66)	66 (66)	49 (49)
DQB1	94 (188)	96 (190)	60 (74)	62 (62)
DPA1	–	–	51 (51)	62 (62)
DPB1	99 (193)	101 (195)	70 (83)	62 (62)

TABLE 2.4 – Nombre d'individus non apparentés (entre parenthèses : total) à chaque locus, pour les deux populations selon les méthodes de typage employées. Les Cham ont été typés seulement par NGS-MiSeq, les Mandenkalu ont été typés avec les trois méthodes. Pour les Mandenka, les loci de classe I et HLA-DPA1 n'ont pas été typés par NGS-454 et HLA-DQA1 et -DPA1 n'ont pas été typés par PCR-SSO.

La Table 2.5 donne les pValeurs des tests d'équilibre de Hardy-Weinberg aux 8 loci testés pour les deux populations.

Locus	**Mandenka**	**Cham**
A	0.331	1
B	0.524	1
C	0.044	1
DRB1	0.089	0.622
DQA1	1	0.244
DQB1	1	0.172
DPA1	**0.006**	1
DPB1	0.055	1

TABLE 2.5 – pValeurs du test d'équilibre de Hardy-Weinberg pour ces populations. Tests basés sur les typages NGS-MiSeq. La valeur en gras indique un rejet de l'hypothèse d'équilibre de Hardy-Weinberg.

Huit loci ayant été testés, le seuil α' choisi pour rejeter l'hypothèse de l'équilibre de Hardy-Weinberg est de $\alpha' = \frac{0.05}{8} = 0.00625$. Pour la population Mandenka, seul HLA-DPA1 rejette l'hypothèse de l'équilibre de Hardy-Weinberg, tous les autres loci étant à l'équilibre. Pour la population Cham, l'ensemble des 8 loci sont à l'équilibre de Hardy-Weinberg.

Les distributions de fréquences alléliques à chacun des loci sont disponibles en annexe S-26.

3.2 Comparaison des techniques de typage

La Table 2.6 fournit les valeurs numériques des comparaisons de typages entre les trois techniques (uniquement pour la population Mandenka) et la Figure 2.3 illustre ces résultat.
Les intervalles de confiances ont été obtenus par 1'000 tirages aléatoires (sans remise) de 66 génotypes (plus petit nombre de génotypes comparés, correspondant au locus HLA-C) pour chaque comparaison.

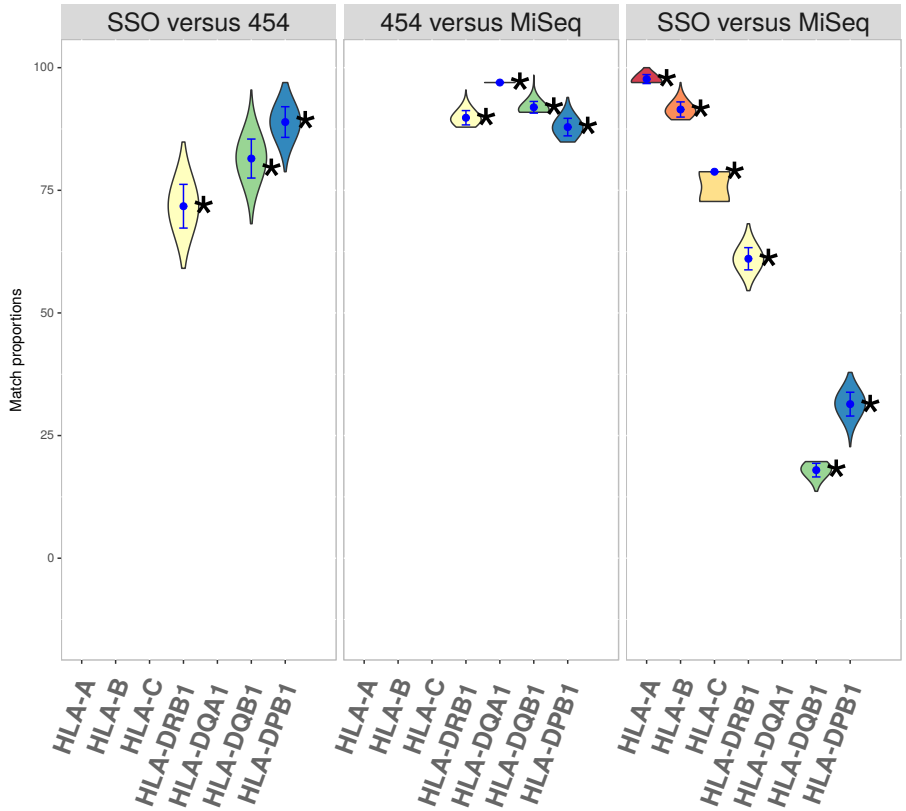


FIGURE 2.3 – Graphique en violons des correspondances de typages obtenues entre les typages PCR-SSO («SSO» sur la Figure), les typages NGS-454 («454») et les typages NGS-MiSeq («MiSeq») chez les Mandenkalu, pour trois à six des loci pouvant être comparés (HLA-DQA1 n'a pas été typé par PCR-SSO et les loci de classe II n'ont pas été typés par NGS-454). Seules les correspondances totales (des deux allèles, pour les deux techniques) sont considérées comme réussies. La valeur observée correspond à l'étoile *, les violons correspondent aux 1'000 tirages aléatoires (sans remise) de 66 génotypes (plus petit nombre de génotypes comparés, correspondant au locus HLA-C) et l'intervalle représente l'intervalle de confiance à 95%. Figure issue de [Goeury et al., 2018a].

Locus		Comparaisons entre les techniques de typage					
		PCR-SSO vs NGS-454		NGS-454 vs NGS-MiSeq		PCR-SSO vs NGS-MiSeq	
HLA-A	N. ind. genotypés					196	87
	N. ind. comparés					85	
	N. correspondances					83 (97.6%)	
	IC95					97.0 - 100.0%	
HLA-B	N. ind. genotypés					198	83
	N. ind. comparés					82	
	N. correspondances					75 (91.5%)	
	IC95					89.4 - 95.5%	
HLA-C	N. ind. genotypés					165	83
	N. ind. comparés					66	
	N. correspondances					48 (72.7%)	
	IC95					72.7 - 72.7%	
HLA-DRB1	N. ind. genotypés	198	194	194	81	198	81
	N. ind. comparés	188		78		77	
	N. correspondances	135 (71.8%)		70 (89.7%)		47 (61.0%)	
	IC95	63.6 – 80.3%		87.9 – 92.4%		57.5 – 65.2%	
HLA-DQA1	N. ind. genotypés			194	85		
	N. ind. comparés			66			
	N. correspondances			64 (97.0%)			
	IC95			97.0%			
HLA-DQB1	N. ind. genotypés	195	196	196	76	195	76
	N. ind. comparés	188		74		72	
	N. correspondances	153 (81.4%)		68 (91.9%)		13 (18.1%)	
	IC95	72.7 – 87.9%		90.9 – 93.9%		15.2 – 19.7%	
HLA-DPB1	N. ind. genotypés	193	199	199	82	193	82
	N. ind. comparés	193		82		79	
	N. correspondances	172 (89.1%)		72 (87.8%)		24 (30.3%)	
	IC95	83.3 – 94.0%		84.9 – 90.9%		25.8 – 36.4%	

TABLE 2.6 – Correspondances entre les typages effectués par PCR-SSO, NGS-454 et NGS-MiSeq chez les Mandenkalu. Détail du nombre total d'individus génotypés à chaque locus par chaque méthode (N. ind. genotypés), du nombre d'individus génotypés par les deux méthodes (N. ind. comparés) et du nombre et pourcentage de correspondances. IC95 correspond à l'intervalle de confiance de 95% autour de la valeur, obtenu par 1'000 tirages aléatoires (sans remise) de 66 génotypes (plus petit nombre de génotypes comparés, correspondant au locus HLA-C).

La forme des violons, ainsi que les positions des valeurs observées dans les intervalles de confiance montrent que les valeurs observées (la valeur observée étant toujours dans l'intervalle de confiance) ne semblent biaisées ni par les différences de tailles d'échantillons comparés, ni par les tailles d'échantillons en elles-mêmes.

Les comparaisons entre les typages PCR-SSO et NGS-454 montrent de bonnes correspondances pour les trois gènes comparés (concordances de 71.8% pour HLA-DRB1 à 89.1% pour HLA-DPB1), principalement dues au fait que les deux techniques ont ciblé la même région (exons 2) et donc la même information génétique. Toutefois, quelques allèles n'ont pas été détectés par PCR-SSO (DRB1*07 et DQB1*02), suggèrant une plus faible résolution des typages PCR-SSO par rapport aux typages NGS-454. Quelques allèles identifiés par PCR-SSO (voir annexe S-27) n'ont pas été détectés par NGS-454, indiquant soit une erreur de typage par PCR-SSO, soit une non détection (*dropout*) de cet allèle par NGS-454.

Concernant les typages NGS-454 et NGS-MiSeq, de bons résultats sont observés pour les quatre loci de classe II comparés (de 88% de concordance pour HLA-DPB1 à 97% pour HLA-DQA1). La principale différence est due au fait que les deux techniques ne ciblent pas les mêmes régions : NGS-MiSeq inclut tous les exons (ou presque, selon les protocoles et les gènes [11]) ce qui implique une assignation à un seul allèle en général (98.7% de typages non ambigus au 3ème champ). NGS-454 ne cible que l'exon 2, ce qui amène à davantage d'ambiguïtés sur les allèles nominaux, notamment ceux définis par des SNP hors des exons 2 (de 2.8 allèles par séquence pour HLA-DPB1 à 12.3 pour HLA-DQB1, voir annexe S-44 pour la liste de tous les allèles possibles par séquence NGS-454). Un résultat particulièrement marquant concerne plusieurs séquences NGS-454 (présentes à haute fréquence chez les Mandenkalu pour certaines) qui se subdivisent en plusieurs allèles avec NGS-MiSeq : DQB1*03:01 devient DQB1*03:19, *03:01:01 et *03:01:04 ; DPB1*17:01 devient DPB1*17:01 et *131:01. DQB1*03:19 (FA [12]=43%) et DQB1*03:01:01 diffèrent d'un unique SNP sur l'exon 3 (codon 185 : A**T**C devient A**C**C). De même, DPB1*17:01 et *131:01 diffèrent de plusieurs SNP, tous en dehors de l'exon 2 (sept SNP sur l'exon 3 et un sur l'exon 4). Les résultats observés (Figure 2.3 et Table 2.6) correspondent en réalité aux correspondances entre les exons 2 obtenus par les deux techniques de séquençage plutôt qu'à la stricte correspondance des allèles nominaux (eu égard au grand nombre d'ambiguïtés générées par la méthode NGS-454).

Les plus importantes différences de typage concernent les typages PCR-SSO et NGS-MiSeq. Si les loci de classe I montrent de bonnes correspondances (de 78.8% de concordances pour HLA-C à 97.6% pour HLA-A) ce n'est pas le cas des classes II, beaucoup plus hétérogènes en termes de correspondances : 18.1% pour HLA-DQB1, 31.3% pour HLA-DPB1 mais 61% pour HLA-DRB1. Pour HLA-DRB1, l'allèle le plus fréquent (DRB1*13:04, FA=28%) a été bien détecté par les deux techniques et ne correspond pas à d'autres allèles par NGS-MiSeq (allèles qui se différenciaient hors de l'exon 2).
On retrouve des différences similaires dans la comparaison des NGS-454 / NGS-Miseq : DQB1*03:01 trouvé par PCR-SSO qui devient DQB1*03:19, et DPB1*17:01 qui devient DPB1*17:01 et *131:01 pour NGS-MiSeq. En plus de la présence de sites polymorphiques hors de portée des typages PCR-SSO, certains des allèles (des fois parmi les plus fréquents)

11. HLA-DRB1 possède un très long intron 1, de plusieurs milliers de paires de bases, généralement mal couvert par les typages NGS.
12. Fréquence allélique

étaient inconnus à l'époque des typages PCR-SSO (aucune sonde moléculaire n'était capable de les identifier), par exemple l'allèle HLA-DQB1 le plus fréquent par PCR-SSO, DQB1*03:01, est donné comme ambigu avec NGS-454 puisque pouvant correspondre soit à DQB1*03:01 soit à DQB1*03:19 (découvert en 2007 [Witter et al., 2007]) et les typages NGS-MiSeq lèvent l'ambiguïté en donnant accès à l'information de ces sites polymorphiques hors de l'exon 2.
Ces résultats expliquent pourquoi les deux techniques NGS montrent des résultats semblables là où la comparaison PCR-SSO / NGS-MiSeq montre de très faibles concordances pour HLA-DQB1 ou -DPB1.

3.3 Profils moléculaires HLA des populations Mandenka et Cham

Les Tables 2.7 et 2.8 décrivent le profil moléculaire des deux populations étudiées. Pour chaque locus séquencé par NGS-MiSeq sont rapportés : le nombre d'allèles observés et la richesse allélique, l'hétérozygotie (indice de diversité génétique), le F50 (proposé ici comme statistique descriptive des profils des distributions alléliques de loci hypervariables) et la liste des allèles les plus fréquents (fréquence seuil de 10% minimum). Les distributions de fréquences alléliques complètes sont disponibles en annexe S-26.

Locus	**K - ar**	**H**	**F50**	**Allèles fréquents**
HLA-A	22 - 20.66	0.920	5	A*23:01:01 (17%) A*33:03:01 (10%) A*30:02:01 (10%)
HLA-B	30 - 27.83	0.936	7	B*35:01:01 (16%)
HLA-C	18 - 17.94	0.910	5	C*04:01:01 (19%) C*16:01:01 (16%)
HLA-DRB1	20 - 16.94	0.876	4	DRB1*13:04 (28%)
HLA-DQA1	14 - 13.05	0.715	1	DQA1*05:05:01 (50%) DQA1*01:02:01 (12%)
HLA-DQB1	13 - 12.78	0.767	2	DQB1*03:19 (44%) DQB1*02:01:01 (10%)
HLA-DPA1	10 - 10	0.719	2	DPA1*02:01:01 (46%) DPA1*01:03:01(19%) DPA1*03:01 (16%)
HLA-DPB1	19 - 16.59	0.863	3	DPB1*17:01 (22%) DPB1*131:01 (19%) DPB1*01:01:01 (14%) DPB1*02:01:02 (14%)

TABLE 2.7 – Profil moléculaire des Mandenka, basé sur les typages NGS-MiSeq. K : nombre d'allèles observés, ar : richesse allélique, H : hétérozygotie, F50 : nombre minimum d'allèles les plus fréquents dont les fréquences cumulées atteignent 50%, Allèles fréquents : allèles de plus de 10% et leur fréquence observée.

Locus	K - ar	H	F50	Allèles fréquents
HLA-A	24 -22.08	0.894	4	A*11:01:01 (21%) A*33:03:01 (15%) A*02:03:01 (12%)
HLA-B	38 - 34.55	0.951	7	B*58:01:01 (11%)
HLA-C	17 - 16.89	0.918	5	C*07:02:01 (15%) C*03:02:02 (12%)
HLA-DRB1	19 - 17.96	0.894	4	DRB1*15:02:01 (21%) DRB1*12:02:01 (16%) DRB1*07:01:01 (12%)
HLA-DQA1	14 - 14	0.901	4	DQA1*01:01:01 (16%) DQA1*01:02:01 (13%) DQA1*02:01 (12%) DQA1*01:02:02 (10%) DQA1*06:01:01 (10%)
HLA-DQB1	14 - 13.49	0.885	4	DQB1*05:01:01 (17%) DQB1*05:02:01 (15%) DQB1*03:01:01 (15%) DQB1*02:01:01 (11%)
HLA-DPA1	7 - 6.37	0.687	2	DPA1*01:03:01 (44%) DPA1*02:02:02 (29%) DPA1*02:01:01 (19%)
HLA-DPB1	18 - 17.05	0.885	3	DPB1*04:01:01 (18%) DPB1*13:01:01 (18%) DPB1*05:01:01 (11%) DPB1*03:01:01 (15%)

TABLE 2.8 – Profil moléculaire des Cham, basé sur les typages NGS-MiSeq. K : nombre d'allèles observés, ar : richesse allélique, H : hétérozygotie, F50 : nombre minimum d'allèles les plus fréquents dont les fréquences cumulées atteignent 50%, Allèles fréquents : allèles de plus de 10% et leur fréquence observée.

Pour les loci de classe I, les deux populations sont similaires en termes d'hétérozygotie et de F50. Ces loci sont toujours les plus diversifiés : hétérozygotie supérieure à 90% (à l'exception de HLA-A pour les Cham), F50 important (5.5 en moyenne) et une importante richesse allélique (de 16.89 à 34.55).
La population Cham se démarque toutefois par l'importante diversité observée au locus HLA-B, ce dernier totalisant 38 allèles détectés (ar=34.55), une hétérozygotie de 95% et un F50 de 7, en lien avec la distribution des fréquences alléliques puisque, sur les 38 allèles HLA-B observés dans cette population, un seul dépasse 10% (B*58:01:01, FA=11%).
Les deux populations se démarquent principalement par leurs loci de classe II et principalement la paire HLA-DQA1 et -DQB1 (codant pour la molécule HLA-DQ). En effet, les Mandenkalu montrent une plus faible diversité à ces deux loci (en termes d'hétérozygotie et de F50), ce qui peut s'expliquer par la présence de deux allèles très fréquents, DQA1*05:05:01 (FA=50%) et DQB1*03:19 (FA=44%).
Parmi les allèles les plus fréquents, A*33:03:01 est le seul allèle d'un gène de classe I retrouvé dans les deux populations. Pour les gènes de classe II, les allèles fréquents et partagés entre les populations sont DQA1*01:02:01, DQB1*02:01:01, DPA1*02:01:01 et DPA1*01:03:01.

3.4 Diversité nucléotidique

Les Figures 2.4 et 2.5 montrent, pour les 8 loci séquencés chez les Mandenkalu et les 11 loci séquencés chez les Cham, respectivement la diversité nucléotidique ($\pi \pm \sigma^2$) pour les exons 2, 3, 4 (loci de classes I et II) et 5 (uniquement loci de classe I) et la diversité en acides aminés, estimée après traduction des chaînes de nucléotides (à l'aide de la table de traduction des codons en acides aminés), pour les domaines $\alpha 1 - \alpha 4$ et $\beta 1 - \beta 3$ des molécules HLA. Les facettes correspondent aux éléments structurellement et fonctionellement comparables (région de liaison peptidique, régions interagissant avec les récepteurs des cellules T CD4+ et CD8+, région transmembranaire). Les autres régions n'ont pas été intégrées au graphique à cause d'une couverture trop faible. Les données de $\pi \pm \sigma^2$ sont disponibles en annexe S-21.

Que ce soit pour les loci de classe I ou de classe II, la diversité nucléotidique observée chez les Mandenkalu est toujours plus importante aux codons ARS par rapport aux autres codons codant aussi pour la région de liaison peptidique (en haut à gauche). Cette différence est d'autant plus marquée si l'on considère la diversité en acides aminés des domaines moléculaires correspondants (en bas à gauche), résultats qui suggèrent une hypothèse d'un avantage sélectif de la diversité moléculaire de ces régions [Hughes and Nei, 1988, Hughes and Nei, 1989c].
Pour les loci de classe I, HLA-A et surtout HLA-C montrent un motif particulier, où la diversité de la région $\alpha 1$ (codée par l'exon 2) est plus faible que celle de la région $\alpha 2$ (codée par l'exon 3). Cela suggère soit 1) une sélection diversifiante plus faible sur les régions $\alpha 1$ de ces molécules, ou 2) une sélection directionnelle sur ces régions. Au contraire, les gènes de classe II HLA-DRB1 et -DQB1 montrent les plus importantes diversités moléculaires à leurs codons ARS (sur l'exon 2).
Les autres régions (régions interagissant avec les cellules T et régions transmembranaires) montrent de beaucoup plus faibles diversités (aussi bien nucléotidiques qu'en acides aminés), similaires aux codons non-ARS des PBR. À noter que HLA-DPA1 montre une diversité du même ordre de grandeur, mais pour ses codons ARS, et c'est le seul locus à être si peu diversifié à ses codons ARS.

FIGURE 2.4 – Diversité (et écart-type) par site pour les nucléotides (figures du haut) et les acides aminés déduis des séquences nucléotidiques (figures du bas), pour chacune des régions fonctionnelles codées par HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1 chez les Mandenkalu. PBR : Peptide binding region – Région de liaison peptidique, TCRIR : T-Cell receptors interacting regions – Régions interagissant avec les récepteurs des cellules T CD4+ et CD8+, TMR : Transmembrane Region – Région Transmembranaire. Les crochets regroupent les paires de loci formant les molécules HLA-DQ et -DP. ARS (Non-ARS) : codons/acides aminés de l'exon 2/PBR et (non) impliqués dans la liaison à l'antigène ; Other : codons/acides aminés hors de la PBR. *Amino-acid* : acides aminé, *diversity* : diversité. Figure tirée de [Goeury et al., 2018a].

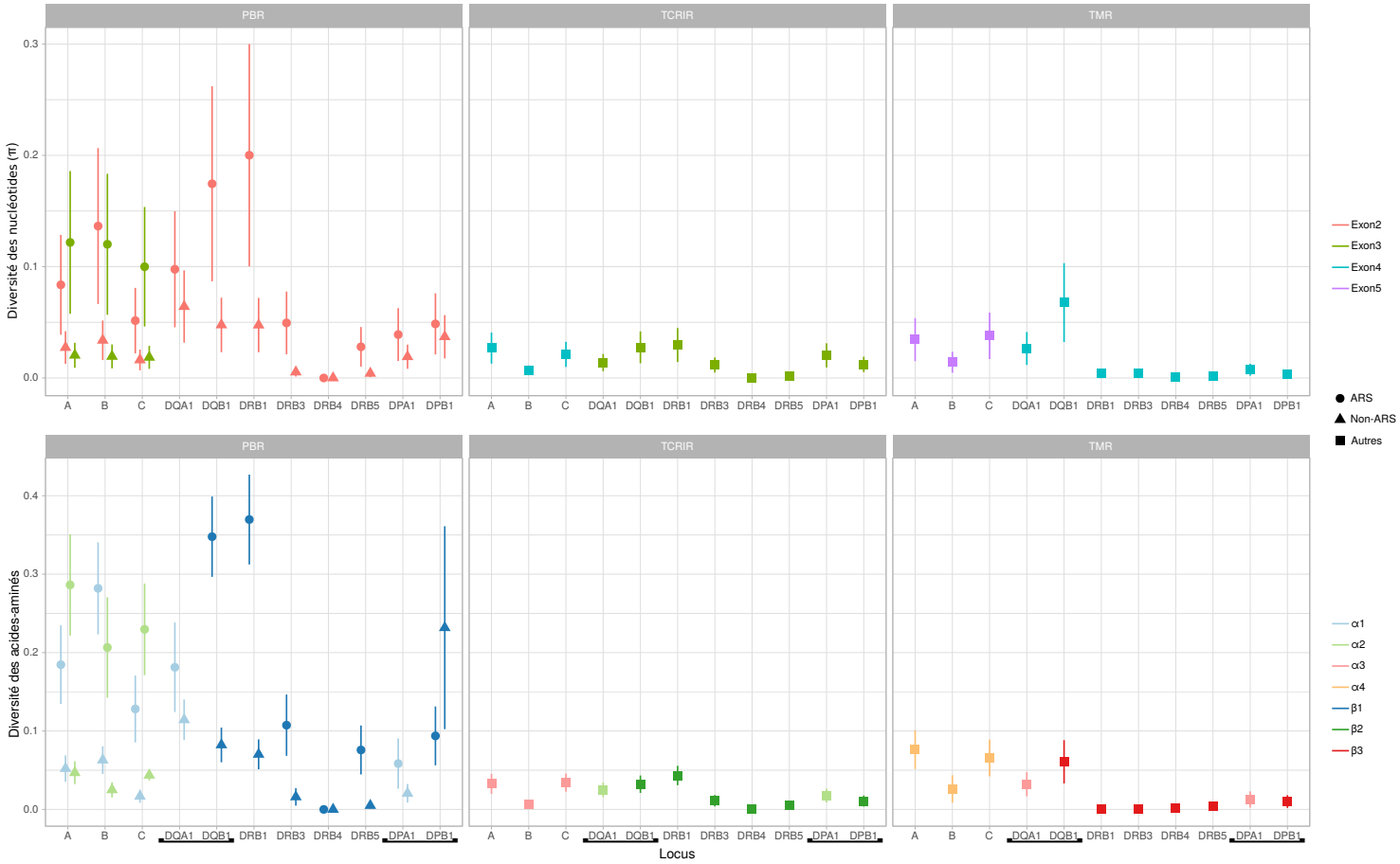


FIGURE 2.5 – Diversité (et écart-type) par site pour les nucléotides (figures du haut) et les acides aminés déduis des séquences nucléotidiques (figures du bas), pour chacune des régions fonctionnelles codées par HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1 chez les Cham. PBR : Peptide binding region – Région de liaison peptidique, TCRIR : T-Cell receptors interacting regions – Régions interagissant avec les récepteurs des cellules T CD4+ et CD8+, TMR : Transmembrane Region – Région Transmembranaire. Les crochets regroupent les paires de loci formant les molécules HLA-DQ et -DP. ARS (Non-ARS) : codons/acides aminés de l'exon 2/PBR et (non) impliqués dans la liaison à l'antigène ; Other : codons/acides aminés hors de la PBR. *Amino-acid* : acides aminé, *diversity* : diversité.

Chez les Cham, les loci de classe I montrent une importante diversité au niveau de l'ARS, notamment en termes d'acides aminés. HLA-DQB1 et -DRB1 ont une diversité très importante de leurs ARS, d'autant plus visible sur la diversité en acides aminés.
Comme pour les Mandenkalu, les codons ARS sont toujours plus diversifiés que les codons non-ARS, à l'exception de la région $\beta 1$ de HLA-DPB1 où les codons non-ARS montrent une diversité en acides aminés bien supérieure (0.23 ± 0.13) à celle observée pour les codons ARS (0.09 ± 0.04).
De la même façon que chez les Mandenka, les codons ARS de DPA1 et DPB1 ont une diversité nucléotidique similaire à celle des codons non ARS des autres loci (à l'exception de HLA-DRB3/4/5). HLA-DPB1, quant à lui, montre une très grande diversité en acides aminés pour les codons non-ARS, ce qui suggère une pression évolutive différente (telle qu'une sélection directionnelle agissant spécifiquement sur le site de reconnaissance de l'antigène).
Les régions ne participant pas à la liaison peptidique (TCRIR et TMR) montrent une faible diversité en général, à l'instar de ce qui est observé chez les Mandenkalu.

La Figure 2.6 met en relation la diversité nucléotidique observée dans les différentes régions des huit gènes chez les Cham (en rouge) et chez les Mandenkalu (bleu), en séparant les codons ARS (les points) des codons non-ARS (les triangles).

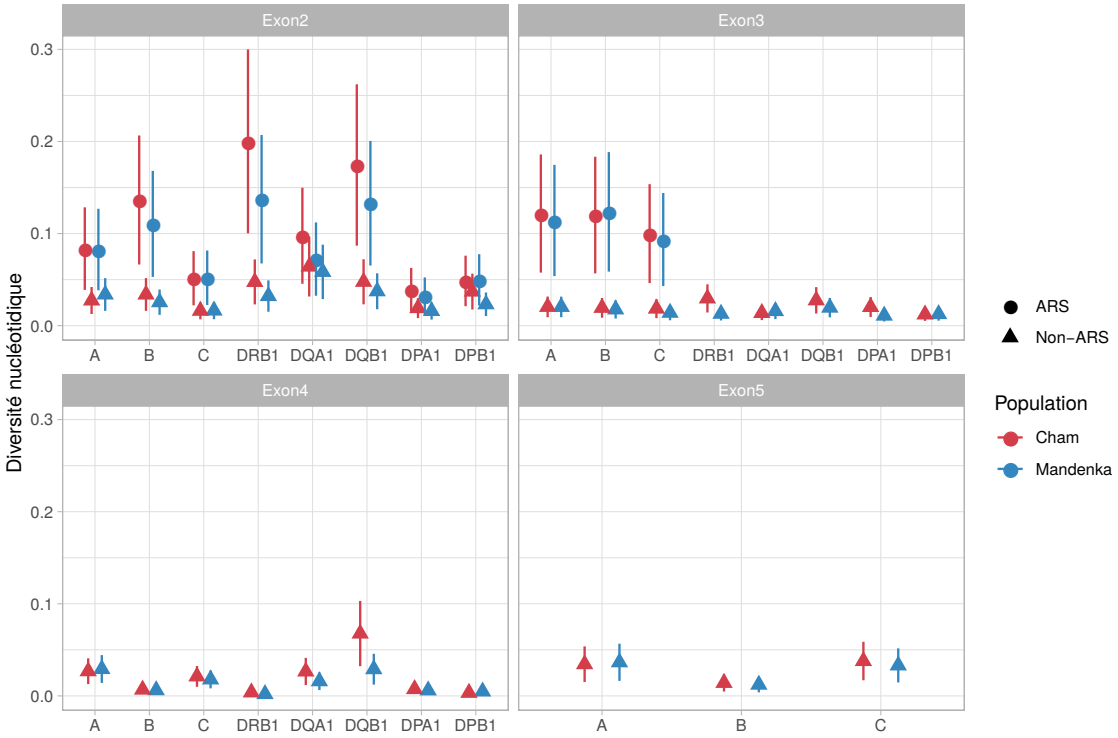


FIGURE 2.6 – Diversité nucléotidique par site ($\pi.n \pm \sigma^2$) pour les Mandenkalu (bleu) et les Cham (rouge) aux régions codant pour la région de liaison peptidique, selon l'exon codant pour ces régions. Les points représentent les codons codant pour le site de reconnaissance de l'antigène et les triangles les codons ne codant pas pour le site de reconnaissance de l'antigène.

La diversité nucléotidique est similaire entre les populations à la plupart des loci, à l'exception de HLA-DRB1 et -DQB1. Pour ces derniers, les Cham montrent une plus im-

portante diversité des ARS (pour DRB1 : $\pi.S = 0.20 \pm 0.10$ et DQB1 : $\pi.S = 0.18 \pm 0.09$) que les Mandenkalu (respectivement 0.14 ± 0.07 et 0.13 ± 0.07). Cette différence s'observe aussi pour les acides aminés de ces mêmes régions (Figures 2.4 et 2.5) où la diversité en acides aminés pour les loci DRB1 et DQB1 est de 0.37 ± 0.06 et 0.35 ± 0.05 tandis que pour les Mandenkalu ces valeurs sont, respectivement, de 0.24 ± 0.05 et 0.28 ± 0.04.

La Table 2.9 donne pour chacun des huit loci typés pour les deux populations les valeurs moyennes de π observées sur les exons 2, 3, 4 et 5 (exon 5 uniquement pour les gènes de classe I). Les autres exons n'ont pas été analysés car présentant une couverture de séquençage incomplète.

Locus	$\pi \pm \sigma^2$ (**Mandenka**)	$\pi \pm \sigma^2$ (**Cham**)
A	35.02 ± 17.97	33.21 ± 17.14
B	26.04 ± 13.63	29.97 ± 15.54
C	23.62 ± 12.54	26.13 ± 13.77
DRB1	20.38 ± 10.54	32.64 ± 16.44
DQA1	22.07 ± 11.46	25.51 ± 13.16
DQB1	25.73 ± 13.21	36.88 ± 18.57
DPA1	8.81 ± 5.12	12.63 ± 6.96
DPB1	12.07 ± 6.63	14.17 ± 7.62

TABLE 2.9 – Diversité nucléotidique moyenne par locus pour les exons 2, 3, 4 et 5 (ce dernier uniquement pour les loci de classe I) pour les Cham du Vietnam et les Mandenkalu du Sénégal.

3.5 Tests de neutralité sélective

La Table 2.10 donne, pour chacune des deux populations, les résultats du test d'Ewen-Watterson-Slatkin.

Locus	**Mandenka**	**Cham**
A	**0.005** - 0.452	0.553 - 0.929
B	**0.013** - 0.684	**0.008** - 0.646
C	**0.002** - 0.272	**0.001** - 0.247
DRB1	0.050 - 0.671	0.093 - 0.670
DQA1	0.783 - 0.935	**0.002** - 0.221
DQB1	0.425 - 0.870	**0.024** - 0.331
DPA1	0.751 - 0.920	0.623 - 0.826
DPB1	0.719 - 0.897	0.238 - 0.782

TABLE 2.10 – Valeurs du test BEWS (Bootstrapped Ewen-Watterson-Slatkin) à chaque locus pour les deux populations. Le test est significatif si l'une des deux valeurs est inférieure à 0.025 (excès d'hétérozygotes) ou supérieure à 0.975 (excès d'homozygotes). Les valeurs significatives sont en gras.

Chez les Mandenkalu, les trois loci de classe I montrent un excès d'hétérozygotes alors que les loci de classe II ne montrent pas de rejet de l'équilibre d'Ewen-Watterson-Slatkins. Les Cham montrent un profil différent. Pour les loci de classe I, HLA-B et HLA-C montrent un excès d'hétérozygotes alors que HLA-A ne montre pas de rejet de l'équilibre, tandis que pour les loci de classe II, seule la paire HLA-DQA1 et -DQB1 (codant

pour la même molécule HLA-DQ) montre un rejet de l'équilibre en faveur d'un excès d'hétérozygotes. Tous les autres loci sont à l'équilibre.

La Figure 2.7 représente les D de Tajima pour les exons 2 et 3 de chacun des huit loci HLA typés pour les deux populations. Les valeurs pour les autres exons ne sont pas significativement différentes de zéro et n'ont pas été représentées (voir annexe S-22).

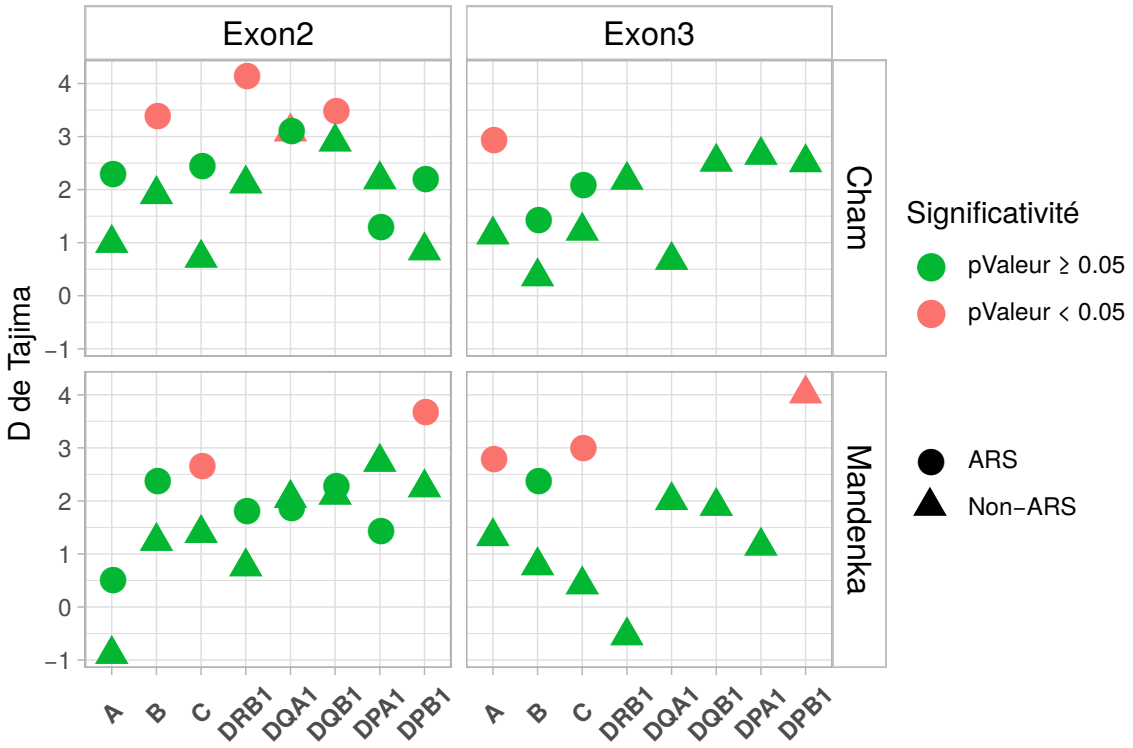


FIGURE 2.7 – Représentation graphique des indices D de Tajima pour les exons 2 et 3 des huit loci pour les deux populations. Les ronds représentent les codons ARS et les triangles les codons non-ARS. La couleur indique la significativité des D de Tajima (après correction pour tests multiples par la méthode *fdr* [Benjamini and Hochberg, 1995]), vert : pValeur > 0.05, rouge : pValeur < 0.05.

Les deux populations montrent des valeurs de D significatives principalement aux codons codant pour le site de reconnaissance de l'antigène, sauf pour les codons non-ARS de HLA-DQA1 qui sont aussi significatifs chez les Cham et l'exon 3 de HLA-DPB1 qui est significatif pour les Mandenkalu.
Les deux populations sont toutefois différentes quant aux régions avec un D significatif. Les Mandenkalu ont un D significatif aux codons ARS de HLA-A (exon 3), HLA-C (exons 2 et 3) et HLA-DPB1 (exon 2) alors que les Cham montrent un D significatif pour les codons ARS de HLA-A (exon 3), à l'instar des Mandenka, mais aussi pour HLA-B (exon 2), ainsi que pour les codons ARS des gènes de classe II HLA-DRB1, -DQA1 (où les codons non ARS sont en limite de significativité avec pValeur=0.05) et -DQB1.
Les deux populations diffèrent sur les exons 3, pour lesquels les Cham montrent un D significatif pour HLA-A alors que les Mandenkalu montrent une valeur de D significative pour HLA-DPB1.

3.6 Déséquilibres de liaison

Déséquilibre de liaison global

Les Tables 2.11 et 2.12 donnent les résultats des tests paramétriques (au-dessus de la diagonale) et non-paramétriques (en-dessous de la diagonale) de déséquilibre de liaison global pour, respectivement, la population Mandenka du Sénégal et Cham du Vietnam.

pValeurs du test paramétrique

	A	**B**	**C**	**DRB1**	**DQA1**	**DQB1**	**DPA1**	**DPB1**
A	—	1	1	1	1	0.99	0.99	1
B	1	—	1	1	1	1	1	1
C	1	1	—	0.99	0.99	0.99	0.99	1
DRB1	1	1	0.37	—	**0.002**	0.07	0.99	1
DQA1	1	1	0.6	**<0.01**	—	**<0.01**	0.99	1
DQB1	0.48	1	0.27	**<0.01**	**<0.01**	—	0.99	1
DPA1	0.79	1	0.56	0.05	0.12	0.17	—	**<0.01**
DPB1	1	1	1	1	1	1	**<0.01**	—
	A	**B**	**C**	**DRB1**	**DQA1**	**DQB1**	**DPA1**	**DPB1**

pValeurs du test non-paramétrique

TABLE 2.11 – Valeurs des tests (paramétriques et non paramétriques) de déséquilibre de liaison global entre les loci (HLA-DRB3/4/5 exclus) estimées par Gene[RATE] [Nunes, 2016] pour les Mandenkalu. Les valeurs significatives sont en gras (p<0.05)

Pour la population Mandenka, peu de loci apparaissent en déséquilibre de liaison. En effet, aucun locus de classe I n'est en déséquilibre et pour les classes II les trois gènes HLA-DRB1, -DQA1 et -DQB1 sont tous les trois en déséquilibre de liaison global (HLA-DRB1∼DQB1 n'est significatif que pour le test non paramétrique). La paire HLA-DPA1∼DPB1 est aussi en déséquilibre de liaison, mais aucun des deux ne présente de déséquilibre avec d'autres loci.

pValeurs du test paramétrique

	A	**B**	**C**	**DRB1**	**DQA1**	**DQB1**	**DPA1**	**DPB1**
A	—	1	0.99	1	1	1	1	1
B	1	—	0.99	1	1	1	1	1
C	**0.01**	**< 0.01**	—	0.98	0.89	0.99	0.98	0.99
DQA1	1	1	**0.01**	—	**< 0.01**	**< 0.01**	0.68	0.98
DQB1	1	1	**< 0.01**	**< 0.01**	—	**< 0.01**	0.99	0.99
DRB1	1	1	0.05	**< 0.01**	**< 0.01**	—	0.99	1
DPA1	1	1	0.17	**0.03**	0.39	0.45	—	**< 0.01**
DPB1	1	1	0.47	**0.01**	0.2	0.1	**< 0.01**	—
	A	**B**	**C**	**DRB1**	**DQA1**	**DQB1**	**DPA1**	**DPB1**

pValeurs du test non-paramétrique

TABLE 2.12 – Valeurs des tests (paramétriques et non paramétriques) de déséquilibre de liaison global entre les loci (HLA-DRB3/4/5 exclus) estimées par Gene[RATE] [Nunes, 2016] pour les Cham. Les valeurs significatives sont en gras (p<0.05)

En comparaison des Mandenkalu, les Cham semblent montrer beaucoup plus de déséquilibre de liaison global, avec une importante différence entre les résultats des tests paramétriques et non paramétriques.
Les tests paramétriques donnent deux groupes de loci en déséquilibre de liaison global, tous de classe II avec, d'un côté, HLA-DRB1∼DQA1∼DQB1 tous les trois en déséquilibre de liaison et, d'un autre côté, la paire HLA-DPA1∼DPB1.
Les tests non paramétriques montrent aussi ces associations, mais rapportent aussi des associations entre les classe I (notamment HLA-C en déséquilibre de liaison avec les deux autres loci de classe I) mais aussi entre un locus de classe I (HLA-C) et trois loci de classe II (HLA-DRB1, -DQA1 et -DQB1). HLA-DQA1 montre aussi du déséquilibre de liaison global avec l'ensemble des autres loci de classe II.

Déséquilibre de liaison haplotypique

La Table 2.13 donne les résultats des tests de déséquilibre de liaison haplotypique pour les six haplotypes en déséquilibre de liaison positif les plus fréquents, ainsi que pour les quatre haplotypes en déséquilibre de liaison négatif chez les Mandenkalu. La liste complète des haplotypes en déséquilibre de liaison significatif est disponible en annexe S-23.1.

Haplotype	F. obs	F. exp	D	stdres
DQA1*05:05:01∼DQB1*03:19	0.430	0.228	0.202	3.963
DQA1*05:05:01∼DRB1*13:04	0.282	0.148	0.140	3.609
DQB1*03:19∼DRB1*13:04	0.259	0.119	0.140	4.092
DPA1*02:01:01∼DRB1*13:04	0.206	0.117	0.090	2.418
DPA1*02:01:01∼DPB1*131:01	0.201	0.097	0.104	3.198
DPA1*02:01:01∼DPB1*17:01	0.201	0.097	0.104	3.198
DPA1*02:01:01∼DQA1*01:02:01	0	0.071	-0.070	-2.531
DPA1*02:01:01∼DPB1*01:01:01	0	0.063	-0.063	-2.459
DPA1*02:01:01∼DPB1*02:01:02	0	0.063	-0.063	-2.459
DQA1*01:02:01∼DQB1*03:19	0	0.056	-0.056	-2.455

TABLE 2.13 – Haplotypes en déséquilibre de liaison positif les plus fréquents (en vert, stdres$>$ 1.960 et D$>$ 0.050) ou négatif (bleu, stdres$<$ -1.960 et D $<$ -0.050) dans la population Mandenka aux loci étudiés. F.obs et F.exp sont respectivement les fréquences observées et attendues (produit des fréquences alléliques individuelles) pour les haplotypes, D représente la valeur du D de déséquilibre de liaison (différence entre les fréquences observées et attendues) et stdres donne les valeurs des résidus standardisés qui doivent être $>$ 1.960 (resp. $<$ -1.960) pour représenter un déséquilibre de liaison positif (resp. négatif) significatif.

Les haplotypes les plus fréquents sont tous des haplotypes de classe II et représentent un seul haplotype étendu, de HLA-DRB1 à HLA-DPA1 : DRB1*13:04∼DQA1*05:05:01∼DQB1*03:19∼DPA1*02:01:01. Deux allèles HLA-DPB1 (HLA-DPB1*17:01 et *131:01) sont aussi en déséquilibre de liaison avec HLA-DPA1*02:01:01, étendant cet haplotype étendu de classe II de HLA-DRB1 à -DPB1. Ces deux allèles HLA-DPB1 partagent un même exon 2, suggérant qu'ils détectent les mêmes peptides antigéniques.
Cet haplotype est composé de chacun des allèles les plus fréquents à chaque locus. Sur les 45 individus non apparentés typés pour les 5 loci de classe II, 14 portent cet haplotype

(soit 30% des individus) alors que la fréquence attendue de cet haplotype n'est que de 1.18% (selon le produit des fréquences individuelles calculées par Gene[Rate] [Nunes, 2016] pour ces 43 individus), soit moins de un individu sur les 45 typés pour les 5 loci de classe II.

La Table 2.14 donne les résultats des tests de déséquilibre de liaison haplotypique pour les six haplotypes en déséquilibre de liaison positif les plus fréquents, ainsi que pour les quatre haplotypes en déséquilibre de liaison négatif chez les Cham. La liste complète des haplotypes en déséquilibre de liaison significatif est disponible en annexe S-23.2.

Haplotype	F. obs	F. exp	D	stdres
DPA1*01:03:01∼DPB1*04:01:01	0.167	0.077	0.090	3.465
DQA1*01:01:01∼DRB1*15:02:01	0.163	0.037	0.127	6.426
DQA1*01:01:01∼DQB1*05:01:01	0.143	0.028	0.115	6.642
DQB1*05:01:01∼DRB1*15:02:01	0.123	0.035	0.088	5.069
DPA1*01:03:01∼DPB1*03:01:01	0.123	0.063	0.060	2.562
DQA1*02:01∼DRB1*07:01:01	0.122	0.015	0.107	8.622
DRB1*15:02:01∼DPA1*01:03:01	0.022	0.087	-0.065	-2.336
DPA1*01:03:01∼DPB1*13:01:01	0	0.077	-0.077	-2.973

TABLE 2.14 – Haplotypes en déséquilibre de liaison positif les plus fréquents (en vert, stdres$>$ 1.960 et D$>$ 0.050) ou négatif (bleu, stdres$<$ -1.960 et D$<$ -0.050) dans la population Cham aux loci étudiés. F.obs et F.exp sont respectivement les fréquences observées et attendues (produit des fréquences alléliques individuelles) pour les haplotypes, D représente la valeur du D de déséquilibre de liaison (différence entre les fréquences observées et attendues) et stdres donne les valeurs des résidus standardisés qui doivent être > 1.960 (resp. < -1.960) pour représenter un déséquilibre de liaison positif (resp. négatif) significatif.

Les haplotypes en déséquilibre de liaison positif peuvent être regroupés dans deux groupes d'haplotypes : un premier groupe d'haplotypes (en gras dans la Table), correspondant à l'haplotype étendu DRB1*15:02:01∼DQA1*01:01:01∼DQB1*05:01:01, et un second groupe d'haplotypes incluant DPA1*01:03:01∼(DPB1*03:01:01 ou DPB1*04:01:01) et DRB1*07:01:01∼DQA1*02:01, les deux derniers allèles étant aussi en déséquilibre de liaison avec DPA1*01:03:01 (résultats non montrés sur la Table 2.14, se référer à l'annexe S-23).

3.7 Analyses en composantes principales

Les Figures 2.8 à 2.10 représentent les analyses en composantes principales réalisées sur les indices de diversité de, respectivement, la population Mandenka, la population Cham et les deux populations ensemble (pour comparaison). La Figure 2.8 est extraite de la publication [Goeury et al., 2018a].

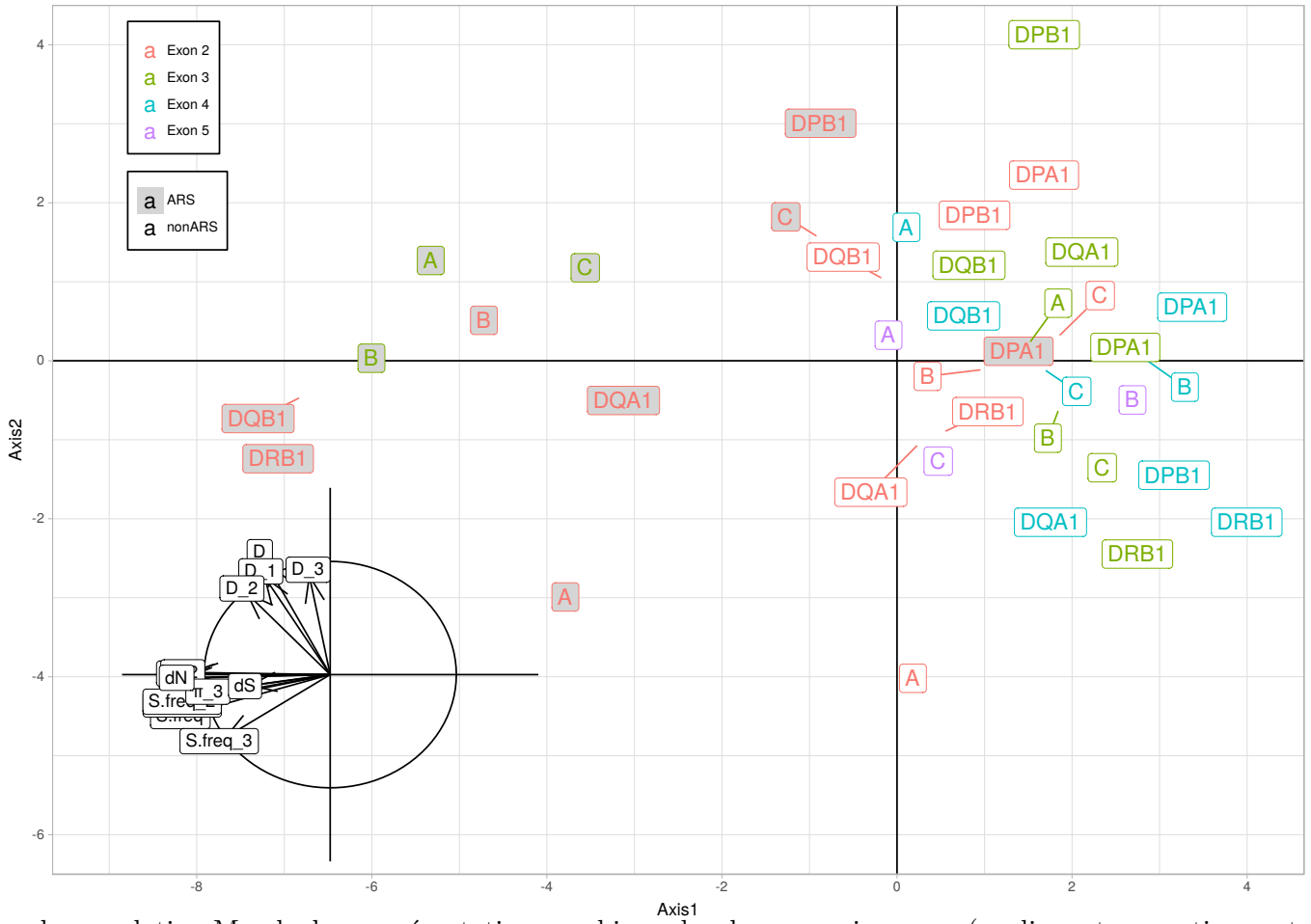


FIGURE 2.8 – Pour la population Mandenka, représentation graphique des deux premiers axes (expliquant respectivement 61% et 19% de la variance totale) de l'analyse en composantes principales basée sur le D de Tajima (D), la diversité nucléotidique par site (π), la fréquence de sites ségrégeants (S.freq), la fréquence des mutations synonymes (dN) et non-synonymes (dS) aux exons 2 (en rouge), 3 (vert) et 4 (bleu) des loci HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1, -DPB1 et exons 5 (violet) des loci HLA-A, -B, -C Les étiquettes sur fond gris correspondent aux codons codant pour les ARS et les étiquettes sur fond blanc aux codons non-ARS. Les notations _1, _2 et _3 correspondent à ces même statistiques mais calculées pour le premier, second et troisième nucléotide de chaque codon. Le graphique en bas à gauche représente les corrélations entre les projections des variables (la corrélation entre deux variables est donnée par le cosinus de l'angle formé par les vecteurs de chaque variable). Figure issue de [Goeury et al., 2018a].

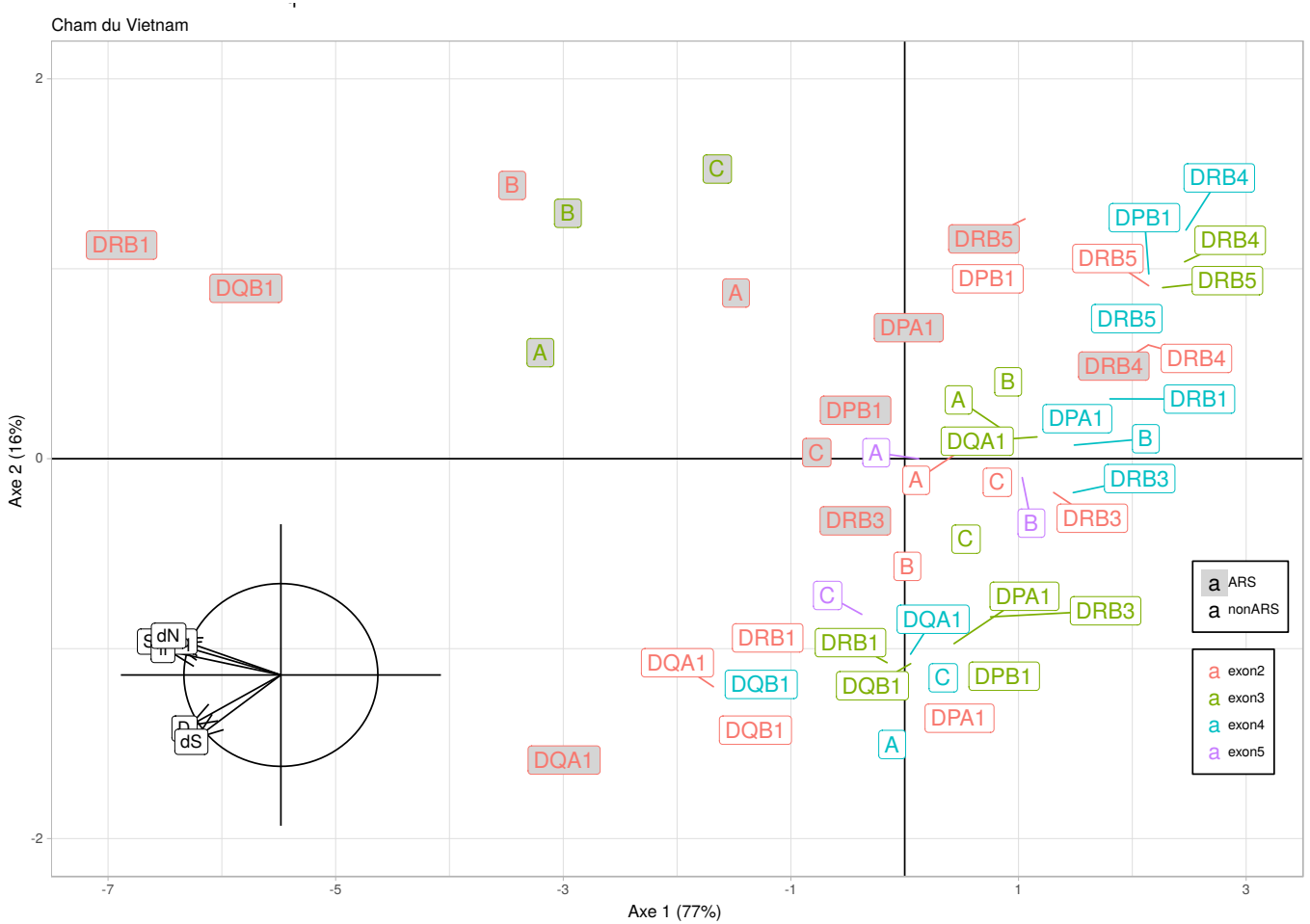


FIGURE 2.9 – Pour la population Cham du Vietnam, représentation graphique des deux premiers axes (expliquant respectivement 77% et 16% de la variance totale) de l'analyse en composantes principales basée sur le D de Tajima (D), la diversité nucléotidique par site (π), la fréquence de sites ségrégeants (S.freq), la fréquence des mutations synonymes (dN) et non-synonymes (dS) aux exons 2 (en rouge), 3 (vert) et 4 (bleu) des loci HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1, -DPB1 et exons 5 (violet) des loci HLA-A, -B, -C. Les étiquettes sur fond gris correspondent aux codons codant pour les ARS et les étiquettes sur fond blanc aux codons non-ARS. Le graphique en bas à gauche représente les corrélations entre les projections des variables (la corrélation entre deux variables est donnée par le cosinus de l'angle formé par les vecteurs de chaque variable).

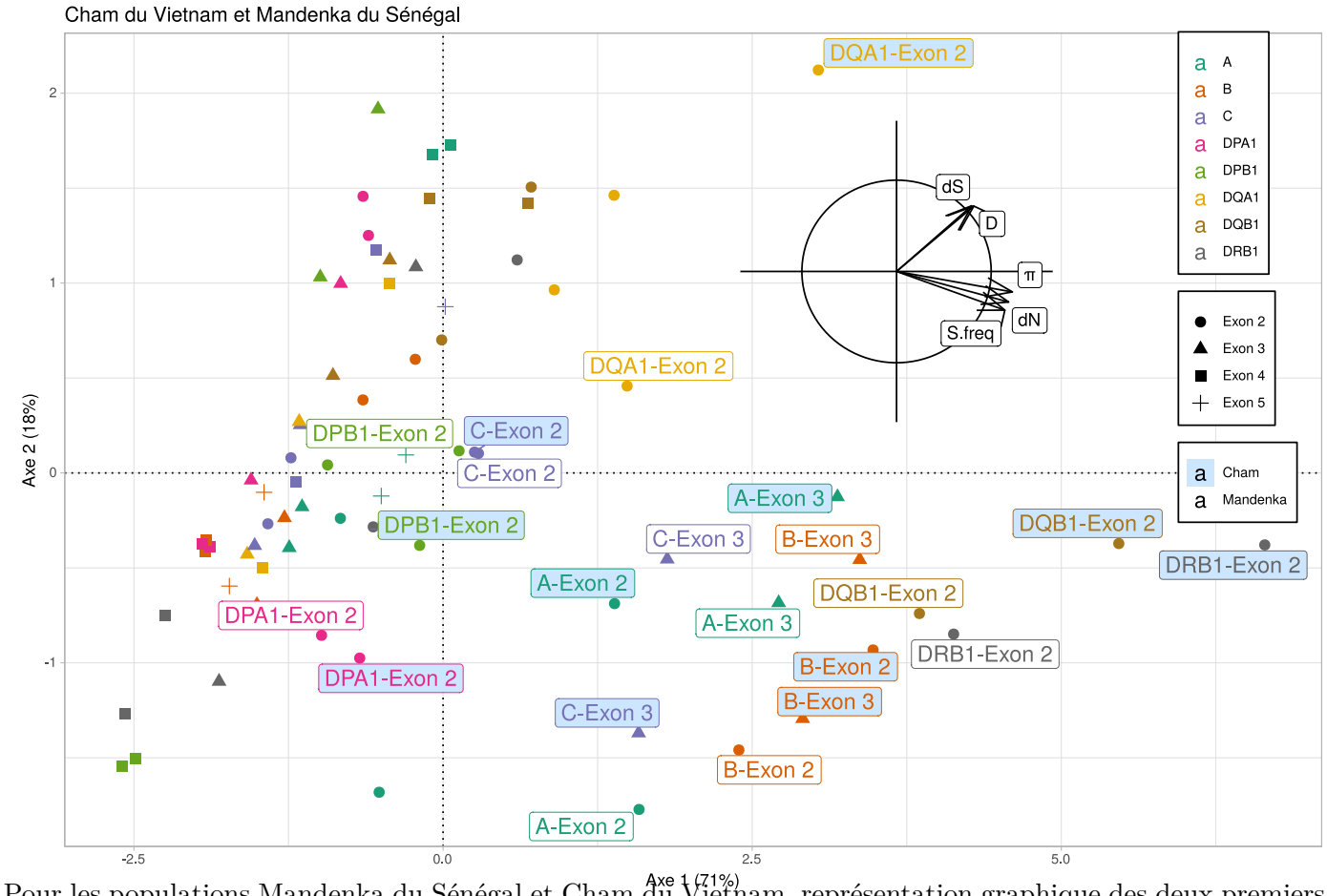


FIGURE 2.10 – Pour les populations Mandenka du Sénégal et Cham du Vietnam, représentation graphique des deux premiers axes (expliquant respectivement 71% et 18% de la variance totale) de l'analyse en composantes principales basée sur le D de Tajima (D), la diversité nucléotidique par site (π), la fréquence de sites ségrégeants (S.freq), la fréquence des mutations synonymes (dN) et non-synonymes (dS) aux exons 2 (ronds), 3 (triangles) et 4 (carrés) des loci HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1, -DPB1 et exons 5 (signe +) des loci HLA-A, -B, -C. Les étiquettes correspondent aux codons codant pour les ARS, les étiquettes à fond bleu correspondent à la population Cham et les étiquettes à fond blanc à la population Mandenka. Le graphique en haut à droite représente les corrélations entre les projections des variables (la corrélation entre deux variables est donnée par le cosinus de l'angle formé par les vecteurs de chaque variable).

Afin d'interpréter les résultats des tests de neutralité sélective et les indices de diversité, trois analyses en composantes principales (ACP, visibles sur les Figures 2.8, 2.9 et 2.10) ont été réalisées sur les exons 2, 3, 4 et 5 (ce dernier uniquement pour les classe I), en séparant les codons ARS des codons non ARS pour les exons 2 (classe I et II) et 3 (uniquement classe I). Cinq statistiques liées à des tests de neutralité sélective ou de la diversité moléculaire ont été utilisées : le D de Tajima, la diversité nucléotidique π (moyenne par site), la fréquence des sites ségrégeants (S.freq), la fréquence des mutations synonymes (dN) et non-synonymes (dS). Le rapport dN/dS n'a pas été utilisé car certaines régions possèdent un dS de zéro (HLA-DPB1-Exon4 pour les deux populations et les codons ARS de HLA-DPB1-Exon2 chez les Mandenka) et parce que le dN et le dS étaient déjà présents dans l'analyse. Les graphiques en bas à gauche (Figures 2.8 et 2.9) et en haut à droite (Figure 2.10) représentent les projections des variables dans le nouveau référentiel de l'ACP, la corrélation entre deux variables étant donnée par le cosinus de l'angle formé par les deux vecteurs. Ces graphiques permettent d'interpréter les différentes composantes principales en termes de variables qui ont participé à leur construction.

Les Tables 2.15 à 2.17 fournissent les valeurs d'inertie de chacune des composantes (ou axes) des ACP. Les deux premiers axes expliquent, respectivement, 79, 93 et 88% des données, tandis que les autres axes n'expliquent que 21, 7 et 12% (dont 11, 5 et 11% pour les seuls axes 3). Ainsi, seuls les trois premiers axes ont été conservés pour chaque analyse. Les axes 1 et 2 de chaque analyse sont représentés ici tandis que les représentations des axes 1 et 3 sont disponibles en annexes S-25.1 à S-25.3.

	Inertie - ACP « Mandenka »		
	Valeur	**Cumulée**	**Cumulée (%)**
Composante 1	8.504	8.504	60.74
Composante 2	2.612	11.116	79.40
Composante 3	1.556	12.672	90.51
Composante 4	0.447	13.119	93.71
Composantes 5 à 14	0.881	14.000	100.00

TABLE 2.15 – Distribution de l'inertie (variance) le long des 14 composantes principales de l'analyse réalisée sur les données des Mandenkalu.Les composantes 5 à 14, étant peu informatives, ont été regroupées dans la ligne « Composantes 5 à 14 ».

	Inertie - ACP « Cham »		
	Valeur	**Cumulée**	**Cumulée (%)**
Composante 1	3.860	3.860	77.19
Composante 2	0.814	4.674	93.47
Composante 3	0.266	4.940	98.80
Composante 4	0.053	4.992	99.85
Composante 5	0.007	5.000	100.00

TABLE 2.16 – Distribution de l'inertie (variance) le long des cinq composantes principales de l'analyse réalisée sur les données des Cham.

	Inertie - ACP « Cham et Mandenka »		
	Valeur	**Cumulée**	**Cumulée (%)**
Composante 1	3.527	3.527	70.54
Composante 2	0.878	4.406	88.10
Composante 3	0.537	4.943	98.85
Composante 4	0.050	4.992	99.84
Composante 5	0.008	5.000	100.00

TABLE 2.17 – Distribution de l'inertie (variance) le long des cinq composantes principales de l'analyse réalisée sur les données des deux populations.

Les Table 2.18 à 2.20 donnent les contributions de chaque variable quant à la construction des axes, c'est-à-dire comment se répartit la variance des variables le long des axes, les axes 4 et 5 n'étant pas montrés. Ces Tables, associées aux graphiques des Figures 2.8 à 2.10, permettent d'interpréter les ACP.

	Composante 1	**Composante 2**	**Composante 3**
S.freq	0.903	0.086	0.000
D	0.199	0.749	0.005
π	0.975	0.001	0.008
$S.freq\_1$	0.880	0.045	0.020
D_1	0.194	0.532	0.001
π_1	0.925	0.000	0.032
$S.freq\_2$	0.882	0.037	0.028
D_2	0.314	0.370	0.116
π_2	0.874	0.000	0.083
$S.freq\_3$	0.496	0.221	0.174
D_3	0.020	0.549	0.250
π_3	0.602	0.015	0.311
dN	0.948	0.001	0.038
dS	0.292	0.007	0.488

TABLE 2.18 – Contribution de chaque variable utilisée dans la construction des trois composantes retenues pour l'ACP réalisée sur les données des Mandenkalu.

	Composante 1	**Composante 2**	**Composante 3**
S.freq	0.838	0.115	0.019
D	0.427	0.328	0.243
π	0.960	0.032	0.001
dN	0.898	0.072	0.011
dS	0.405	0.332	0.263

TABLE 2.19 – Contribution de chaque variable utilisée dans la construction des trois composantes retenues pour l'ACP réalisée sur les données des Cham.

	Composante 1	**Composante 2**	**Composante 3**
S.freq	0.879	0.085	0.005
D	0.631	0.220	0.148
π	0.946	0.048	0.000
dN	0.858	0.118	0.001
dS	0.545	0.342	0.112

TABLE 2.20 – Contribution de chaque variable utilisée dans la construction des trois composantes retenues pour l'ACP réalisée sur les données des Mandenkalu et des Cham.

L'ACP représentée sur la Figure 2.10 étant la plus complète et intégrant les données des deux populations, elle servira de point de référence pour les interprétations et comparaisons des résultats pour les deux populations.

La première composante principale (71% de la variance expliquée) est corrélée à la diversité moléculaire (π, S.freq et dN). Ainsi les régions à droite du graphique sont caractérisées par une plus grande diversité nucléotidique (π et S.freq) mais aussi fonctionnelle (dN), tandis que les régions à gauche du graphique sont nettement moins diversifiées. La seconde composante principale (18% le la variance expliquée) semble surtout liée aux D de Tajima et devrait donc discriminer les régions évoluant sous sélection balancée (en haut) des régions évoluant sous sélection purificatrice (en bas). Cette interprétation est à nuancer, car le D de Tajima est aussi influencé par la démographie et diminue en cas d'expansion démographique [Tajima, 1989b, Tajima, 1989a].
Il apparaît ainsi que le premier axe sépare principalement les codons ARS (à l'exception de ceux des exons 2 de HLA-C, -DQA1 et -DPA1) des deux populations sur la droite, des autres régions sur la gauche. Ces résultats confirment que la plus grande part de la variabilité moléculaire HLA est concentrée dans les ARS, cette variabilité étant directement liée à la contrainte fonctionelle de présentation des peptides antigéniques. Les régions les plus à droite sur ce graphique sont les ARS portés par les exons 2 de HLA-DRB1 et -DQB1, similaire à la Figure 2.6. Il apparaît que ces deux régions sont plus diversifiées chez les Cham que chez les Mandenkalu.
Les régions les plus extrêmes (en haut) sur le second axe présentent toutes un D de Tajima significativement supérieur à 0 (voir Figure 2.7). La position de l'exon 2 de HLA-A (aussi bien les codons ARS que non ARS) pour les Mandenkalu traduit un D très faible (D=0.57 pour les codons ARS et -0.87 pour les codons non ARS), ce qui pourrait être interprété comme de la sélection purificatrice, mais aussi une marque d'expansion démographique.

4 Discussion

4.1 Résumé de l'étude effectuée

Le but de cette étude était de caractériser et comparer la diversité génétique aux loci HLA des Mandenkalu du Sénégal et des Cham du Vietnam en utilisant des typages à haute résolution (séquençage des gènes complets).
Les mêmes individus Mandenka ayant déjà été génotypés pour les mêmes loci dans de précédentes études utilisant des sondes PCR-SSO ou le séquençage des exons 2, les typages obtenus par les trois méthodes ont été comparés afin d'évaluer ce que les nouvelles techniques de typage HLA apportent aux études de génétique des populations.
Dans une seconde partie, les résultats de séquençage haute résolution de 8 à 11 loci HLA pour ces deux populations ont été analysés dans un contexte de génétique des populations, afin de caractériser les processus évolutifs et évènements démographiques qui ont façonné leur diversité moléculaire.

4.2 Apports de 25 ans d'évolution de techniques de typage

Des précédentes études comparant des techniques de typage ont été réalisées, notamment en 2011 par Erlich *et al.* [Erlich et al., 2011], comparant des typages PCR-SSO et NGS-454, puis en 2013 par Major *et al.* [Major et al., 2013] où des typages PCR-SSO ont été évalués contre des typages NGS-MiSeq.
En appliquant les mêmes critères de comparaison (concordance entre les typages uniquement si les deux allèles d'un individu sont identiques, ou du moins compatibles selon la résolution), la dernière étude ([Major et al., 2013]) a montré des résultats légèrement différents de la présente étude (84.7%, 93.9% et 85,6% pour HLA-A, -B et -C, respectivement, contre 97.6%, 91.5% et 78.8% dans cette étude, Table 2.6).
Plusieurs allèles n'avaient pas été découverts à l'époque des typages PCR-SSO et n'étaient donc pas identifiables. Parmi eux, A*11:50Q, 01:11N et 03:21N, rapportés pour la première fois entre 2005 et 2009, ont été identifiés par Major *et al.* avec les typages NGS-MiSeq mais par PCR-SSO, les auteurs se basant sur les données de HapMap [The International HapMap Consortium, 2003] (décrites en 2003). Cela pourrait expliquer pourquoi la proportion de correspondances trouvées dans leur étude est plus faible que pour la comparaison ici effectué pour les Mandenka.
Chez les Mandenkalu, C*07:18 (rapporté pour la première fois en 2003 [Delfino et al., 2003]) a été identifié à tort comme C*07:01 par PCR-SSO (FA=13% selon les typages PCR-SSO), ce qui explique que cette étude trouve de plus bas scores dans la comparaison des typages PCR-SSO/NGS-MiSeq pour HLA-C par rapport à l'étude menée par Major *et al.*.
Ces premiers résultats illustrent le fait qu'à l'époque des typages PCR-SSO, les amorces et les sondes utilisées pour identifier les allèles étaient majoritairement développées pour des allèles fréquents en Europe (et les populations d'ancestralité européenne) et se montraient alors moins adaptées pour des populations africaines.

Les allèles HLA de classe II sont maintenant définis en prenant en compte la variabilité nucléotidique hors de l'exon 2 (notamment l'exon 3, voir Chapitre 5). Les typages PCR-SSO réalisés dans les années 90 ne ciblaient que l'exon 2 (pour les gènes de classe II), ceci explique le nombre d'allèles qui restaient à découvrir à l'époque (en prennant en compte la variabilité en dehors des exons 2). La Figure 2.11 montre, pour une partie des allèles

les plus fréquents à chacun des trois loci co-typés par PCR-SSO et NGS-MiSeq les dates de découverte de ces allèles, en les situant par rapport aux trois techniques de typage employées dans cette étude.

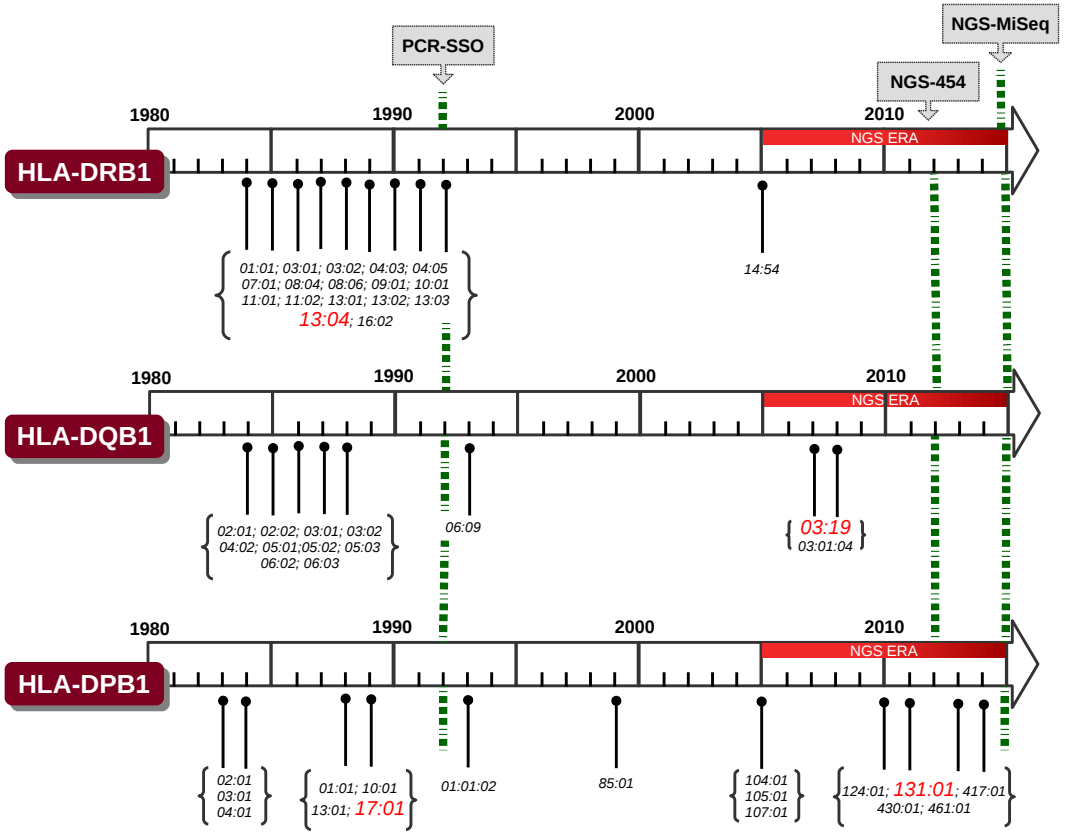


FIGURE 2.11 – Frise chronologique représentant les dates (source : `http://hla.alleles.org`) de découverte (premiers signalements) des allèles de HLA-DRB1, -DQB1 et -DPB1 et identifiés chez les Mandenkalu (quelques allèles de très faibles fréquences ne sont pas montrés). Les boîtes grises et les lignes en pointillés verts rappellent les différentes dates et méthodes de typage employées dans cette étude, les allèles en rouge étant les plus fréquents à chacun de ces trois loci. Les bandeaux rouges « NGS-ERA » marquent l'arrivée des séquenceurs ADN à haut débit.

Par exemple, l'allèle HLA-DQB1 le plus fréquent chez les Mandenkalu, HLA-DQB1*03:19 (FA=44%), a seulement été décrit pour la première fois en 2007 [Witter et al., 2007] et diffère de -DQB1*03:01 (identifié par PCR-SSO) d'un SNP sur l'exon 3.
Concernant les faibles scores de correspondance pour HLA-DPB1, une explication est que la moitié des allèles identifiés par NGS-MiSeq chez les Mandenkalu (incluant HLA-DPB1*131:01, FA=19%) étaient inconnus avant 1999, leurs fréquences cumulées atteignant 33%. Au contraire, sur les 21 allèles HLA-DRB1 identifiés par NGS-MiSeq, seuls deux allèles de faibles fréquences (HLA-DRB1*12:10 et -DRB1*14:54:01, fréquence cumulée de 1.15%) n'étaient pas encore découverts lors des typages PCR-SSO.
En conclusion, ces résultats montrent que les typages NGS-MiSeq (et par extension tout type de typage basé sur du séquençage de gènes complets) donnent accès à une plus grande information moléculaire, en particulier pour les loci de classe II HLA-DQB1

et HLA-DPB1 qui possèdent une part non négligeable de leur diversité hors de l'exon 2. Cette étude montre qu'une attention particulière doit être portée à l'interprétation de résultats basés sur des méthodes de typage limitées, telles que la PCR-SSO ou le séquençage des exons 2, la dernière permettant toutefois de ne pas perdre une quantité d'information trop significative si l'on considère l'ensemble des allèles possibles pour une séquence d'exons 2 au lieu de chercher à assigner un seul allèle nominal. Cela peut par contre générer des ambiguïtés de typage, qui doivent être précisées dans la communication des résultats et peuvent être prises en compte par des logiciels tels que Gene[Rate] et sa structure de données *UNIFORMATE* [Nunes, 2007].

4.3 Comparaison des populations

Les deux populations comparées ont été choisies car elles ont une origine et une histoire démographique très différentes. Les Mandenkalu représentent un important groupe ethnique d'Afrique de l'ouest (de 10 à 20 millions d'individus) parlant une langue de la famille niger-congo, tandis que les Cham représentent une plus petite population (autour de 500'000 individus) parlant une langue de la famille austronésienne.
Dans notre étude, plusieurs différences génétiques sont observées entre ces populations. Si les deux populations montrent une plus grande diversité des codons ARS, celle-ci est plus importante chez les Cham au niveau des exons 2 de HLA-DRB1 et -DQB1, et une part de cette diversité nucléotidique se retrouve aussi d'un point de vue fonctionnel, avec une plus grande diversité des acides-aminés de ces deux régions. Ce premier résultat est particulièrement intéressant, les Mandenkalu étant une population africaine, une plus grande diversité aurait pu être attendue chez cette dernière [Tishkoff et al., 2009, Choudhury et al., 2018].

Ces deux populations montrent aussi une différence dans les résultats des tests d'Ewen-Watterson-Slatkin et du D de Tajima. Le premier test (Table 2.10), basé sur les distributions de fréquences alléliques, montre un excès d'hétérozygotes aux trois loci de classe I chez les Mandenkalu et une neutralité aux loci de classe II, tandis que les Cham montrent un excès d'hétérozygotes à deux loci de classe I (HLA-B et -C) mais aussi deux loci de classe II HLA-DQA1 et -DQB1, codant tous les deux pour la molécule HLA-DQ. Quant au test du D de Tajima, (Figure 2.7) basé sur l'information moléculaire, il montre des valeurs de D très élévées chez les Cham pour les codons ARS de HLA-A (exon 3) -B (exon 2), -DRB1, -DQA1 et -DQB1, mais aussi chez les Mandenkalu aux codons ARS de HLA-A et -B (exon 3), -C (exon 2) et -DPB1 .

Des différences sont aussi retrouvées au niveau du déséquilibre de liaison où les Mandenkalu montrent moins de déséquilibre de liaison (haplotypique et global, voir Tables 2.11 et 2.13), résultat attendu pour une population africaine de grande taille et ne montrant pas de traces de décroissance démographique (voir page 53) et le déséquilibre global ne concerne que des gènes de classe II. Au contraire, la population Cham montre plus de paires d'allèles en déséquilibre de liaison, incluant des allèles de gènes de classe I (Table 2.14). Sachant que les Cham sont une population de taille plus réduite que les Mandenkalu (autour de 500'000 individus selon [Eberhard et al., 2019]) et montrant un signal de contraction démographique il y a ∼700 ans (selon une étude de 2017 basée sur 55'000 SNP sur l'ensemble du génome [Pischedda et al., 2017]), ce déséquilibre de liaison pourrait donc être expliqué par la dérive génétique plus rapide de cette population due à leur plus petit effectif.
Toutefois, les tests de déséquilibre de liaison global (DLg) pour les Cham (Table 2.12)

montrent des différences entre les types de tests. Onze paires de loci (sur les 28 testées) sont en DLg avec les tests paramétriques, tandis que les tests non-paramétriques ne montrent que quatre paires de loci en DLg (aucun locus de classe I n'étant impliqué). Cette différence peut s'expliquer par la façon dont ces tests sont réalisés, le test paramétrique montrant une perte de puissance statistique lorsque le nombre d'haplotypes testés est élevé (voir Chapitre 1, page 40). Le nombre d'haplotypes testés étant de 51 pour HLA-DQB1∼HLA-DPA1 à 161 pour HLA-B∼HLA-DPB1, la différence entre les tests paramétriques et non-paramétriques est alors probablement due à cette perte de puissance des tests paramétriques.

Des similarités sont toutefois retrouvées entre ces populations, notamment au niveau des allèles fréquents dont cinq sont partagés : HLA-A*33:03:01, -DQA1*01:02:01, -DQB1*02:01:01, -DPA1*02:01:01 et -DPA1*01:03:01. La Figure 2.12 montre la distribution des trois premiers de ces allèles (la source ne disposant pas de données pour HLA-DPA1) à l'échelle mondiale.

Même si les données de certains loci sont incomplètes pour les deux zones d'étude (Sénégal et Vietnam), il apparaît que ces allèles communs sont des allèles fréquement retrouvés à l'échelle mondiale (exception faite des Amériques). Cette observation permet donc d'exclure l'hypothèse d'une origine commune ou de flux géniques entre ces populations. Toutefois, et étant donné la fonction première d'immunité adaptative des molécules HLA (voir le Chapitre d'introduction, page 14), on ne peut exclure l'hypothèse d'une convergence évolutive due à des pressions de pathogènes similaires dans l'environnement. En effet, chacun de ces allèles est parmi les plus fréquemment observés à ces loci, avec une fréquence de 10.00% pour HLA-DQB1*02:01:01 chez les Mandenkalu à 46.08% pour -DPA1*02:01:01 chez les Mandenkalu (voir l'annexe S-26 pour les distributions de fréquences alléliques complètes pour les deux populations). Ces fréquences élevées pourraient alors être expliquées par un effet de sélection due à un ou plusieurs pathogènes communs.
L'étude menée par Sanchez-Mazas *et al.* en 2017 sur la corrélation entre la prévalence de la malaria (due au parasite *P. falciparum*) et la fréquence des allèles de classe I en Afrique a identifié, outre les deux allelèles HLA-B*53:01 et HLA-B*78:01, HLA-A*74:01 comme potentiel allèle protecteur vis-à-vis de la malaria [Sanchez-Mazas et al., 2017]. La molécule HLA-A*33:01 montre une similarité de liaison peptidique avec -A*74:01, liaison d'une arginine dans la poche 9 de l'ARS, suggérant une fonction similaire. Or, le Vietnam est une zone géographique touchée par la malaria (présence des parasites *P. falciparum* et *P. vivax* [Bhatt et al., 2015]). Il est donc possible que HLA-A*33:01 soit sous sélection par la malaria chez les Cham du Vietnam, expliquant alors sa fréquence élevée (15%).

4.4 Signatures de la sélection naturelle et de la démographie sur les régions géniques

Les Mandenkalu vivant dans une région où plusieurs maladies infectieuses telles que la malaria sont fortement prévalentes, il était attendu que cet environnement pathogénique influence le polymorphisme de leurs loci HLA. Une précédente étude menée par Sanchez-Mazas *et al.* [Sanchez-Mazas et al., 2017] a mis en évidence l'effet de l'allèle B*53:01:01 comme protecteur vis-à-vis du *Plasmodium falciparum*. Cet allèle n'est étonnamment pas retrouvé à de très hautes fréquences chez les Mandenkalu (FA [13]=6%), mais d'autres

13. Fréquence Allélique

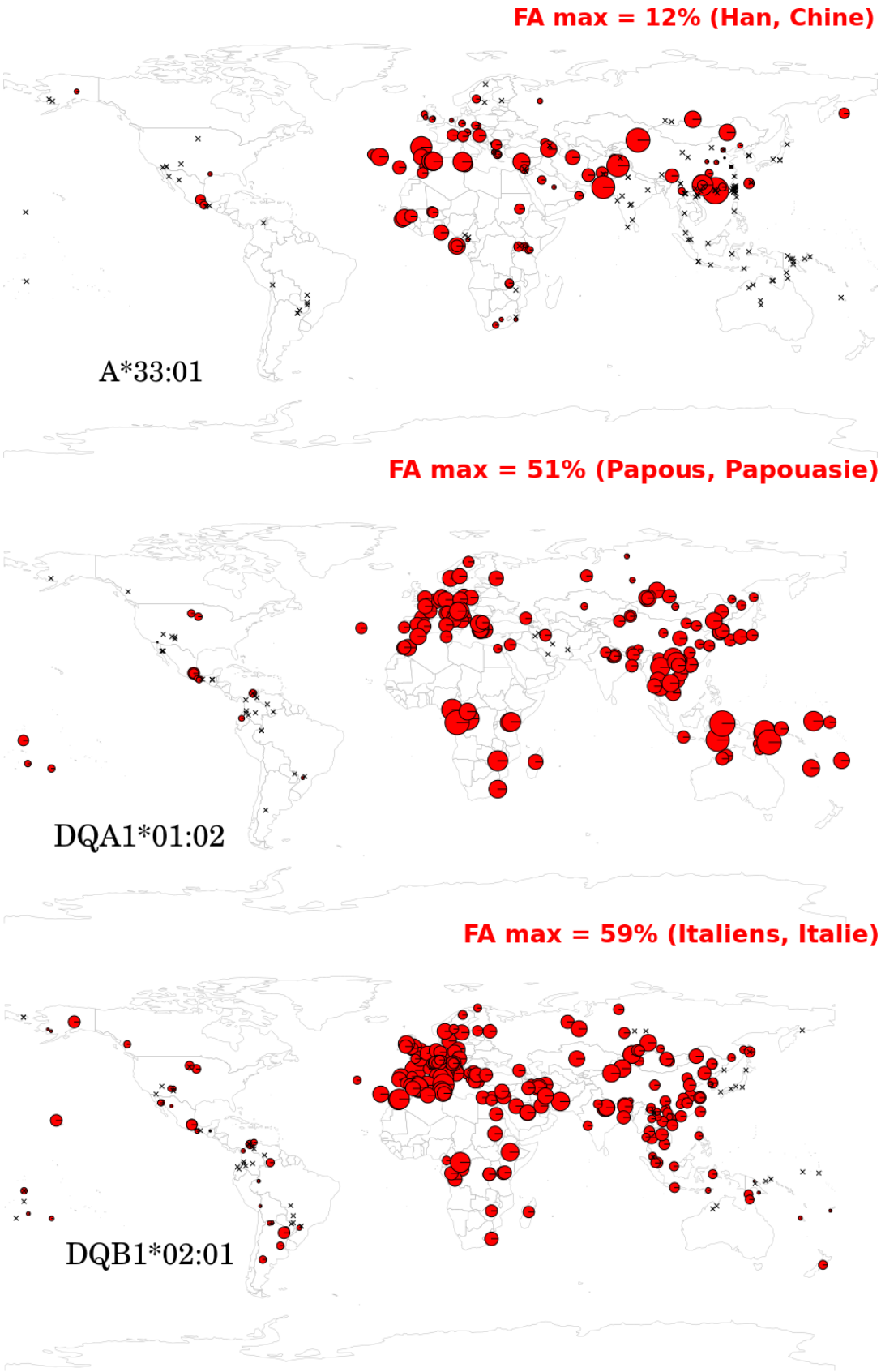


FIGURE 2.12 – Cartes mondiales des allèles HLA fréquents (>10%) et partagés entre les Mandenkalu du Sénégal et les Cham du Vietnam dans cette étude. De haut en bas : HLA-A*33:01, DQA1*01:02 et DQB1*02:01. Les tailles des cercles sont proportionnelles à la fréquence maximale de l'allèle (indiqué par FA max), les X indiquent les populations échantillonnées chez lesquelles cet allèle n'a pas été detecté. Source : `https://hla-net.eu/interactive/HLA_map/`.

allèles observés chez les Mandenkalu, incluant B*35:01:01 (l'allèle HLA-B le plus fréquent, FA=16%), a été rapporté comme protecteur de la malaria dans une étude menée au Ghana [Yamazaki et al., 2011] et possède un profil de liaison peptidique similaire à B*53:01:01 [Sanchez-Mazas et al., 2017], suggérant un probable effet protecteur de cet allèle. Deux autres allèles HLA-B parmi les plus fréquents chez les Mandenkalu (B*15:03:01 et B*78:01:01, tous deux à 8% de fréquence) posséderaient de même un profil de liaison aux peptides similaire, indiquant que ces allèles seraient interchangeables en termes de protection contre les pathogènes.

Un résultat particulièrement intéressant concerne l'unique haplotype de classe II DRB1*13:04∼DQA1*05:05:01∼DQB1*03:19 (auquel on pourrait rajouter ∼DPB1*17:01 ou DPB1*131:01). Cet haplotype de classe II, fréquent et en fort déséquilibre de liaison chez les Mandenkalu, est étonnant dans une population n'ayant pas subi de dérive génétique rapide (voir page 53).
DRB1*13:04, déjà identifié comme un allèle prédominant en Afrique de l'ouest (voir [Tiercy et al., 1992, Hill et al., 1992a] et le Chapitre 4), a été décrit comme pouvant résulter d'une conversion allélique à partir de DRB1*11:02.
Sur les 1'913 allèles HLA-DRB1 étudiés (résolution au troisième champ, provenance : IPD-IMGT/HLA v3.24.0), 86 ne montraient de différences avec DRB1*13:04 que sur l'exon 2 (de 4 à 22 substitutions). La taille du fragment portant ces substitutions (de la première à la dernière) s'étend de 30 à 241 paires de bases, sauf pour DRB1*11:02:01 où les cinq substitutions identifiées étaient localisées sur un fragment de six paires de bases « AGCGCC ». Ce fragment est retrouvé dans 195 allèles HLA-DRB1 (voir la liste en annexe S-24) et les comparaisons avec les exons 2 de DRB1*13:04 et DRB1*11:02:01 montrent 32±28 nucléotides conservés en 5' et 32±16 nucléotides conservés en 3' (l'hypothèse ici étant que les allèles possédant des segments conservés plus longs soient plus probables comme donneurs dans la conversion allélique). DRB1*08:03, retrouvé parmi ces allèles candidats, est l'allèle initialement proposé comme donneur [Lee et al., 1990] mais n'est pas observé chez les Mandenkalu (c'est un allèle fréquent en Asie du sud-est mais pas en Afrique, à l'opposé de DRB1*13:04).
Par contre, sur les 195 allèles candidats, trois sont observés chez les Mandenkalu : DRB1*04:05:01 (FA= 1%), DRB1*08:06 (FA = 5%) et DRB1*13:03:01 (FA = 2%). Il est donc probable que HLA-DRB1*13:04 soit le résultat d'une conversion génique ayant pris place en Afrique de l'ouest, avec -DRB1*11:02:01 comme receveur et un des trois allèles précédemment cités comme donneurs.

La Figure 2.13 illustre la conversion allélique proposée. Les conversions alléliques étant des évènements fréquemment observés dans la région HLA [von Salomé et al., 2007], ces résultats illustrent les apports des nouvelles technologies de séquençage, où une hypothèse de conversion allélique basée sur des profils de réactivité sérologique est affinée grâce à l'information des séquences nucléotidiques.

Reste toutefois à expliquer la forte fréquence de DRB1*13:04 (FA=28%) chez les Mandenkalu, une telle fréquence étant plutôt rare au locus HLA-DRB1 dans une population n'ayant pas subi de dérive génétique rapide.
Une récente étude (communication personnelle de A. Sanchez-Mazas, Mai 2020), a étudié les prédictions *in silico* de liaison peptidique des allèles HLA-DRB1 vis-à-vis de *P. falciparum*. Pour ce faire, les auteurs ont analysé 63'363 peptides de 15 acides-aminés de long, obtenus par une fenêtre glissante de 15 acides-aminés sur les 56 protéines du

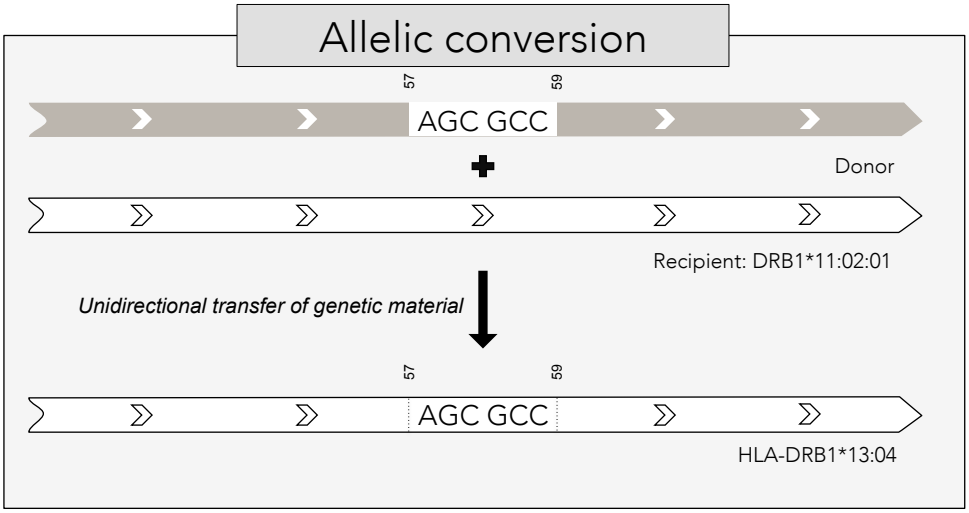


FIGURE 2.13 – Mécanisme de la conversion allélique proposée dans cette étude, suggérant un transfert unidirectionnel de matériel génétique (incluant le fragment «AGCGCC») d'un allèle donneur (probablement HLA-DRB1*04:05:01, -DRB1*08:06 ou -DRB1*13:03:01 chez les Mandenkalu) vers l'allèle receveur DRB1*11:02:01 et amenant à la création de -DRB1*13:04. Figure issue de [Goeury et al., 2018a].

parasite. La molécule HLA-DRB1*13:04 est apparue comme l'un des meilleurs ligands de ces peptides (les meilleurs ligands sont ceux prédits comme liant plus de 1% des peptides présentés). Il semblerait donc que HLA-DRB1*13:04 possède une fonction protectrice contre la malaria.

DQB1*03:19 a seulement été identifié pour la première fois en 2007 [Witter et al., 2007], isolé de DQB1*03:01 duquel il ne diffère que d'une substitution (non synonyme) sur l'exon 3 (position 554 : C $\rightarrow$ T), conservant ainsi un même exon 2 et donc une fonctionnalité similaire. L'étude a aussi mis en évidence un fort déséquilibre de liaison entre DQB1*03:01 et DRB1*11:02:01 (potentiel receveur de la conversion allélique, voir Figure 2.13) ou bien même DRB1*13:04.

L'haplotype HLA-DQ formé par DQA1*05:01 (possédant les mêmes capacité de liaison peptidique que DQA1*05:05:01, FA=50% chez les Mandenkalu) et DQB1*03:01 a été rapporté en 1994 [Meyer et al., 1994] comme étant plus fréquent chez les individus immunisés contre l'onchocercose, suggérant que ces deux allèles jouent un rôle dans la défense immunitaire contre le ver parasite *Onchocerca volvulus*. L'onchocercose est une maladie parasitique fortement prévalente en Afrique de l'ouest (incluant la zone d'échantillonnage des Mandenkalu) [O'Hanlon et al., 2016]. Bien que non mortelle directement, cette maladie, qui cause avant tout la cécité, pourrait réduire l'espérance de vie des personnes touchée en diminuant leur immunité [Miller and Love, 1989]. Dans ce cas, elle pourrait constituer une pression sélective au niveau génétique.

HLA-DQB1*03:19 possédant les mêmes capacités de détection antigénique que DQB1*03:01 (la seule différence entre les deux protéines étant dans le domaine β2), une hypothèse est que -DQB1*03:19 soit protecteur contre *O. volvulus*, cette protection ayant amené à un balayage sélectif de ce dernier, mais par extension de tout l'haplotype HLA-DRB1*13:04 ~DQA1*05:05:01~DQB1*03:19 chez les Mandenkalu.

Le fait que les Mandenkalu vivent dans une zone de forte prévalence de *O. volvulus* [O'Hanlon et al., 2016] mais aussi de la malaria (voir le Chapitre 4, page 227) soutien cette hypothèse d'un balayage sélectif de l'haplotype et apporte une explication plus précise quant aux hautes fréquences de HLA-DRB1*13:04 aussi bien chez les Mandenkalu que chez les Gambiens étudiés par Hill *et al.* en 1992 [Hill et al., 1992a].

Prises individuellement, plusieurs régions ARS portées par des exons 2 (HLA-B, -C, -DQB1 et -DPB1), ou portées par des exons 3 (HLA-A, -B et -C), ainsi que plusieurs exons 3 (HLA-DQA1, -DQB1 et -DPB1), voire des exons 4 (HLA-A et -DQB1) montrent des D de Tajima significativement supérieurs à 0 (voir annexe S-22) chez les Mandenkalu. Ces résultats étaient attendus si l'on considère les résultats d'études précédentes ayant montré de forts signaux de sélection balancée aux exons 2 (voire les exons 3) sur la plupart des loci HLA étudiés dans d'autres populations [Buhler and Sanchez-Mazas, 2011]. Toutefois, après correction pour 86 tests multiples (considérant tous les exons, introns et régions ARS/non-ARS), seuls HLA-A (codons ARS de l'exon 3), -C (codons ARS des exons 2 et 3) et -DPB1 (codons ARS de l'exon 2 ainsi qui l'exon 3) restent significatif (Figure 2.7).
Pour HLA-DPB1 exons 2 et 3, un déséquilibre de liaison existe entre ces deux régions du fait de leur proximité physique (l'intron 2 mesurant au maximum 4014pb [Mack, 2015]), alors la sélection balancée ne s'applique pas forcément aux deux exons en même temps. En effet, de la même façon qu'un balayage sélectif sur un locus va diminuer la diversité au voisinage de ce dernier, la sélection balancée à un locus peut, elle, augmenter la diversité dans le voisinage par un mécanisme appelé «sélection balancée associative» [Ohta and Kimura, 1970, Slatkin, 1995, Sanchez-Mazas, 2007]. Il est donc possible que la sélection balancée ne s'applique que sur l'un des exons, l'autre exon ne montrant un signal de sélection balancée que par déséquilibre de liaison avec le premier.
De la même façon, le ratio dN/dS était significatif seulement pour les codons ARS de quatre régions : les exons 2 de HLA-B et -DPB1 et les exons 3 de HLA-A et -B. Toutefois, le test du dN/dS a été développé initialement pour détecter de la sélection en inter-spécifique et est donc faiblement adéquat pour détecter des pressions de sélection en intra-spécifique [Kryazhimskiy and Plotkin, 2008].

L'analyse en composantes principales (ACP, Figure 2.10) basée sur les indices de diversité moléculaire (D de Tajima, π, fréquence des sites polymorphiques, fréquence des mutations synonymes et non-synonymes) estimés sur les huit gènes (HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1) des deux populations montre d'importantes différences entre les régions et les loci, suggérant des pressions évolutives différentes entre ces loci et à l'intérieur de ces loci.

La variance de l'axe 1 (71%) et les variables associées montrent l'importance de la diversité en acides aminés (c'est-à-dire les mutations non-synonymes) des régions impliquées dans la liaison au peptides (ARS), diversité aussi illustrée par les Figures 2.4, 2.5 et 2.6. Tous les exons 2 et les exons 3 (pour les classe I) ne semblent toutefois pas affectés de la même façon par la sélection. Les trois gènes de classe I montrent une plus grande diversité à leurs codons ARS comparés aux codons non-ARS, résultat attendu et expliqué par l'avantage d'une forte diversité du site de reconnaissance de l'antigène, en lien avec son rôle de liaison des peptides antigéniques.
Les codons ARS des exons 2 (et exons 3 dans une plus petite mesure) de HLA-C sont toujours moins diversifiés que leurs homologues sur les autres loci HLA-A et -B, une explication plausible étant le rôle des domaines α1 codés par HLA-C comme ligand des KIR

(*Killer-cell Immunoglobulin-like Receptors*) [Winter and Long, 1997, Hilton et al., 2015] où une trop grande diversité de ces régions deviendrait préjudiciable à ces régions. Cette hypothèse rejoint les observations de Bitarello *et al.* d'une plus faible diversité moléculaire des ARS de HLA-C [Bitarello et al., 2016].

Parmi les cinq gènes de classe II, HLA-DPA1 montre les plus faibles signaux de diversité pour des codons ARS (en bas à gauche de la Figure 2.10), suggérant donc une plus faible contribution de ce locus à la détection des peptides antigéniques, en lien avec le faible nombre d'allèles HLA-DPA1 rapportés dans la base de données IMGT/HLA (161 allèles codant pour 62 molécules différentes [14]), tandis que les exons 2 de HLA-DQB1 et -DRB1 sont, pour chacune des deux populations, les plus diversifiés de tous les loci HLA étudiés.

Les exons 2 de HLA-A (aussi bien les codons ARS que non-ARS) des Mandenkalu sont retrouvés à une position extrême sur l'axe 2, lié à un faible D de Tajima (-0.89 pour les codons non-ARS et 0.57 pour les codons ARS). Bien que ces valeurs ne soient pas significativement différentes de 0, elles pourraient être la marque d'une expansion démographique plutôt que d'une sélection positive, un tel signal étant attendu chez les Mandenkalu au vu de leur histoire démographique (voir page 53). Ce résultat supporte l'hypothèse que le polymorphisme de l'exon 2 de HLA-A serait plus proche de la neutralité et plus enclin à révéler les signaux démographiques [Di et al., 2015, dos Santos Francisco et al., 2015, Inotai et al., 2015, Sanchez-Mazas et al., 2017].

À l'opposée de HLA-A exon 2, sur l'axe 1, sont retrouvés les codons ARS de l'exon 2 de HLA-DQA1 chez les Cham. HLA-DQA1 exon 2 montre un important D de Tajima aussi bien à ses codons ARS que non-ARS (voir Figure 2.7), le locus HLA-DQA1 montrant dans cette population un excès d'hétérozygotes au test d'Ewen-Watterson-Slatkin. Ce signal peut s'interpréter comme une forte sélection balancée (de type avantage hétérozygote), cette hypothèse n'excluant pas non plus celle d'un gain d'hétérozygotie lié à un mélange entre deux populations. Chez les Mandenkalu, c'est HLA-DPB1 exon 2 (codons ARS) et exon 3 qui montrent un fort D de Tajima, expliquant notamment la position extrême de DPB1-Exon 3 sur l'axe 2 (position la plus élevée sur l'axe 2 pour les Mandenka). Les deux allèles HLA-DPB1 les plus fréquents chez les Mandenkalu (DPB1*17:01 et *131:01) partageant un même exon 2, une pression de sélection similaire impliquant une résistance à un pathogène (indéfini) pourrait expliquer leur importante fréquence chez les Mandenkalu (fréquence cumulée de 42%, proche de celle de DQB1*03:19). La distribution des fréquences alléliques de DPB1 montre quatre allèles à fortes fréquences ($> 10\%$) : DPB1*02:01:01 (FA=14%), DPB1*01:01:01 (FA=14%) en plus de DPB1*17:01 et DPB1*131:01. Cela expliquerait pourquoi les exons 2 et 3 de DPB1 montrent un D de Tajima significatif (même après correction pour tests multiples). Ce résultat contraste toutefois avec les résultats obtenus pour d'autres études, où HLA-DPB1 montrait une distribution de fréquences alléliques plus proches de la neutralité que les autres loci [Solberg et al., 2008, Buhler and Sanchez-Mazas, 2011]. Néanmoins, puisque HLA-DRB1, -DQA1 et -DQB1 (en déséquilibre de liaison global) ont possiblement été la cible d'un balayage sélectif lié à *P. falciparum* et/ou *O. volvulus* (ou d'autres pathogènes) chez les Mandenkalu, la diminution de la diversité consécutive à ce balayage a donc pu être compensée par le polymorphisme plus important de HLA-DPB1, conférant alors une protection aux autres pathogènes auxquels cette population est exposée. Cette explication

14. IPD-IMGT/HLA, v3.38

soutiendrais alors l'hypothèse de la sélection asymétrique divergente conjointe [15], proposée en 2016 par Buhler *et al.* pour les loci HLA de classe I, en l'étendant aux loci de classe II [Buhler et al., 2016].

4.5 Apports de l'étude des gènes HLA à la compréhension de l'origine de la population Cham

Comme avancé précédemment, une plus petite diversité génétique (par rapport aux Mandenkalu) était attendue pour les Cham, qui sont une plus petite population non africaine. Or ce n'est pas du tout ce qui est observé, puisque l'ensemble des régions étudiées sont soit au moins aussi diversifiées, soit plus diversifiées (HLA-DRB1 et -DQB1, voir Figures 2.10 et 2.6, ainsi que la Table 2.9). Cette diversité provient 1) soit d'une grande diversité en pathogènes qui, via la sélection naturelle, induirait une forte sélection balancée diversifiante, 2) soit d'une expansion démographique, 3) soit d'un flux génique ayant apporté à la population une diversité extérieure.
Cette diversité s'accompagne d'un important déséquilibre de liaison, qui peut être causé par la petite taille de population (la dérive induite par la taille de la population créant ou entretenant le déséquilibre), une pression environmentale pathogénique favorisant certains haplotypes (comme c'est le cas pour les Mandenkalu) ou une fusion avec une autre population de profil génétique différent.

L'hypothèse d'une forte pression en pathogènes est peu probable pour expliquer le déséquilibre de liaison observé. En effet, il ne s'agit pas d'un seul haplotype avec un déséquilibre fort, tel qu'observé chez les Mandenkalu (22 haplotypes en déséquilibre de liaison, avec $D = 0.084 \pm 0.036$, annexe S-23.1), mais un ensemble d'haplotypes présentant en moyenne un déséquilibre moins fort (76 haplotypes, $D = 0.057 \pm 0.021$, annexe S-23.2). De même, l'hypothèse d'une expansion démographique est peu probable, puisqu'une étude menée en 2017 sur des génomes mitochondriaux de Cham ont mis en évidence une contraction démographique de cette population il y a 700 ans [Pischedda et al., 2017].

La Figure 2.14 représente graphiquement, sous la forme d'un réseau, le déséquilibre de liaison haplotypique observé chez les Cham. La largeur des segments reliant les haplotypes reflétant la force (D) du déséquilibre de liaison entre les allèles (sans être toutefois strictement proportionnelle). Ce graphique montre l'existence de deux grands groupes d'haplotypes, l'un en bleu en haut à gauche et l'autre en rouge en bas (portant l'allèle DRB1*15:02). Il n'est toutefois pas possible de parler ici d'haplotypes étendus puisque les allèles ne sont pas tous en déséquilibre de liaison et que certains de ces groupes incluent plusieurs allèles d'un même gène, ce qui est incompatible avec le concept d'haplotype.
À des fin de comparaison, le réseau d'haplotypes en déséquilibre de liaison de la population Mandenka est représenté sur la Figure 2.15 selon la même représentation que pour la Figure 2.14.

Il apparaît alors que la population Mandenka montre moins de déséquilibres de liaison haplotypiques, cohérent avec les études montrant un plus faible déséquilibre de liaison dans les populations africaines [Reich et al., 2001, Tishkoff and Kidd, 2004].

15. « *Joint divergent asymmetric selection* », expliquant les importantes différences d'hétérozygotie parfois observées aux loci de classe I.

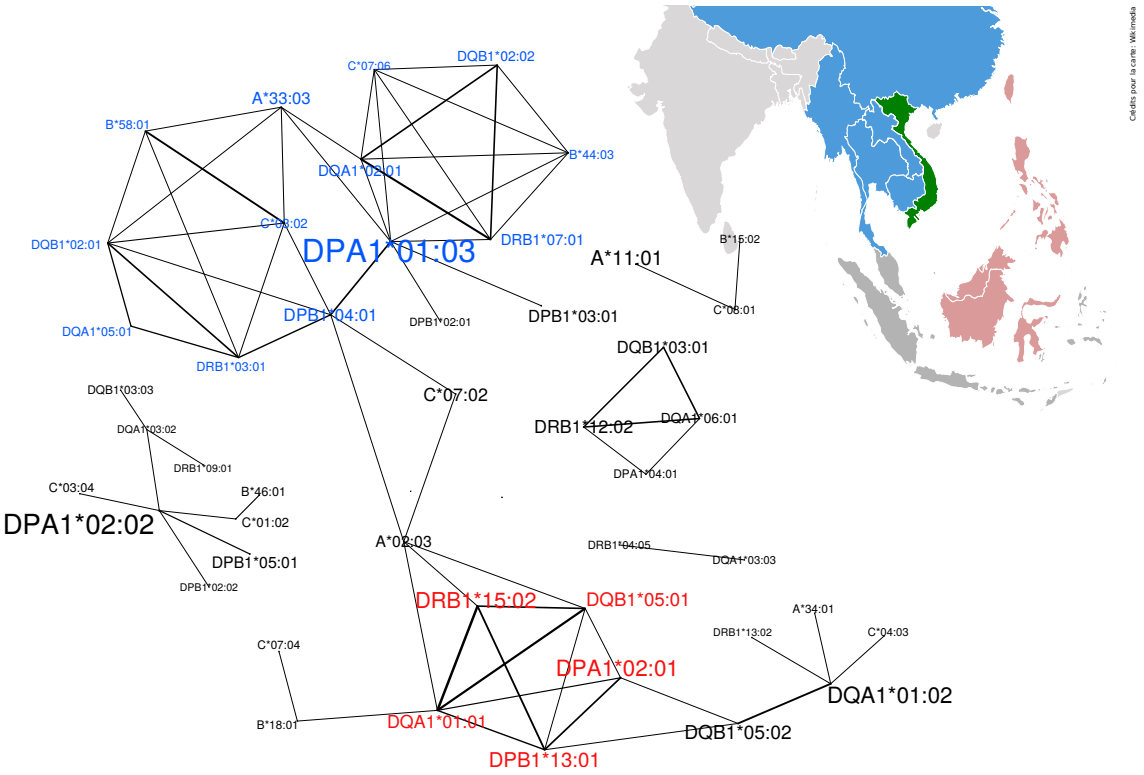


FIGURE 2.14 – Représentation graphique du réseau de déséquilibres de liaison haplotypiques chez les Cham du Vietnam. Seuls les haplotypes en déséquilibre de liaison positif et significatif sont représentés ici, la largeur des segments entre deux allèles traduisant la force (D de LD) du déséquilibre de liaison entre ces deux allèles (segments plus épais signifiant un D plus important). De même, la taille des noms des allèles est proportionnelle à leur fréquence dans la population Cham. Les couleurs illustrent les régions dans lesquelles ces haplotypes sont les plus fréquents, mais ne représentent pas fidèlement la distribution géographique de ces haplotypes. Source de la carte (en haut à droite) : CIA-World Fact Books, adapté de Wikimédia (`https://commons.wikimedia.org/wiki/File:LocationAsia.png`, domaine public).

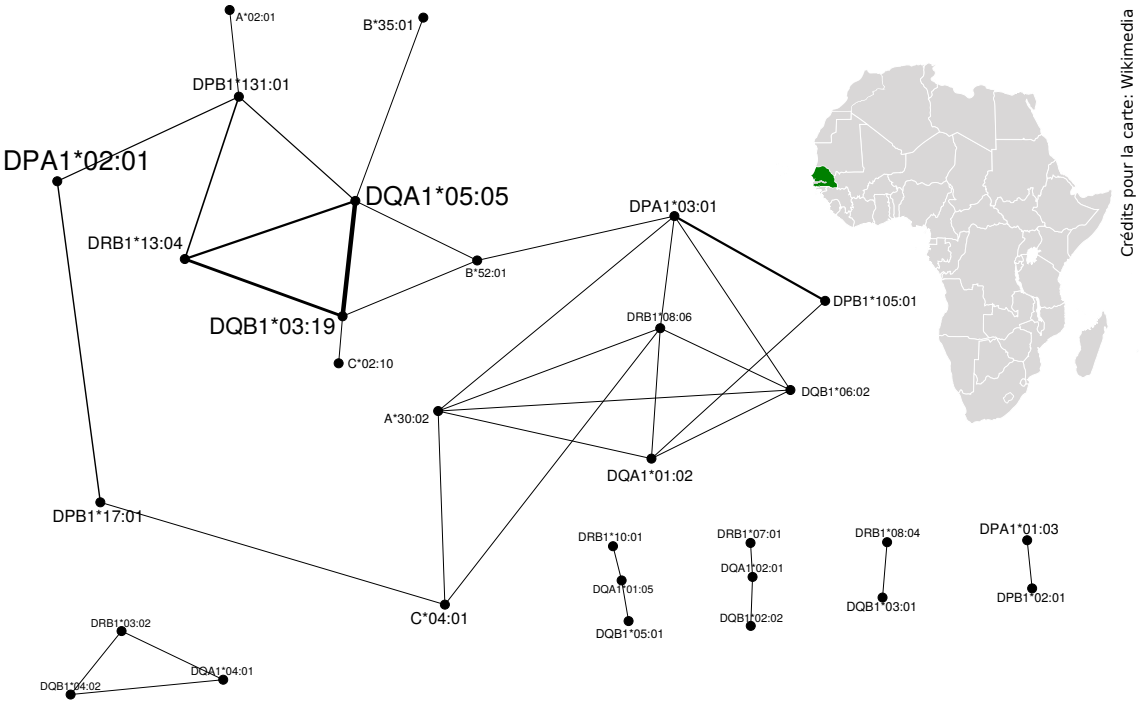


FIGURE 2.15 – Représentation graphique du réseau de déséquilibres de liaison haplotypiques chez les Mandenkalu du Sénégal. Seuls les haplotypes en déséquilibre de liaison positif et significatif sont représentés ici, la largeur des segments entre deux allèles traduisant la force (D de LD) du déséquilibre de liaison entre ces deux allèles (segments plus épais signifiant un D plus important). De même, la taille des noms des allèles est proportionnelle à leur fréquence dans la population Mandenka. Source de la carte (en haut à droite) : CIA-World Fact Books, adapté de Wikimédia (`https://en.wikipedia.org/wiki/File:Location_Senegal_AU_Africa.svg`, domaine public).

Une ACP a été menée sur les génotypes individuels des Cham, comprenant les loci HLA-A, -B, -C, -DRB3/4/5, -DQA1, -DQB1, -DPA1 et -DPB1. Les génotypes obtenus pour HLA-DRB1 ont été retirés pour éviter un sur-paramétrage, puisque ces derniers ont été utilisés pour catégoriser les individus en plusieurs groupes sur la base des résultats des analyses de déséquilibres de liaisons haplotypiques. Cette analyse vise à démontrer l'existence l'une structure génétique chez les Chams. Les deux premiers axes de cette ACP sont représentés sur la Figure 2.16.

Cette ACP montre une différence nette en termes de génotypes entre les individus porteurs de HLA-DRB1*15:02 ou -DRB1*12:02 (présents dans l'haplotype rouge sur la Figure 2.14), et les individus porteurs de l'allèle -DRB1*07:01 (haplotype bleu sur la Figure 2.14). Il apparaît ainsi que ces deux groupes d'allèles identifient une importante structuration en sous-populations, sans qu'il soit question pour autant de sous-populations isolées puisque des individus portant des allèles appartenant aux deux groupes existent et que la population dans son ensemble est à l'équilibre de Hardy-Weinberg, dont l'une des conditions est la panmixie.

La Figure 2.17, provenant de [Sanchez-Mazas et al., 2005], montre la distribution de certains des allèles HLA-DRB1 les plus fréquents dans 48 populations asiatiques. Elle montre que les allèles HLA-DRB1*12:02 et -DRB1*15:02 sont majoritairement retrouvés dans des populations austronésiennes provenant des îles d'Asie du sud-est.

La Figure 2.18 représente la distribution géographique des allèles HLA-DRB1*07:01 et -DRB1*15:02 en Asie (source : `https://hla-net.eu/interactive/HLA_map/`).

L'allèle HLA-DRB1*07:01 est fréquemment retrouvé dans les populations de l'est de l'Asie continentale mais peu retrouvé dans les populations des îles d'Asie du sud-est. Au contraire, HLA-DRB1*15:02, bien que présent sur le continent asiatique, est plus fréquemment retrouvé dans les populations des îles d'Asie du sud-est.

Il est alors possible d'émettre une hypothèse sur l'origine des haplotypes identifiés dans la Figure 2.14, les allèles des haplotypes en bleu étant retrouvés majoritairement chez les populations d'Asie continentale tandis que les allèles des haplotypes en rouge sont retrouvés majoritairement dans les populations habitant les îles d'Asie du sud-est, de Taïwan à l'Indonésie en passant par les Phillipines.

Des ACP ont été réalisées en utilisant les fréquences alléliques aux loci HLA-A, -B, -C et -DRB1 des Cham et en les comparant avec les populations est-asiatiques de la base de données Gene[Rate] [16]. Les Figures 2.19, 2.20, 2.21 et 2.22 représentent respectivement les ACP réalisée sur les loci HLA-A, -B, -C et -DRB1.

16. Grâce à l'outil créé par Di Da, du Laboratoire anthropologie, génétique et peuplements de Genève : `https://hla-net.eu/interactive/HLA_PCA_new/`

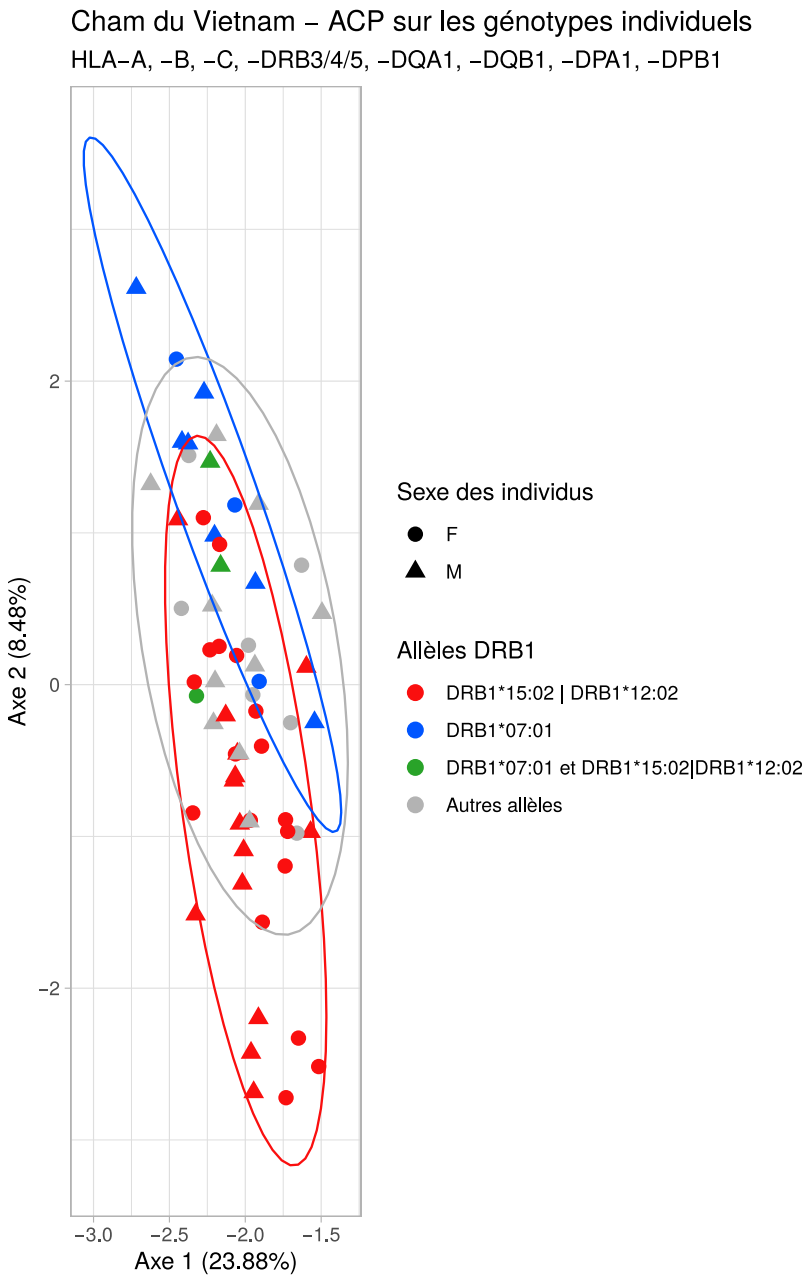


FIGURE 2.16 – ACP basée sur les génotypes individuels des Cham, comprenant les loci HLA-A, -B, -C, -DRB3/4/5, -DQA1, -DQB1, -DPA1 et -DPB1. Les deux premiers axes expliquent 32% de la variance totale. Les points rouges représentent les individus porteurs de HLA-DRB1*15:02 ou -DRB1*12:02, tandis que les points bleus représentent les individus porteurs de -DRB1*07:01, les points verts les individus portant -DRB1*07:01 et -DRB1*15:02 ou -DRB1*12:02, et les points gris les individus ne portant aucun des deux. Les ronds représentent les femmes et les triangles les hommes, il n'y a pas de différence significative de sexe entre les différents groupes (test de Kruskal-Wallis, $pValeur = 0.445$). Les ellipses en bleu, rouge et gris représentent les intervalles de confiance à 95% (calculées sous l'hypothèse d'une distribution de Student multivariée), respectivement pour les individus porteurs de HLA-DRB1*07:01, HLA-DRB1*15:02 ou -DRB1*12:02, et des autres allèles.

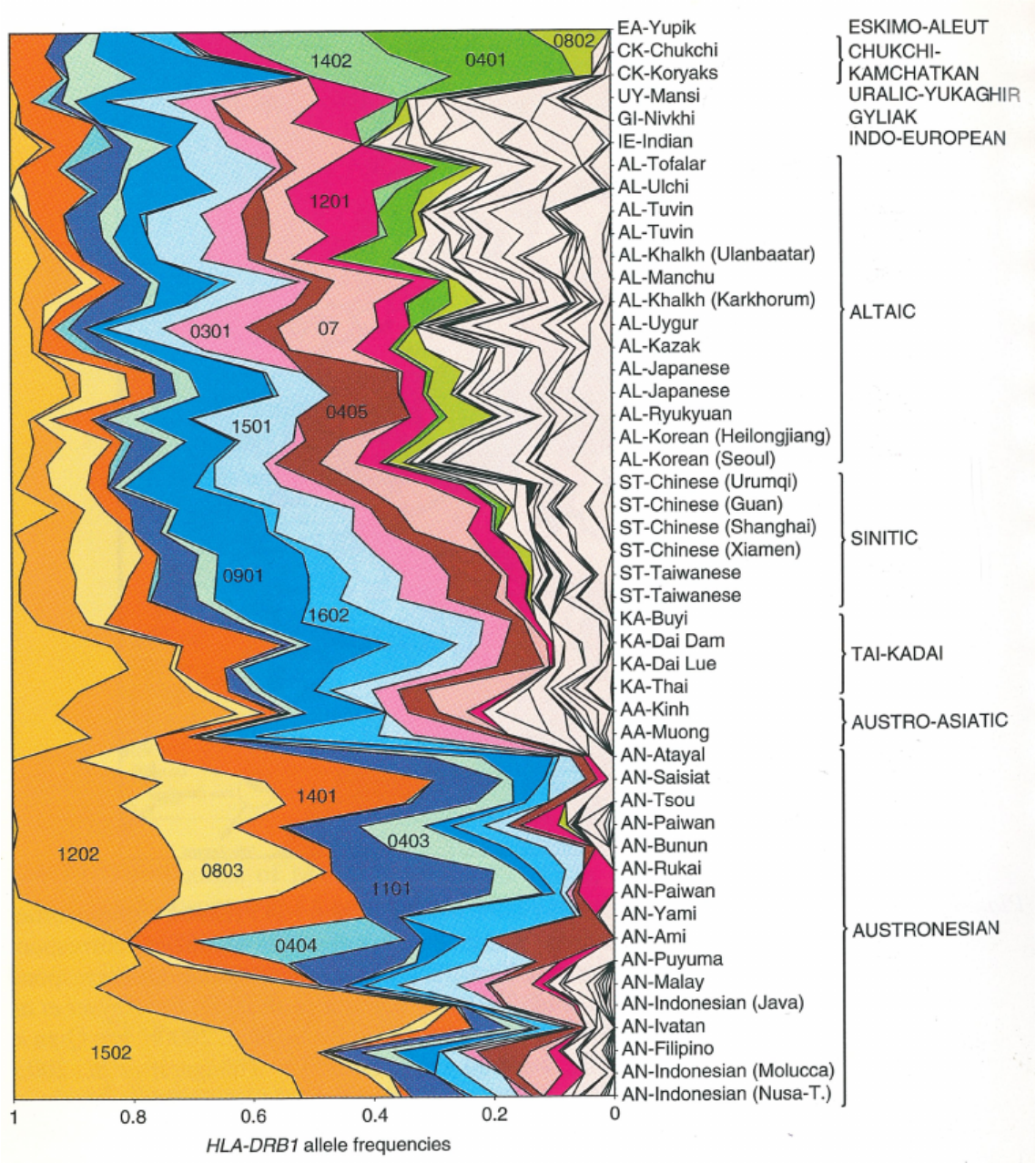


FIGURE 2.17 – Fréquences des allèles HLA-DRB1 dans 48 populations asiatiques, triées par familles linguistiques. Les allèles les plus fréquents (FA>14% dans au moins une population) sont représentées par les couleurs vives. AN :Austronésien ; AL : Altaïque ; AA : Austro-Asiatique ;CK : Chukchi-Kamchatka ; KA : Tai-Kadai, GI : Gilyak ; IE : Indo-Européen ; UY : Ouralique-Yukaghir ; ST : Sino-Tibétain (seulement du Sinitique ici) ; EA : Eskimau-Aléout. Source : [Sanchez-Mazas et al., 2005].

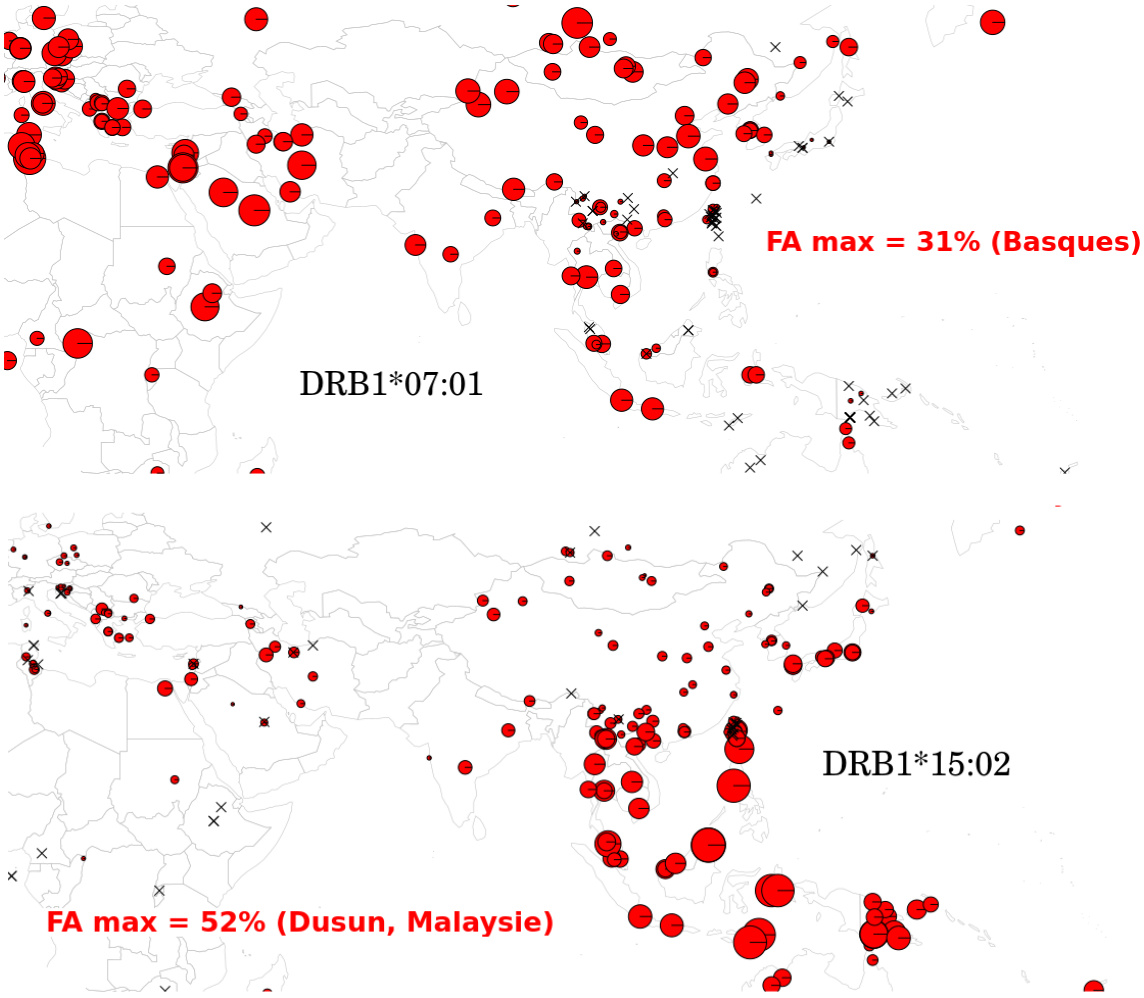


FIGURE 2.18 – Cartes de la répartion des allèles HLA-DRB1*07:01 (en haut) et -DRB1*15:02 (en bas). Les tailles des cercles sont proportionnelles à la fréquence maximale de l'allèle (indiquée par FA max), les X indiquent les populations chez lesquelles cet allèle n'a pas été detecté. Source : `https://hla-net.eu/interactive/HLA_map/`.

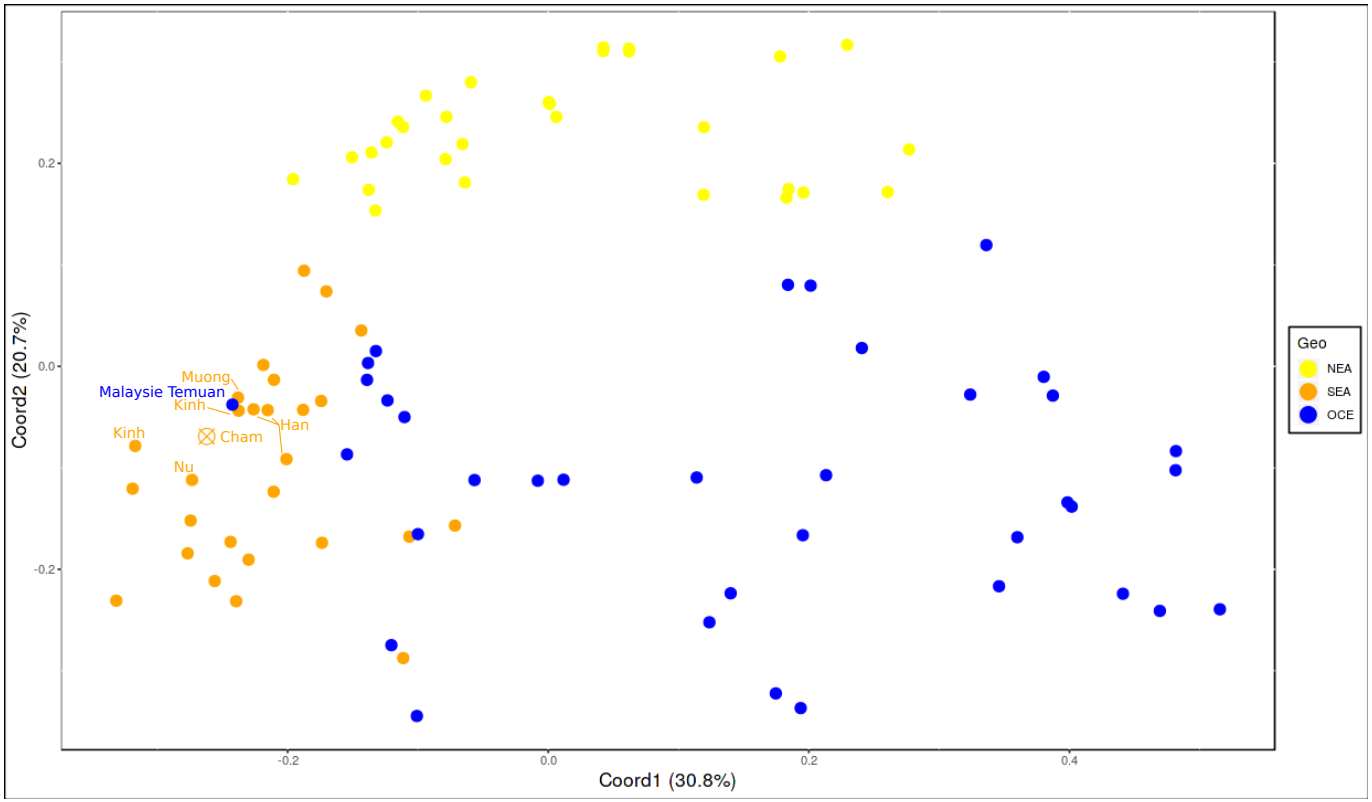


FIGURE 2.19 – Représentation graphique des deux premiers axes de l'ACP réalisée sur les fréquences alléliques des Cham et des populations est-asiatiques, pour le locus HLA-A. Les proportions de variance expliquée sont indiquées entre parenthèses. Les points jaunes représentent les populations du nord-est de l'Asie, les points oranges les populations du sud-est de l'Asie et les points bleus les populations d'Océanie. Les Cham sont indiqués par la croix cerclée et les noms des populations les plus proches des Cham sur la projection ont été indiqués.

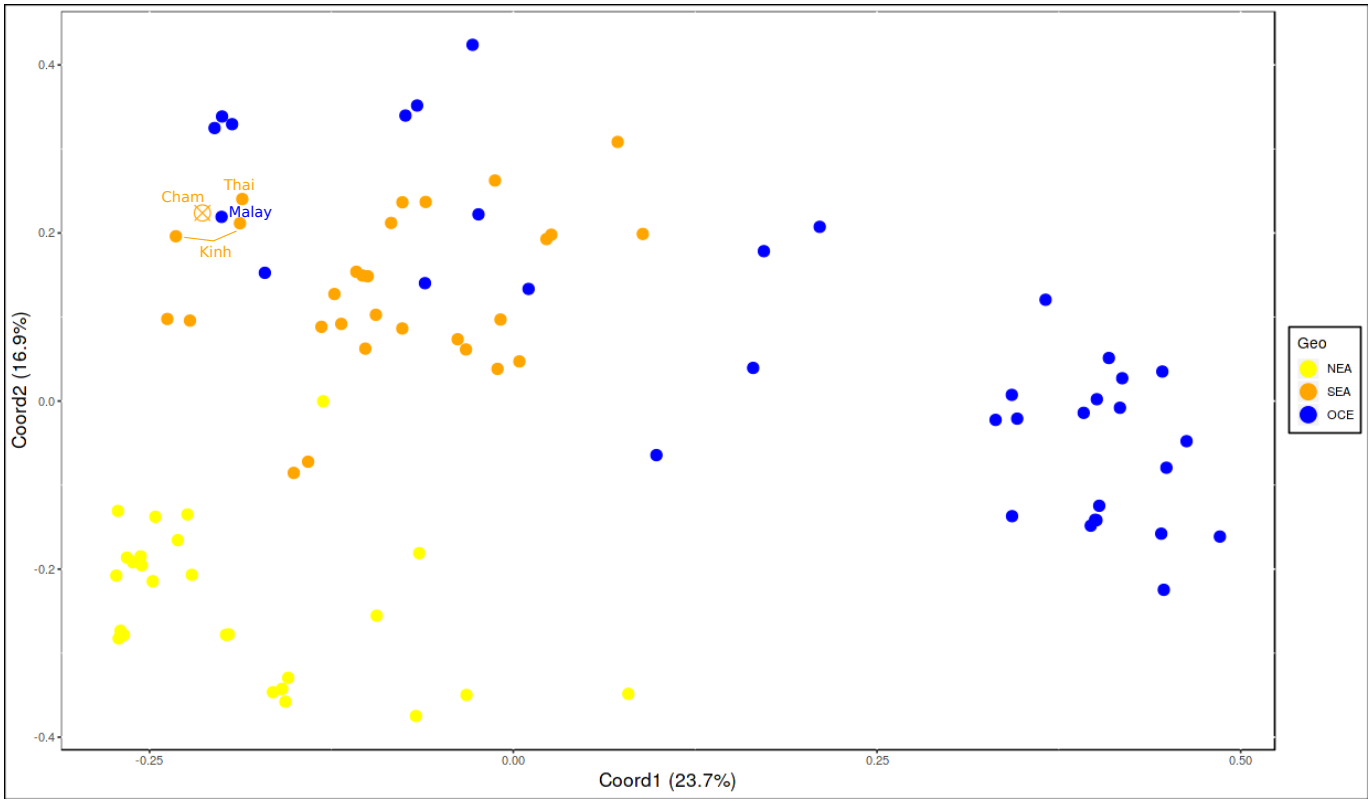


FIGURE 2.20 – Représentation graphique des deux premiers axes de l'ACP réalisée sur les fréquences alléliques des Cham et des populations est-asiatiques, pour le locus HLA-B. Les proportions de variance expliquée sont indiquées entre parenthèses. Les points jaunes représentent les populations du nord-est de l'Asie, les points oranges les populations du sud-est de l'Asie et les points bleus les populations d'Océanie. Les Cham sont indiqués par la croix cerclée et les noms des populations les plus proches des Cham sur la projection ont été indiqués.

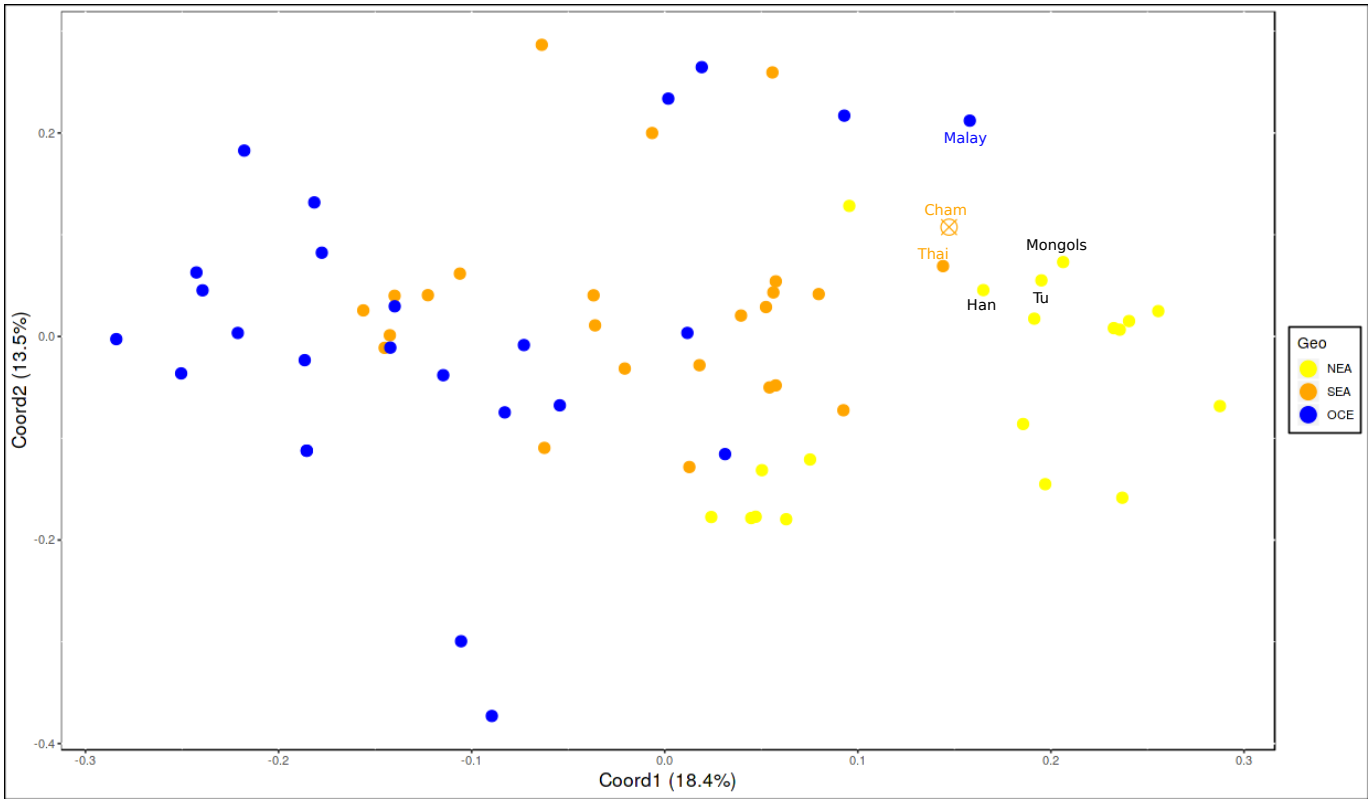


FIGURE 2.21 – Représentation graphique des deux premiers axes de l'ACP réalisée sur les fréquences alléliques des Cham et des populations est-asiatiques, pour le locus HLA-C. Les proportions de variance expliquée sont indiquées entre parenthèses. Les points jaunes représentent les populations du nord-est de l'Asie, les points oranges les populations du sud-est de l'Asie et les points bleus les populations d'Océanie. Les Cham sont indiqués par la croix cerclée et les noms des populations les plus proches des Cham sur la projection ont été indiqués.

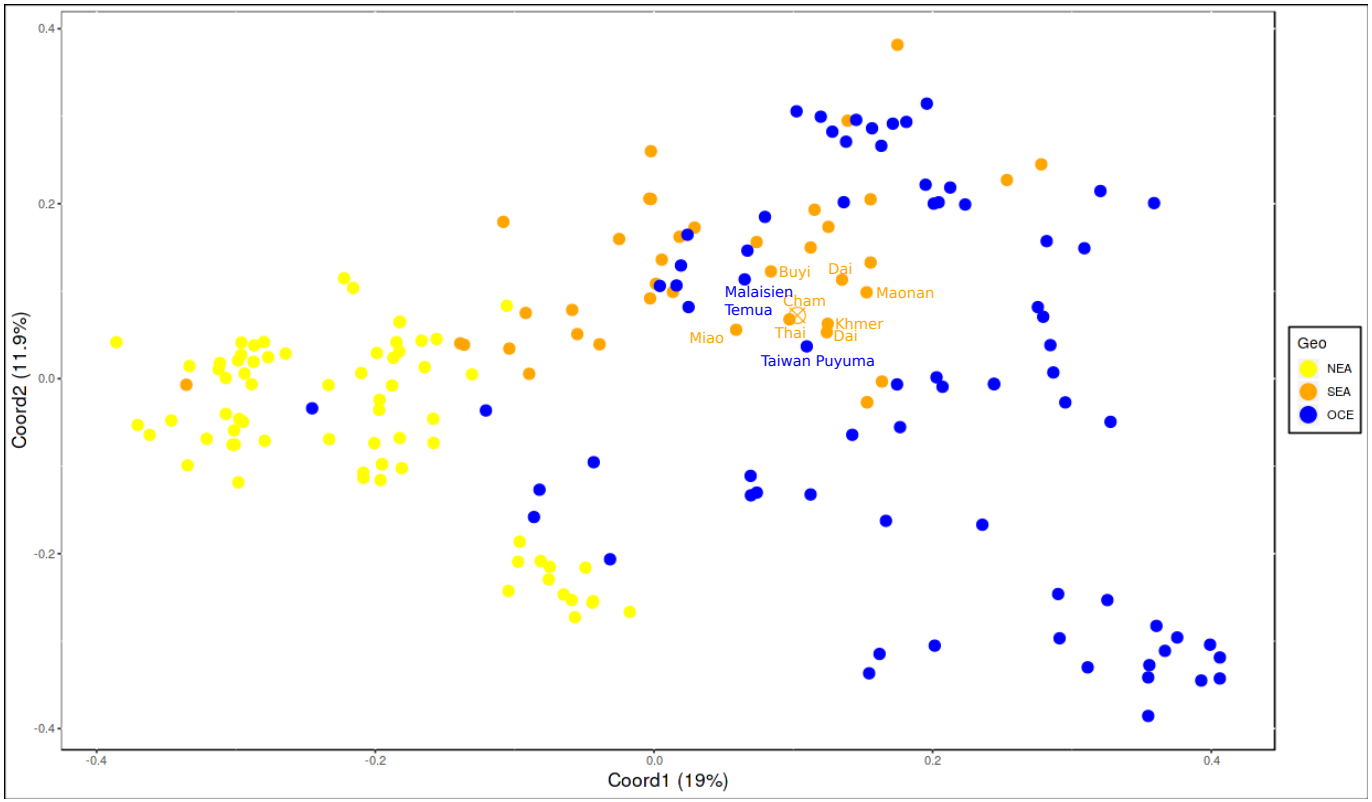


FIGURE 2.22 – Représentation graphique des deux premiers axes de l'ACP réalisée sur les fréquences alléliques des Cham et des populations est-asiatiques, pour le locus HLA-DRB1. Les proportions de variance expliquée sont indiquées entre parenthèses. Les points jaunes représentent les populations du nord-est de l'Asie, les points oranges les populations du sud-est de l'Asie et les points bleus les populations d'Océanie. Les Cham sont indiqués par la croix cerclée et les noms des populations les plus proches des Cham sur la projection ont été indiqués.

Il apparaît que les Cham sont proches des autres populations vietnamiennes, notamment des Kinh et des Khmer (Figure 2.19, 2.20 et 2.22) ainsi que d'autres populations géographiquement voisines telles que les Thaï (Figure 2.21) ou des populations du nord-est de l'Asie (Han, Mongols et Tu). De plus, la Figure 2.22, représentant l'ACP basée sur les fréquences alléliques de HLA-DRB1, montre qu'à ce locus les Cham sont aussi similaires aux populations océaniennes de Malaisie mais aussi à des populations natives de Taïwan (les Puyuma).

Les résultats obtenus par cette étude pour la population Cham du Vietnam tendent à confirmer l'hypothèse selon laquelle la population Cham actuelle serait le résultat d'un ou plusieurs évènements de mélange génétique entre une (ou des) population(s) du sud-est de l'Asie et une (ou des) population(s) austronésienne(s). Ces résultats ne permettent pas de rejeter l'hypothèse démique, selon laquelle les Cham possèderaient une double origine, avec un important flux génique entre une population habitant l'Asie continentale et une population austronésienne migrante [Thurgood, 1999, Higham, 2002, Southworth et al., 2004, Bellwood, 2007, Peng et al., 2010]. Les hauts taux de diversité (hétérozygotie, π et diversité en acides aminés) observés aux loci HLA-DRB1 et -DQB1 seraient alors le résultat de ce flux génique.
Ces résultats ne permettent toutefois pas d'appuyer ou d'infirmer l'hypothèse d'un sexe-ratio biaisé, puisqu'aucune différence génétique n'a pu être observée entre les femmes et les hommes sur les loci étudiés (voir Figure 2.16), au contraire des études menées par Peng *et al.* [Peng et al., 2010] et He *et al.* [He et al., 2012]. Les résultats de cette étude ne permettent pas non plus d'exclure totalement l'hypothèse d'une diversité due à des pressions pathogéniques pour expliquer la diversité observée.

5 Conclusion

Cette étude montre l'avantage procuré par les nouvelles technologies de séquençage. Ces dernières permettent non seulement de réduire le nombre d'erreurs et les ambiguïtés de typage, mais aussi d'analyser finement la diversité nucléotidique des différents loci HLA et des régions qui les composent, allant jusqu'à séparer, au sein d'un même exon, les différents codons qui le composent (codons ARS et non-ARS). Ce gain de résolution se traduit déjà, depuis quelques années, par une augmentation du nombre d'allèles HLA connus (52% des allèles [17] ont été rapportés dans les cinq dernières années), notamment par la description de la variabilité des introns et régions non traduites. L'étude de ces régions encore mal couvertes par les séquençages permettra de mieux comprendre plusieurs phénomènes tels que la régulation de l'expression de ces gènes et ainsi d'affiner la connaissance des mécanismes immunitaires liés à HLA. L'explosion du nombre d'allèles nécessitera probablement une refonte de la nomenclature HLA de la part du comité de nomenclature HLA, mais aussi une adaptation des personnes travaillant avec le HLA.

Outre l'avantage apporté par les NGS, cette étude a aussi apporté des éclaircissements sur les phénomènes évolutifs et démographiques des populations étudiées. Chez les Mandenkalu l'existence d'un haplotype de classe II étendu trouve alors son explication dans un processus de sélection naturelle liée à certaines maladies infectieuses prévalentes en Afrique, comme la malaria et/ou l'onchocercose, mais aussi à un évènement de conversion allélique expliquant l'origine de DRB1*13:04.

Cette étude soutient aussi l'hypothèse d'une origine double des Cham, ces derniers étant alors le résultat d'un mélange entre une population locale et une population austronésienne migrante. Une couverture plus approfondie de la diversité génomique des populations vivant dans ces régions devrait permettre d'identifier plus clairement les régions d'origine de ces deux populations.

17. IPD-IMGT/HLA, v3.38

Received: 22 May 2017 | Revised: 1 November 2017 | Accepted: 15 November 2017

DOI: 10.1111/tan.13180

WILEY HLA Immune Response Genetics

ORIGINAL ARTICLE

Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa

T. Goeury[1,2] | L. E. Creary[3] | L. Brunet[1,4] | M. Galan[5] | M. Pasquier[1,2] | B. Kervaire[1,4] | A. Langaney[1] | J.-M. Tiercy[2,4] | M. A. Fernández-Viña[3] | J. M. Nunes[1,2] | A. Sanchez-Mazas[1,2]

[1]Laboratory of Anthropology, Genetics and Peopling History, Department of Genetics and Evolution - Anthropology Unit, University of Geneva, Geneva, Switzerland

[2]Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland

[3]Department of Pathology, Stanford University School of Medicine, Palo Alto, California

[4]Transplantation Immunology Unit and National Reference Laboratory for Histocompatibility (UIT/LNRH), Geneva University Hospital, Geneva, Switzerland

[5]INRA, UMR 1062 CBGP, avenue du Campus Agropolis, Montferrier sur Lez, France

Correspondence
Thomas Goeury, Laboratory of Anthropology, Genetics and Peopling History, Department of Genetics and Evolution - Anthropology Unit, University of Geneva, Sciences II, 30 quai Ernest-Ansermet, CH-1211 Geneva 4, Switzerland.
Email: thomas.goeury@unige.ch
Alicia Sanchez-Mazas, Laboratory of Anthropology, Genetics and Peopling History, Department of Genetics and Evolution - Anthropology Unit, University of Geneva, Sciences II, 30 quai Ernest-Ansermet, CH-1211 Geneva 4, Switzerland.
Email: alicia.sanchez-mazas@unige.ch

Funding information
Swiss National Science Foundation, Grant/Award number: 31003A_144180

With the aim to understand how next-generation sequencing (NGS) improves both our assessment of genetic variation within populations and our knowledge on HLA molecular evolution, we sequenced and analysed 8 HLA loci in a well-documented population from sub-Saharan Africa (Mandenka). The results of full-gene NGS-MiSeq sequencing compared with those obtained by traditional typing techniques or limited sequencing strategies showed that segregating sites located outside exon 2 are crucial to describe not only class I but also class II population diversity. A comprehensive analysis of exons 2, 3, 4 and 5 nucleotide diversity at the 8 HLA loci revealed remarkable differences among these gene regions, notably a greater variation concentrated in the antigen recognition sites of class I exons 3 and some class II exons 2, likely associated with their peptide-presentation function, a lower diversity of HLA-C exon 3, possibly related to its role as a KIR ligand, and a peculiar molecular diversity of HLA-A exon 2, revealing demographic signals. Based on full-length HLA sequences, we also propose that the most frequent DRB1 allele in the studied population, *DRB1*13:04*, emerged from an allelic conversion involving 3 potential alleles as donors and *DRB1*11:02:01* as recipient. Finally, our analysis revealed a high occurrence of the *DRB1*13:04-DQA1*05:05:01-DQB1*03:19* haplotype, possibly resulting from a selective sweep due to protection to *Onchorcerca volvulus*, a prevalent pathogen in West Africa. This study unveils highly relevant information on the molecular evolution of HLA genes in relation to their immune function, calling for similar analyses in other populations living in contrasting environments.

KEYWORDS
allelic conversion, balancing selection, full-length HLA genes, HLA nucleotide diversity, *Onchocerciasis*, population genetics, selective sweep, West Africa

1 | INTRODUCTION

Due to the extreme polymorphism of the HLA genomic region[1] (www.ebi.ac.uk/ipd/imgt/hla/stats.html), the application of next-generation sequencing (NGS) to HLA genes has been particularly challenging in the last decades, requiring careful tests and comparisons among different competing technologies.[2–8] However, thanks to tremendous efforts motivated by the need of tissue-typing laboratories to improve the accuracy and throughput of HLA genotyping of potential donors and patients, new HLA sequencing platforms are currently being implemented in most countries, leading to both a

rapid discovery of new alleles (eg, 17% increase of the total number of alleles and 119% of the total number of alleles with complete sequences between April 2016 and April 2017, www.ebi.ac.uk/ipd/imgt/hla/stats.html, see also, for example, Reference [9]) and a better characterisation of multi-locus haplotypes which are expected to improve the HLA data quality of hundreds of donor registries throughout the world.

Besides its potential benefits for histocompatibility, DNA sequencing of the HLA region also opens new perspectives in the area of human population genetics by permitting direct analyses of nucleotide variation within a molecular evolutionary genetics framework.[10–12] Although stimulating results on human populations' molecular diversity were obtained previously by inferring sequence genotypes to large sets of population samples thanks to the molecular information stored in the IMGT-HLA database,[13] such approaches could only use data defined at the second field level of resolution, thus ignoring the information provided by synonymous substitutions and by regions located outside exons 2 and 3. As NGS established itself very recently as the next HLA typing standard,[14] we thus decided to use NGS techniques to decipher the fine nucleotide diversity of 8 HLA genes in a well-documented human population.

Our primary goal in this study was to understand better the relationships between HLA nucleotide variation and different demographic and selective forces which could have driven the molecular evolution of the HLA loci. We thus chose to analyse a population that was already well-known from both an anthropological (*lato sensu*) and a genetic point of view. The Mandenkalu (plural of Mandenka) were sampled during a field study undertaken in the 1990s in Eastern Senegal, West Africa, after several previous expeditions which documented the demography and peopling history of this region (see Reference [15] for a review). Several genetic polymorphisms were then analysed, among which immunoglobulin markers,[16–18] HLA by both serological[16] and Polymerase Chain Reaction-Sequence Specific Oligonucleotide (PCRSSO)[19] methods, mtDNA,[20] genome-wide Restriction Fragment Length Polymorphisms (RFLPs),[21] alpha-[22] and beta-globins[23] and N-acetyltransferase 2.[24] Based on these different sources of independent information, the Mandenka population (which is also since many years a reference population in the HGDP-CEPH Database, www.cephb.fr/hgdp/main.php), is known to exhibit a very high level of genetic diversity, probably as a result of population expansion,[25] and is considered to be representative of a larger population group of Western Africa.[23]

Interestingly, these results allow the exclusion of genetic drift as a major evolutionary factor shaping the molecular profile of this population, thus providing an ideal framework to explore the effects of natural selection on the HLA genes. Such effects, which result from the crucial immunological function played by the HLA molecules, are indeed generally difficult to disentangle from those of demography.[26] Although linear modelling or comparisons with simulated data may be successfully used in this perspective[27–29], such approaches need large sets of HLA-typed population samples, which are currently only available in the form of first-field HLA allelic frequencies. Moreover, different types of selective pressures have been invoked to affect the HLA region (see Reference [30] for a review). Although balancing selection in the form of heterozygous advantage (more particularly under the *divergent allele advantage* model[31]) is generally considered as the main mechanism explaining the very high level of polymorphism observed at these genes in most human populations,[32] positive selection increasing the frequency of alleles putatively protective to specific pathogens[29,33–36] as well as *joint divergent asymmetric selection* involving simultaneous contributions of several HLA loci[37] have also been proposed to explain the lower diversity observed in some populations at specific HLA loci. In this context, the analysis of the Mandenka population is particularly motivating because it lives in a region where several infectious diseases are highly prevalent. In addition, although sub-Saharan African populations, which represent more than 2000 ethno-linguistic groups,[38] have been extensively studied for different genetic markers[39–44] including HLA,[29,45–48] to our knowledge their HLA molecular variability has not been analysed so far at its greatest degree of detail, that is, the nucleotide level. In this study, we thus investigated the nucleotide polymorphism of 8 extended HLA genes in a sample of the Mandenka population by using full-gene NGS-MiSeq high-throughput sequencing. To evaluate the advantages of using this technology, we first compared the NGS-MiSeq results with those of a traditional typing technique (PCR-SSO on exons 2/3) and a limited sequencing strategy (NGS-454 on exon 2) applied to the same population sample. For the purpose of comparing the typing technologies we used all individuals that were typed with those techniques, hence using the maximum amount of data available, and for the subsequent population genetics analyses we used only subsets of unrelated individuals. In this context, we estimated several molecular population diversity indexes and compared them among different gene regions within and between 8 HLA loci by using both the nucleotide and inferred amino acid information. We finally discussed our results in relation to several hypotheses of natural selection affecting different HLA gene regions and loci.

2 | MATERIALS AND METHODS

2.1 | Population sample

The Western African Mandenka population speaks a language belonging to the Mande branch of the Niger-Congo family (widely represented among African-Americans nowadays[49]). During a sampling campaign performed in January to February 1990, 20 mL of peripheral blood was taken

from 205 informed volunteers living in 5 Mandenka villages in the Bandafassi district near Kédougou (Supplementary Information S01). Pedigree relationships between individuals were also recorded, and 101 of them could be considered as unrelated.

2.2 | HLA typings

In a time span of about 25 years, the same Mandenka population sample was HLA typed by using 3 different techniques: a traditional typing technology (PCR-SSO on exons 2/3), whose results were reported between 1992 and 2007,[19,50,51] a limited sequencing strategy (NGS-454 on exon 2, only applied to class II loci) and full-gene NGS-MiSeq sequencing, the results of both NGS-454 and NGS-MiSeq being reported here for the first time. Of the initial set of 205 sampled individuals, 165 to 198 individuals (depending on the locus) were successfully typed by PCR-SSO for the 6 loci A, B, C, DRB1, DQB1 and DPB1, 194 to 199 individuals by NGS-454 for the 4 class II loci DRB1, DQA1, DQB1 and DPB1 and 51 to 86 individuals by NGS-MiSeq for the 8 loci A, B, C, DRB1, DQA1, DQB1, DPB1 and DPB1. These HLA-typings, for which complete descriptions can be found in Supplementary Information S01, are summarized below.

2.3 | PCR-SSO typings of class I (exons 2 and 3) and class II (exon 2) genes

HLA-DRB1, DQB1 and DPB1 typings were performed by locus- and group-specific PCR amplification, followed by direct hybridization with sequence-specific oligonucleotide (SSO) probes on nylon filters. For HLA class I DNA typings, samples were first tested by direct PCR-SSO hybridization using locally designed probes[52–54] and were further retyped by using the reverse PCR-SSO hybridization protocol of the 13th International Histocompatibility and Immunogenetics Workshop.[51]

2.4 | NGS-454 sequencing of class II genes (exon 2)

Tagging and multiplexing methods designed by Galan et al[55] for Roche NGS-454 sequencing were used. For this high-throughput sequencing, adaptors are required for the emPCR and 454 GS-FLX pyrosequencing using Lib-L Titanium Series reagents. Four HLA Class II genes were amplified and sequenced using this method (DRB1, DQA1, DQB1, DPB1). For each gene, 245 samples (+12 control H2O) were amplified. A total of 46 replicates (19%) were used to confirm the protocol. Sequencing was performed by Beckman Coulter Genomics (Genomic Services, Danvers, Massachuetts). Reads were filtered using Mothur[56] with a minimal PhredScore of 30. These data were explored using SESAME Barcode.[57]

2.5 | NGS-MiSeq sequencing of full class I and class II genes

Two complementary techniques were used in 2 different laboratories to sequence full HLA class I and class II genes:

a. *Geneva University Hospital*: Fifty-four individuals were typed by using the Holotype HLA X2 kit (Omixon Biocomputing Ltd, Budapest, Hungary) in combination with the Illumina MiSeq platform to type the 7 HLA loci A, B, C, DRB1, DQA1, DQB1 and DPB1. Sequences generated by MiSeq were processed by the software HLA Twin v1.1.1 (Omixon, Biocomputing Ltd, Budapest, Hungary).
b. *Stanford University*: Sixty-five individuals (25 among the 54 individuals sequenced at Geneva University Hospital to confirm uncertain sequences, plus 40 additional individuals) were typed for the 8 HLA loci A, B, C, DPA1, DPB1, DQA1, DQB1, DRB1 using the MIA FORA NGS typing method developed by Sirona Genomics (Immucor, Inc, Norcross, Georgia) and performed following the manufacturer's instructions. Alleles were assigned using the Sirona Genomics NGS alignment software which uses 2 complementary informatic strategies to make genotyping calls. The first strategy utilizes Expectation Maximization to rank computed allele candidates based on mapping metrics. Coverage is calculated from competitive alignment of paired-end NGS reads with all HLA reference sequences in the IMGT/HLA database 3.22.0[58] and internal references generated by cloning and sequencing. The second strategy utilizes a dynamic phasing algorithm to assemble reads and construct one or 2 phased consensus sequences by de novo assembly of mapped paired-end sequences. These consensus sequences are then aligned to the HLA allele database to find the best fit.

Unambiguous genotypes composed of 2 phased sequences were finally obtained for 51 to 86 different individuals (as mentioned above in HLA typings), depending on the locus (the precise numbers per locus are given in Tables 2 and 3). Only a few alleles (1.28%) were found to be ambiguous. In those cases, we kept the list of all possible alleles to perform the analyses.

2.6 | NGS-MiSeq sequence alignments

Consensus sequences for each individual and each locus were retrieved from NGS-MiSeq sequencing. Sequences were aligned against a reference sequence (from IMGT/HLA[58] v3.25.00) using MAFFT[59] and exons were mapped onto the alignments using MAFFT "*--add*" option. Table 1 lists the different reference sequences used. Alignments were checked individually to remove divergent and badly aligned sequences. From these alignments, gene regions (corresponding to the "gene features" defined by Mack 2015,[60]

TABLE 1 Reference sequences used for the alignments of the 8 HLA loci and their gene regions

Locus	Reference gene	Reference exon	Number of exons	N Seqs
A	*A*01:01:01:01*	*A*01:01:01:01*	8	174
B	*B*07:02:01*	*B*07:02:01*	7	166
C	*C*01:02:01*	*C*01:02:01*	8	166
DRB1	*DRB1*01:01:01*	*DRB1*01:01:01*	6	160
DQA1	*DQA1*01:01:02*	*DQA1*01:01:01:01*	4	158
DQB1	*DQB1*02:01:01*	*DQB1*05:03:01:01*	6	154
DPA1	*DPA1*01:03:01:01*	*DPA1*01:03:01:01*	4	102
DPB1	*DPB1*02:01:02*	*DPB1*01:01:01*	5	166

N Seqs, number of consensus sequences retrieved after the filtering steps.

and ranging from 5'UTR to 3'UTR when available) were extracted separately for each sequenced individual.

2.7 | Population genetics analyses

2.7.1 | Comparison of molecular typing strategies

For each locus, the HLA genotype distributions obtained for the Mandenka individuals by using different molecular typing strategies (PCR-SSO, NGS-454 and NGS-MiSeq) were compared 2 by 2. To that aim, we considered the subsamples of individuals typed by each pair of techniques. In order to be conservative, a good match between 2 techniques was only reported when the 2 alleles of the compared genotype were identical (ie, if at least one allele was different, the comparison was reported as a mismatch).

We also used random sampling to estimate a possible variability in the number of genotype matches due to unequal numbers of individuals used to compare different pairs of techniques (either at the same locus or at different loci). As the smallest number of individuals compared between 2 techniques was 66 (both at HLA-C when comparing PCR-SSO and NGS-MiSeq and at HLA-DQA1 when comparing NGS-454 and NGS-MiSeq), we generated 1000 pairs of random samples of 66 individuals taken without replacement among those who were genotyped by 2 different techniques at each locus and we calculated their match. For each pair of techniques compared at each locus, we then plotted in the same graph the observed number of genotype matches and the distribution of genotype matches generated by this resampling procedure.

2.7.2 | Allele and haplotype frequencies and linkage disequilibrium

We then used the subsamples of unrelated individuals typed at different loci and by different typing strategies to estimate allele and haplotype frequencies as well as other basic statistics (ie, number of alleles observed (k), allelic richness (*ar*), heterozygosity (*H*) and number of most frequent alleles reaching a cumulated frequency of at least 50% (*F50*) at each studied locus). HLA allele and haplotype frequencies were estimated by using an Expectation-Maximisation (EM) algorithm. Global linkage disequilibrium between each pair of loci was tested by both parametric and non-parametric approaches and individual linkage disequilibrium between the alleles of each haplotype by means of standardised residuals. All analyses were performed with the GENE [RATE] computer tools.[61,62] Allelic richness was computed with the rarefaction method,[63] estimating the number of alleles that would be detected if all sample sizes were as small as the smallest sample size used in the study.

2.7.3 | Neutrality tests

For each exon taken separately, Tajima's *D*,[64,65] the ratio of non-synonymous to synonymous substitutions d*N*/d*S*[66] and the nucleotide diversity π indexes were computed (Tajima's D and π with Arlequin v3.5 software[67] and d*N*/d*S* with the MEGA software,[68] using the Nei-Gojobori method[69]) both on the whole exon sequence (obtained with NGS-MiSeq) and on the first, second and third nucleotides of each codon to explore the nucleotide variation across all non-degenerate and degenerate nucleotides. Codons encoding the antigen recognition site (ARS)[70] were analysed separately from the other codons (non-ARS) of exons 2 (for class I and II) and 3 (for class I). Note that, as it was impossible to distinguish putative indels from non-sequenced positions, all gene regions' sequences with gaps were discarded from the analysis (the only exception being for HLA-DQA1 exon 2, which includes a well-known indel tri-nucleotide polymorphism at codon 56), thus focusing the study only on nucleotide substitutions' variability. A missing level of 5% per site was allowed in Arlequin's computations, whereas MEGA considered the codons with gaps as missing information.

As we performed Tajima's test for selective neutrality, which is 2-sided, we used "*p-Adj = 2*(1 – p-Value)*" for *P*-Values above .5 and "*p-Adj* = 2*p-Value" for *P*-Values below .5. The Benjamini-Hochberg[71] correction for multiple testing was then applied, as implemented in R. *dN/dS* significance was assessed with the *Z* test, where "*Z = (dN-dS)/(Var(dN) + Var(dS))*" follows a standard normal distribution under the null hypothesis H_0.

As several correlated statistics were computed for several gene regions at several HLA loci (Tajima's D and number of segregating sites S, d*N* and d*S*), we compared the molecular diversity of all sequenced exons 2, 3, 4 and 5 (the latter for class I genes), which were well covered at all studied loci, globally by performing a Principal Component Analysis (PCA) on these statistics.

3 | RESULTS

3.1 | Genotype matches between the 3 HLA typing strategies

Table 2 gives the proportions of genotype matches observed between each pair of typing strategies used to analyse the

TABLE 2 Number of genotype matches between PCR-SSO, NGS-454 and NGS-MiSeq

Locus		Pairwise comparisons between different typing techniques					
		PCR-SSO vs NGS-454		NGS-454 vs NGS-MiSeq		PCR-SSO vs NGS-MiSeq	
HLA-A	No. of genotyped individuals					*196*	*87*
	No. of compared individuals					*85*	
	No. of genotype matches (%)					*83 (97.6%)*	
	CI95					97.0–100.0%	
HLA-B	No. of genotyped individuals					*198*	*83*
	No. of compared individuals					*82*	
	No. of genotype matches (%)					*75 (91.5%)*	
	CI95					89.4–95.5%	
HLA-C	No. of genotyped individuals					*165*	*83*
	No. of compared individuals					*66*	
	No. of genotype matches (%)					*52 (78.8%)*	
	CI95					72.7–72.7%	
HLA-DRB1	No. of genotyped individuals	*198*	*194*	*194*	*81*	*198*	*81*
	No. of compared individuals	*188*		*78*		*77*	
	No. of genotype matches (%)	*135 (71.8%)*		*70 (89.7%)*		*47 (61.0%)*	
	CI95	63.6–80.3%		87.9–92.4%		57.5–65.2%	
HLA-DQA1	No. of genotyped individuals			*194*	*82*		
	No. of compared individuals			*66*			
	No. of genotype matches (%)			*64 (97.0%)*			
	CI95			97.0%			
HLA-DQB1	No. of genotyped individuals	*195*	*196*	*196*	*76*	*195*	*76*
	No. of compared individuals	*188*		*74*		*72*	
	No. of genotype matches (%)	*153 (81.4%)*		*68 (91.9%)*		*13 (18.1%)*	
	CI95	72.7–87.9%		90.9–93.9%		15.2–19.7%	
HLA-DPB1	No. of genotyped individuals	*193*	*199*	*199*	*82*	*193*	*82*
	No. of compared individuals	*193*		*82*		*79*	
	No. of genotype matches (%)	*172 (89.1%)*		*72 (87.8%)*		*24 (30.3%)*	
	CI95	83.3–94.0%		84.9–90.9%		25.8–36.4%	

No., Number; CI95, 95% confidence interval of the percentage of matches estimated by considering that the number of individuals compared between 2 different techniques is as small as the smallest number of individuals actually compared (ie, 66 individuals which is the number compared both for HLA-C between PCR-SSO and NGS-MiSeq and for HLA-DQA1 between NGS-454 and NGS-MiSeq); these intervals were obtained by drawing 1000 random samples of 66 individuals without replacement, see section 2.

HLA polymorphism in the Mandenka population. These values are also shown in Figure 1, together with violin plots representing the distributions of the proportion of genotype matches estimated on 1000 random samples of identical size to account for a possible variability due to differences in the number of individuals compared for each pair of techniques (see section 2). The shape and size of these violin plots as well as their position relatively to the observed values indicate that differences in sample sizes occur as expected from a well-behaved normal sampling distribution, hence ensuring the absence of sampling bias in our results. As a further evaluation of the relative importance of the typing mismatches on the characterisation of the HLA genetic profile of the Mandenka population, we also estimated HLA allele frequencies and basic statistics on the subsamples of unrelated individuals typed at each locus by each technique (Table 3 and Supplementary Information S02).

Relatively good matches were found between PCR-SSO and NGS-454 ("SSO versus 454" on Figure 1) for the 3 class II genes considered (from 71.8% for DRB1 to 89.1% for DPB1). Although both typing techniques were applied to exon 2 for these loci and were thus expected to detect identical genetic information, some groups of alleles were not detected at all by PCR-SSO (eg, *DRB1*07* and *DQB1*02* alleles) and a few alleles reported with PCR-SSO (eg, *DRB1*08:02*) were not confirmed by NGS-454 (Supplementary Information S02). These results suggest a lack of resolution of PCR-SSO compared with NGS-454.

Very good matches were found between the 2 NGS techniques ("454 versus MiSeq" on Figure 1) for the 4 class II loci compared (from 88% for DPB1 to 97% for DQA1) despite the fact that they do not target the same DNA regions. NGS-MiSeq sequences include all exons and are generally assigned to one unique HLA allele (98.7% of unambiguous typings were obtained at the third field level of resolution). By contrast, NGS-454 sequences are only defined at exon 2 here, and are thus often ambiguously assigned to several alleles (ranging from a mean of 2.8

TABLE 3 Statistics describing the genetic diversity of the Mandenka population based on HLA molecular typings obtained by 3 different techniques, PCR-SSO, NGS-454 and NGS-MiSeq

Locus	*N* (total)			*N* (unrelated)			*k/ar*			*H* (%)			*F50*		
	PCR-SSO (second -field)	NGS-454	NGS-MiSeq	PCR-SSO (second -field)	NGS-454	NGS-MiSeq	PCR-SSO (second -field)	NGS-454	NGS-MiSeq	PCR-SSO (second -field)	NGS-454	NGS-MiSeq	PCR-SSO (second -field)	454	NGS-MiSeq
A	196	–	87	72	–	72	23/21.6	–	22/20.7	92.2	–	92.0	5	–	5
B	198	–	83	67	–	67	30/27.4	–	30/27.8	93.6	–	93.6	6	–	7
C	165	–	83	54	–	54	15/15	–	18/17.9	89.4	–	91.0	4	–	5
DRB1	198	194	81	96	96	65	22/19.0	23/20.3	20/16.9	87.7	87.6	87.6	3	4	4
DQA1	–	194	82	–	66	66	–	9/9	14/13.0	–	60.7	71.5	–	1	1
DQB1	195	196	76	94	96	60	12/10.9	11/10.5	13/12.8	66.2	68.2	76.7	1	1	2
DPA1	–	–	51	–	–	51	–	–	10/10	–	–	71.9	–	–	2
DPB1	193	199	82	99	101	70	18/14.4	14/12.7	19/16.6	80.0	80.7	86.3	2	2	3

N (total), total number of individuals analysed for which the typing yielded usable results; *N* (unrelated), number of unrelated individuals analysed for which the typing yielded usable results; *k*, number of alleles detected; ar, allelic richness or number of alleles expected in a population whose size is equal to the smallest *N* used with a given technique (ie, 54 for PCR-SSO, 66 for NGS-454 and 54 for NGS-MiSeq); *H*, heterozygosity; *F50*, number of most frequent alleles whose cumulated frequency reaches at least 50%. Class I loci were not typed with NGS-454, DQA1 was not typed with PCR-SSO, and DPA1 was only typed with NGS-MiSeq.

alleles per sequence for DPB1, to 12.3 for DQB1, see Supplementary Information S03 for a detailed list of possible alleles per NGS-454 sequence) due to segregating sites located outside this exon. Indeed, our results show that a number of NGS-454 alleles, some of which reach very high frequencies in the Mandenka population (Supplementary Information S02), are split through NGS-MiSeq, namely *DQB1*03:01* into *DQB1*03:19, 03:01:01 and 03:01:04*, and *DPB1*17:01* into *DPB1*17:01* and *131:01. DQB1*03:19* (allele frequency [AF] = 0.43) is discriminated from *DQB1*03:01:01* (AF = 0.07) through a segregating site located in exon 3 (codon 185: ATC → ACC); likewise, *DPB1*17:01* and *DPB1*131:01* (AF = 0.22 and 0.20, respectively) differ by 7 substitutions located in exon 3, plus 1 in exon 4. The very good matching scores shown in Figure 1 between NGS-454 and NGS-MiSeq thus correspond to genotypes with compatible allele assignments between exon 2 sequences rather than to genotypes with real allelic matches.

Finally, PCR-SSO and NGS-MiSeq genotypes were compared for both class I and II genes ("SSO versus MISeq" on Figure 1), revealing major differences among the loci. Fairly good concordances were achieved at the class I loci (from 78.8% for HLA-C to 97.6% for HLA-A), whereas very low concordances were obtained for some class II loci (18.1% for HLA-DQB1 and 31.3% for HLA-DPB1). A less dramatic situation was found for locus DRB1 (61% of matches), where the most frequent allele, *DRB1*13:04* (frequency of almost 30% in the Mandenkalu), was detected by both techniques and was not split with NGS-MiSeq. By contrast, at the other 2 loci the most frequent allele(s) differ(s) between PCR-SSO and NGS-MiSeq: with MiSeq, *DQB1*03:19* was found in place of *DQB1*03:01*, and both *DPB1*17:01* and *DPB1*131:01* were found in place of *DPB1*17:01.* As mentioned above, these alleles differ by substitutions located outside exon 2, but these polymorphisms were unknown at the time of the PCR-SSO typings and could thus not be reported as ambiguities. For example, the most frequent DQB1 allele found with PCR-SSO was reported as 03:01 whereas it was ambiguously defined as 03:01 *or* 03:19 with NGS-454 after the discovery of 03:19 in 2007,[72] and was finally found to be 03:19 with MiSeq. This explains why NGS-454 vs NGS-MiSeq gave relatively high matching scores whereas PCR-SSO vs NGS-MiSeq gave low scores for DQB1. A similar explanation holds true for DPB1.

3.2 | HLA molecular profile of the Mandenka population

3.2.1 | Allele and haplotype frequencies

HLA allele frequencies (Supplementary Information S02) and basic summary statistics at each locus (Table 3) as well as the results of global linkage disequilibrium (LD) tests between each pair of loci and the list of 2-locus haplotypes

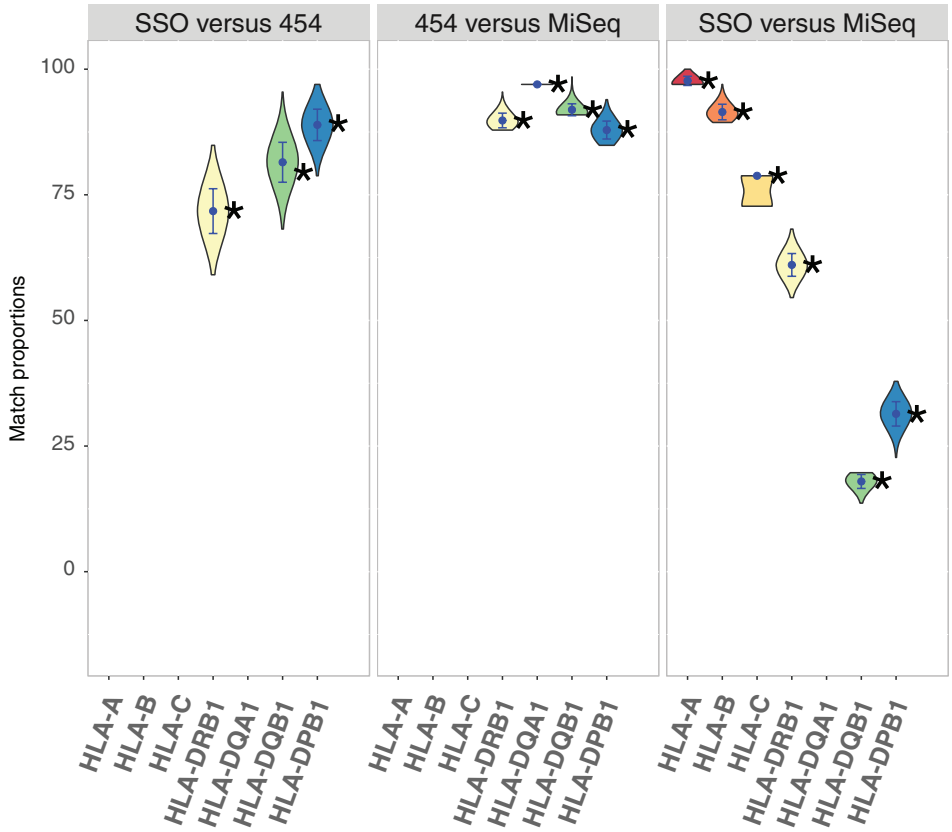


FIGURE 1 Proportions of genotype matches (or « matching scores ») obtained between PCR-SSO ("SSO" in the Figure), NGS-454 ("454") and NGS-MiSeq ("MiSeq") at 3 to 6 HLA loci (DQA1 was not typed with PCR-SSO and class I genes were not typed with NGS-454). For the comparisons involving NGS-454, as the sequences obtained with this technique were limited to exon 2 and may thus correspond to different alleles, we reported a match when the allele found with the other technique was *compatible* with the NGS-454 sequence. Only perfect matches (i.e. for any compared genotype, when both alleles found with the 2 techniques were either compatible or identical) were counted. The comparisons were replicated 1′000 times each on random samples of 66 individuals (corresponding to the lowest sample size of the observed data) drawn without replacement to assess the variability of genotype matches due to sampling size

in significant LD (Supplementary Information S04) estimated in the Mandenka population on the basis of the most precise typing technique that we used, NGS-MiSeq, are summarized in our companion *Population Report.* [73] Relevant differences were observed between class I and class II allele frequencies, as the former are all below 20% whereas the latter always exhibit at least one allele with a frequency above 20%, that is, *DRB1*13:04* (28.5%), *DQA1*05:05:01* (50%), *DQB1*03:19* (44.2%), *DPA1*02:01:01* (46.1%) and *DPB1*17:01* (21.8%, the frequency of another allele, *DPB1*131:01*, being very close: 18.9%). This is reflected by the contrasting diversity indexes (allelic richness, heterozygosity and *F50*) values given in Table 3, in particular between the A and B loci, on one side ($ar \geq 21$, $H \geq 92\%$, $F50 = 5$), and the DQ and DP loci, on the other side ($ar \leq 16$, $H \leq 86\%$, $F50 \leq 3$).

At first sight, our analyses also suggest the presence of an extended class II haplotype, *DRB1*13:04~DQA1*05:05:01~DQB1*03:19~DPB1*131:01*, as all allelic pairs of this putative haplotype are in significant LD (Supplementary Information S04). However, if we consider the DPA1 locus, *DPA1*02:01:01* is in LD with *DRB1*13:04* but not with either *DQA1*05:05:01* or *DQB1*03:19*, and is also significantly associated with both *DPB1*131:01* and *DPB1*17:01*. Therefore, our results merely support the existence of 3 frequent class II haplotypes, that is, *DRB1*13:04~DQA1*05:05:01~DQB1*03:19, DPA1*02:01:01~DPB1*131:01* and *DPA1*02:01:01~DPB1*17:01*, rather than an extended haplotype across the 5 class II loci. In addition, we find no global LD between the DP loci and the other class II genes, in agreement with the existence of (a) recombination hotspot(s) between DQB1 and DPA1 near the TAP2 genes[74,75] whereas DRB1, DQA1 and DQB1 are strongly associated with each other, as are DPA1 and DPB1 (Supplementary Information S04).

3.2.2 | Nucleotide diversity

Figure 2 shows the nucleotide diversity per site ($\pi \pm \sigma$) at exons 2, 3, 4 (for class I and class II) and 5 (for class I) of the 8 HLA loci (with distinction between ARS and non-ARS codons for the peptide binding region) and the amino acid diversity per site (estimated after translating all codons into amino acids) at the corresponding α1-α4 and β1-β3 domains of the HLA molecules, arranged by structurally comparable elements (peptide binding region [PBR], T-cell receptors interaction region [TCRIR] and trans-membrane region [TMR]), in the Mandenka population. Other regions were discarded from the analysis because they were insufficiently covered by sequencing.

At both class I and class II genes, the nucleotide diversity appears to be greater at ARS than at non-ARS codons of exons 2 (for class I and II) and 3 (for class I) encoding the PBR (top left of Figure 2). The differences between ARS and non-ARS are even more pronounced (ie, most often significant) if we consider the amino-acid diversity of the corresponding domains of the HLA molecules (bottom left of Figure 2), in support of an advantage of diversity at the sites involved in peptide presentation.[76,77] We also

FIGURE 2 Nucleotide (top) and inferred amino acid (bottom) diversity per site (±σ) at exons encoding the peptide-binding region (left, with a distinction between antigen-recognition sites (ARS) and non-antigen-recognition sites (non-ARS) sites); the domains interacting with CD4+ and CD8+ T-cell receptors (middle); and the trans-membrane region (right) of the HLA-A, -B, -C, -DRB1, DQA1, DQB1 and DPB1 molecules in the Mandenka population. Brackets remind the chains forming the HLA-DQ (DQA1 and DQB1) and HLA-DP (DPA1 and DPB1) dimers

TABLE 4 Results of Tajima's *D* and *dN/dS* selective neutrality tests for the 7 (among 86) regions rejecting significantly the null hypothesis of selective neutrality according to one or both tests (see Supplementary Information S05 for the results relative to the other regions)

Gene region	B exon 2 (ARS)	A exon3 (ARS)	B exon 3 (ARS)	DPA1 intron 1	DPB1 exon 2 (ARS)	DPB1 intron 2	DPB1 exon 3
Size (bp)	66	54	54	3584	75	4014	282
S	18	17	12	106	8	133	7
Tajima's D	2.4	2.9	2.3	**3.5**	**3.7**	**4.0**	**4.0**
Adj. p-value	.07	.06	.08	**0**	**0**	**0**	**0**
dN/dS	**6.6**	**3.7**	**8.0**	-	**dN = .07, dS = 0**	-	.2
Z value	**2.91**	**3.02**	**2.29**	-	**2.29**	-	-1.17

Size (bp), length of the region in base pairs; S, number of segregating sites; Adj. p-value, p value corrected for multiple testing according to Benjamini Hochberg (fdr), $\alpha = .05$. *Z* value is significant when outside the [-1.96:1.96] interval. Values in bold are significant.

observe a lower ARS diversity at the α1 (encoded by exon 2) than at the α2 (encoded by exon 3) domain of HLA-C and (to a lesser degree) HLA-A, suggesting that these molecules (particularly HLA-C) are less prone to diversifying selection at their α1 domains. By contrast, the ARS diversity is particularly high at the β1 domain (encoded by exon 2) of HLA-DQB1 and -DRB1 compared with the α1 and/or β1 domains of other class II molecules. Finally, both the nucleotide and amino-acid diversities observed within the other structural elements (TCRIR and TMR) are low (below 0.05 and 0.1, respectively), with little differences among the loci, and of the same range as those observed at non-ARS sites within the PBR. The only locus showing such a low diversity at ARS sites is HLA-DPA1.

3.3 | Selective neutrality across the HLA regions

3.3.1 | Selective neutrality tests

We applied Tajima's test of selective neutrality on all gene regions (ie, exons and introns) sufficiently covered by

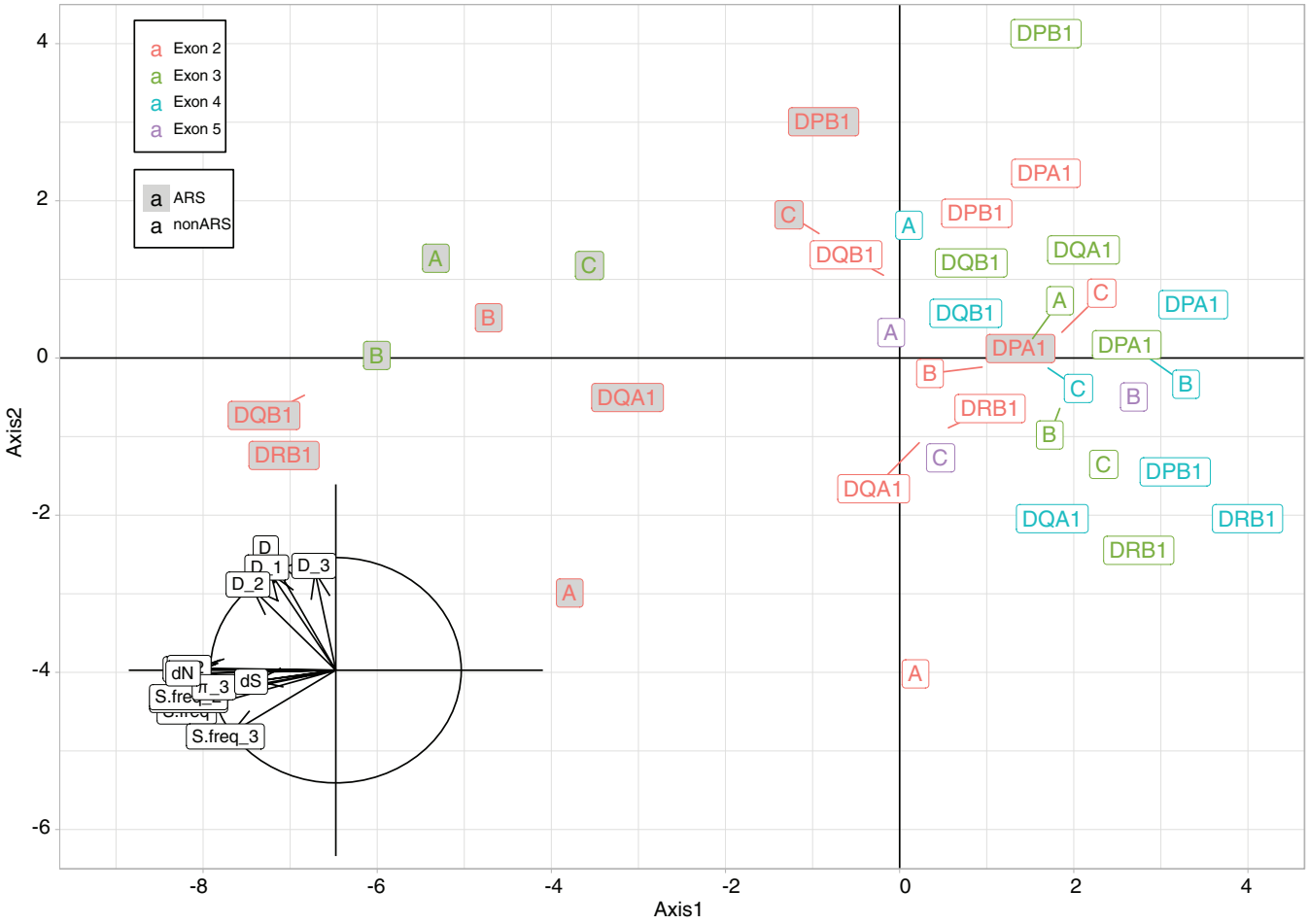


FIGURE 3 Principal component analysis (axes 1 and 2, explaining respectively 60% and 20% of the total variance) based on Tajima's *D*, nucleotide diversity π, frequency of segregating sites S.freq, number of non-synonymous *dN* and synonymous *dS* nucleotides at exons 2, 3 and 4 of loci HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1, -DPB1 and exons 5 of loci HLA-A, -B, -C. Symbols D, D_1, D_2, D_3 represent Tajima's *D* estimated on the whole gene region and at the first, second and third nucleotide of each codon, respectively. Similarly, π, π_1, π_2 and π_3 as well as S.freq, S.freq_1, S.freq_2 and S.freq_3 represent the nucleotide diversity π and the frequency of segregating sites S estimated on the whole gene region and at nucleotide positions 1, 2 and 3 of each codon, respectively. Grey boxes correspond to ARS codons, white boxes to non-ARS codons. The inset graph at the bottom left represents the correlations between the projections of the variables (for each pair of variables, the correlation is measured by the cosine of the angle of the 2 variable vectors) on the plan of the PCA

sequencing of the 8 studied HLA loci (Supplementary Information S05). Most Tajima's *D* values are positive and sometimes significant, but only 4 regions still exhibit a significant value after correction for multiple tests (Table 4), 3 of them being neighbouring regions (in significant linkage disequilibrium, results not shown) within the DPB1 locus [DPB1 exon 2 (ARS)—DPB1 intron 2—DPB1 exon 3]. We also observe a much higher proportion of segregating sites (S) at DPB1 exon 2 (ARS) (S = 8 for 75 nucleotides, that is, 10.7%) than at DPB1 exon 3 and intron 2 (2.5% and 3.3%, respectively, Table 4).

Because genomic regions evolving under a neutral model of molecular evolution are expected to exhibit similar proportions of non-synonymous (*dN*) and synonymous (*dS*) nucleotide substitutions,[78] we tested a putative deviation from 1 of the *dN/dS* ratio at each HLA exon of each locus (Supplementary Information S05). We found 4 regions with a *dN/dS* ratio significantly greater than 1, that is, the ARS codons of HLA-A and -B exons 3 and of HLA-B and -DPB1 exons 2.

3.3.2 | PCA of molecular diversity indices

To better understand the results of our nucleotide diversity analyses and selective neutrality tests, we compared more in depth the molecular diversity of exons 2, 3 (split into ARS and non-ARS codons for class I exons 2 and 3 and class II exons 2) and 4 of the 8 sequenced HLA genes and of exon 5 of HLA class I genes by including all estimated indexes related to genetic selection (ie, Tajima's *D*, nucleotide diversity *π*, frequency of segregating sites *S.freq*, number of non-synonymous *dN* and synonymous *dS* nucleotide substitutions, computed both across the whole gene region and at the first, second (non-degenerate) and third (degenerate) nucleotide of each codon, respectively) in a principal component analysis (PCA, Figure 3). The ratio *dN/dS* was not used in this study because both *dN* and *dS* values were

already included and because HLA-DPB1 ARS and exon 4 exhibit *dS* values of 0, therefore making *dN/dS* undetermined.

The inset graph at the bottom left of Figure 3 represents the correlations between the projections of the variables (for each pair of variables, the correlation is measured by the cosine of the angle of the 2 variable vectors) on the plan of the PCA. The first 2 principal components (PC), axes 1 and 2, explain together 80% of the total variance, and the third PC adds another 11% (see Supplementary Information S06). The first PC (60%) appears to be (inversely) correlated to the amount of molecular diversity described by π, *S.freq* and *dN* indexes (left side of the graph). Therefore, the left part of the graph mostly indicates non-synonymous molecular variation (which fits with the left position of *dN*), whereas the right part indicates low molecular variability. The second PC seems to be directly correlated with Tajima's *D* values (both global and by positions). It is thus expected to discriminate regions evolving under balancing selection (top) and purifying selection (bottom). Note, however, that demographic effects may also affect *D* values,[64,65] that is, population expansion towards negative *D* values (bottom of the graph).

Interestingly, the first PC discriminates most ARS codons (except that of HLA-DPA1), on the left, from non-ARS codons (except those of HLA-DQA1 and DQB1), on the right, all other codons (exons 4 and 5) falling also on the right side of the graph. This confirms that the greatest amount of HLA molecular variation is generally concentrated within the ARS, and is thus related to the peptide presentation function of the HLA molecules. Interestingly, the most extreme positions on the left are those of the ARS of DQB1, DRB1 and B exons 2 and of A exon 3. Along the second PC, most regions projected at the top of the graph deviate significantly from selective neutrality according to individual Tajima's D tests (Supplementary Information S05). However, DPB1 exon 2 regions (both ARS and non-ARS, extreme top) show the strongest signals of balancing selection (Tajima's $D >> 0$ remaining significant after correction for multiple testing) despite their lower nucleotide diversity compared with most other exons 2 and 3. Interestingly, this axis also discriminates HLA-A exon 2 (both ARS and non-ARS) in its bottom part due to very low (partly negative) Tajima's *D* values (Supplementary Information S05), which may be interpreted either as purifying selection or, more likely, as population expansion (see section 4).

4 | DISCUSSION

In this study, our objective was to provide insights into the molecular diversity and evolution of the HLA genomic region by deciphering the fine nucleotide diversity of 8 HLA loci in a well-characterized human population. To that aim, we used 2 NGS techniques to sequence HLA genes in a West-African population sample that we formerly typed by PCR-SSO for several HLA loci within the frame of a wider research project investigating the genetic diversity of many independent loci of the genome in this population (see references in the Introduction). This allowed us, first, to evaluate, on the same population sample, the advantage acquired by using NGS rather than classical molecular technologies (eg, PCR-SSO) for the characterisation of HLA population diversity and, second, to explore in detail the nucleotide variation within and among distinct gene regions of the 8 HLA loci.

4.1 | Contribution of NGS to the assessment of HLA population diversity

Previous studies have evaluated the accuracy of different HLA-typing methods, for example, by comparing PCR-SSO to NGS-454[79] or PCR-SSO to paired Illumina MiSeq short-reads[80] typings for HLA class I. The proportion of matches between PCR-SSO and NGS-MiSeq found by Major et al in the latter study (after applying on those data a similar way to count the matches as we did in our study, that is, by counting a match only when the 2 alleles of an individual were compatible or identical between the 2 methods, and a mismatch when at least one of the 2 alleles was discordant) is slightly different to ours (84.7%, 93.9% and 85,6% at HLA-A, -B and -C, respectively, vs 97.6%, 91.5% and 78.8% in our study). Actually, some alleles were unknown (and thus not detectable) at the time of the PCR-SSO studies, for example, *HLA-A*11:50Q*, *01:11N* and *03:21N*, first reported between 2005 and 2009, were found with NGS-MiSeq by Major et al[80] but mistyped with PCR-SSO according to the HapMap data (described in 2003[81]) used by these authors, which mainly explains why the proportion of matches is lower at HLA-A in their study compared with ours; also, *HLA-C*07:18*, first reported in 2002, was often mistyped by PCR-SSO (in the late 1990s) as *07:01* in the Mandenka where this allele is frequent, explaining in part why the proportion of matches is lower at HLA-C in our study compared with Major et al. This also suggests that by the time of our PCR-SSO study, primers were used to be developed for alleles frequent in Europe and North-America, and were thus likely to be less adapted to type African samples. Nowadays, class II alleles are also defined by taking into account nucleotide substitutions at exon 3, which explains why, by the time of the first PCR-SSO typings, many HLA alleles had still to be discovered. Indeed, taking into account the molecular information of exon 3 did allow us to better identify the most common DQB1 allele found in the Mandenka population, *DQB1*03:19*, which was first defined in 2007.[72] Also, half of the DPB1 alleles observed in this study (among which *DPB1*131:01*, described in 2010) were not reported before 1999, while their cumulated frequency reaches 33% in the

Mandenka population, thus probably explaining the low matching score between PCR-SSO and NGS-MiSeq typings obtained for DPB1. By contrast, of the 21 alleles observed at the DRB1 locus, only 2 minor (ie, less frequent) alleles were unknown before 1992 (*DRB1*12:10* and *DRB1*14:54:01*). These results confirm that we acquired a significant amount of information on HLA class II, and more particularly on DQB1 and DPB1 diversity, by improving our typing strategies. The kind of typing mismatches found in this study has to be kept in mind when using traditional typing techniques (like PCR-SSO) or limited sequencing strategies (like single exon NGS-454 sequencing) rather than full-gene sequencing for the characterisation of HLA genetic profiles. This also suggests that care should always be taken in interpreting results depending on the technique used. In particular, potential ambiguities that are present at the time of the analyses should always be reported rather than "solved" by arbitrary or *ad hoc* procedures, even when using sequence-based typings.[82] This can be easily achieved using UNIFORMAT[83] or alternatives like Genotype List strings[84] or MIRING.[85]

4.1.1 | Gene conversion and pathogen-driven selection (selective sweep) at HLA loci

The HLA allelic profile characterising the West African Mandenka population (the basic statistics of which are given in our companion *Population Report*[73]) brings us new insights into the mechanisms that drove the evolution of this polymorphism, mostly because this population lives in a region where infectious diseases, like malaria, are highly prevalent. Surprisingly, the Mandenkalu do not exhibit a particularly high frequency of *B*53:01:01* (Allele Frequency [AF] = 0.06)—the most commonly recognised HLA class I allele protective to this disease—contrary to what is observed in most West-African regions as a putative result of resistance to *Plasmodium falciparum*.[29] However, the most frequent HLA-B allele observed in the Mandenkalu, *B*35:01:01* (AF = 0.16), has been reported as protective to malaria in Ghana[86] and its predicted peptide-binding profile is similar to that of *B*53:01:01*,[29] which also suggests a protective effect to this disease (this is also the case, although to a lesser extent, for the following 2 most common alleles found in this population, *B*15:03:01* (AF = 0.08) and *B*78:01:01* (AF = 0.08)). This indicates that distinct HLA alleles may be interchangeable in terms of protection to given pathogens.

On the other hand, we also wondered why a unique HLA class II haplotype in strong linkage disequilibrium, *DRB1*13:04~DQA1*05:05:01~DQB1*03:19*, was so frequent in the Mandenkalu, a large population that was not particularly prone to rapid genetic drift (see section1). Interestingly, *DRB1*13:04*, which was previously identified as a predominant allele in West Africa,[19,87] was proposed to originate from an allelic conversion from DRB1*11:02 based on both serological[88] and RFLP1[87] analyses. We investigated more in depth this hypothesis comparing the sequences of the *DRB1*13:04* exons to those of the 1913 other DRB1 alleles taken from the IMGT/HLA database 3.25.0. As the allelic conversion is supposed to have happened on exon 2, we expected the 5 other exons to be identical between *DRB1*13:04* and the putative recipient allele of the allelic conversion.

Of the 1913 DRB1 alleles studied (defined at the third field level of resolution in the database), 86 had differences (ranging from 4 to 22 substitutions) with *DRB1*13:04* only in exon 2. The span from the first to the last substitution ranged from 30 to 241 base pairs (bp), apart from *DRB1*11:02:01* for which the 5 substitutions were on a 6 bp-long fragment, AGCGCC. Of the 1913 alleles, 195 - (Supplementary Information S07) had the AGCGCC pattern, with (compared with *DRB1*11:02:01/DRB1*13:04* exons 2) 32±28 conserved nucleotides before and 32±16 conserved nucleotides after the fragment (the underlying hypothesis here being that the longer the fragment, the more likely the recombination could happen), among which *DRB1*08:03*, which was formerly proposed as a donor for this gene conversion.[88] However, *DRB1*08:03* was not detected in the Mandenkalu and this allele is frequent in South-East Asia, but not in Africa (at the opposite of *DRB1*13:04*). On the other hand, among the 195 potential donor alleles, 3 are observed in the Mandenkalu, that is, *DRB1*04:05:01* (AF = 0.008), *DRB1*08:06* (AF = 0.054) and *DRB1*13:03:01* (AF = 0.023). We thus find very likely that a gene conversion generating *DRB1*13:04* occurred (likely in West Africa) with *DRB1*11:02:01* as recipient and with 1 of the 3 alleles listed above as possible donors. A scheme of the proposed allelic conversion is shown in Figure 4. Allelic conversion being a frequent event in the HLA region,[89] this illustrates very well the progress brought by NGS: a former hypothesis of allelic conversion based on serological and RFLP typings has been refined thanks to detailed DNA sequence information.

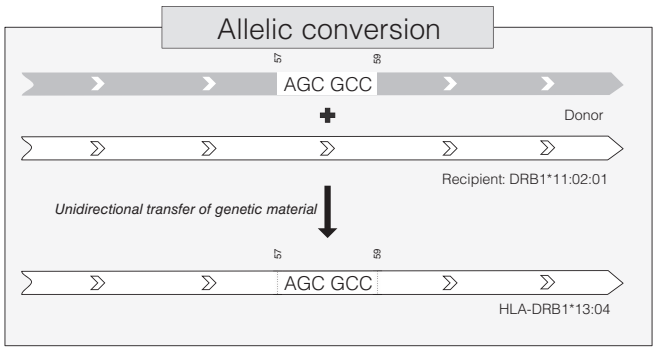


FIGURE 4 Putative mechanism of the allelic conversion mentioned in this study, which suggests an unidirectional transfer of genetic material including the « AGCGCC » pattern from a donor allele (potentially *DRB1*04:05:01, DRB1*08:06 or DRB1*13:03:01* in the Mandenka) to the recipient allele *DRB1*11:02:01*, leading to the creation of the *DRB1*13:04* allele

A further challenge was to understand the very high frequency of the *DRB1*13:04* allele (AF = 0.284) in the Mandenkalu, such a high allelic frequency being a rare situation at locus DRB1 in populations which have not been submitted to rapid genetic drift, like the Mandenka. To our knowledge, *DRB1*13:04* has not been reported as protective for any infectious disease so far, whereas the alleles found in linkage disequilibrium at the other class II loci, namely *DQA1*05:05:01* and *DQB1*03:19*, are potential targets of positive selection, as explained hereafter. Actually, *DQB1*03:19* was only discovered in 2007, when Witter et al[72] found that it differed from *DQB1*03:01* by a single non-synonymous substitution in the third exon (position 544 C → T), whereas exon 2 is identical. Interestingly, the combination of *DQB1*03:01* and *DQA1*05:01* (whose peptide-binding region is also identical to that of *DQA1*05:05*) was reported by Meyer et al[90] to be more frequent in individuals putatively immune to onchocerciasis disease, suggesting a role of these 2 alleles in the immune defence against the filarial parasitic worm *Onchocerca volvulus* (*O. volvulus).* Onchocerciasis, or "river blindness," is a highly prevalent disease in West Africa, including Eastern Senegal.[91] Although it does not directly kill the infected people, it leads to severe disabilities and decreases life expectancy due to a reduced immunity. As *DQB1*03:19* probably behaves like *DQB1*03:01* regarding peptide presentation (their unique amino acid difference being in the β2 domain), a likely hypothesis is thus to consider *DQB1*03:19* as a protective allele to *O. volvulus*, which would have led to a strong selective sweep increasing the frequency of the whole haplotype *DRB1*13:04~DQA1*05:05:01~DQB1*03:19* in the Mandenkalu. The fact that this West African population lives in an area where *O. volvulus* is highly prevalent[91] strongly supports this hypothesis. The hypothesis formerly proposed by Hill et al in 1992[87] of recent directional selection to account for the high frequency of *DRB1*13:04* in Gambia finds here a more precise explanation.

4.1.2 | Signatures of natural selection and demography on distinct HLA gene regions

Taken individually, several exons 2 ARS regions (B, C, DPB1, DQB1), all exons 3 ARS regions of class I (A, B, C) loci, several exons 3 of class II (DQA1, DQB1, DPB1) loci and a few exons 4 (A, DQB1) generated positive and significant Tajima's *D* values in the Mandenkalu (Supplementary Information S05), as previously found for exon2/exon 3 sequences at most HLA loci in other populations and explained by balancing selection.[13] However, only 2 of the 49 exons included in our analyses, i.e. DPB1 exon 2 ARS and exon 3, still reached significance towards an excess of heterozygotes after correction for multiple tests (86 tests, considering all exons, introns and ARS/non-ARS regions). Moreover, because these 2 regions are in significant linkage disequilibrium, heterozygous advantage did not necessarily affect both of them, as "associative balancing selection" may have also occurred.[92–94] Also, *dN/dS* ratio were only significant at ARS codons of 4 regions, that is, B and DPB1 exons 2, and A and B exons 3 (note, however, that this latter result must be taken with caution as d*N*/d*S* ratio tests on molecular data were originally developed to detect selection by comparing different species and may not be adequate to infer selective pressures for samples drawn from a single population[95]). We thus find limited evidence of natural selection by means of neutrality tests applied to our data, which might be due either to a lack of statistical power or to the interplay of multiple selective forces (eg, both positive and balancing selection).

Actually, the PCA (Figure 3) based on different molecular diversity indices estimated on the 8 HLA genes (Tajima's *D*, nucleotide diversity π, frequency of segregating sites, number of non-synonymous d*N* and synonymous d*S* nucleotide substitutions) revealed remarkable differences between distinct gene regions and loci, suggesting signatures of diverse evolutionary pressures acting across the HLA loci. In the PCA, the weight of the first PC (60% of the total variance) highlights the importance of amino acid diversity (ie, non-synonymous substitutions) within the HLA peptide-binding region, also revealed in Figure 2.

Interestingly, however, not all exons 2 and (for class I) exons 3 seem to be equally targeted by this kind of selection. At exon 3, the 3 class I genes exhibit, as expected, a significantly greater level of diversity at their ARS (the highest being at locus A) than at their non-ARS codons, which is easily explained by the advantage of amino acid variation in the α2 domain of the HLA class I molecules involved (with α1) in peptide presentation. Interestingly, at ARS codons locus C is less diverse than loci A and B, and more particularly at exon 2, a plausible explanation being the main role of the corresponding α1 domain (eg, amino acid position 80) of the HLA-C molecule as a KIR ligand,[96,97] which would weaken the advantage of amino acid diversity related to peptide binding. This hypothesis is in line with Bitarello et al 2016's[11] observations of a lower molecular diversity of HLA-C at the antigen recognition site (ARS) although these authors did not report detailed results on exons 2 and 3. Among the 5 class II genes, DPA1 has the less variable ARS at exon 2, suggesting a marginal contribution of this locus to peptide binding, in agreement with the low number of DPA1 alleles reported in the IMGT/HLA database, whereas B, DRB1 and DQB1 exons 2 are the most variable of all HLA loci.

To our surprise, the PCA also revealed an outlier position of locus A exon 2 (both ARS and non-ARS codons) towards the bottom of the second PC axis, indicating reduced Tajima's *D*'s, actually a very large negative value (−0.89) at its non-ARS codons probably explaining the negative *D* value (−0.42) found for HLA-A exon 2 as a whole (Supplementary Information S05). Although these negative

values are not significantly different from 0, they might represent a signal of demographic expansion (rather than purifying selection), as such a signal is expected in the Mandenka population based on both its known demographic history (see section 1) and other genetic studies (eg, Reference [25]). This would also support the hypothesis that the evolution of the HLA-A polymorphism (at least at the level of its exon 2 region, as suggested by the present study) is closer to neutrality and more prone to reveal demographic signals.[28,29,98,99] At the opposite of locus A exon 2 along the second axis of the PCA lie DPB1 exons 2 and 3. As the 2 most common alleles of this locus, *DPB1*17:01* and *DPB1*131:01*, share an identical exon 2 (they differ at exon 3), a similar selective pressure involving an adaptive resistance to a specific pathogen (to define) would explain their high frequency in the Mandenkalu (cumulated frequency of 42%, close to that of *DQB1*03:19*). Actually, the DPB1 frequency distribution is much more even than that observed at other loci (as shown in Supplementary Information S02) due to high frequencies (>10%) of 4 different alleles (ie, *DPB1*02:01:02* and *DPB1*01:01:01*, in addition to *DPB1*17:01* and *DPB1*131:01*), explaining why DPB1 exon 2 deviates significantly from neutrality, even after correction for multiple tests, towards an excess of heterozygotes. At this locus, the significant excess of heterozygotes also observed at exon 3 would then be due to linkage disequilibrium with exon 2. This result contrasts with previous studies where DPB1 was found to exhibit a more "L-shaped" (ie, neutral-like) distribution than the other loci.[13,100] However, as DQA1, DQB1 and DRB1 (in linkage disequilibrium) have possibly been submitted to a strong selective sweep due to resistance to *O. volvulus* in this population, the resulting loss of diversity at these genes may have been compensated by a greater (and significant) heterozygous advantage at DPB1 (not affected by the selective sweep) conferring protection to other pathogens also present in the environment. Interestingly, this explanation fits with the model of *joint divergent asymmetric selection* recently proposed by Buhler et al in 2016[37] to explain the sharp differences of heterozygosity sometimes found among distinct class I loci. The putative existence of such similar mechanisms at both class I and class II genes makes this model appear even more robust and consistent in the evolution of the HLA region.

5 | CONCLUDING REMARKS

The recent development of high-throughput sequencing technologies applied to HLA can be considered as a major upheaval: as shown in the present study, not only the typing errors dropped, but the deciphering of the fine nucleotide diversity of different HLA gene regions opens new ways to explore the evolution of this exceptional polymorphism and to understand better the mechanisms of our immune defences. With the generalization of these methodologies, the number of HLA named alleles is expected to explode in the next years, notably at the fourth field level of resolution describing the variability of introns and other untranslated regions. While this will probably represent a turning point in the way the HLA polymorphism is reported—a real challenge for the nomenclature committee and a delicate shift for HLA researchers and clinicians in histocompatibility—it will also certainly catalyse our understanding of the hidden face of the HLA genomic region: how such crucial genes are regulated. In the meanwhile, we can already say that a new and promising period for researchers in HLA molecular population genetics just started.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (SNSF), grant #31003A_144180, to ASM. We warmly thank David Glauser for drawing the map of the Kédougou region, and the Mandenka people for their contribution to this study. We also thank 2 anonymous reviewers for their useful comments which substantially helped to improve the first version of this manuscript.

Conflict of interest

The authors have declared no conflicting interests.

ORCID

A. Sanchez-Mazas http://orcid.org/0000-0002-7714-2432

REFERENCES

1. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*. 2009;54(1):15-39. https://doi.org/10.1038/jhg.2008.5.
2. Bentley G, Higuchi R, Hoglund B, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens*. 2009;74(5):393-403. https://doi.org/10.1111/j.1399-0039.2009.01345.x.
3. Shiina T, Suzuki S, Ozaki Y, et al. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers: Super high-resolution DNA typing of HLA loci. *Tissue Antigens*. 2012;80(4):305-316. https://doi.org/10.1111/j.1399-0039.2012.01941.x.
4. Gabriel C, Fürst D, Faé I, et al. HLA typing by next-generation sequencing - getting closer to reality: HLA typing by NGS. *Tissue Antigens*. 2014;83(2):65-75. https://doi.org/10.1111/tan.12298.
5. Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. *J Hum Genet*. 2015;60(11):665-673. https://doi.org/10.1038/jhg.2015.102.
6. Barone JC, Saito K, Beutner K, et al. HLA-genotyping of clinical specimens using Ion Torrent-based NGS. *Hum Immunol*. 2015;76(12):903-909. https://doi.org/10.1016/j.humimm.2015.09.014.
7. Cereb N, Kim HR, Ryu J, Yang SY. Advances in DNA sequencing technologies for high resolution HLA typing. *Hum Immunol*. 2015;76(12):923-927. https://doi.org/10.1016/j.humimm.2015.09.015.
8. Carapito R, Radosavljevic M, Bahram S. Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies. *Hum Immunol*. 2016;77:1016-1023. https://doi.org/10.1016/j.humimm.2016.04.002.

9. Hernández-Frederick CJ, Cereb N, Giani AS, et al. Detection of 549 new HLA alleles in potential stem cell donors from the United States, Poland and Germany. *HLA*. 2016;87(1):31-35. https://doi.org/10.1111/tan.12721.
10. Sanchez-Mazas A, Meyer D. The relevance of HLA sequencing in population genetics studies. *J Immunol Res*. 2014;2014:1-12. https://doi.org/10.1155/2014/971818.
11. Bitarello BD, Francisco R Dos S, Meyer D. Heterogeneity of dN/dS ratios at the classical HLA class I genes over divergence time and across the allelic phylogeny. *J Mol Evol*. 2016;82(1):38-50. https://doi.org/10.1007/s00239-015-9713-9.
12. Meyer D, Aguiar VRC, Bitarello BD, Brandt DYC, Nunes K. A genomic perspective on HLA evolution. *Immunogenetics*. July 2017. https://doi.org/10.1007/s00251-017-1017-3.
13. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PLoS One*. 2011;6(2):e14643. https://doi.org/10.1371/journal.pone.0014643.
14. Erlich HA. HLA typing using next generation sequencing: An overview. *Hum Immunol*. 2015;76(12):887-890. https://doi.org/10.1016/j.humimm.2015.03.001.
15. Blanc M, Sanchez-Mazas A, Van Blyenburgh NH, Sevin A, Pinson G, Langaney A. Interethnic genetic differentiation: GM polymorphism in Eastern Senegal. *Am J Hum Genet*. 1990;46:383-392.
16. Dard P, Schreiber Y, Excoffier L, et al. Polymorphism of HLA class I loci HLA-A, -B, -C, in the Mandenka population from eastern Senegal. *C R Acad Sci III*. 1992;314(13):573-578.
17. Dard P, Sanchez-Mazas A, Dugoujon J-M, et al. DNA analysis of the immunoglobulin IGHG loci in a Mandenka population from eastern Senegal: correlation with Gm haplotypes and hypotheses for the evolution of the Ig CH region. *Hum Genet*. 1996;98(1):36-47. https://doi.org/10.1007/s004390050156.
18. Dard P, Huck S, Frippiat JP, et al. The IGHG3 gene shows a structural polymorphism characterized by different hinge lengths: sequence of a new 2-exon hinge gene. *Hum Genet*. 1997;99(1):138-141.
19. Tiercy JM, Sanchez-Mazas A, Excoffier L, et al. HLA-DR polymorphism in a Senegalese Mandenka population: DNA oligotyping and population genetics of DRB I Specificities. *Am J Hum Genet*. 1992;51:592-608.
20. Graven L, Passarino G, Semino O, et al. Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol Biol Evol*. 1995;12(2):334-345.
21. Poloni ES, Excoffier L, Mountain JL, Langaney A, Cavalli-Sforza LL. Nuclear DNA polymorphism in a Mandenka population from Senegal: comparison with eight other human populations. *Ann Hum Genet*. 1995;59(1):43-61. https://doi.org/10.1111/j.1469-1809.1995.tb01605.x.
22. Martinson JJ, Excoffier L, Swinburn C, et al. High diversity of alpha-globin haplotypes in a Senegalese population, including many previously unreported variants. *Am J Hum Genet*. 1995;57(5):1186-1198.
23. Currat M, Trabuchet G, Rees D, et al. Molecular analysis of the β-globin gene cluster in the Niokholo Mandenka Population reveals a recent origin of the βs senegal mutation. *Am J Hum Genet*. 2002;70(1):207-223. https://doi.org/10.1086/338304.
24. Sabbagh A, Langaney A, Darlu P, Gérard N, Krishnamoorthy R, Poloni ES. Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. *BMC Genet*. 2008;9(1):21. https://doi.org/10.1186/1471-2156-9-21.
25. Excoffier L, Schneider S. Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc Natl Acad Sci USA*. 1999;96(19):10597-10602.
26. Meyer D. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics*. 2006;173(4):2121-2142. https://doi.org/10.1534/genetics.105.052837.
27. Nunes JM, Riccio ME, Buhler S, et al. Analysis of the HLA population data (AHPD) submitted to the 15th International Histocompatibility/Immunogenetics Workshop by using the Gene[rate] computer tools accommodating ambiguous data (AHPD project report). *Tissue Antigens*. March 2010;76:18-30. https://doi.org/10.1111/j.1399-0039.2010.01469.x.
28. Di D, Sanchez-Mazas A, Currat M. Computer simulation of human leukocyte antigen genes supports two main routes of colonization by human populations in East Asia. *BMC Evol Biol*. 2015;15(1):240. https://doi.org/10.1186/s12862-015-0512-0.
29. Sanchez-Mazas A, Černý V, Di D, et al. The HLA-B landscape of Africa: signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. *Mol Ecol*. 2017;26(22):6238-6252.
30. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a review. *Ann Hum Genet*. 2001;65:1-26.
31. Wakeland EK, Boehme S, She JX, et al. Ancestral polymorphisms of MHC class II genes: divergent allele advantage. *Immunol Res*. 1990;9(2):115-122.
32. Sanchez-Mazas A, Fernandez-Viña M, Middleton D, et al. Immunogenetics as a tool in anthropological studies. *Immunology*. 2011;133(2):143-164. https://doi.org/10.1111/j.1365-2567.2011.03438.x.
33. Hill AVS, Greenwood BM. Common West African HLA antigens are associated with protection from severe Malaria. *Nature*. 1991;352:595-600.
34. Meyer CG, Kremser PG. Malaria and onchocerciasis: on HLA and related matters. *Parasitol Today*. 1996;12(5):179-186.
35. Hammer C, Begemann M, McLaren PJ, et al. Amino acid variation in HLA class II proteins is a major determinant of humoral response to common viruses. *Am J Hum Genet*. 2015;97(5):738-743. https://doi.org/10.1016/j.ajhg.2015.09.008.
36. Cangussu LOF, Teixeira R, Campos EF, et al. HLA Class II Aaleles and chronic hepatitis C virus infection. *Scand J Immunol*. 2011;74(3):282-287. https://doi.org/10.1111/j.1365-3083.2011.02568.x.
37. Buhler S, Nunes JM, Sanchez-Mazas A. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*. 2016;68(6–7):401-416. https://doi.org/10.1007/s00251-016-0918-x.
38. Simons GF, Fennig CD. Ethnologue: Languages of the World. 20th ed. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.
39. Tishkoff SA, Reed FA, Friedlaender FR, et al. The genetic structure and history of Africans and African Americans. *Science*. 2009;324(5930):1035-1044. https://doi.org/10.1126/science.1172257.
40. Poloni ES, Naciri Y, Bucho R, et al. genetic evidence for complexity in ethnic differentiation and history in East Africa. *Ann Hum Genet*. 2009;73(6):582-600. https://doi.org/10.1111/j.1469-1809.2009.00541.x.
41. Ranciaro A, Campbell MC, Hirbo JB, et al. Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am J Hum Genet*. 2014;94(4):496-510. https://doi.org/10.1016/j.ajhg.2014.02.009.
42. Podgorná E, Diallo I, Vangenot C, et al. Variation in NAT2 acetylation phenotypes is associated with differences in food-producing subsistence modes and ecoregions in Africa. *BMC Evol Biol*. 2015;15(1):263. https://doi.org/10.1186/s12862-015-0543-6.
43. Triska P, Soares P, Patin E, Fernandes V, Cerny V, Pereira L. Extensive admixture and selective pressure across the Sahel Belt. *Genome Biol Evol*. 2015;7(12):3484-3495. https://doi.org/10.1093/gbe/evv236.
44. Beltrame MH, Rubel MA, Tishkoff SA. Inferences of African evolutionary history from genomic data. *Curr Opin Genet Dev*. 2016;41:159-166. https://doi.org/10.1016/j.gde.2016.10.002.
45. Sanchez-Mazas A. African diversity from the HLA point of view: influence of genetic drift, geography, linguistics, and natural selection. *Hum Immunol*. 2001;62(9):937-948. https://doi.org/10.1016/S0198-8859(01)00293-2.
46. Cao K, Moormann AM, Lyke KE, et al. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens*. 2004;63(4):293-325. https://doi.org/10.1111/j.0001-2815.2004.00192.x.
47. Testi M, Battarra M, Lucarelli G, et al. HLA-A-B-C-DRB1-DQB1 phased haplotypes in 124 Nigerian families indicate extreme HLA diversity and low linkage disequilibrium in Central-West Africa: HLA-phased haplotypes in Nigerians. *Tissue Antigens*. 2015;86(4):285-292. https://doi.org/10.1111/tan.12642.
48. Tshabalala M, Mellet J, Pepper MS. Human leukocyte antigen diversity: a Southern African perspective. *J Immunol Res*. 2015;2015:1-11. https://doi.org/10.1155/2015/746151.
49. Schaffer M. Bound to Africa: the Mandinka Legacy in the New World. *Hist Afr*. 2005;32:321-369. https://doi.org/10.1353/hia.2005.0021.
50. Sanchez-Mazas A, Steiner Q-G, Grundschober C, Tiercy J-M. The molecular determination of HLA-Cw alleles in the Mandenka (West Africa) reveals a close genetic relationship between Africans and Europeans. *Tissue Antigens*. 2000;56(4):303-312. https://doi.org/10.1034/j.1399-0039.2000.560402.x.

51. Sanchez-Mazas A, Kervaire B, Tiercy J-M. Population: Mandenka from Senegal. In: Hansen JA, ed. Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol 1. Victoria, CA; Seattle, WA 12-22 May: IHWG Press; 2006, 2002.

52. Andrien M, Tiercy JM, Defleur V, et al. HLA-B locus DNA typing: detection of B*7801 and seven additional alleles by BW6-specific exon 2 amplification. *Tissue Antigens*. 1993;42(5):480-487.

53. Tiercy JM, Djavad N, Rufer N, Speiser DE, Jeannet M, Roosnek E. Oligotyping of HLA-A2, -A3, and -B44 subtypes. Detection of subtype incompatibilities between patients and their serologically matched unrelated bone marrow donors. *Hum Immunol*. 1994;41(3):207-215.

54. Grundschober C, Rufer N, Sanchez-Mazas A, et al. Molecular characterization of HLA-C incompatibilities in HLA-ABDR-matched unrelated bone marrow donor-recipient pairs. *Tissue Antigens*. 1997;49(6):612-623. https://doi.org/10.1111/j.1399-0039.1997.tb02809.x.

55. Galan M, Guivier E, Caraux G, Charbonnel N, Cosson J-F. A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*. 2010;11(1):296. https://doi.org/10.1186/1471-2164-11-296.

56. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75(23):7537-7541. https://doi.org/10.1128/AEM.01541-09.

57. Méglecz E, Piry S, Desmarais E, et al. SESAME (SEquence Sorter & AMplicon Explorer): genotyping based on high-throughput multiplex amplicon sequencing. *Bioinformatics* 2011;27(2):277–278. doi:https://doi.org/10.1093/bioinformatics/btq641

58. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015;43(D1):D423-D431. https://doi.org/10.1093/nar/gku1161.

59. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30(14):3059-3066. https://doi.org/10.1093/nar/gkf436.

60. Mack SJ. A gene feature enumeration approach for describing HLA allele polymorphism. *Hum Immunol*. 2015;76(12):975-981. https://doi.org/10.1016/j.humimm.2015.09.016.

61. Nunes JM, Buhler S, Roessli D, Sanchez-Mazas A, the HLA-net 2013 Collaboration. The HLA-net GENE[RATE] pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. *Tissue Antigens*. 2014;83(5):307-323. https://doi.org/10.1111/tan.12356.

62. Nunes JM. Using UNIFORMAT and GENE[RATE] to analyze data with ambiguities in population genetics. *Evol Bioinform Online*. 2016;19-26. https://doi.org/10.4137/EBO.S32415.

63. El Mousadik A, Petit RJ. High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theor Appl Genet*. 1996;92(7):832-839. https://doi.org/10.1007/BF00221895.

64. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123:585-595.

65. Tajima F. The effect of change in population size on DNA polymorphism. *Genetics*. 1989;123(3):597-601.

66. Li W-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol*. 1993;36(1):96-99. https://doi.org/10.1007/BF02407308.

67. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 2010;10(3):564-567. https://doi.org/10.1111/j.1755-0998.2010.02847.x.

68. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016;33(7):1870-1874. https://doi.org/10.1093/molbev/msw054.

69. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 1986;3(5):418-426. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

70. Reche PA, Reinherz EL. Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *J Mol Biol*. 2003;331(3):623-641.

71. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300.

72. Witter K, Mautner J, Albert T, Zahn R, Kauke T. HLA-DQB1*0319, a novel HLA-DQB1 allele, shows strong haplotype association to HLA-DRB1*1102. *Tissue Antigens*. 2007;70(1):73-75. https://doi.org/10.1111/j.1399-0039.2007.00848.x.

73. Goeury T, Creary LE, Fernández-Viña MA, Tiercy J-M, Nunes JM, Sanchez-Mazas A. Population Report : Mandenka from Senegal, NGS typings reveal very high frequencies of particular HLA class II alleles and haplotypes. *HLA*. In Press.

74. Cullen M, Noble J, Erlich H, et al. Characterization of recombination in the HLA class II region. *Am J Hum Genet*. 1997;60(2):397-407.

75. Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*. 2001;29:217-222.

76. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*. 1988;335(8):167-170.

77. Hughes AL, Nei M. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci USA*. 1989;86(3):958-962.

78. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 1977;267(5608):275-276.

79. Erlich RL, Jia X, Anderson S, et al. Next-generation sequencing for HLA typing of class I loci. *BMC Genomics*. 2011;12(1):42. https://doi.org/10.1186/1471-2164-12-42.

80. Major E, Rigo K, Hague T, Berces A, Juhos SHLA. Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data. *PLoS One*. 2013;8(11):e78410. https://doi.org/10.1371/journal.pone.0078410.

81. The International HapMap Consortium, Gibbs RA, et al. The International HapMap Project. *Nature*. 2003;426(6968):789-796. https://doi.org/10.1038/nature02168.

82. Nunes JM, Buhler S, Sanchez-Mazas A. NO to obsolete definitions: YES to blanks: Letter to the Editor. *Tissue Antigens*. 2014;83(2):119-120. https://doi.org/10.1111/tan.12276.

83. Nunes JM. Tools for analysing ambiguous HLA data. *Tissue Antigens*. 2007;69((suppl 1)):203-205. https://doi.org/10.1111/j.1399-0039.2006.00808.x.

84. Milius RP, Mack SJ, Hollenbach JA, et al. Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string: Genotype List String. *Tissue Antigens*. 2013;82(2):106-112. https://doi.org/10.1111/tan.12150.

85. Mack SJ, Milius RP, Gifford BD, et al. Minimum information for reporting next generation sequence genotyping (MIRING): guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Hum Immunol*. 2015;76(12):954-962. https://doi.org/10.1016/j.humimm.2015.09.011.

86. Yamazaki A, Yasunami M, Ofori M, et al. Human leukocyte antigen class I polymorphisms influence the mild clinical manifestation of Plasmodium falciparum infection in Ghanaian children. *Hum Immunol*. 2011;72(10):881-888. https://doi.org/10.1016/j.humimm.2011.06.007.

87. Hill AVS, Allsopp CEM, Greenwood BM. Extensive genetic diversity in the HLA class II region of Africans, with a focally predominant allele, DRB1*1304. *Proc Natl Acad Sci USA*. 1992;89:2277-2281.

88. Lee KW, Hurley CK, Hartzman R, Johnson AH. The complexity of DRw6 and DR5 haplotypes in American blacks demonstrated by serology, cellular typing, and restriction fragment length polymorphism analysis. *Hum Immunol*. 1990;29:202-219.

89. von Salomé J, Gyllensten U, Bergström TF. Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. *Immunogenetics*. 2007;59(4):261-271. https://doi.org/10.1007/s00251-007-0196-8.

90. Meyer CG, Gallin M, Erttmann KD, et al. HLA-D alleles associated with generalized disease, localized disease, and putative immunity in *Onchocerca volvulus* infection. *Proc Natl Acad Sci USA*. 1994;91:7515-7519.

91. O'Hanlon SJ, Slater HC, Cheke RA, et al. Model-based geostatistical mapping of the prevalence of *Onchocerca volvulus* in West Africa. *PLoS Negl Trop Dis*. 2016;10(1):e0004328. https://doi.org/10.1371/journal.pntd.0004328.

92. Ohta T, Kimura M. Development of associative overdominance through linkage disequilibrium in finite populations. *Genet Res*. 1970;16(2):165-177.

93. Slatkin M. Hitchhiking and associative overdominance at a microsatellite locus. *Mol Biol Evol*. 1995;12(3):473-480.

94. Sanchez-Mazas A. An apportionment of human HLA diversity. *Tissue Antigens*. 2007;69((suppl 1)):198-202. https://doi.org/10.1111/j.1399-0039.2006.00802.x.

95. Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS. *PLoS Genet*. 2008;4(12):e1000304. https://doi.org/10.1371/journal.pgen.1000304.

96. Winter CC, Long EO. A single amino acid in the p58 killer cell inhibitory receptor controls the ability of natural killer cells to discriminate between the two groups of HLA-C allotypes. *J Immunol*. 1997;158(9):4026-4028.

97. Hilton HG, Guethlein LA, Goyos A, et al. Polymorphic HLA-C receptors balance the functional characteristics of KIR haplotypes. *J Immunol*. 2015;195(7):3160-3170. https://doi.org/10.4049/jimmunol.1501358.

98. Dos Santos Francisco R, Buhler S, Nunes JM, et al. HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. *Immunogenetics*. 2015;67(11–12):651-663. https://doi.org/10.1007/s00251-015-0875-9.

99. Inotai D, Szilvasi A, Benko S, et al. HLA genetic diversity in Hungarians and Hungarian Gypsies: complementary differentiation patterns and demographic signals revealed by HLA-A, -B and -DRB1 in Central Europe. *Tissue Antigens*. 2015;86(2):115-121. https://doi.org/10.1111/tan.12600.

100. Solberg OD, Mack SJ, Lancaster AK, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol*. 2008;69(7):443-464. https://doi.org/10.1016/j.humimm.2008.05.001.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Goeury T, Creary LE, Brunet L, et al. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa. *HLA*. 2018;91:36–51. https://doi.org/10.1111/tan.13180

Received: 8 December 2017 | Accepted: 19 December 2017

DOI: 10.1111/tan.13197

WILEY HLA Immune Response Genetics

POPULATION REPORT

Mandenka from Senegal

Next Generation Sequencing typings reveal very high frequencies of particular HLA class II alleles and haplotypes

T. Goeury[1,4] | L. E. Creary[2] | M. A. Fernandez-Vina[2] | J.-M. Tiercy[3,4] | J. M. Nunes[1,4] | A. Sanchez-Mazas[1,4]

[1]Department of Genetics and Evolution - Anthropology Unit, AGP Laboratory, University of Geneva, Geneva, Switzerland

[2]Stanford University, School of Medicine, Palo Alto, California

[3]Transplantation Immunology Unit and National Reference Laboratory for Histocompatibility, Geneva University Hospital, Geneva, Switzerland

[4]Institute of Genetics and Genomics in Geneva (IGE3), Geneva, Switzerland

Correspondence

* Alicia Sanchez-Mazas, Department of Genetics and Evolution-Anthropology Unit, University of Geneva, Sciences II, 30 quai Ernest-Ansermet, 1205 Geneva, Switzerland.

Email: alicia.sanchez-mazas@unige.ch

Funding information

Swiss National Science Foundation (SNSF), Grant/Award number: 31003A_144180

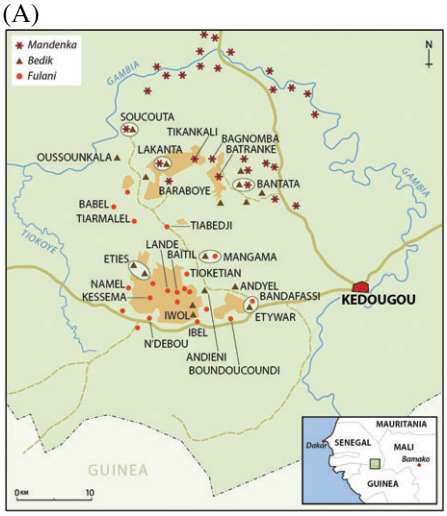
A total of 72 unrelated Mandenka individuals from Eastern Senegal, Niokholo region, were typed using Next Generation Sequencing (*library preparation with the Holotype HLA X2 and MIA FORA NGS HLA Typing kits, sequencing with Illumina MiSeq, and bio-informatic processing with HLA Twin v1.1.1 (Omixon) and MIA FORA NGS software)* and yielded reliable genotypes for 8 HLA loci, namely A, B, C, DRB1, DQA1, DQB1, DPA1 and DPB1. We used the recommendations of the *HLA-net* consortium (http://hla-net.eu) to characterise the population under study (Figure 1A) and the HLA-net GENE[RATE] tools[1] to compute 1- and 2-locus frequencies and the main population genetics statistics (summarised in Figure 1B,C).

The Mandenka population is considered at Hardy-Weinberg equilibrium, as the statistical tests indicate non-significant *P*-values (after correction for multiple tests) at all HLA loci except HLA-DPA1 but the adjusted *P*-value for DPA1 (.044) is very close to the 5% threshold and not significant at the 1% level with a sample size substantially smaller at this locus.

Class I and class II loci differ regarding their most frequent alleles, the frequencies of which are all below 20% for the former (from 17% for *A*23:01:01* to 18% for *B*35:01:01* and *C*04:01:01*) and above 20% for the latter (from 22% for *DPB1*17:01* to 50% for *DQA1*05:01:01*). This explains the contrasted heterozygosities observed at class I (above 90%) and class II (between 72% and 88%). As expected based on the IPD-IMGT/HLA[2] data, DPA1 is the less diversified locus in the Mandenka, with only 10 third field level alleles detected (and 5 second field level alleles representing the functional diversity) and the 3 most common alleles representing as much as 80% of the total allele frequencies.

Actually, the 3 HLA class I loci reject the assumption of neutrality towards an excess of heterozygotes, whereas none of the class II loci do. This suggests different types of selection acting on class I and class II loci in this population: class I loci would mainly evolve under balancing selection favouring heterozygous individuals; regarding class II loci, as a purely neutral evolution is unlikely due (like class I) to their immune function, our results merely suggest a combination of both balancing (heterozygous advantage, leading to a higher diversity) and positive (due to some specific pathogen resistance leading to a lower diversity) selection resulting in apparently neutral distributions. However, we do not exclude pathogen-driven selection at class I loci either, as the most common HLA-B allele found in the Mandenka, *B*35:01:01*, has been identified as a resistance allele against *Plasmodium falciparum* (the most prevalent malaria pathogen in West Africa) in a Ghanaian population[3] and could have thus reached its high frequency in the Mandenka through positive selection.

(A)



Information on the population under study

Names: Mandenka, Mandinka, Mandinko, Mandingo, Malinke
Language spoken: Mandinka
Linguistic family: Niger-Congo (NC)
Geographic location: Western Africa, Senegal, Niokholo region. *12°41'N, 12°21'W.*
Geographic region: Western Sub-Saharan Africa (WSAFR)
Ethnological information at the time of blood sampling (1991):
Monotheist pagans, patrilocal and usually exogamous according to the patrilineage; shared geographical area and historical relationships with Bedik and Fulani populations through ancient slavery or marriage [4, 5].

Map: sampling location in Eastern Senegal, around Kedougou, showing the close geographical proximity between Bedik▲, Fulani ∘ and Mandenka*villages.

(B) **Allele frequencies and main summary statistics**

Locus	N	k	H	HWE	BEWS	Most frequent alleles (>10%)
A	72	22	0.92	0.33	**6e-3 – 0.34**	***23:01:01 (17%)***; *30:02:01 (10%); 33:03:01 (10%)*
B	68	29	0.93	0.57	**0.01 – 0.63**	***35:01:01 (18%)***
C	68	18	0.91	0.11	**1e-3 – 0.18**	***04:01:01 (18%)***; *16:01:01 (16%)*
DRB1	67	21	0.88	0.15	0.14 – 0.52	**13:04** ***(28%)***
DQA1	68	14	0.72	1	0.78 – 0.93	***05:05:01 (50%)***; *01:02:01 (13%)*
DQB1	61	13	0.77	1	0.40 – 0.81	***03:19 (43%)***; *05:01:01 (10%)*
DPA1	51	10	0.72	**0.005***	0.74 – 0.94	***02:01:01 (46%)***; *01:03:01 (19%); 03:01 (16%)*
DPB1	69	21	0.86	0.03**	0.55 – 0.81	***17:01 (22%); 131:01 (20%)***; *01:01:01 (15%)*

N: sample size; k: number of alleles (max. 3rd field level of resolution); H: heterozygosity; HWE: p-Value for Hardy-Weinberg equilibrium test; BEWS: p-Value range for two-tailed Bootstrapped Ewens-Watterson-Slatkin test; the significant p-Values (after correction for multiple tests) and the most frequent alleles of each locus are shown in bold; * p =0.044 after correction for multiple tests; ** p=0.12 after correction for multiple tests. Supplementary Information S01 gives the complete list of allele frequencies.

(C) **Most common two-locus haplotypes in linkage disequilibrium in the Mandenka population**

Most frequent two-loci haplotypes (> 10%)	Observed frequency	LD (D coefficient)	Standardized residual
DQA1*05:05:01~DQB1*03:19	0.43	0.20	3.96
DQA1*05:05:01~DRB1*13:04	0.28	0.13	3.61
DQB1*03:19~DRB1*13:04	0.26	0.14	4.09
DPA1*02:01:01~DRB1*13:04	0.21	0.09	2.42
DPA1*02:01:01~DPB1*131:01	0.20	0.10	3.20
DPA1*02:01:01~DPB1*17:01	0.20	0.10	3.20
DPB1*131:01~DQA1*05:05:01	0.17	0.07	2.62
DPB1*131:01~DRB1*13:04	0.15	0.10	4.50
DPB1*131:01~DQB1*03:19	0.15	0.06	2.18
DPA1*01:03:01~DPB1*02:01:02	0.11	0.08	5.13
DPA1*03:01~DPB1*105:01	0.11	0.09	7.00

Haplotypes were considered in positive linkage disequilibrium (LD) if their standardized residuals were above 2; Supplementary Information S02 gives the complete list of two-locus haplotypes with significant LD; pairs of loci in global linkage disequilibrium (GLD) are DPA1~DPB1, DQA1~DQB1, DRB1~DQA1 and DRB1~DQB1 (see Supplementary Information S03).

FIGURE 1 Summary of ethnographical and genetic data in the Mandenka population from Eastern Senegal

Two-locus analyses show that several HLA class II (but no class I) loci pairs are in significant global linkage disequilibrium (GLD) according to both parametric and non-parametric tests: DPA1~DPB1, DQA1~DQB1, DQA1~DRB1 and DQB1~DRB1. Accordingly, whereas no class I haplotypes reach a frequency higher than 7%, several 2-locus class II haplotypes (in positive linkage disequilibrium) are observed at a very high frequency (up to 43% for

*DQA1*05:01:01~DQB1*03:19*). Actually, the results indicate the existence of one very frequent 3-locus haplotype, *DRB1*13:04~DQA1*05:05:01~DQB1*03:19*, that may have increased in frequency through selective sweep[5]. Whereas no relevant information is currently available for the 2 DQ alleles, *DRB1*13:04* has formerly been reported as frequent in Gambia.[6] By contrast, several DQA1~DQB1 and DPA1~DPB1 haplotypes are found in negative linkage disequilibrium, suggesting a negative selective pressure against the corresponding HLA-II dimers (putative dimer instability).

Overall, this study reveals particularly high frequencies of specific HLA class II alleles and haplotypes defined at the highest level of resolution by next generation sequencing in the Mandenka population.

ACKNOWLEDGMENTS

This study was supported by Swiss National Science Foundation (SNSF) grant #31003A_144180. We also thank André Langaney, Laurent Excoffier and Alain Epelbouin for the sampling, Lydie Brunet for technical help and David Glauser for the detailed map of the Niokholo region. We are particularly grateful to the Mandenka people for their informed consent and contribution to this study.

Conflict of interest

The authors have declared no conflicting interests.

ORCID

A. Sanchez-Mazas http://orcid.org/0000-0002-7714-2432

REFERENCES

1. Nunes JM. Using uniformat and gene[rate] to analyze data with ambiguities in population genetics. *Evol Bioinform Online*. 2016;11:19-26.
2. Robinson J, Halliwell JA, Hayhurst JH, Flicek P, Parham P, Marsh SGE. The IPD and IPD-IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015;43:D423-D431.
3. Yamazakia A, Yasunamia M, Oforib M, et al. Human leukocyte antigen class I polymorphisms influence the mild clinical manifestation of *Plasmodium falciparum* infection in Ghanaian children. *Hum Immunol*. 2011;72:881-888.
4. Blanc M, Sanchez-Mazas A, Hubert Van Blyenburgh N, Sevin A, Pison G, Langaney A. Interethnic genetic differentiation: GM polymorphism in eastern Senegal. *Am J Hum Genet*. 1990;46:383-392.
5. Goeury T, Creary LE, Brunet L, et al. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa. *HLA*. 2018;91:36-51.
6. Hill AVS, Allsopp CEM, Kwiatkowski D, et al. Extensive genetic diversity in the HLA class II region of Africans, with a focally predominant allele, DRB1*1304. *Proc Natl Acad Sci USA*. 1992;89:2277-2281.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Goeury T, Creary LE, Fernandez-Vina MA, Tiercy J-M, Nunes JM, Sanchez-Mazas A. Mandenka from Senegal: Next Generation Sequencing typings reveal very high frequencies of particular HLA class II alleles and haplotypes. *HLA*. 2018;91:148–150. https://doi.org/10.1111/tan.13197

Chapitre 3

Présentation de `MADaM` : *Multiplexed Amplicon Data Miner*

1 Introduction

Le génotypage de loci hyper-variables tels que le MHC/HLA et la découverte de nouveaux allèles sont des tâches complexes, qui ont grandement progressé avec l'aide des séquenceurs à haut débit (*NGS - Next Generation Sequencing*) et d'outils bio-informatiques adaptés.
Ces méthodes, de par l'important volume de séquences générées (c'est-à-dire la profondeur de séquençage) ont permis de réduire les coûts de séquençage, permettant de séquencer en parallèle un grand nombre d'individus (des centaines voire des milliers) sur plusieurs loci. Ces avancées ont grandement bénéficié aux études de génétique des populations, permettant d'étudier un grand nombre de populations et de loci en simultané.

Toutefois, ces méthodes étant sujettes à de nombreuses erreurs (soit lors de la préparation des librairies de séquençage, soit lors du séquençage lui-même), le défi est alors de filtrer ces erreurs.
Ce défi est d'autant plus ardu lors du séquençage de régions génomiques hautement variables, telles que le complexe HLA chez l'humain (ou le MHC en général), au sein duquel 26'373 allèles ont été recensés pour les 12 gènes les plus polymorphes [Robinson et al., 2015] [1] et pour lequel plus de 2'400 allèles ont été découverts par an sur les cinq dernières années. Cette diversité est générée par des processus évolutifs, décrits dans le Chapitre 1 (page 7), qui vont eux aussi rendre difficile la tâche de distinguer les vrais allèles des artefacts : des duplications de gènes menant à l'apparition de gènes paralogues qui co-amplifieront lors de la préparation de la librairie, des échanges de matériel génétique (transposons, conversions alléliques ou géniques) pour lesquels le signal sera difficile à distinguer d'une chimère (lors des PCR préséquençage), des polymorphismes de taille ou des différences d'un seul nucléotide (sur des centaines de nucléotides) entre deux vrais allèles, similaires à des artefacts issus de mutations/erreurs ponctuelles de lecture.

Ce chapitre présente `MADaM`, pour *Multiplexed Amplicon Data Miner* (Explorateur de Données Multiplexées d'Amplicons), un algorithme développé pour traiter des données de séquençage d'amplicons pour des régions habituellement difficiles à traiter, telles que le

1. Base de données IPD-IMGT/HLA version 3.40, Avril 2020.

HLA. Après un état de l'art des méthodes déjà existantes, le fonctionnement détaillé de `MADaM` sera présenté. Puis, son fonctionnement sera illustré au travers d'une application concrète sur un jeu de données de séquençage d'exons 2 de la région HLA-DRB (l'un des jeux de données générés au laboratoire AGP [2] pour lesquels `MADaM` a été développé). Une application de `MADaM` à un autre jeu de données externe au laboratoire sera ensuite présenté, afin de valider les méthodes employées dans l'algorithme. Finalement, les limites des autres méthodes déjà existantes seront évaluées au travers d'exemples issus de l'application de `MADaM` sur le jeu de données HLA-DRB.

1.1 Jeux de données disponibles

Jeu de données « 454 »

`MADaM` a été initialement développé pour traiter les résultats de séquençage des exons 2 de HLA-DRBx [3], HLA-DQA1, HLA-DQB1 et HLA-DPB1 obtenus pour 3'455 humains issus de 47 populations. Ces séquençages, réalisés dans le cadre d'une étude sur la diversité des gènes HLA de classe II de populations principalement africaines, ont été réalisés par pyroséquençage (Roche-454) par la société Beckman Coulter Genomics et sont décrits en détail dans le Chapitre 4.
Le jeu de données utilisé ici se compose donc de résultats de séquençage des exons 2 des gènes HLA-DRBx chez 3'455 humains. Ces données présentent la particularité de représenter plusieurs loci HLA-DRB (à savoir HLA-DRB1, -DRB3, -DRB4 et -DRB5) co-amplifiés à cause d'une spécificité non totale des amorces PCR utilisées.
En effet, les gènes HLA-DRB étant des gènes paralogues (résultats de plusieurs évènements de duplication de gènes [Satta et al., 1996a]), les régions auxquelles s'hybrident les amorces PCR sont similaires (conservées) entre les différents gènes HLA-DRB, le résultat étant une co-amplification de plusieurs loci.
Les amorces PCR ayant étés conçues pour amplifier spécifiquement HLA-DRB1 et -DRB3, et non les autres loci co-amplifiés, ces derniers montrent un biais d'amplification. Un filtre spécifiquement développé à cet effet, le filtre markovien, sera utilisé ici afin de retirer de l'étude les séquences ne correspondant pas à HLA-DRB1 ou -DRB3.
La structure particulière de la région HLA-DRB, illustrée sur la Figure 1.8, ne permet pas de déterminer à l'avance combien de vrais variants seront à identifier pour chaque individu. Le locus HLA-DRB1 est systématiquement présent, mais le locus HLA-DRB3 montre un polymorphisme de présence/absence. Ainsi, le nombre de séquences attendues pour un individu va de 1 (absence de locus HLA-DRB3 et HLA-DRB1 homozygote) à 4 (HLA-DRB1 et -DRB3 tous les deux hétérozygotes).

Jeu de données « Glouton »

Le second jeu de données provient de l'étude de [Oomen et al., 2013] et [Rico et al., 2015]. Il s'agit des données de séquençage d'une étude basée sur le pyroséquençage (technologie Roche-454) des exons 2 de la région MHC-DRB chez 89 *Gulo gulo* (communément appelé Carcajou ou Glouton), ainsi que chez 2 *Procyon lotor* (Raton-Laveur), cette dernière espèce étant utilisée dans l'étude en tant que groupe externe (*outgroup*).

2. Laboratoire d'Anthropologie, génétique et peuplements, Université de Genève, Suisse.
3. Le x indique que plusieurs loci de la région HLA -DRB, de HLA-DRB1 à -DRB5 ont été co-amplifiés.

Deux loci, MHC-DRB1 et MHC-DRB2 et un pseudogène MHC-DRB3 ont été séquencés. Ce dernier a été retrouvé chez tous les individus, mais ne présente aucun polymorphisme.

1.2 Etat de l'art des techniques d'« Amplicon processing »

Ne seront abordées ici que les techniques relatives au traitement de données NGS ne nécessitant pas d'assemblage de fragments et étant hautement variables, telles que les données HLA. Avant de rentrer dans le détail des techniques déjà existantes, il convient de définir 1) les termes utilisés dans ce chapitre, 2) les principaux défis à résoudre, 3) les buts à atteindre.

Afin d'éviter toute confusion liée aux significations parfois multiples de certains termes dans le domaine de la bio-informatique, les termes spécifiques employés dans ce chapitre empruntent les définitions suivantes :

- **Amplicon** : résultat d'une extraction et amplification par PCR d'un ou plusieurs loci, habituellement pour un même individu ;
- **Lecture** : enchaînement de nucléotides inférés par un séquenceur ADN à partir d'un fragment d'ADN extrait du génome d'un individu ;
- **Séquence** : *lecture* assignée à un individu sur la base de tags oligonucléotidiques. Une même séquence peut être retrouvée de multiples fois au sein d'un même amplicon ;
- **Variant** : séquence dérépliquée, c'est-à-dire association entre l'enchaînement des nucléotides d'une séquence et l'occurrence de cette séquence dans un ensemble donné (généralement l'ensemble des séquences assignées à un individu) ;
- **Artefact** : variant issu d'une (ou plusieurs) erreur(s), lors de la PCR ou du séquençage, sur un (ou plusieurs) variant(s) ;
- **Vrai Variant** : variant dont la séquence nucléotidique correspond à celle portée, sur le locus ciblé, par l'individu génotypé.

Les principaux problèmes rencontrés avec le génotypage *de-novo* de loci hyper variables sont :

- des erreurs de séquençage non aléatoires (homopolymères fréquents avec le pyroséquençage 454 ou un plus grand taux d'insertions/délétions pour la technologie Illumina) ;
- des artefacts à haute fréquence (générés lors des premiers cycles de PCR)
- des fortes similarités entre allèles séquencés (de l'ordre d'une différence pour 250-300 nt [4]) ;
- des biais d'amplification liés à une hybridation imparfaite des amorces (à cause de la grande diversité génétique des loci) ;
- des contaminations ;
- une amplification et un séquençage de gènes paralogues (dus à une spécificité non totale des amorces) ;
- certaines fois, une absence d'informations (*priors*) quant au système génétique étudié (par exemple des variations du nombre de copies chez des espèces non modèles) ;
- une présence de séquences chimériques (issues de la PCR) ;
- une forte hétérozygotie, augmentant le nombre d'allèles totaux à détecter dans l'analyse ;

4. Nucléotides

— une variabilité de ces biais entre amplicons et une synergie de ces biais (un individu hétérozygote portant deux allèles ne différant que d'une paire de bases et dont l'un des allèles est sous-amplifié).

Dans leur article de 2014, Lighten *et al.* définissent deux buts que chaque méthode de génotypage (de la préparation de la librairie au traitement bio-informatique) doit viser [Lighten et al., 2014a] :

— réduire le risque d'inclure des artefacts dans les génotypes finaux (erreurs de génotypage de type I) ;
— réduire le risque d'exclure de véritables allèles des génotypes (erreurs de génotypage de type II).

Les quatre méthodes existantes peuvent être regroupées en deux catégories : deux basées sur les occurences des variants (méthodes dites de l'*Allele Validation Threshold* et du *Degree of Change*), et deux basées sur un partitionnement des variants selon plusieurs critères (méthodes dites de *Clustering*).

Allele Validation Threshold

Cette technique est basée sur l'hypothèse que les artefacts sont générés de manière aléatoire lors d'un séquençage et donc que ces artefacts doivent être retrouvés de manière nettement moins fréquente que les allèles dont ils découlent. Ainsi il doit être possible de séparer les artefacts des allèles sur la base de leurs occurrences au sein d'un amplicon, ou parmi plusieurs amplicons [Zagalska-Neubauer et al., 2010].

Ainsi, Zagalska-Neubaueur définit deux critères de décision :

1. le **premier critère** est un nombre minimum de copies au sein d'un amplicon pour accepter un variant comme vrai variant ;
2. le **deuxième critère** est un nombre minimum d'amplicons au sein desquels sera retrouvée cette séquence : si les artefacts sont générés de manière aléatoire, il est peu probable d'observer le même artefact au sein de plusieurs amplicons indépendants.

Ensuite, les variants ayant passé le premier filtre sont triés selon leur MPAF (*Maximum Per Amplicon Frequency*, fréquence maximale par amplicon) et en partant du variant avec la plus faible MPAF, les trois amplicons dans lesquels ce variant est le plus fréquent sont examinés pour déterminer si ce variant peut être expliqué comme une chimère issue de variants plus fréquents.
Cette technique permet d'établir un seuil de fréquence par amplicon au-dessus duquel tous les variants sont des vrais variants et d'un autre seuil en dessous duquel tous les variants sont des artefacts. Mais entre les deux seuils existe une zone grise où peuvent se retrouver artefacts et vrais variants, nécessitant une observation manuelle.

Clustering Method I

Cette méthode se base sur un arbre de décision assez complexe, qui commence, au sein de chaque amplicon, par trier les variants par occurrence. Le variant le plus fréquent

de chaque amplicon est automatiquement déclaré comme vrai variant et les singletons [5] comme artefacts [Sommer et al., 2013].

Pour les variants restants, une série de critères est ensuite appliquée sous la forme d'un arbre de décision binaire : présence dans d'autres amplicons/réplicats, différence (en nombre de nucléotides) avec des variants plus fréquents, différents critères de fréquences, statut du variant dans d'autres amplicons (vrai variant, artefact, chimère...).

Cet arbre de décision aboutit à la classification des variants en vrais variants, artefacts et variants non classifiés, ces derniers correspondant à une zone grise, dans laquelle seront retrouvés aussi bien des artefacts à haute fréquence que des vrais variants ayant sous amplifié.

Clustering Method II

La deuxième méthode de partitionnement est basée sur la similarité entre les vrais variants et leurs artefacts : pour un individu avec N allèles, les séquences devraient pouvoir être regroupées en N groupes sur la base de leurs similarités [Stutz and Bolnick, 2014].

Il s'agit d'un algorithme itératif, dont la première étape regroupe les variants par similarité jusqu'à obtenir des groupes de séquences similaires qui doivent valider deux conditions :

1. critère de taille : un groupe doit contenir suffisamment de séquences, ce qui permet d'éviter des groupes de variants rares, mais très différents (telles que les chimères) ;
2. critère de dominance : un groupe ne doit contenir qu'un seul variant fréquent (le vrai variant) et une majorité de variants à faible fréquence (les artefacts).

Si ces conditions ne sont pas validées, l'algorithme est à nouveau appliqué avec un critère de similarité minimale entre séquences plus strict : un critère de similarité pas assez strict augmentera le nombre de groupes ne validant par le critère de dominance, tandis qu'une valeur trop stricte augmentera le nombre de groupes rejetés par le critère de taille.
Lors de la seconde étape, les groupes rejetés par le critère 1 sont recoupés avec les allèles validés à la première étape afin de s'assurer de ne pas rejeter des allèles qui auraient sous-amplifié.

Degree Of Change

Cette dernière technique, basée sur les fréquences des variants au sein de chaque amplicon, commence par trier les variants par fréquences décroissantes, puis calcule une fréquence cumulée pour chaque variant, du plus fréquent au moins fréquent [Lighten et al., 2014b].
Un taux de changement (*ROC - Rate Of Change*), correspondant à la dérivée des fréquences cumulées, est ensuite calculé puis la dérivée de ce ROC, appelée ici DOC (*Degree Of Change*), est calculée.

5. Variants retrouvés une seule fois, donc d'occurence 1.

Si un point d'inflexion est détecté, il correspond au seuil entre les vrais variants et les artefacts. Dans le cas contraire, l'amplicon est déclaré de mauvaise qualité et retiré de l'étude.

Critiques des techniques d'Amplicon processing

Plusieurs critiques peuvent être adréssées aux méthodes précédemment décrites.

La méthode de l'*Allele Validation Threshold* possède une importante zone grise (une zone d'incertitude), et l'expérimentateur devra manuellement départager les variants ou ajouter d'autres critères de décision qui augmenteront la complexité de la méthode.

La première méthode de *Clustering* comprend un arbre de décision complexe, comprenant beaucoup de paramètres (avec un risque de sur-paramétrage rendant difficile l'application à d'autres technologies de séquençage), et dont un tiers des noeuds terminaux aboutissent à des variants non classifiés. De plus, cette méthode utilise un critère de similarité (ou différence) entre séquences qui est arbitraire. Par exemple, le choix du nombre de différences de nucléotides entre variants est totalement dépendant de la taille des variants, de la technologie de séquençage employée et de l'importance du polymorphisme du locus, il faut donc trouver une valeur qui serait plus indépendante de la taille des lectures.

La seconde méthode de *Clustering* n'est pas capable de distinguer deux vrais variants ne différant que d'une seule base. Ces variants seront regroupés dans un même groupe, qui ne passera pas le critère de dominance.

La méthode du *Degree Of Change* est directement dépendante de la bonne amplification des allèles. Si un allèle est sous amplifié, le point d'inflexion de la courbe du DOC ne sera pas visible.

L'utilisation d'un seuil de fréquence doit être adapté en fonction du nombre de loci co-étudiés : un seuil valable pour un locus diploïde ne sera plus du tout valable pour une région où plusieurs loci co-amplifieraient (telle que la région HLA-DRB), ce seuil doit être adapté à chaque situation.
De plus, plusieurs techniques ne se basent que sur une seule statistique pour départager les variants (*Allele Validation Threshold* et l'occurence des variants, le *Degree of Change* et la fréquence de ces variants).
Finalement, chaque méthode possède une zone grise directement liée à la statistique utilisée pour départager les vrais variants des artefacts, que ce soit un intervalle de fréquence/occurrence où les variants sont indéterminés (*Allele Validation Threshold*), ou une différence de nucléotide entre deux variants (*Clustering Method I & II*).

Une possible solution, si l'on considère que les zones grises de chaque statistique sont indépendantes les unes des autres, est d'utiliser conjointement plusieurs de ces statistiques.
C'est à partir de ce constat qu'a été développé `MADaM`.

1.3 But du nouvel algorithme (MADaM) développé dans ce travail

Le présent chapitre décrit le fonctionnement de MADaM, un algorithme conçu pour traiter les résultats de séquençage NGS sur des amplicons 1) ne nécessitant pas d'étape d'assemblage (tels que le séquençage 454 des exons 2 et 3 des gènes HLA classiques), 2) issus de régions hyper-variables pour lesquelles les méthodes traditionnelles montrent de faibles résultats, 3) utilisant le multiplexage des séquences d'un grand nombre d'individus au sein d'un même séquençage. S'appuyant sur les méthodes déjà existantes pour en identifier les limites, cet algorithme vise à améliorer les procédés de génotypage.
En termes bio-informatiques, MADaM vise à identifier les vrais variants parmi des résultats de séquençage haut-débit et en termes statistiques (et d'apprentissage machine), il s'agit d'une classification non supervisée de ces lectures en artefacts et vrais variants, classification réalisée sur des classes d'effectifs biaisés.

2 Fonctionnement de l'algorithme

2.1 Pré-traitement des données

Initialisation

Chaque projet est initialisé par le chargement de son fichier de configuration. Il précise les différents paramètres utilisés par l'algorithme tout au long de son fonctionnement. Chaque projet peut contenir plusieurs séries [6], correspondant à des données de séquençage supplémentaires pour le projet.
Ensuite, chaque série est initialisée en chargeant dans la base de données associée les données de séquençage (lectures au format `fasta`) ainsi que les données relatives aux individus (tags oligonucléotidiques, amorces PCR spécifiques, identifiants et éventuelles données populationnelles) à l'aide d'un fichier `csv` (*Comma Separated Values*).

Filtres et assignation des lectures

Après le chargement en mémoire des données, l'étape qui vient consiste tout d'abord en un tri sur la taille des séquences. Si la taille des lectures est nettement plus petite ou plus grande que le locus ciblé, cela signifie une probable co-amplification d'un ou plusieurs autres loci (due à une spécificité non totale des amorces PCR), dont les lectures seront retirées à cette étape.

Les lectures restantes sont ensuite filtrées à l'aide du programme BLAST (*Basic Local Alignment Search Tool*, [Altschul et al., 1990]) et d'une séquence de référence. Seules les lectures avec une `e-Value` plus petite qu'un seuil donné (spécifié dans le fichier de configuration) sont conservées. Cette étape permet de retirer des lectures contaminantes passées au travers du filtre basé sur la taille.

```
>DRB
TTTCCTGTGGCAGGGTAAGTATAAGTGTCATTTCTTCAACGGGACGGAGCGGGTGCAGTTCCTGGAAAGAC
TCTTCTATAACCAGGAGGAGTTCGTGCGCTTCGACAGCGACGTGGGGGAGTACCGGGCGGTGACGGAGCTA
GGGCGGCCTGTCGCCGAGTCCTGGAACAGCCAGAAGGACATCCTGGAGGACAGGCGGGGCCAGGTGGACAC
CGTGTGCAGACACAACTACGGGGTTGGT
```

FIGURE 3.1 – Exemple d'une séquence de référence. Séquence de DRB1-Exon2, fournie par M. Galan.

Les séquences sont ensuite démultiplexées (assignées aux individus sur la base des tags oligonucléotidiques) puis les amorces PCR, les tags et les éventuels adaptateurs sont retirés des séquences.

Il est alors possible à cette étape d'appliquer un filtre supplémentaire, appelé filtre markovien car fonctionnant sur la base de chaînes de Markov. Ce filtre, beaucoup plus discriminant que la taille ou le BLAST, a été ici développé pour séparer des lectures provenant de loci très similaires (par exemple des gènes paralogues) telles que rencontrées dans le jeu de données pour lequel `MADaM` a initialement été développé (voir Section 1.1

6. *Sequencing runs* en anglais.

pour les informations relatives à ces données).

Ce filtre markovien est un algorithme de classification à apprentissage supervisé.
Étape 1 - Apprentissage : des séquences du locus cible et des autres loci co-amplifiés sont fournies et, à partir des catégories associées à chaque séquence, l'algorithme va extraire les probabilités de transition [7] d'états pour chaque catégorie.
Étape 2 - Classification : pour chaque séquence à tester, la matrice de transition est calculée et comparée avec les matrices de transition dérivées à l'étape précédente. La comparaison est réalisée à l'aide de la statistique du $\chi^2 = \frac{(observée\ -\ théorique)^2}{théorique}$ et la séquence d'intérêt se voit attribuer la catégorie correspondant à la plus petite valeur de χ^2 calculée pour chacune des catégories.

Une fois les lectures assignées, les amorces, tags et adaptateurs retirés et le filtre markovien éventuellement appliqué, les séquences sont alignées à l'aide du programme MAFFT (*Multiple Alignment Fast Fourier Transform* [8] [Katoh, 2002]), et les séquences identiques sont dérépliquées afin d'extraire les variants (et le nombre de séquences associées) de chaque individu.

Une fois ces étapes effectuées, chaque individu a ses propres séquences assignées et dénombrées. Seuls les individus avec un nombre suffisant de séquences (correspondant au seuil $T1_{Galan}$ décrit dans [Galan et al., 2010] et spécifié dans le fichier de configuration) seront retenus pour la suite des analyses.

2.2 Extraction des variables descriptives

Afin de transformer les enchaînements de nucléotides de chaque variant en données numériques utilisables par les algorithmes d'apprentissage machine (*Machine Learning*) utilisés après, il est nécessaire d'effectuer une étape d'extraction des variables descriptives (*Features Engineering*).

Cette étape représente l'étape-clef du fonctionnement de `MADaM`. En effet, si l'on veut réduire la zone grise entre les artefacts et les vraies variants (voir Section 1.2), il est nécessaire d'extraire des variables pertinentes des données.

Afin d'obtenir les variables qui serviront à classifier chaque variant en artefact ou vrai variant, l'extraction des variables se calque sur les processus qui ont mené à la création de l'ensemble des variants observés : chaque variant observé peut s'expliquer soit comme étant une lecture fidèle du locus ciblé, soit comme une déformation de ce locus liée à une ou plusieurs erreurs (parfois en cascade) dont la cause est à rechercher soit dans l'amplification PCR effectuée lors de la préparation de la librairie de séquençage (insertions/délétions/mutations ponctuelles ou chimères), soit lors de la lecture de la séquence par le séquenceur (dont le type et la fréquence varient selon la technologie de séquençage utilisée).

Ainsi, il est possible de dériver quelques règles :

7. Probabilité d'observer un nucléotide spécifique à une position N en fonction du nucléotide à la position N-1. Ne pas confondre avec le processus moléculaire de transition des purines dans l'ADN.

8. Qui a montré de meilleurs résultats en terme d'alignement que MUSCLE [Edgar, 2004] sur ce type de données.

— **principe d'antériorité** : chaque artefact résultant d'une erreur stochastique sur un autre variant, ce dernier doit être présent de manière plus fréquente au sein d'un individu que les artefacts qui en découlent ;
— **principe de causalité** : chaque artefact doit pouvoir s'expliquer comme le résultat d'une suite d'opérations sur un ou plusieurs (dans le cas des chimères) variants.

Ainsi, l'extraction des variables descriptives dans `MADaM` peut s'apparenter à la réalisation d'un modèle statistique : après avoir trié les variants par fréquence descendante (principe d'antériorité) et arrangé les **séquences** en un ensemble, chaque variant, du plus fréquent au moins fréquent, est ajouté à une liste de *variants explicatifs* et à chaque ajout d'un variant v_k, chaque position n_i sur chaque séquence s_j peut alors être expliquée ou non comme découlant d'une lecture fidèle de cette même position n_i sur le variant v_k. Si cette position ne peut être expliquée, c'est que le variant qui expliquerait cette position n'a pas encore été ajouté à la liste des variants explicatifs.
Cette position non expliquée est appelée ici « résidu », par similarité avec la différence entre la valeur prédite et observée dans un modèle linéaire. Pour continuer l'analogie, ce que le modèle chercherait à expliquer est l'ensemble des nucléotides observés à chaque position et les variables explicatives seraient les différents variants déjà ajoutés à la liste des variants explicatifs. La Figure 3.2 illustre ce processus sur les trois premières itérations de l'algorithme d'extraction des variables.

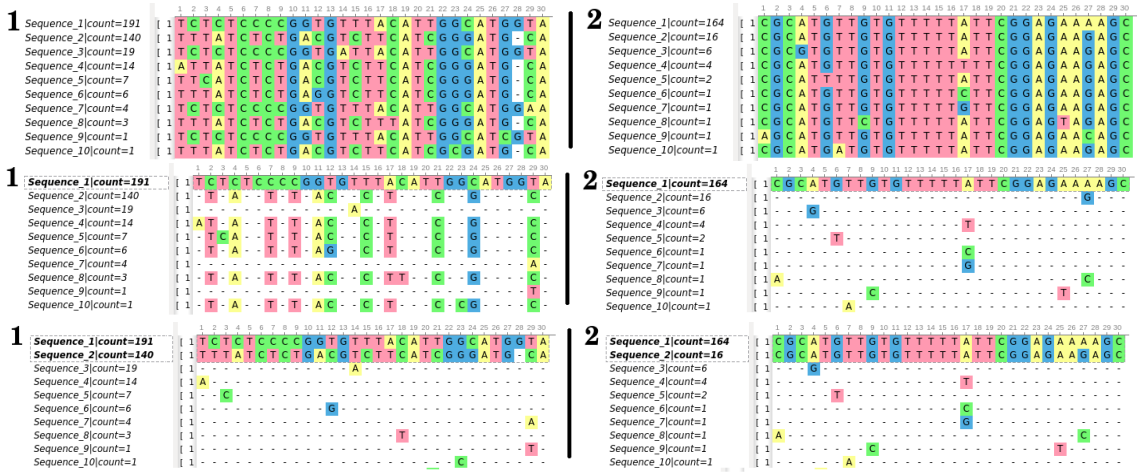


FIGURE 3.2 – Illustration, sur des variants (simulés), du processus d'extraction des variables explicatives. Les variants de gauche (1) représentent un individu hétérozygote, tandis que les variants de droite (2) représentent un individu homozygote. La première étape (en haut) représente l'initialisation, où les variants sont triés par fréquences décroissantes. À la deuxième étape (au milieu) le variant le plus fréquent est ajouté à la liste des variants explicatifs et les positions non expliquées (appelées résidus) sur les autres variants sont laissées telles quelles, les positions expliquées étant représentées par des "-". À la troisième étape le deuxième variant le plus fréquent est lui aussi ajouté à la liste des variants explicatifs et ne subsistent dans les autres variants que les positions non expliquées par les deux premiers variants. Le processus est répété jusqu'à épuisement des variants ou jusqu'à ce qu'un nombre pré-défini de variants aient été ajoutés.

Ensuite, à chaque étape, la distribution des résidus est calculée : le nombre de résidus

par séquence (toutes les positions de la séquence) ou par position (une même position n_i au sein de toutes les séquences) est calculé et l'écart-type (par séquence et par position) de ces distributions est calculé. La moyenne est aussi calculée de la même façon mais n'est pas retenue pour la suite des analyses, puisqu'elle n'a pas montré être un indicateur stable du statut de vrai variant ou artefact des séquences.
Cela donne, pour chaque variant ajouté à la liste des variants explicatifs, deux valeurs (écart-type par séquence et par position) en plus de la fréquence du variant, auxquels s'ajoutent, dès le deuxième variant ajouté à la liste, la différence (δ) entre ces valeurs calculées pour ce variant et les valeurs calculées pour le variant précédemment ajouté.
L'ajout des variants s'interrompt soit quand tous les variants ont été ajoutés, soit quand un nombre pré-déterminé de variants a été ajouté.

Le pseudo-code associé à cette étape est le suivant, pour un individu :

N séquences alignées, taille L
K ($< N$) variants, taille L
M1 ← Matrice des séquences alignées, dimension : N x L
M2 ← Matrice ne contenant que des 1, dimension : N x L
V ← Variants ordonnés par fréquences décroissantes

Pour chaque v_k dans V :
    Pour i, j dans $1 : L$, $1 : N$:
        si V_k[i] = M1[j,i] :
            M2[j,i] = 0
    $\mathrm{SD}_{pos}^{v_k} \leftarrow \sigma^2(\sum_{colonnes}$ M2)
    $\mathrm{SD}_{seq}^{v_k} \leftarrow \sigma^2(\sum_{lignes}$ M2)
    $\delta\ \mathrm{SD}_{pos}^{v_k} \leftarrow \mathrm{SD}_{pos}^{v_{k-1}}$ - $\mathrm{SD}_{pos}^{v_k}$
    $\delta\ \mathrm{SD}_{seq}^{v_k} \leftarrow \mathrm{SD}_{seq}^{v_{k-1}}$ - $\mathrm{SD}_{seq}^{v_k}$
    $f^{v_k} \leftarrow$ fréquence(v_k)

FIGURE 3.3 – Algorithme d'extraction des variables descriptives de `MADaM`. *pos* : Position, *seq* : Séquence

L'un des principaux avantages de cette méthode est le problème des chimères, qui est géré automatiquement : les chimères étant le résultat d'une fusion entre au moins deux variants, elles devraient se soumettre au principe de causalité dès lors que chacun de ces variants sera ajouté à la liste des variants explicatifs.

2.3 Réduction de la dimensionalité

L'extraction des variables explicatives permet à chaque variant d'être décrit par 5 variables (fréquence, SD_{pos}, SD_{seq}, δSD_{pos}, δSD_{seq}) et, positionnant ainsi le variant dans un espace à cinq dimensions, l'étape suivante de l'algorithme est de réduire ce nombre de dimensions par une projection de ces variants dans un espace de plus faible dimension (dans notre cas, deux dimensions).

Algorithme t-SNE

La technique utilisée à cette étape est une t-SNE (*t-Distributed Stochastic Neighbors Embedding* [van der Maaten and Hinton, 2008]). La t-SNE est une méthode non-linéaire de réduction de dimensionalité, dans laquelle les observations similaires (présentant un contenu similaire) dans l'espace à haute dimension, restent proches lors de la projection sur un espace de plus petite dimension.

L'algorithme de la t-SNE comprend deux étapes (décrites en détail dans à la page 49) :

1. Construction d'une **distribution de probabilité** entre chaque paire d'observations, permettant d'établir une similarité entre les observations de chaque paire ;
2. Projection sur un espace de plus faible dimension (habituellement 2 ou 3) :
   (a) Projection des observations aléatoirement dans l'espace de faible dimension (effet stochastique) ;
   (b) Construction d'une seconde distribution de probabilités basée sur la précédente projection ;
   (c) Application du *gradient descent* pour minimiser la divergence de Kullback-Leibler entre les deux distributions.

L'un des paramètres importants d'une t-SNE est le facteur de perplexité. Ce paramètre peut être vu comme le nombre de voisins que chaque observation est censée avoir. Habituellement entre 5 et 50, ce paramètre définit le type de structures observées dans l'espace à faibles dimensions : une petite valeur de perplexité fera apparaître des structures très locales, tandis que des valeurs élevées feront apparaître des structures plus globales.

De plus, l'effet stochastique dû à la projection pseudo-aléatoire [9] peut conduire, lors de l'étape de *gradient descent*, à un minimum **local** de la fonction à optimiser. Ainsi, il est courant de réaliser plusieurs t-SNE avec une même valeur de perplexité et de ne conserver que celle présentant la plus faible valeur de divergence de Kullback-Leibler.

Intégration à `MADaM`

Lors de l'étape de réduction de la dimensionalité, plusieurs t-SNE sont réalisées en faisant varier le paramètre de perplexité, afin de capturer au mieux la structure totale des données (structures locales et structure globale). La Figure 3.4 illustre l'effet de différents facteurs de perplexité sur le résultat de la t-SNE. Une faible valeur de perplexité (à gauche sur la figure) fait ressortir des structures locales, avec les points relativement espacés

9. Car habituellement reproductible à l'aide d'une graine aléatoire, (*random seed)*.

(groupés par points possédant des variables descriptives très similaires). Les points présentant une fréquence élevée ne sont pas clairement isolés des points de fréquences réduites. Au contraire, pour des valeurs de perplexité plus élevées (au milieu et surtout à droite), les groupes de points sont plus denses et la structure globale (vrais variants d'un côté et artefacts d'un autre) est plus visible.
Pour circonvenir à l'effet stochastique, plusieurs t-SNE de même perplexité sont réalisées à chaque fois, et la t-SNE présentant la plus faible valeur de divergence de Kullback-Leibler est retenue.

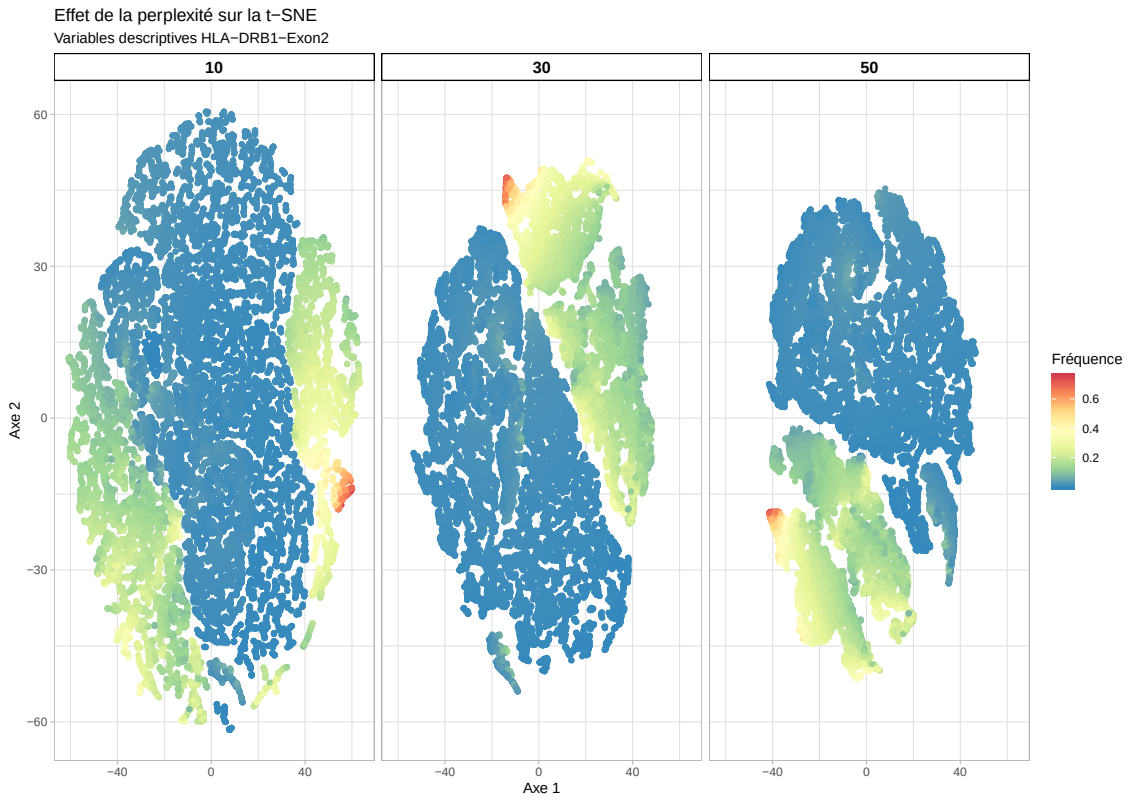


FIGURE 3.4 – Effet de différentes valeurs du facteur de perplexité sur le résultat d'une t-SNE. De gauche à droite, t-SNE réalisées sur le même jeu de données (les variables descriptives de DRB1-Exon2, décrit dans la Section 1.1) avec une valeur de perplexité de 10 (gauche), 30 (milieu) et 50 (droite).

2.4 Classification en vrais séquences/artefacts

La dernière étape réalisée par `MADaM` est d'identifier, à l'aide de l'algorithme DBSCAN, des groupes au sein des projections réalisées par la t-SNE, puis de détecter les groupes composés de vrais variants.

L'algorithme DBSCAN

L'algorithme DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*, [Ester et al., 1996]) est une méthode de partitionnement des données (*clustering*) basée sur la densité des points. L'idée de cette méthode est de regarder, pour un point donné,

combien de points sont présents dans le voisinage (dans un rayon ϵ autour du point) et de considérer ces points voisins comme appartenant au même groupe que le point initial si une densité minimale (un nombre de points voisins) est atteinte.

Cet algorithme présente l'avantage de pouvoir identifier des groupes non séparables linéairement, sans a priori sur le nombre de groupes à identifier, et capable de gérer des observations aberrantes (s'il n'y a pas suffisamment de points voisins d'un autre, ce dernier est exclu et déclaré comme du bruit). Toutefois, les densités des différents groupes doivent être similaires (puisque le paramètre de densité reste inchangé tout au long de l'étape de partitionnement).

La Figure 3.5 illustre l'effet de différentes valeurs de ϵ sur les groupes identifiés par DBSCAN. Cette figure représente un partionnement DBSCAN sur la projection t-SNE de perplexité 30 présentée dans la Figure 3.4 en utilisant trois valeurs de ϵ.

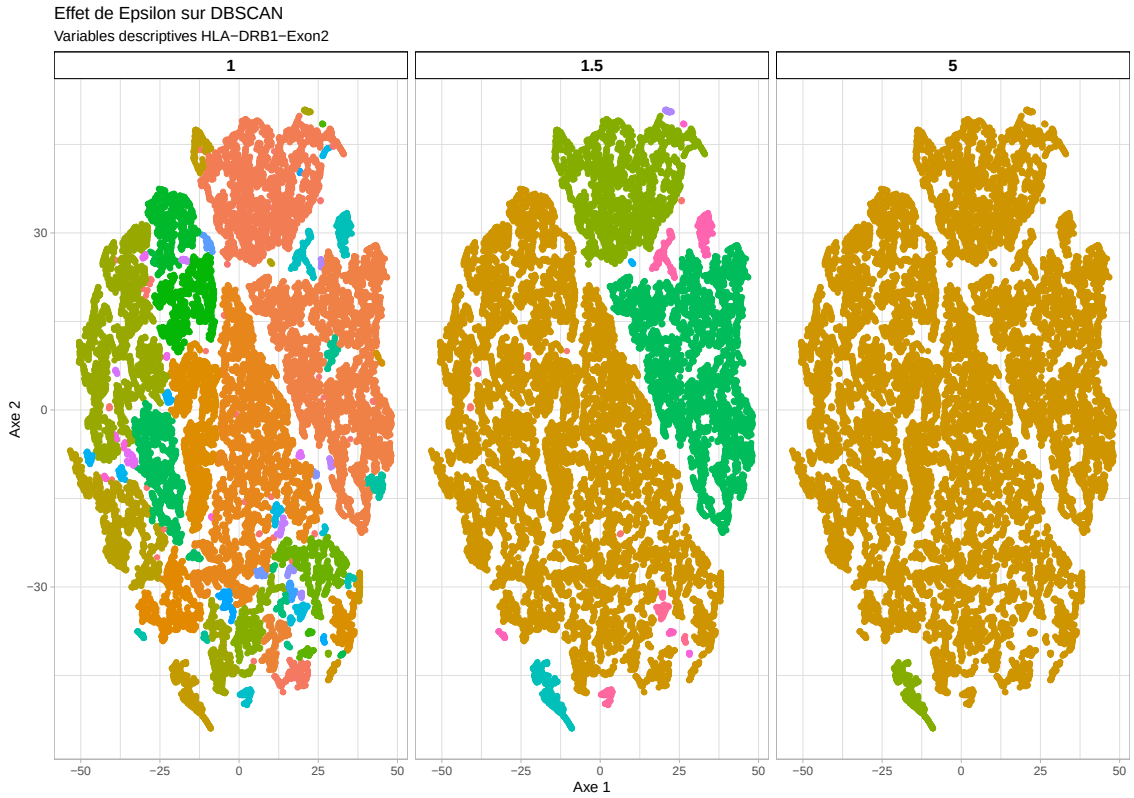


FIGURE 3.5 – Exemple de trois partitionnements DBSCAN réalisés sur le même jeu de données, mais avec des valeurs de ϵ différentes (resp. 1, 1.5 et 5). Les couleurs des points illustrent les groupes attribués par l'algorithme DBSCAN.

Une petite valeur de ϵ (1, à gauche sur la Figure) permet d'identifier 73 groupes différents, tandis que des valeurs plus élevées (respectivement 1.5 et 5, au milieu et à droite) conduisent à l'identification de respectivement 18 et 2 groupes.

Application

Une fois les différentes projections réalisées à l'aide des t-SNE, une gamme de valeurs pour ϵ est testée.

Pour chaque projection et chaque valeur de ϵ, les groupes identifiés par DBSCAN sont évalués comme étant des groupes de vrais variants si la fréquence moyenne du groupe est supérieure à un seuil fixé (habituellement 5%). Chaque variant se voit assigner un score correspondant au nombre de fois qu'il est retrouvé dans un groupe de vrais variants.

Les distributions des scores de chaque variant sont ensuite mises en commun, réduites entre 0 et 100 afin d'obtenir (en pourcentage) une probabilité d'être un vrai variant et la méthode d'Otsu [Otsu, 1979] est utilisée afin de trouver un seuil départageant les vrais variants des artefacts.
Initialement developpée pour le seuillage automatique d'image en nuances de gris vers une image en noir et blanc, la méthode d'Otsu consiste à trouver, au sein d'une distribution, une valeur seuil telle que la somme des variances intra-classes soit minimale (ce qui revient à maximiser la variance inter-classes).
Ici, la méthode d'Otsu est appliquée afin de trouver un seuil qui sépare les artefacts des vrais variants, basé sur la distribution des probabilités préalablement calculée.
`MADaM` renvoie aussi un intervalle de confiance, correspondant à l'intervalle de part et d'autre du seuil, allant de la plus haute valeur de probabilité pour les artefacts à la plus faible valeur de probabilité pour les vrais variants. Si cet intervalle de confiance est réduit, cela signifie que les groupes déterminés auparavant ne discriminent pas parfaitement les variants. Il est alors nécessaire d'explorer manuellement les variants de part et d'autre de la valeur seuil afin de réduire les faux positifs et faux négatifs.
Au contraire, si l'intervalle de confiance est assez large, cela signifie que l'algorithme a su regrouper les vrais variants dans un ou plusieurs groupes suffisamment distincts du (ou des) groupe(s) d'artefacts.

Le pseudo-code correspondant aux étapes de réduction de la dimensionalité, du regroupement et de la classification des variants est représenté sur la Figure 3.6.

2.5 Avantages des méthodes employées

Dans sa version actuelle, `MADaM` possède plusieurs avantages sur les autres méthodes (voir Section 1.2).

Le seuillage est toujours basé sur un critère de fréquence, mais de fréquence moyenne. Ainsi un vrai variant, mais sous-amplifié (et donc de fréquence inférieure au seuil fixé), peut quand même, dans une certaine mesure, être validé si le groupe dans lequel il se trouve a une fréquence moyenne supérieure à ce seuil.

L'extraction des variables descriptives présente deux avantages majeurs :
Il n'y a pas d'a priori sur le nombre de différences entre un artefact et le variant dont il découle. Ces différences étant traduites dans `MADaM` par la proportion de positions de l'artefact expliquée par le variant parent, elles ne dépendent pas d'un seuil fixé à l'avance et, de par l'utilisation de l'écart-type et non de la moyenne, elles ne dépendent pas de la taille des variants.
Le problème des chimères est réglé automatiquement. Dès lors que les deux variants à

N_{var} : Nombre de variants
F1 ← Matrice des variables explicatives, dimension : N_{var} x 5 variables
P ← Gamme de valeurs de perplexité pour la t-SNE
n ← Nombre de répétition des t-SNE
E ← Gamme de valeurs ϵ pour DBSCAN
S ← Score des variants : vecteur de 0, de taille N_{var}
t ← Seuil de fréquence

Pour chaque p dans P :
    Faire n t-SNE sur F1 de perplexité p
    F2 ← projection t-SNE retenue (critère : plus faible divergence Kullback-Leibler)
    Pour chaque ϵ dans E :
        Appliquer DBSCAN(ϵ) à F2
        Pour chaque groupe identifié :
            Si $\overline{fréquence_{groupe}} \geq t$:
                S[variants ∈ groupe] ++

S ← $100 \cdot \frac{S}{max(S)}$

Appliquer Otsu sur S → prédictions = vecteur de 0 ou 1, taille N_{var}

FIGURE 3.6 – Pseudo-code résumant les étapes de réduction de dimensionalité, de regroupement et de classification de `MADaM`.

l'origine de la chimère ont été ajoutés à la liste des variants explicatifs, l'ensemble des positions de la chimère sont expliquées et lorsque cette dernière sera ajoutée à la liste des variants explicatifs, la faible quantité d'information supplémentaire qu'elle apportera (notamment le δSD_{pos} et δSD_{seq}) fera qu'elle sera classifiée comme artefact.
Il n'y a pas non plus de conditions fixes pour classifier un variant comme un vrai variant puisque l'utilisation de suffisamment de statistiques descriptives permet de réduire les zones grises propres à chaque statistique.

La réduction de dimensionalité par t-SNE a montré sa capacité à regrouper, de manière non supervisée, les vrais variants ensemble et les artefacts ensemble sur la base des similarités de leurs variables descriptives. De plus, le fonctionnement de la t-SNE, en regroupant les observations de manière relative (les unes par rapport aux autres), permet de mettre dans un même ensemble des observations en apparence différentes. Par exemple, dans un groupe de vrais variants, tous les variants ne présentent pas une fréquence élevée qui permettrait de les classifier en vrais variants, mais les variants de fréquences moins élévées possèdent quand même des caractéristiques de vrais variants (tels qu'un δSD élevé, indiquant que ce variant explique à lui seul une part importante des artefacts) et de proche en proche, le long d'un gradient de fréquence, ces variants sont quand même regroupés avec les vrais variants.

L'utilisation d'un classificateur à chaînes de Markov permet de différencier de manière efficace (plus rapide qu'un BLAST) des séquences fortement similaires (voir page 146 pour une preuve de fonctionnement de la méthode) et ce avec une complexité d'ordre $O(n)$, que ce soit dans la phase d'apprentissage ou dans la phase de classification.

2.6 Implémentation de l'algorithme

Une implémentation, sous la forme d'un set de scripts Python3, est disponible pour `MADaM` à l'adresse suivante `https://gitlab.unige.ch/Thomas.Goeury/madam-software`.

La persistance des données utilise le système de gestion de bases de données (*SGDB*) SQLite3.
Les scripts Python3 font appel aux librairies `SciKit-Learn` [Pedregosa et al., 2011] (notamment pour l'algorithme DBSCAN), `Numpy` [Oliphant, 2006] afin d'optimiser les calculs numériques sur les matrices, `sqlite3` pour les requêtes à la base de données et `Matplotlib` [Hunter, 2007] pour la réalisation de graphiques diagnostiques.
Pour la recherche des couples amorces PCR/tags oligonucléotidiques, il est possible d'utiliser les ambiguïtés de bases nucléotidiques telles que définies par la norme IUPAC [Cornish-Bowden, 1985].

Les scripts font plusieurs fois appel à des programmes extérieurs, via des appels systèmes :

- Le programme `grep` [Foundation, 2020] pour la recherche des couples de tags oligonucléotidiques au sein des lectures. Cette méthode a été choisie car rapide à mettre en place et montrant des performances (en termes d'utilisation de la puissance de calcul et de temps de calcul) satisfaisantes. L'utilisation de ce programme oblige toutefois l'utilisation d'un environnement UNIX ;
- Le programme BLAST [Altschul et al., 1990], exécuté en local contre une base de données créée à l'aide de la séquence de référence ;

— Les t-SNE sont réalisées avec l'environnement R [R Core Team, 2020], via le package `RTSNE` [Krijthe, 2015], l'implémentation en Python disponible avec la librairie `SciKit-Learn` montrant une utilisation trop importante de la mémoire lorsqu'un grand nombre de données sont utilisées.

Le choix de certains programmes externes déjà optimisés et l'optimisation de certaines parties du code parmi les plus consommatrices de ressources permet à l'implémentation de `MADaM` d'être (relativement) peu consommatrice en ressources et de fonctionner sur des ordinateurs du bureau. En effet, le temps pour traiter les données décrites dans la partie 3.1 est d'environ 6h30 heures sur un ordinateur de bureau (testé sur un Dell Optiplex 9020, 16GB de RAM et processeur i7-4770).

3 Application à des données réelles

Comme indiqué dans la Section 1.1, `MADaM` a été initalement développé pour traiter un jeu de données de séquençage Roche-454 (appelé « Jeu de données 454 »), où les exons 2 de quatre loci HLA de classe II ont été séquençés chez 3'455 individus en trois séries [10] chacun. Nous présenterons ici le traitement par `MADaM` des séquençages du locus HLA-DRB. Ce locus a été choisi car il regroupe en réalité plusieurs gènes, HLA-DRB1, -DRB3, -DRB4, -DRB5, -DRB7, tous des gènes paralogues (issus de duplications sucessives d'un même gène ancestral [Satta et al., 1996a]) et ayant co-amplifié à cause d'une spécificité non totale des amorces par similarité des régions d'hybridation de ces amorces. Ces résultats de séquençage sont donc particulièrement intéréssants car ils permettent de présenter l'ensemble des fonctionalités de `MADaM`, notamment le filtre markovien.
Une illustration du fonctionnement détaillé de l'algorithme est donc présentée dans la section suivante, utilisant ces données de séquençage.

De plus, la Section 3.2 présente une application de `MADaM` à un jeu de données issu de recherches déjà publiées, consistant en des séquençages Roche-454 des exons 2 de MHC-DRB1 et MHC-DRB2 chez 91 Carcajous (*Gulo gulo*) [Oomen et al., 2013, Rico et al., 2015].

3.1 Jeu de données 454

Initialisation

Les exons 2 de HLA-DRB1 pour 3'455 individus ont été séquencés en 3 séries, comprenant au total 1,77 million de lectures. Les lectures ont ensuite été filtrées sur la base de leur PhredScore. Le PhredScore est relié à la probabilité d'une erreur de lecture sur une base selon la formule suivante :

$$P = 10^{\frac{-Q}{10}} \tag{3.1}$$

Équation 3.1 – Calcul du PhredScore P d'un nucléotide d'une lecture de séquençage à partir de la probabilité d'une erreur de lecture Q.

Le tri des lectures basé sur leur PhredScore a été réalisé à l'aide de la commande `trim.seqs` du logiciel Mothur [Schloss et al., 2009], et d'un PhredScore de minimum 30. Ce seuil a été choisi car correspondant à une probabilité d'erreur de lecture d'une base pour 1'000, valeur bien supérieure à la taille attendue des lectures (entre 230 et 250 nucléotides pour HLA-DRB).
Toute lecture dont la qualité moyenne est inférieure à 30 a été rejetée par la commande `trim.seqs`. Cela a représenté une faible fraction des lectures totales (entre 0.27 et 2.21 %), suggérant qu'un filtre un peu moins strict a déjà été appliqué en amont (avant la réception des données au laboratoire).

Les lectures ont ensuite été chargées en mémoire ainsi que les données individuelles, puis trois filtres ont été appliqués :

10. En anglais : *runs*.

1. **filtre sur la taille** : n'ont été conservées que les lectures dont la taille était comprise entre 200 et 400 nt. Un total de 151'793 lectures ont été écartées par ce filtre. La Figure 3.7 montre les distributions de tailles des différentes lectures de HLA-DRB ;
2. **BLAST** contre une séquence de référence HLA-DRB1 (HLA-DRB1*07:01:01, séquence disponible en annexe S-31) qui a exclu 344'745 lectures ;
3. **filtre markovien** : entraîné avec un jeu de données contenant, respectivement, 2'038, 50, 43, 3, 2 séquences d'exons 2 de HLA-DRB1/3, -DRB4, -DRB5, -DRB7 [11]. Ce filtre a exclu 108'179 lectures ne correspondant pas à HLA-DRB1/3.

Les 1'077'852 de lectures restantes ont été assignées aux individus sur la base de leurs tags oligonucléotidiques, puis les amorces PCR et les tags ont été retirés et les séquences ont été alignées.

Seul les individus ayant un minimum de 50 séquences ont été retenus pour la suite des analyses (seuil $T1_{Galan}$), ce qui représente 3'011 individus [12]. La Figure 3.8 montre la relation entre le nombre de variants et de séquences par individu ayant plus de 50 séquences.

La Table 3.1 résume le nombre de lectures retenues après les différents filtres, le nombre de lectures assignées, le nombre de variants identifiés et le nombre d'individus considérés avant et après l'application des différents filtres.

11. Correspond à l'ensemble des séquences d'exons 2 connues pour ces gènes lors du traitement des données le 23/11/2018.
12. Incluant les réplicats.

	Série 1	**Série 2**	**Série 3**
Nb. d'individus	1'152	1'152	1'151
Nb. lectures totales	676'345	539'290	557'613
Après QC	671'088	527'383	556'098
Après filtre sur la taille	670'529	525'110	407'137
Après BLAST	552'962	409'897	295'172
Après filtre markovien	476'676	358'063	243'113
Assignées	400'863	293'566	198'996
Nb. de variants	99'103	127'063	91'899
Nb. individus avec séquences	1'142	1'148	1'097
Nb. individus avec 50+ séquences	975	1'108	928

TABLE 3.1 – Description numérique des 3 séries de séquençage des loci HLA-DRBx. Nb. d'individus : nombre d'individus initialement séquencés ; Nb. lectures totales : nombre de lectures de séquençage obtenues ; Après QC : nombre de lectures restantes après filtre sur la base du PhredScore (QC : *Quality-Check*) ; Après filtre sur la taille : nombre de lectures restantes après un filtre basé sur des tailles minimales et maximales de lectures ; Après BLAST : nombre de lectures restantes après comparaison avec une séquence de référence à l'aide de BLAST ; Après le filtre markovien : nombre de lectures restantes après l'application du filtre markovien, censé ne conserver que les lectures correspondant à HLA-DRB1 ou -DRB3 ; Assignées : nombre de lectures assignées à des individus sur la base de leurs tags oligonucléotidiques ; Nb. de variants : nombre de variants (séquences uniques chez un individu) ; Nb. individus avec séquences : nombre d'individus auxquels des séquences ont pu être assignées ; Nb individus avec 50+ séquences : nombre d'individus auxquels plus de 50 séquences (seuil $T1_{Galan}$) ont pu être assignées.

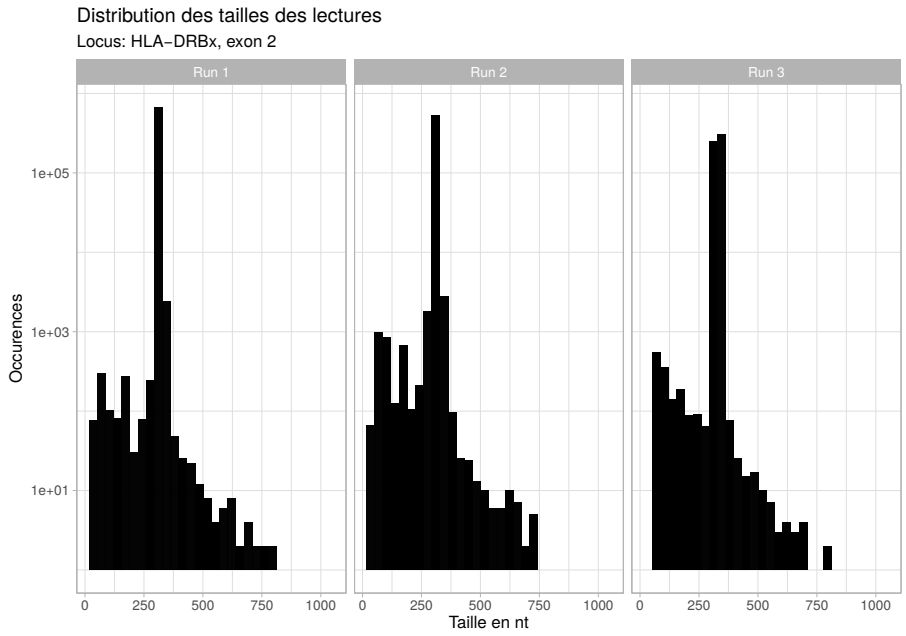


FIGURE 3.7 – Histogrammes représentant les distributions observées des tailles des lectures pour les 3 séries de séquençage de HLA-DRB. L'axe Y est en base log_{10}.

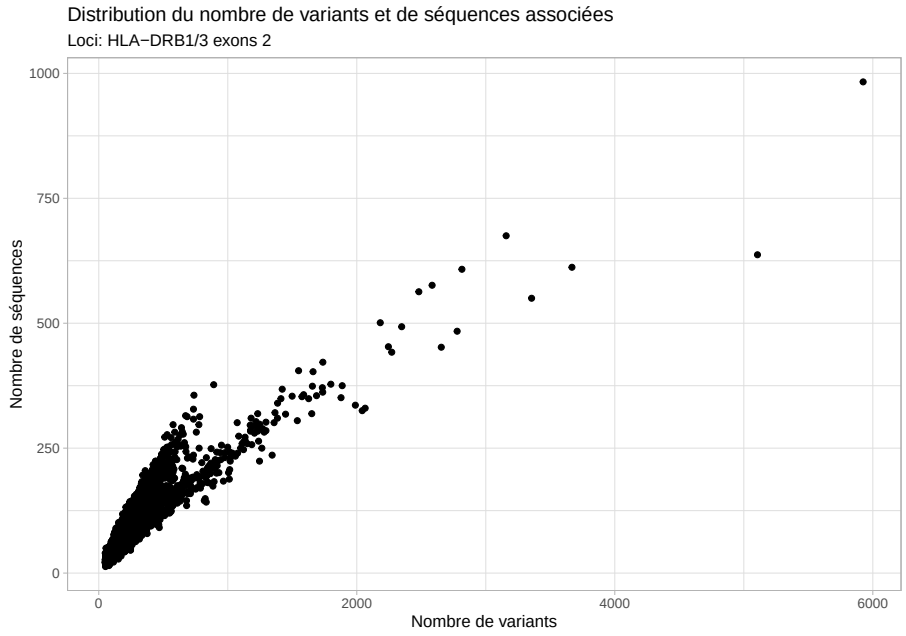


FIGURE 3.8 – Distribution du nombre de variants et de séquences par individu pour les exons 2 de HLA-DRB1/3. Ne sont affichés que les variants des individus ayant plus de 50 séquences (seuil $T1_{Galan}$).

Démonstration du filtre markovien

Respectivement 2'581, 304, 151, 121 séquences d'exons 2 de HLA-DRB1, -DRB3, -DRB4 et -DRB5 ont été utilisées [13]. Les capacités d'apprentissage et de prédiction ont

13. Téléchargées le 25/09/2019 sur IMGT-HLA

été évaluées. Certaines séquences d'exons 2 de HLA-DRB1 et HLA-DRB3 étant très similaires, il est impossible au filtre markovien de les départager avec précision, c'est pour cela que les séquences de ces deux loci ont été regroupées dans le même groupe DRB1/3.

Capacités d'apprentissage :
La capacité d'apprentissage du filtre de Markov a été évaluée en étudiant l'évolution des fréquences des dinucléotides pour les séquences des exons 2 de HLA-DRB1, selon le nombre de séquences considérées. Les fréquences des dinucléotides correspondent aux fréquences des états de transition des chaînes de Markov calculées.

De 5 à 99 séquences ont été échantillonnées aléatoirement (sans remise) parmi les 2'581 séquences d'exons 2 de HLA-DRB1 disponibles et l'écart-type a été évalué par 100 répétitions à chaque fois. Dans le cas d'un apprentissage rapide du classificateur, on s'attend à ce que les fréquences se stabilisent rapidement (faible écart-type) dès un faible nombre de séquences. Les résultats de ce test sont représentés sur la Figure 3.9.

L'analyse montre que les fréquences des dinucléotides se stabilisent rapidement : pour le dinucléotide `AA`, pour 5 séquences analysées la fréquence moyenne est de 0.0273 ± 0.0022, tandis que pour 99 séquences analysées, la moyenne est de 0.0272 ± 0.0005.
L'écart-type relatif (le coefficient de variation) décroît lui aussi rapidement avec le nombre de séquences : la valeur la plus importante observée est de 8.1% (pour le dinucléotide `AA` et 5 séquences) et, dès 16 séquences, plus aucun coefficient de variation n'est supérieur à 5%.

Précision et Rappel :
Les performances en termes de précision et de rappel[14] du filtre markovien ont étés évaluées. La précision, correspond ici au nombre de séquences de DRB1/3 correctement attribuées et le rappel au nombre de séquences prédites comme étant DRB1/3 qui sont réellement des séquences DRB1/3. Ces deux indicateurs se calculent de la façon suivante :

$$\begin{aligned} Precision &= \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Négatifs} \\ Rappel &= \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Positifs} \end{aligned} \tag{3.2}$$

ÉQUATION 3.2 – Calcul de la précision et du rappel d'un classificateur.

L'apprentissage supervisé s'est fait sur 30% des données de chaque catégorie et les 70% de données restantes ont été utilisées pour les prédictions. Le but étant de discriminer les séquences HLA-DRB1/3 des autres, les premières ont été regroupées au sein d'une même catégorie.
Ce processus a été réalisé 100 fois afin de tenir compte de la variabilité apportée par l'échantillonnage aléatoire 30%/70%. Les résultats des prédictions du filtre markovien sont présentés dans la Figure 3.10.

14. *Recall* en anglais.

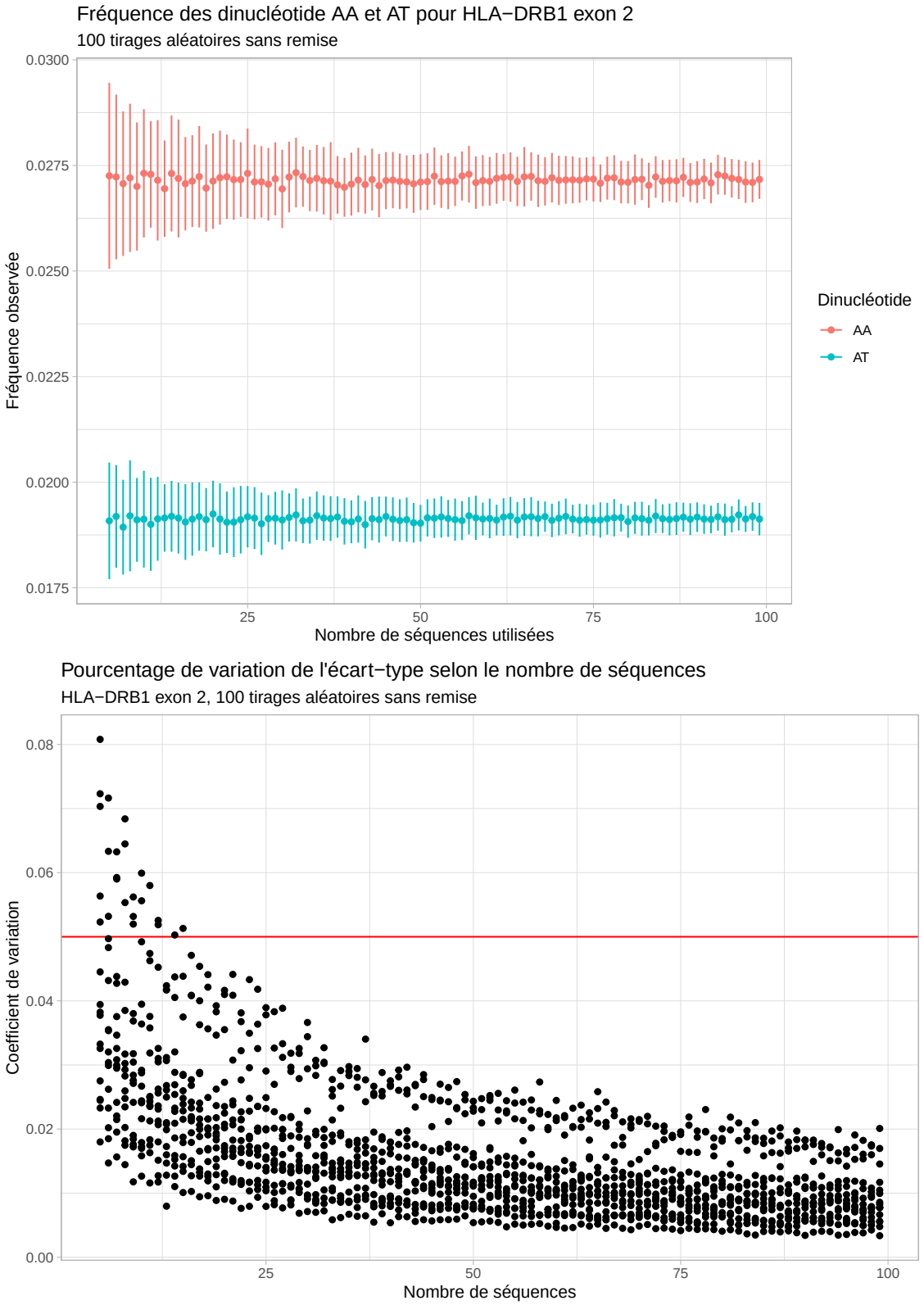


FIGURE 3.9 – Graphique du haut : évolution des fréquences observées des dinucléotides AA et AT, pour l'exon 2 de DRB1, selon le nombre de séquences utilisées (de 5 à 99). L'écart-type a été obtenu par 100 répétitions. Les 14 autres dinucléotides ne sont pas affichés par souci de clarté, voir l'annexe S-32 pour le graphique complet.
Graphique du bas : coefficients de variation des écart-types (100 répétitions à chaque fois), selon le nombre de séquences utilisées, calculés pour l'ensemble des 16 dinucléotides. La ligne rouge correspond au seuil de 5%.

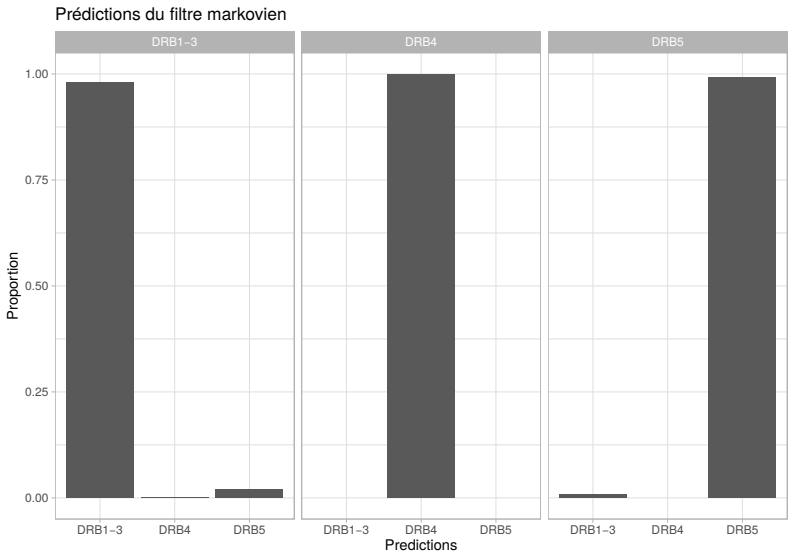


FIGURE 3.10 – Histogramme des prédictions du filtre markovien. Chaque facette représente le vrai locus des séquences à prédire et l'axe des abscisses représente les prédictions réalisés par le classificateur. L'axe des ordonnées représente la proportion de chaque prédiction au sein de chaque classe.

Sur un total de 218'889 prédictions réalisées, 97.95% des séquences DRB1/3 sont correctement prédites (**précision**), et 99.97% des séquences prédites en tant que DRB1/3 le sont réellement (**rappel**).

Conclusion :
Le classificateur à chaînes de Markov montre de très bonnes capacités d'apprentissage et de prédiction : les valeurs des états de transition se stabilisent dès 16 séquences utilisées et montrent une robustesse pour des valeurs plus faibles (8% de coefficient de variation obtenu pour 5 séquences). De plus, il est capable de réaliser des prédictions fiables à 97.95% et montre un taux de faux positifs de 0.03%.

Réduction de la dimensionalité, partitionnement et classification

Une fois les variables explicatives extraites pour les 10 variants les plus fréquents de chaque individu [15], les t-SNE ont été réalisées en considérant 3 valeurs de perplexité différentes : 30, 40 et 50 [16]. Pour chaque valeur de perplexité, afin de réduire les risques de tomber dans un minimum local non optimal, 5 t-SNE indépendantes ont été réalisées et celle qui présentait la plus faible valeur de divergence de Kullback-Leibler a été retenue à chaque fois. Le choix de 5 t-SNE est un compromis entre le temps de calcul et le risque d'obtenir un minimum local non optimal. Un test réalisé avec 30 t-SNE sur le même jeu de données et un facteur de perplexité de 50 n'a montré que 1.09% de variation ($KL = 1.56 \pm 0.017$) de la divergence de Kullback-Leibler, montrant la robustesse de cette méthode vis-à-vis des minima locaux.

Les Figures 3.11 à 3.14 sont produites automatiquement par le programme à des fins

15. Limite fixée afin de réduire la complexité des calculs à venir.
16. Valeurs de perplexité habituelles pour un jeu de données de cette taille.

de visualisation et de diagnostic.

La Figure 3.11 est le résultat de la t-SNE de perplexité 50 ayant la plus faible valeur de divergence de Kullback-Leibler. Chaque point est un variant, et la couleur représente la fréquence de ce variant chez l'individu qui le porte.
Le partitionnement DBSCAN a été appliqué avec une gamme de paramètres ϵ allant de 1 à 10 avec un pas de 0.1 (ce qui fait 100 valeurs de ϵ testées). La fréquence minimale pour considérer un groupe comme étant un groupe de vrais variants est de 3%. Cette valeur seuil est arbitraire, mais elle a été choisie car les 2/3 des variants ont une fréquence inférieure à 3%. La Figure 3.12 est le résultat du partitionnement par DBSCAN sur les résultats de cette t-SNE.
Les scores des variants (Figure 3.13) ont ensuite étés utilisés pour classifier les variants en vrais variants (en vert sur l'histogramme) ou en artefacts (en rouge).

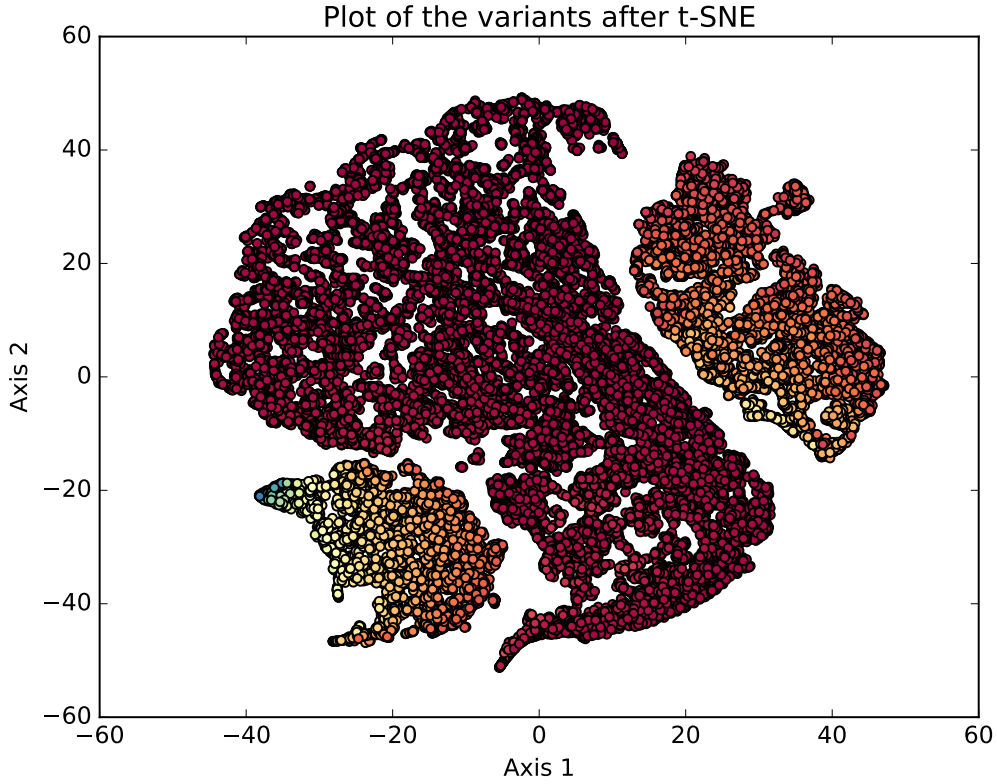


FIGURE 3.11 – Résultat de la t-SNE (perplexité 50) sur les données de DRB1/3. Les variants sont colorés par leur fréquence (rouge : fréquence réduite, bleu : fréquence élevée). *Plot of the variants after t-SNE* : graphique des variants après la t-SNE ; *Axis* : Axe.

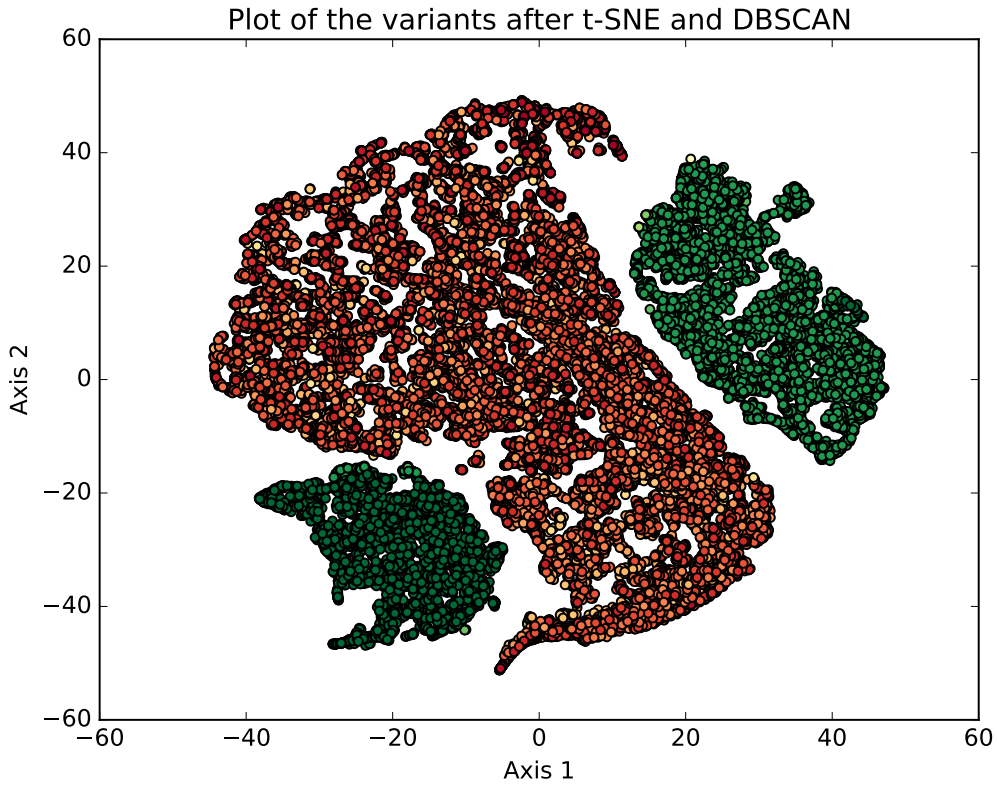


FIGURE 3.12 – Résultat du partitionnement DBSCAN et de la classification sur les données de DRB1/3. Les variants sont colorés par leur score (rouge : score faible, vert : score élevé. *Plot of the variant after t-SNE and DBSCAN* : graphique des variants après la t-SNE et le partitionnement par DBSCAN ; *Axis* : Axe.

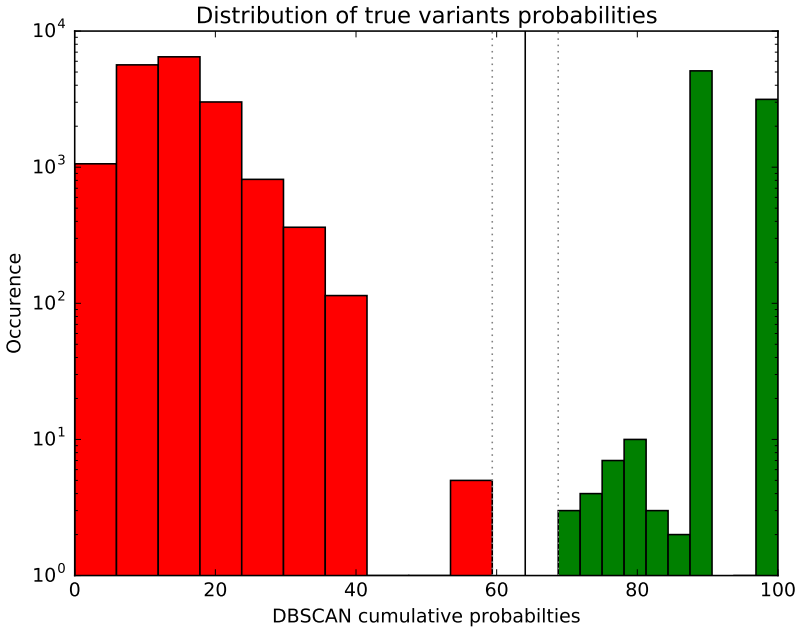


FIGURE 3.13 – Distribution des scores de chaque variant suite au partitionnement DBSCAN. La ligne continue verticale correspond au seuil identifié par la méthode d'Otsu, et les lignes pointillées de part et d'autre l'intervalle de confiance autour de ce seuil. La distribution des probabilités des variants identifiés comme vrais variants est en vert et celle des variants identifiés comme artefacts est en rouge. L'axe Y suit une échelle logarithmique. *Distribution of true variants probabilites* : distribution des probabilités des vrais variants ; *DBSCAN cumulative probabilities* : probabilité d'être un vrai variant.

La ligne verticale continue correspond au seuil trouvé par la méthode d'Otsu pour départager les deux classes de variants et les lignes en pointillés correspondent à l'intervalle de confiance autour de cette valeur, c'est-à-dire la marge entre la valeur seuil et les premières valeurs adjacentes.
À gauche de cette marge, on observe 5 variants de score 59.38, classifiés comme artefacts, mais proches de la valeur seuil. L'analyse manuelle montre que ces variants possèdent des fréquences élevées (de 0.10 à 0.13) mais au sein d'individus avec peu de lectures (ces 10 à 13% de fréquences représentent en réalité 6 à 7 séquences par variant). Ces quelques variants possèdent alors peu d'artefacts, ce qui se traduit par des statistiques descriptives de faible amplitude et a mené `MADaM` à les classifier comme des artefacts, tout en indiquant à l'utilisateur un potentiel risque de mauvaise classification (à l'aide du score).

Une fois les scores des variants seuillés, le résultat graphique est visible sur la Figure 3.14, qui correspond à une projection réalisée par t-SNE, de paramètre de perplexité 50. On observe trois principaux groupes : en bas à gauche (centroïde autour de [-20, -30]) et en haut à droite (centroïde [40, 10]) se trouvent les groupes correspondant aux vrais variants, et le groupe central (en rouge) correspond aux artefacts.

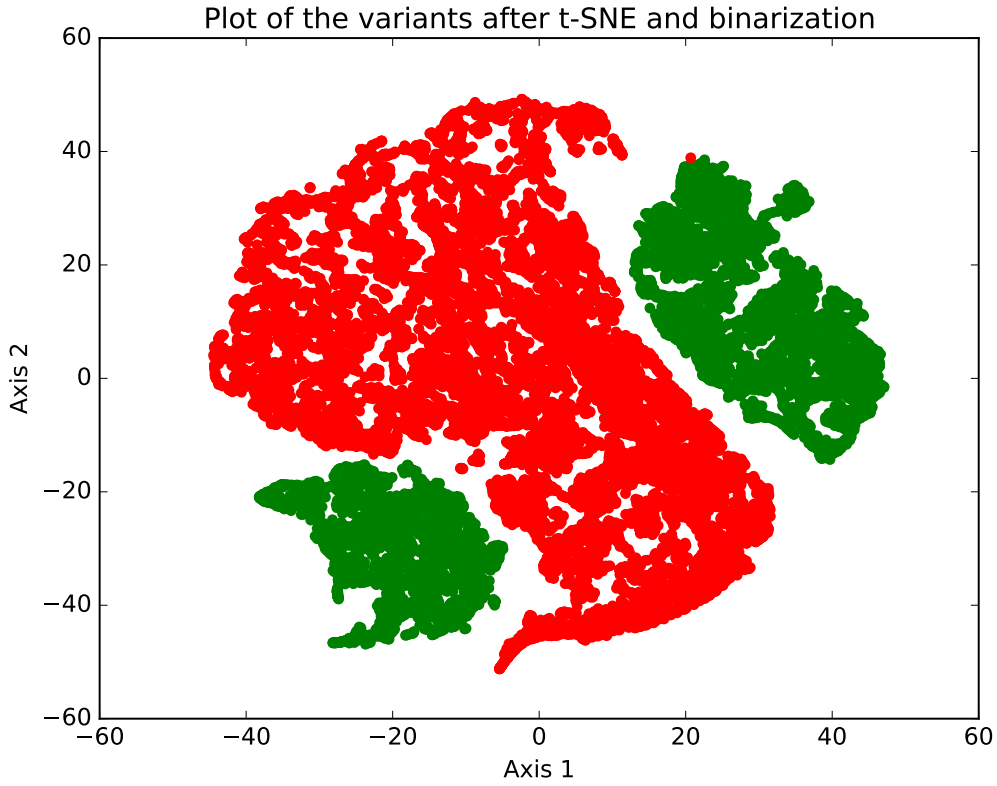


FIGURE 3.14 – Résultat du seuillage par la méthode d'Otsu sur les données de DRB1/3. Les artefacts sont représentés en rouge et les vrais variants en vert. *Plot of the variant after t-SNE and binarization* : graphique des variants après la t-SNE et le seuillage par méthode d'Otsu ; *Axis* : Axe.

Résultats finaux

Après le seuillage par la méthode d'Otsu, le programme a identifié 8'316 vrais variants pour 3'011 individus, avec un nombre moyen de vrais variants par individu de 2.76 ± 0.80. La fréquence moyenne d'un vrai variant est de 0.195 ± 0.093 et celle d'un artefact est de 0.014 ± 0.008.

Parmi les 3'011 individus pour lesquels le programme a identifié les vrais variants, 267 étaient des réplicats, utilisés pour une évaluation finale de la précision de la méthode. L'analyse préliminaire de ces réplicats a identifié 255 correspondances, et 12 différences, soit une précision de 95.51%. La liste complète des réplicats est disponible en annexe S-35. L'analyse approfondie des paires de réplicats ne correspondant pas entre eux a identifié trois sources d'erreurs :

— Quatre paires de réplicats montrent un contenu en variants différents, dont la cause

peut être des contaminations ou des problèmes lors de l'extraction de l'ADN ou de la préparation de la librairie de séquençage. Ces erreurs, probablement en amont du traitement informatique, sont indépendantes de l'algorithme ;
— Trois paires de réplicats ne correspondent pas à cause de la présence des autres loci co-amplifiés (exons 2 de HLA-DRB4/5/6/7), problème dont l'origine se situe dans le pré-traitement des données, lors des différents filtres censés exclure ces séquences ;
— Cinq paires de réplicats sont discordants à cause d'une importante différence de fréquence entre le vrai variant le plus fréquent et le vrai variant le moins fréquent (identifié alors à tort comme artefact). Ces différences de fréquences vont de (fréquence du vrai variant le plus fréquent / fréquence du vrai variant le moins fréquent) : 0.569/0.048 à 0.629/0.033, soit jusqu'à un facteur 18 entre les deux vrais variants. Le problème a une origine double : une faible amplification de ces vrais variants lors de la préparation de la librairie de séquençage et un échec de l'algorithme à identifier ces vrais variants à cause de leur faible fréquence.

En excluant les erreurs totalement indépendantes de `MADaM`, il reste 8 paires de réplicats discordants, ce qui correspond à une précision allant de 97.00 % (considérant les deux individus de chaque paire comme faux) à 98.50 % (considérant un seul individu faux au sein de chaque paire).
Les 3% (maximum) d'erreur se répartissent ici en 1.13 % dus à des erreurs dans le pré-traitement des données (filtres) et 1.87 % dus à des erreurs dans les étapes de classification, au cœur de l'algorithme.

3.2 Application au jeu de données « Glouton »

Traitement des données

Un total de 621'821 lectures ont été chargées en mémoire et réparties en 7 séries (découpage initial de la publication). Après filtre (uniquement sur la taille, pas de BLAST ni filtre markovien) et assignation, il reste 412'163 séquences pour 91 individus.
Ces séquences correspondent à 87'083 variants, au sein desquels 1'820 ont été utilisés dans le processus de classification (les autres étant des variants de trop faible occurrence, principalement des singletons).

Une fois les variables explicatives extraites, 4 valeurs de perplexité [17] (30, 40, 50, 60) ont été utilisées pour les t-SNE, avec à chaque fois 5 t-SNE indépendantes (dont une seule sera retenue à chaque fois sur la base de la plus petite valeur de divergence de Kullback-Leiber).
La Figure 3.15 représente les quatre meilleures t-SNE obtenues pour chaque valeur de perplexité.

Le partitionnement par DBSCAN a été réalisé avec une gamme de valeurs d'ϵ allant de 0.1 à 13 et un pas de 0.1. La Figure 3.16 illustre l'évolution du nombre de groupes identifiés par DBSCAN, selon la projection t-SNE utilisée et la valeur de ϵ considérée.

17. Afin de couvrir une large gamme de facteurs de perplexité.

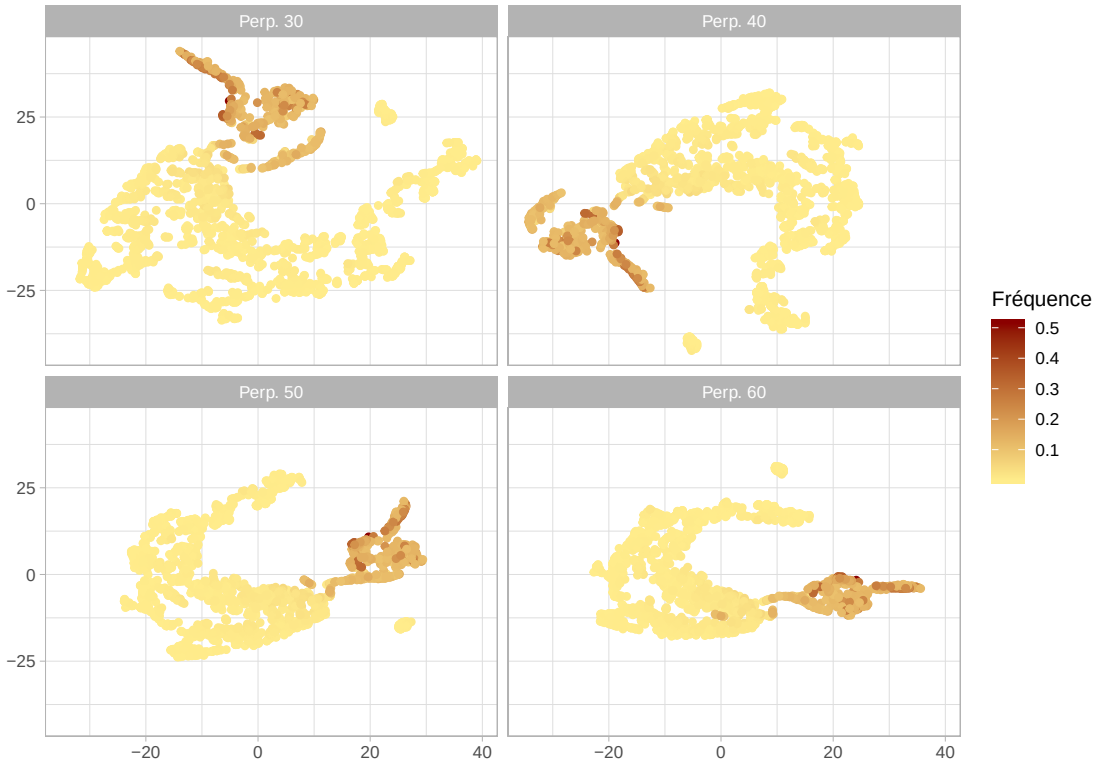


FIGURE 3.15 – Représentation des quatre t-SNE retenues lors de la phase de réduction de la dimensionalité sur le jeu de données « Glouton ». Les variants sont colorés selon leur fréquence, donnée par l'échelle à droite.

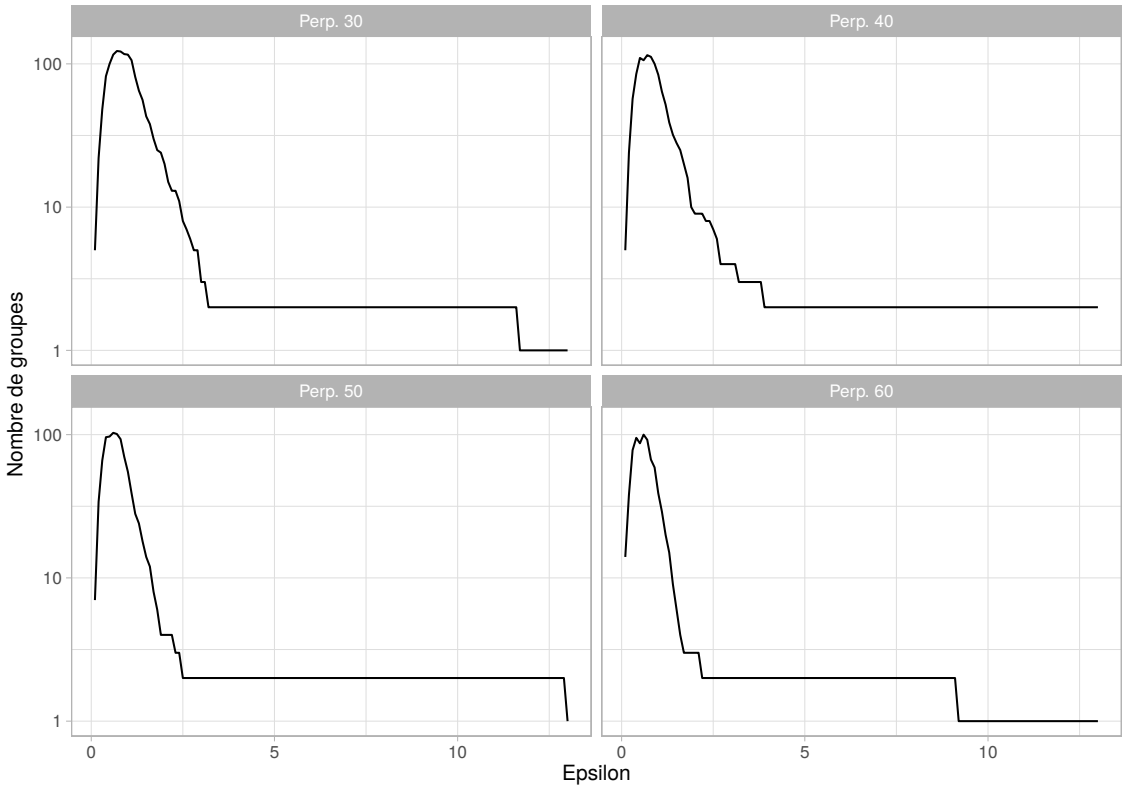


FIGURE 3.16 – Évolution du nombre de groupes identifiés par DBSCAN en fonction de la valeur d'ϵ considérée, pour chacune des 4 projections t-SNE retenues.

L'étape de classification a identifié 18 vrais variants différents (pour les trois loci et les deux espèces).

Assignation des noms d'allèles aux variants

Les séquences trouvées dans l'étude originale ont été téléchargées à partir des références NCBI [18]. Ces séquences ont été utilisées pour associer un nom d'allèle aux variants identifiés par `MADaM`.
La Table 3.2 résume les noms d'allèles associés aux variants identifiés, ainsi que l'occurrence de ces variants et l'espèce associée. Les variants 11 à 15 correspondent aux *P. lotor* séquencés et ne seront pas utilisés pour la suite des analyses (les génotypes pour cette espèce ne sont pas disponibles dans l'étude originale). Parmi les 13 variants restants, les variants 8 et 9 n'ont été identifiés qu'une seule fois chacun et ne correspondent à aucune séquence de l'étude originale (respectivement 3 et 2 différences par rapport à DRB2*10 et DRB2*08). Ces variants ont été retirés de l'étude par manque de fiabilité sur le génotypage.

18. Se référer à l'annexe S-33 pour les références.

Id. variant	Occurrence	Allèle	Espèce
1	89	DRB3*03	*G. gulo*
2	21	DRB2*06	*G. gulo*
3	65	DRB1*04	*G. gulo*
4	41	DRB2*01	*G. gulo*
5	31	DRB2*08	*G. gulo*
6	70	DRB1*02	*G. gulo*
7	26	DRB2*09	*G. gulo*
8	1	-	*G. gulo*
9	1	-	*G. gulo*
10	26	DRB2*05	*G. gulo*
11	1	-	*P. lotor*
12	2	-	*P. lotor*
13	1	-	*P. lotor*
14	1	-	*P. lotor*
15	1	-	*P. lotor*
16	2	DRB2*11	*G. gulo*
17	3	DRB2*10	*G. gulo*
18	3	DRB2*07	*G. gulo*

TABLE 3.2 – Vrais variants identifiés par `MADaM`. Id. variant : numéro unique assigné au variant par `MADaM` ; Occurence : nombre d'individus chez lesquels ce variant est retrouvé ; Allèle : désignation officielle de ce variant selon la nomenclature ; Espèce : espèce à laquelle appartient cet allèle (*G. gulo* : *Gulo gulo* et *P. lotor* : *Procyon lotor*).

Comparaison avec l'étude originale

L'étude originale a identifié 11 allèles différents. Après éviction des variants 8 et 9, ainsi que ceux correspondant à *P. lotor*, la présente étude identifie aussi 11 allèles différents.
Parmi les génotypes disponibles dans l'étude originale, 88 individus ont pu être comparés (génotypes disponibles dans les deux études). Les résultats de l'étude originale ne mentionnent pas les génotypes DRB3 (locus monomorphe), c'est donc une comparaison des allèles trouvés aux loci DRB1 et DRB2 pour 88 individus qui a été réalisée.

Sur 88 individus, 5 sont discordants et correspondent à chaque fois à un allèle non détecté par `MADaM` (*dropout*). L'analyse des variants des individus concernés a montré que ces 5 erreurs ont une cause commune : la présence d'un second variant, similaire (1-2 nt de différence) et de plus haute fréquence, qui conduit systématiquement au classement du premier variant en tant qu'artefact. Ces variants « oubliés » ont toutefois un score très proche de la limite basse de l'intervalle de confiance calculé après seuillage par la méthode d'Otsu. Cela démontre l'intérêt de cet intervalle de confiance et de l'étude manuelle systématique des variants dont le score est proche de cet intervalle si l'on veut réduire les taux de non-détection en cas de situation complexe (variants très similaires, dont l'un serait détecté comme artefact de l'autre).
Ainsi, en considérant quatre variants par individu (deux loci chez des individus diploïdes), sur 352 variants[19] à identifier, `MADaM` a correctement identifié 347/352 variants, soit une précision de 98.58%.

19. 4 variants x 88 individus.

Conclusion

Montrant un taux d'erreur de seulement 1.42% pour 88 individus d'une espèce non modèle (dont le génome a peu été étudié), `MADaM` montre de très bonnes performances quant à la classification d'un grand nombre de variants en vrais variants et artefacts et a identifié de manière robuste les génotypes des individus.

3.3 Commentaires sur les autres méthodes

Si `MADaM` a été developpé pour traiter le jeu de données « 454 », c'est parce qu'aucune des méthodes précédemment citées ne montrait suffisament de fiabilité pour traiter ce jeu de données. En effet, les échantillons de ce jeu de donnés proviennent de plusieurs campagnes d'échantillonnage importantes menées par plusieurs équipes de recherche et représentent donc des données de grande valeur scientifique, qui ne sont pas disponibles facilement. Il fallait donc une méthode permettant de s'assurer que ces données auraient été utilisées au mieux. Or les principales méthodes existantes possèdent plusieurs limitations que `MADaM` a pu dépasser. Cette section présente les limites des méthodes présentées dans la Section 1.2 lorsqu'appliquées au jeu de données « 454 ».

La première technique présentée, celle de l'*Allele Validation Threshold* (page 128) repose sur l'utilisation d'un seuil de fréquence par amplicon au-dessus duquel tous les variants sont des vrais variants et un autre seuil en-dessous duquel tous les variants sont des artefacts. La Figure 3.17 montre la distribution, pour les séquençages de HLA-DRB1 précédement présentés, des fréquences des artefacts (en rouge) et des vrais variants (en bleu). Le graphique de gauche est un agrandissement de la zone de fréquence située entre 0 et 15%, représentant la zone où l'on observe l'ensemble des artefacts mais aussi les vrais variants les moins fréquents.
Cette figure illustre le problème posé par l'utilisation d'un seuil de fréquence par variant pour classifier les variants en vrais variants ou en artefacts (tel que proposé par [Zagalska-Neubauer et al., 2010]) ; puisque peu importe la valeur seuil de fréquence utilisée, cela conduit systématiquement à des faux positifs (artefacts de fréquence supérieure au seuil) et à des faux négatifs (vrais variants de fréquence inférieure au seuil). De plus, l'utilisation de deux seuils de fréquences (un au dessus duquel on ne retrouve plus aucun artefact et un en dessous duquel on ne retrouve plus aucun vrai variant) créerait ici une « zone grise » de 2'474 [20] variants à classifier manuellement.

20. Sur les 25'766 variants traités par `MADaM`.

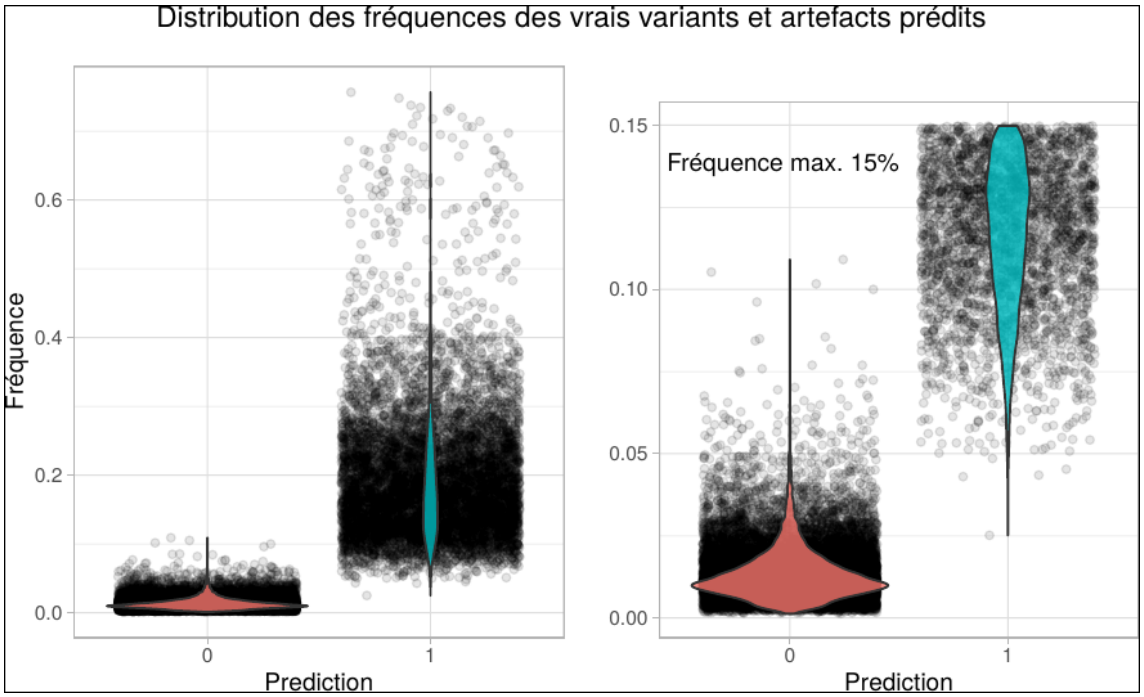


FIGURE 3.17 – Distribution des fréquences observées des variants de DRB1/3-Exon2 selon la prédiction réalisée par `MADaM`. Graphique de gauche : ensemble de variants, graphique de droite : uniquement les variants de moins de 15% de fréquence. 0 signifie que le variant a été prédit comme artefact et 1 que le variant a été prédit comme vrai variant.

Les deux méthodes de *Clustering* (présentées page 128) reposent toutes les deux sur un nombre de différences entre les séquences pour départager les artefacts des vrais variants. Or sur l'analyse des 3'011 individus pour lesquels `MADaM` a obtenu un génotype, 16 individus hétérozygotes possèdent des vrais variants ne différant que d'une seule paire de base (et 63 qui ne diffèrent que de deux paires de bases). Or, sans considérer d'autres facteurs que la simple différence entre deux variants, il est impossible de différencier des artefacts de vrais variants.

La dernière technique, celle du *Degree Of Change* présentée à la page 129, repose sur une différence importante des fréquences entre les vrais variants et les artefacts pour détecter le point d'inflexion dans la courbe du *Degree Of Change*. Cette méthode nécessite que tous les variants amplifient de manière similaire, puisque si un variant sous-amplifie, le point d'infexion de la courbe peut alors ne plus être détectable. La Figure 3.18 donne la distribution des fréquences de chacun des variants HLA-DRB les plus fréquemment observés (ne sont représentés que les variants présents dans au moins 50 individus, dans un but de lisibilité de la figure).

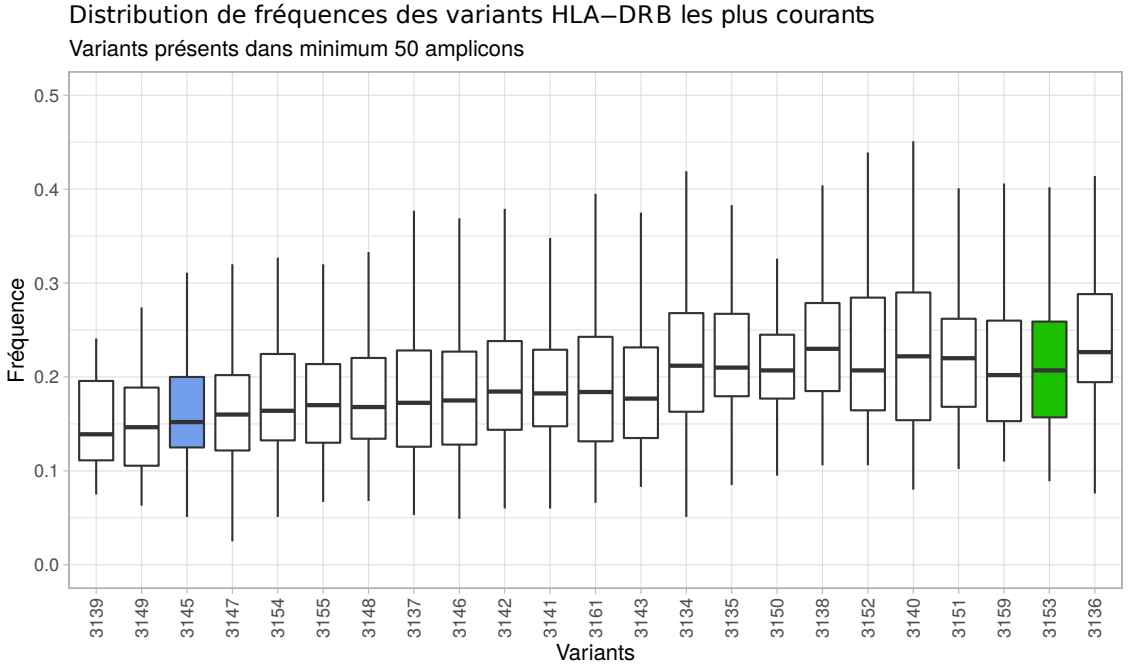


FIGURE 3.18 – Distribution des fréquences des variants HLA-DRB1/3 les plus fréquemment observés (observés chez minimum 50 individus). Les variants sont triés par fréquence moyenne croissante. Les distributions en bleu et en vert identifient les deux variants discutés ci-après, dont les distributions de fréquences sont significativement différentes (Kruskal-Wallis, $pValeur < 2.2e-16$).

On observe alors que tous les variants ne sont pas observés à la même fréquence au sein des individus. Par exemple la fréquence moyenne du variant 3145 (retrouvé chez 215 individus), en bleu sur la Figure 3.18, est de $16.9 \pm 6.8\%$ tandis que celle du variant 3153 (retrouvé chez 273 individus), en vert sur la Figure 3.18 est de $25.8 \pm 14.0\%$. Selon le test de Kruskal-Wallis, ces différences de fréquences sont significatives ($pValeur < 2.2e-16$). Ces résultats indiquent que l'hypothèse d'une amplification similaire de tous les allèles est fausse, certains allèles amplifiant significativement moins que d'autres.
De plus, l'analyse des individus répliqués (dont le génotype, lorsqu'il est identique entre les deux individus répliqués, est fiable) montre que certains individus ont une différence de fréquence entre le vrai variant le plus fréquent et le vrai variant le moins fréquent qui peut aller jusqu'à un facteur 3.32 [21]. Ces mêmes variants, dans l'autre réplicat [22], ne montrent pas de différence aussi importante (différence d'un facteur 1.5, voir annexe S-34).
Ainsi, l'hypothèse d'une amplification similaire de tous les allèles et entre les différents individus n'est pas pertinente, illustrant encore une fois l'importance d'utiliser plusieurs variables pour identifier les vrais variants et les artefacts.

21. Individu #14551, fréquence du variant le plus (respectivement le moins) fréquent de 41.5% (resp. 12.5%).
22. individu #14550

4 Discussion

Ce chapitre a présenté le développement et le fonctionnement de `MADaM`, un algorithme visant à classifier des lectures de séquençage NGS d'amplicons, se basant sur un ensemble original de variables descriptives de ces séquences, d'une réduction de dimensionalité par t-SNE puis un regroupement à l'aide de DBSCAN suivi d'un seuillage par méthode d'Otsu. En termes statistiques et d'apprentissage machine, `MADaM` est un algorithme de classification non supervisé travaillant sur des données présentant une distribution de classe biaisée, puisqu'en fixant par exemple le nombre de variants les plus fréquents à traiter pour un individu (diploïde) à 10, on attend entre un et deux vrais variants pour huit à neuf artefacts.
Cet algorithme a été appliqué à un jeu de données complexe de séquençage Roche-454, composé de deux loci ciblés (HLA-DRB1 et -DRB3), ainsi que de trois autres loci co-amplifiés, et a montré un taux d'erreur (qui lui était attribuable) de 3% sur 3'455 individus. Il a été aussi appliqué à un jeu de données issu du séquençage 454 de trois loci MHC-DRB de 89 *Gulo gulo* (une espèce non modèle) et a alors montré un très faible taux d'erreur de 1.42%.
Outre son très faible taux d'erreur, `MADaM` présente l'avantage de pouvoir fonctionner sur un ordinateur de bureau de manière relativement rapide (de l'ordre de 6 à 7h pour traiter les 3'455 individus du premier jeu de données) et autonome, en nécessitant peu d'a priori de la part de l'utilisateur.

Concernant les variables utilisées par `MADaM` pour réaliser la classification, ces dernières présentent l'avantage d'être indépendantes de la taille des séquences (les valeurs retenues étant des écart-types) et de résoudre automatiquement le problème des chimères, puisque dès que les deux séquences parentales ont été traitées, les éventuelles séquences chimériques sont automatiquement expliquées.

L'analyse des erreurs d'assignation de `MADaM` a permis de mettre en évidence ses principales limitations et d'en identifier dans certains cas la source.
Premièrement, `MADaM` n'est adapté qu'à des séquençages incluant un grand nombre d'individus. En effet, l'algorithme a aussi été testé pour un jeu de données consistant en un séquençage Illumina-MiSeq de cinq lignées cellulaires humaines pour les exons 2 et 3 de HLA-A et HLA-B [Bai et al., 2014] et n'a pas réussi à classifier correctement les variants à cause d'un trop petit nombre d'individus (et donc de variants), visible par le peu de points sur la Figure 3.19, représentant le résultat de la t-SNE de perplexité 50 sur ce jeu de données.

La cause de ce problème se situe lors de l'étape de réduction de la dimensionalité et de regroupement puisque les deux algorithmes alors utilisés, t-SNE et DBSCAN, sont empruntés au domaine de l'apprentissage machine et sont prévus pour fonctionner sur des jeux de données d'une taille importante. L'algorithme t-SNE inclut un paramètre appelé facteur de perplexité, qui doit être bien en *deçà* du nombre d'observations. La perplexité peut être définie comme le nombre effectif de voisins les plus proches pour chaque point et va ainsi permettre de définir le type de structure observée : de faibles valeurs de perplexité vont faire apparaître dans l'espace à faible dimension des structures (de l'espace à haute dimension) très locales tandis que des plus fortes valeurs de perplexité feront apparaître des structures plus globales. Or la structure qui est recherchée dans l'étape de réduction de la dimensionalité est une structure très globale (les artefacts et les vrais variants),

nécessitant une valeur de perplexité élevée.

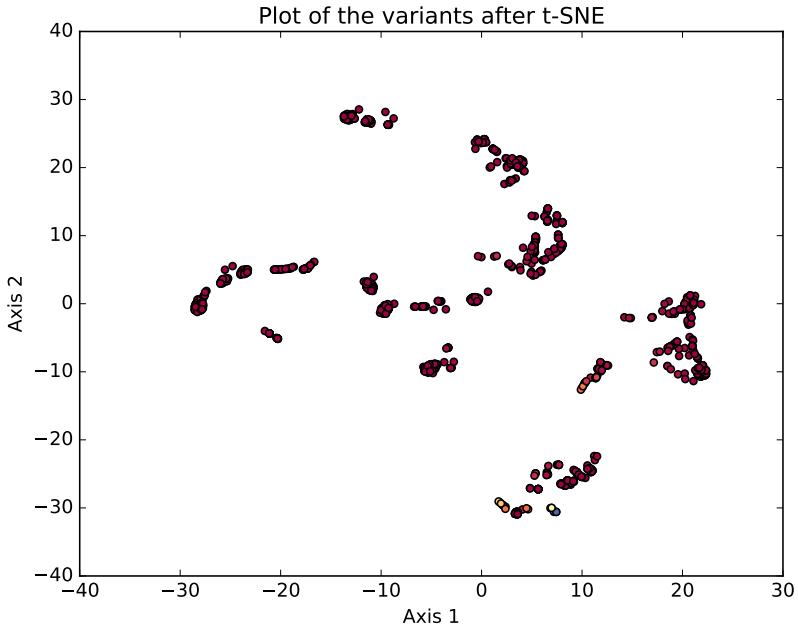


FIGURE 3.19 – Résultat de la t-SNE (perplexité 50) sur les données de [Bai et al., 2014]. Les variants sont colorés par leur fréquence (rouge : fréquence réduite, bleu : fréquence élevée). *Plot of the variants after t-SNE* : graphique des variants après la t-SNE ; *Axis* : Axe.

Ainsi, dans le jeu de données de Bai *et al.* 2014 intégrant seulement cinq individus, `MADaM` va devoir identifier (pour chaque exon et chaque locus) au maximum 10 vrais variants (dans le cas d'individus tous hétérozygotes), voire huit selon le résultat de l'étude originale qui a identifié deux individus homozygotes (lignées cellulaires *Raji* et *C1Rneo*). Le groupe correspondant alors aux vrais variants serait un groupe de maximum 10 points, ne permettant pas d'utiliser un facteur de perplexité supérieur à 10, incompatible avec les valeurs élevées de perplexité nécessaires à `MADaM`.
L'algorithme DBSCAN, quant à lui, bien que capable d'identifier des groupes de différentes formes et de différentes tailles, nécessite toutefois des groupes de densités similaires. Si le nombre d'individus est suffisant (comme dans le cas des séquençages HLA-DRB1/3), alors les groupes des vrais variants présentent une densité similaire à ceux des artefacts et DBSCAN peut alors être appliqué sans contraintes, mais, dans le cas de l'étude de Bai *et al*, le trop petit nombre de vrais variants cause une différence de densité trop importante avec les groupes d'artefacts pour que DBSCAN soit efficace.

Le second problème concerne le cas des variants très similaires de fréquences très différentes, voire sous amplifiés. Ce problème n'est pas lié directement à `MADaM`, mais il en représente une limitation. En effet, bien qu'il n'y ait pas de seuil fixe concernant la différence minimale entre un artefact et un vrai variant, il est facile de concevoir que si deux vrais variants sont similaires mais que l'un des deux est largement sous-amplifié par rapport à l'autre, alors `MADaM` aura du mal à identifier le deuxième variant comme un vrai variant puisque les statistiques descriptives calculées sur ce deuxième variant le feront

ressembler fortement à un artefact qui aurait une fréquence anormalement haute.
Cette sous-amplification est généralement due à des effets aléatoires lors de la PCR qui induisent un biais d'amplification entre les allèles, voire pour certains allèles une sous-amplification très forte, avec une fréquence du variant résultant proche des fréquences des artefacts.
En temps normal si les statistiques calculées sur les résidus d'un vrai variant (voir page 133) sont faibles (et ressemblent donc à un artefact), c'est sa fréquence plus élevée qui évite à ce dernier de se retrouver regroupé avec les artefacts, mais si cette fréquence est elle aussi faible, `MADaM` ne donne pas une bonne classification. Toutefois, la pratique a montré que ces variants obtiennent un score relativement élevé (comme sur la Figure 3.13), permettant leur identification et une décision manuelle de l'utilisateur.

Le dernier problème n'est pas vraiment un problème, mais une spécificité de conception de l'algorithme, puisque ce dernier a été conçu pour fonctionner sur des séquençages où le locus ciblé était séquencé en une seule fois, sans assemblage préalable (`MADaM` ne réalisant pas d'étape d'assemblage, par exemple avec des graphes de de Bruijn [Bruijn, de and Erdös, 1948]), mais il a été aussi conçu pour travailler sur des régions habituellement difficiles à génotyper (à cause d'un important polymorphisme et d'allèles inconnus), notamment la région MHC/HLA, alors que pour des régions moins complexes les méthodes déjà existantes sont efficaces (telles que [Galan et al., 2010, Goecks et al., 2010, Boyer et al., 2015]).

La principale critique qui peut alors être formulée contre `MADaM` concerne le manque de test contre des jeux de données issus d'autres séquençages que des séquençages Roche-454. En effet, bien que ces séquençages se soient généralisés ces dernières années [23] il a été difficile de trouver des études correspondant aux pré-requis de `MADaM` et dont les données de séquençage brutes étaient disponibles. `MADaM` a toutefois été appliqué sur un jeu de données issus de l'étude [Grogan et al., 2016], où 302 *Lemur catta* ont été séquencés par Roche-454 et IonTorrent (à des fins de comparaison des techniques) sur la région MHC-DRB, l'analyse originale utilisant une version modifiée de la technique de [Sommer et al., 2013] pour assigner les génotypes. `MADaM` n'a pas été en mesure de reproduire complètement les résultats obtenus. Bien que les lectures Roche-454 aient donné de très bons résultats avec `MADaM` (vrais variants et artefacts clairement identifiés), ces derniers sont pourtant différents de ceux de l'étude originale. `MADaM` a n'a par contre pas été en mesure d'identifier clairement les vrais variants parmi les séquençages IonTorrent. Les Figures 3.20 et 3.21 représentent les résultats des quatre t-SNE de perplexité 20, 30, 40 et 50 pour les données obtenues par, respectivement, séquençages 454 et Ion-Torrent.

23. Une recherche (menée le 09/01/2020) sur `https://www.ncbi.nlm.nih.gov/pubmed/` utilisant les termes «Illumina» ou «IonTorrent» ou «PacBio» montre une évolution de 397 références en 2010 contre 3'142 en 2019.

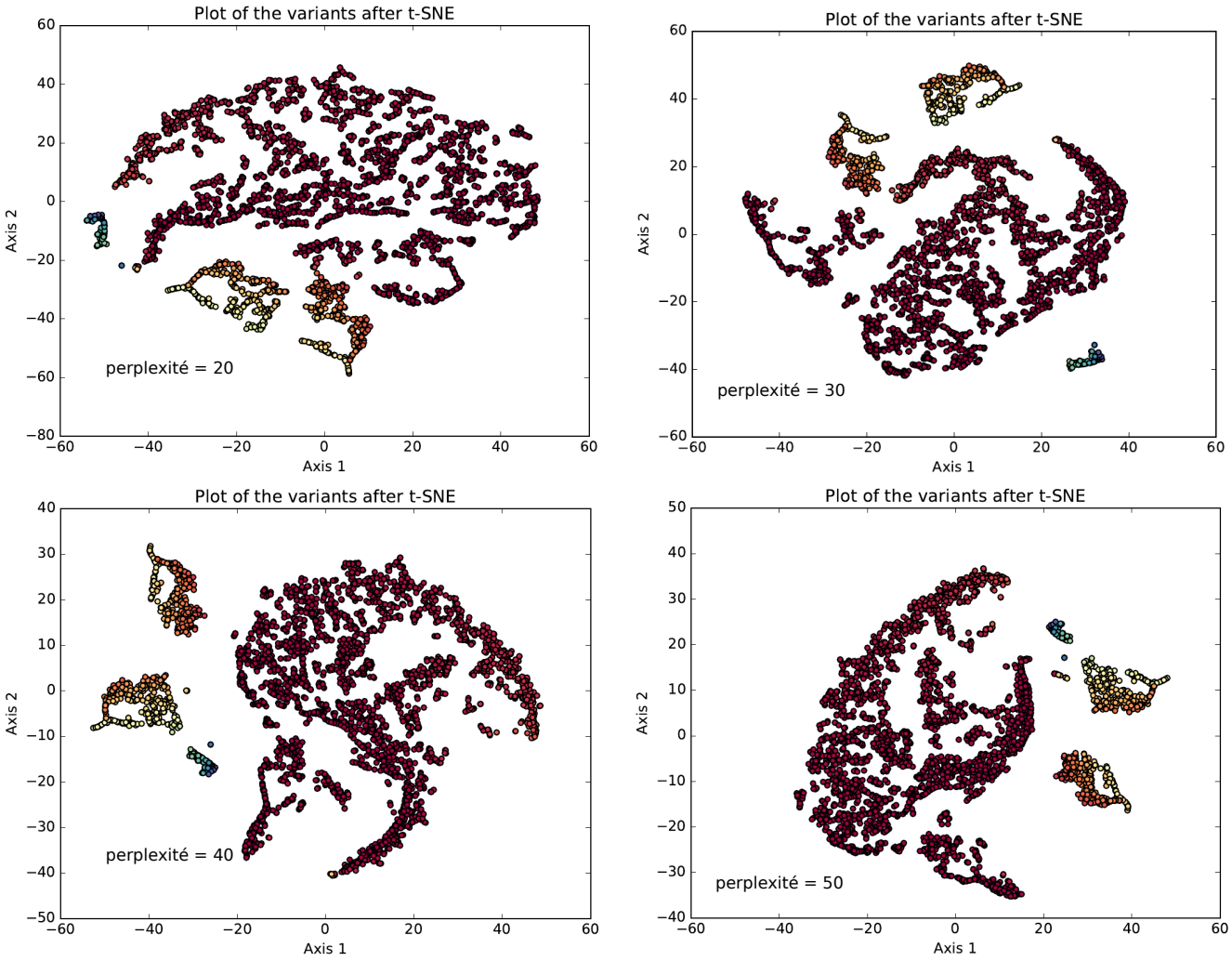


FIGURE 3.20 – Résultat des t-SNE (valeurs de perplexité de 20, 30, 40 et 50) sur les données de [Grogan et al., 2016] obtenues par séquençages 454. Les variants sont colorés par leur fréquence (rouge : fréquence réduite, bleu : fréquence élevée). *Plot of the variants after t-SNE* : graphique des variants après la t-SNE ; *Axis* : Axe.

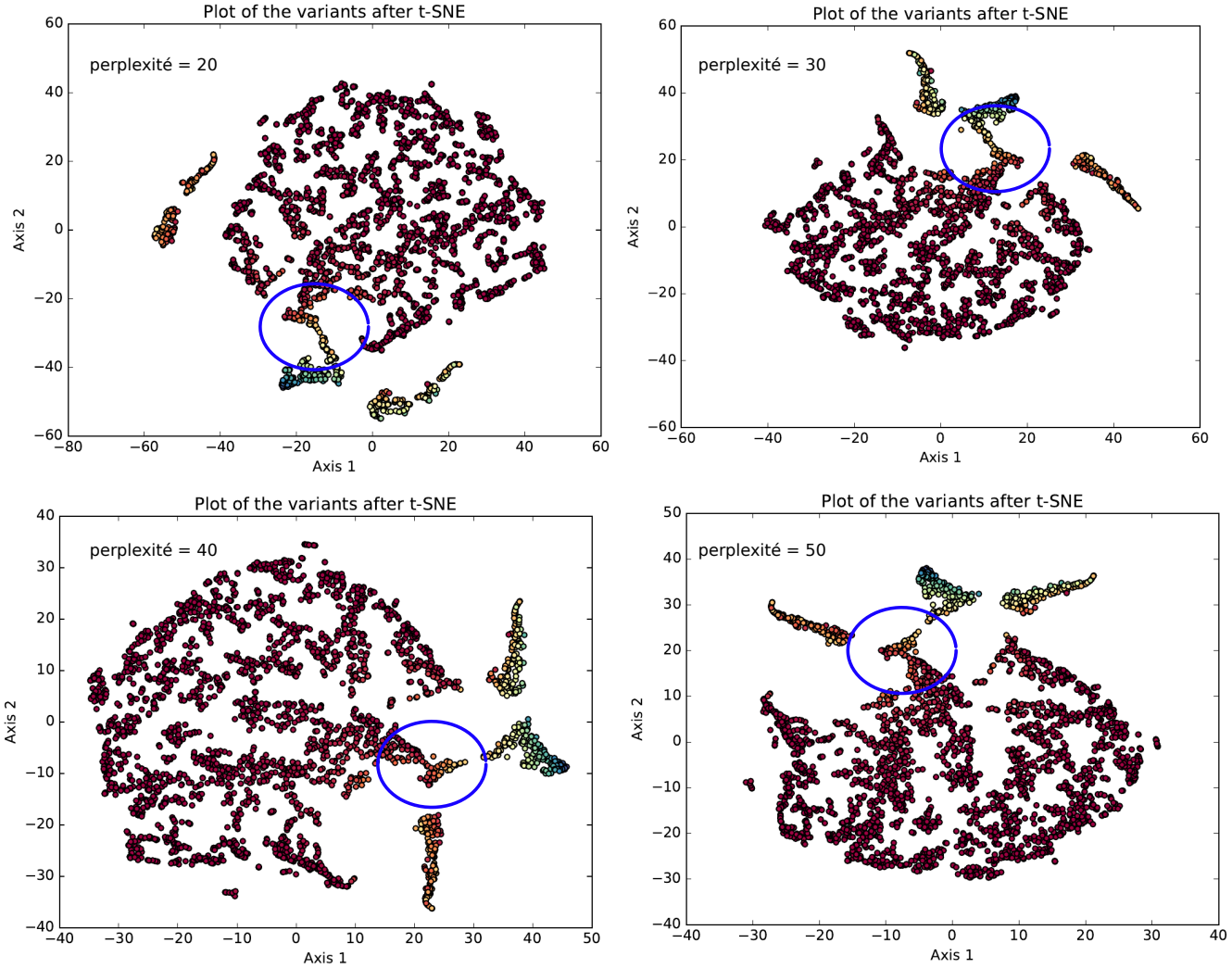


FIGURE 3.21 – Résultat des t-SNE (valeurs de perplexité de 20, 30, 40 et 50) sur les données de [Grogan et al., 2016] obtenues par séquençages Ion-Torrent. Les variants sont colorés par leur fréquence (rouge : fréquence réduite, bleu : fréquence élevée). Les cercles bleu indiquent les zones où la différence entre les vrais variants et les artefacts n'est pas suffisament importante pour permettre de les distinguer. *Plot of the variants after t-SNE* : graphique des variants après la t-SNE ; *Axis* : Axe.

Ces figures illustrent la raison des différences de résultats entre les données obtenues par séquençages 454 et Ion-Torrent. Si, pour les données de 454, les vrais variants sont nettement différenciés des artefacts, ce n'est pas le cas pour les données Ion-Torrent. Pour ces dernières, on observe des zones de continuité entre les vrais variants et les artefacts, représentées par les cercles bleu sur la Figure 3.21 et ce peu importe la valeur de perplexité utilisée pour les t-SNE. Cette continuité, dont la cause n'est pas identifiée, fait que l'algorithme DBSCAN n'arrive pas à détecter (à cet endroit) la limite entre les vrais variants et les artefacts.

5 Conclusion

`MADaM` présente donc de bons résultats sur les jeux de données pour lesquels il a été développé (séquençages des exons 2 de HLA-DRB1/3 présentés dans ce chapitre et les séquençages des exons 2 de HLA-DQA1, -DQB1 et -DPB1 du Chapitre 4) ainsi que sur un jeu de données externe (voir page 154). L'utilisation de plusieurs variables descriptives et l'absence de seuils pré-determinés permettent de dépasser une partie des limitations des autres méthodes.
Les bons résultats obtenus ici pour `MADaM` valident donc la méthode employée. Cet algorithme a alors été utilisé pour traiter les résultats des séquençages Roche-454 présentés au chapitre suivant (rappelons que c'était la raison pour laquelle `MADaM` a été développé).

Il est maintenant nécessaire de tester `MADaM` avec d'autres jeux de données issus de technologies de séquençage plus récentes (Illumina, Ion Torrent, PacBio, Nanopore) afin de valider la méthode et justifier un usage courant pour les études de génétique des populations.
Les différentes méthodes spécifiquement développées pour `MADaM` (filtre markovien et l'ensemble des variables descriptives des séquences) pourront trouver d'autres applications dans le domaine de la bio-informatique. C'est le cas pour la décomposition en chaînes de Markov des séquences nucléotidiques qui servent de base à l'une des analyses du Chapitre 5.

Chapitre 4

Analyse fine de la diversité moléculaire des exons 2 des gènes HLA de classe II des populations du Sahel en Afrique

1 Introduction

L'Afrique est habitée par 1,2 milliards d'individus parlant plus de 2'000 langues différentes, soit un tiers de toutes les langues parlées sur la planète [Eberhard et al., 2019]. Ces 1,2 milliards d'humains habitent un continent de plus de 30 millions de km^2 présentant un environnement très varié : climat méditerranéen sur les zones côtières d'Afrique du nord et du sud, environnement tropical le long de l'équateur, de part et d'autre de cette bande équatoriale s'étendent des zones de savane, tels que la bande du Sahel, mais aussi déserts chauds, notamment au sud le désert du Kalahari et de Namibie et au nord le désert du Sahara, plus grand désert chaud au monde et limite septentrionale du Sahel. L'Afrique présente aussi une diversité en termes d'altitude de ces écosystèmes puisqu'à part le mont Cameroun et le massif de l'Atlas en Afrique du nord, toutes les chaînes de montagnes et plateaux d'altitude se situent à l'est d'une ligne allant de l'Éthiopie au massif du Drakensberg en Afrique du Sud.

La diversité génétique en Afrique, aussi bien entre les populations qu'à l'intérieur des populations, est plus élevée que dans n'importe quelle autre région du monde [Cann et al., 1987, Reed and Tishkoff, 2006, Tishkoff et al., 2009]. De récentes analyses sur l'ensemble du génome ont mis en évidence la présence de 15 à 24% de variants en plus dans les génomes africains que non-africains [The 1000 Genomes Project Consortium, 2015, Gurdasani et al., 2015].

Cette importante diversité génétique est le fruit tout d'abord d'une présence humaine plus longue que dans n'importe quelle autre région du monde. Les restes fossiles les plus anciens d'*Homo sapiens* proviennent d'Afrique et sont datés de -200'000 à -300'000 ans [McDougall et al., 2005, Hublin et al., 2017]. En comparaison, les populations non-africaines sont issues de l'évènement appelé « sortie d'Afrique », ayant pris place il y a 100'000 à 150'000 ans [Scally and Durbin, 2012, López et al., 2015] et représentant un sous-ensemble de la diversité africaine de l'époque puisque la taille efficace du

groupe ayant quitté l'Afrique à cette époque est estimée entre 1'000 et 1'500 individus [Liu et al., 2006, Garrigan et al., 2007].

La diversité génétique africaine résulte aussi d'une structuration très ancienne des populations [Choudhury et al., 2018], puisque des récentes études menées sur des génomes anciens sud-africains font remonter la première divergence entre les groupes d'humains archaïques il y a -260'000 à -350'000 ans [Schlebusch et al., 2017], tandis qu'une révision des taux de mutations du génome humain place la divergence entre les Khoi-sans et les autres groupes africains entre -250'000 à -300'000 ans [Scally and Durbin, 2012].

Cette diversité est aussi le résultat de flux géniques entre les populations africaines : une étude menée en 2018 sur des séquences du chromosome Y a montré d'importants flux géniques à travers le Sahara à l'époque du Sahara vert [1], accompagnés d'expansions démographiques des populations de cette région [D'Atanasio et al., 2018].

Mais cette diversité est aussi le résultat de la sélection naturelle, notamment liée à la malaria (principalement due, en Afrique, à *Plasmodium falciparum* et *Plasmodium vivax*), dont plusieurs mécanismes génétiques ont été identifiés, entre autres, des résistances conférées par la mutation *HbS* à l'état hétérozygote [Allison, 1954, Pasvol et al., 1978, Aidoo et al., 2002], des mutations du gène *DARC* (certains allèles, tel que le SNP rs2814778, ayant été fixées [2] dans plusieurs populations d'Afrique de l'ouest et de l'est) [Choudhury et al., 2018], des CNV [3] et ré-arrangements structuraux des gènes GYPA, GYPB et GYPC [Leffler et al., 2017], ou la mise en évidence d'un lien entre certains allèles des gènes HLA de classe I (HLA-B*53) et la prévalence de la malaria [Hill et al., 1991, Hill et al., 1992b, Sanchez-Mazas et al., 2017].

Malgré cette importante diversité génétique, cette dernière reste majoritairement inexplorée [Campbell and Tishkoff, 2008, Martin et al., 2018, Erlich et al., 2018], d'autant plus pour les loci HLA (principalement explorés pour les gènes de classe I), la majeure partie des études portant jusqu'alors sur des populations d'Afrique du nord (*e.g.* [Hajjej et al., 2015, Messoussi et al., 2019], ou d'Afrique du sud (*e.g.* [Lombard et al., 2006, Thorstenson et al., 2018]).

1.1 Buts de l'étude

Cette étude cherche à apporter des éléments de réponse aux questions relatives, 1) aux processus sélectifs qui agissent sur les exons 2 des quatres gènes étudiés, indépendament des populations, et 2) aux processus évolutifs (c'est-à-dire les forces sélectives et les forces démographiques) qui agissent sur les populations.

Pour répondre à ces questions, cette étude analysera les résultats des estimations de diversité allélique (distributions de fréquences alléliques, hétérozygotie, nombres d'allèles et richesse allélique) et moléculaire (diversité nucléotidique π et nombre de sites variables S).

1. Le Sahara vert est une époque s'étendant de -12'600 à -3'000 avant notre ère et caractérisée par un climat plus humide sur la zone de l'actuel Sahara et une végétation de type savane herbeuse avec présence d'arbres et de buissons [Brooks et al., 2005].

2. La fréquence de cet allèle a atteint 100%

3. *Copy Number Variation* : Variations du nombre de copies.

L'existence ou non et, le cas échéant, la nature des pressions de sélection naturelle seront analysées à l'aide des tests de neutralité sélective d'Ewens-Watterson-Slatkin (basé sur les distributions de fréquences alléliques) et du D de Tajima (basé sur les données moléculaires).
La relation entre les distances géographiques et les distances génétiques évolutives de Reynolds [Reynolds et al., 1983] entre populations sera quant à elle analysée par un test de Mantel [Mantel, 1967].
La distribution de la variance moléculaire entre et à l'intérieur des populations, selon des facteurs géographiques, linguistiques ou de style de vie (nomade, semi-nomade ou sédentaire) sera étudiée à l'aide de l'analyse AMOVA [Excoffier et al., 1992] [4].
Les relations entre les populations seront étudiées à l'aide d'analyses d'échelonnement multi-dimensionnel (MDS), basées sur les distances génétiques entre les populations (Θ_w) estimées à chacun des loci, mais aussi sur une analyse factorielle des correspondances [Benzécri, 1973, Greenacre, 1984] réalisée sur l'ensemble des fréquences alléliques estimées à chaque locus pour chacune des populations.
Finalement, cette étude analysera la relation entre la prévalence de la malaria due à *Plasmodium falciparum* et les distributions de fréquences alléliques pour les populations africaines, afin d'identifier des allèles potentiellement sous sélection naturelle liée à ce pathogène.

Cette étude porte sur l'analyse de la diversité moléculaire des exons 2 de quatre gènes de classe II (HLA-DRB1, -DQA1, -DQB1 et -DPB1) au sein d'échantillons provenant de 42 populations (voir la carte en Figure 4.3) de la ceinture du Sahel (Afrique de l'ouest, centrale et de l'est) et d'Afrique du nord, mais aussi, à titre de comparaison, des populations non-africaines de continents voisins, provenant de Syrie (Asie de l'ouest) et de Slovaquie (Europe centrale).
Les échantillons ont été obtenus à l'aide de collaborations internationales, dans le cadre du projet FNRS #31003A-144180 visant à étudier le rôle de l'histoire du peuplement humain et des facteurs environnementaux sur l'évolution et la diversité du système HLA. Les collaborateurs internationaux de ce projet sont :

— Hacene Brouk : Université Badji Mokhtar, Annaba, Algérie ;
— Viktor Černý : Institut d'Archéologie, Prague, République tchèque ;
— Eric Crubézy : Université Paul Sabatier Toulouse III, Toulouse, France ;
— Jacques Chiaroni : Établissement français du sang, Marseille, France ;
— Jean-Michel Dugoujon : Université Paul Sabatier Toulouse III, Toulouse, France.

Le séquençage des exons 2 (et non des gènes complets) a été choisi pour permettre de typer un grand nombre d'individus, tout en ciblant les régions fonctionnelles impliquées dans la présentation des peptides antigéniques (voir Chapitre 1) et parce qu'ils présentent la majeure partie de la variabilité génétique des gènes de classe II (voir Chapitre 5). L'utilisation de typages basés sur le séquençage de l'ADN permet d'obtenir une meilleure résolution (à l'échelle nucléotidique) que les méthodes traditionnelles telles que la PCR-SSO.

4. *Analyse of MOlecular VAriance* : Analyse de variance moléculaire.

2 Matériel et Méthodes

2.1 Echantillons de populations

La Table 4.1 fournit, pour chaque population de l'étude, son nom et son collecteur, sa localisation géographique (région et coordonnées du lieu d'échantillonnage), des informations linguistiques (langue parlée et famille linguistique de rattachement) et sa taille d'échantillon à chacun des quatre loci.

Pour définir les régions d'où proviennent les populations, la définition des régions géographiques de l'Organisation des Nations Unies a été utilisée (Figure 4.1), en utilisant une version francisée des noms de régions de [Nunes et al., 2014]. Les populations du Soudan (Arabes Rashaida, Nubiens, Beja Hadendoa et Arabes soudanais) ont été ainsi placées dans la région Afrique du nord malgré leur plus grande proximité géographique avec l'Afrique de l'est.

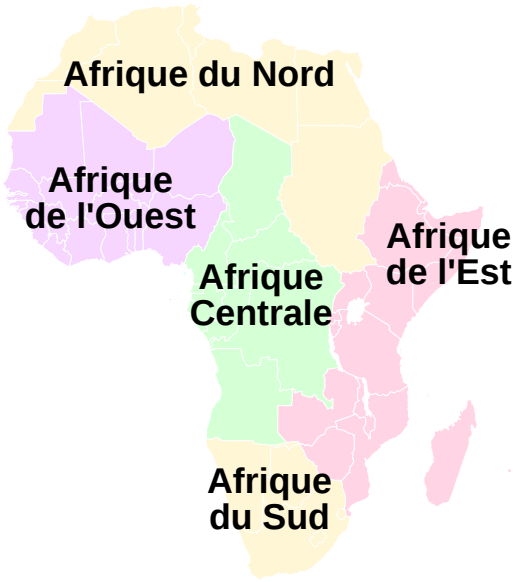


FIGURE 4.1 – Carte des régions géographiques africaines telles que définies par les Nations Unies. Adapté de `https://en.wikipedia.org/wiki/File:United_Nations_geographical_subregions.png`, Crédits : T. Seppelt, CC BY-SA.

À l'exception des Mandenkalu, les ADN de toutes les populations ont été extraits dans les laboratoires collaborant à l'étude avant d'être envoyés à Genève pour séquençage. Les échantillonnages des Mandenkalu ont été réalisés dans le cadre d'une campagne menée par le Département d'Anthropologie de l'Université de Genève en 1990. L'ADN des Mandenkalu a été extrait à Genève par le LGB [Tiercy et al., 1992] (Laboratoire de Génétique et Biométrie, maintenant AGP : Laboratoire d'Anthropologie, génétique et peuplements, Université de Genève) en collaboration avec le LNRH (Laboratoire National de Référence pour l'Histocompatibilité, Genève, Suisse).

		Géographie			Style de vie	Tailles d'échantillons			
Collecteur	Population	Région	Localisation	Langue		DRB1	DQA1	DQB1	DPB1
HB	Algérie-(Annaba)	AFR-N	36.8, 7.7	Arabe (AA)	S	108	113	109	111
HB	Algérie-(Constantine)	AFR-N	36.2, 6.6	Arabe (AA)	S	42	41	44	42
HB	Algérie-(ElOued)	AFR-N	33.3, 6.8	Arabe (AA)	S	2*	2*	2*	2*
HB	Algérie-(Ghardaia)	AFR-N	32.4, 3.6	Tamazight (AA)	S	83	80	82	82
HB	Algérie-(Oran)	AFR-N	35.7, 0.9	Arabe (AA)	S	6*	6*	6*	6*
HB	Algérie-(Tamanrasset)	AFR-N	22.7, 5.5	Tamasheq (AA)	N	28	32	30	32
HB	Algérie-(Tebessa)	AFR-N	35.4, 8.1	Arabe (AA)	S	14*	18*	17*	18*
HB	Algérie-(TiziOuzou)	AFR-N	36.7, 4.1	Tamazight (AA)	S	15*	18*	17*	18*
JMD	Maroc-Amazigh-(Amizmiz)	AFR-N	31.2, -8.2	Tamazight (AA)	SN	50	49	25	48
JMD	Maroc-Amazigh-(Asni)	AFR-N	31.2, -8.0	Tamazight (AA)	SN	48	50	0*	45
JMD	Maroc-Amazigh-(Figuig)	AFR-N	32.1, -1.2	Tamazight (AA)	SN	97	92	79	94
VC	Soudan-BejaHadendoa	AFR-N	15.6, 36.3	Beja (AA)	N	47	45	48	48
VC	Soudan-Nubien	AFR-N	20.8, 30.5	Nubien (NS)	S	49	52	51	54
VC	Soudan-ArabeRashaida	AFR-N	15.3, 36.2	Arabe (AA)	SN	39	50	45	48
VC	Sudan-ArabeSoudanais	AFR-N	18.4, 30.8	Arabe (AA)	S	37	46	46	45
VC	BurkinaFaso-Gourmantché	AFR-O	11.2, 0.7	Gourmantché (NC)	S	32	37	36	34
VC	BurkinaFaso-Gourounsi	AFR-O	11.2, -1.1	Gourounsi (NC)	S	32	33	33	33
VC	BurkinaFaso-Mossi	AFR-O	12.6, -1.3	Mossi (NC)	S	34	35	35	34
JC	Mali-Bambara	AFR-O	14.3, -3.6	Bambara (NC)	S	11*	12*	9*	11*
JC	Mali-Dogon	AFR-O	14.3, -3.6	Dogon (NC)	S	147	150	143	151
JC	Mali-Peul	AFR-O	14.3, -3.1	Fulfulde (NC)	N	67	51	18*	61
JC	Mali-Tamasheq	AFR-O	14.3, -3.6	Tamasheq (NC)	N	3*	3*	1*	3*
VC	Sénégal-Bedik	AFR-O	12.3, -12.2	Bedik (NC)	S	45	39	3*	48
VC	Sénégal-Peul	AFR-O	15.3, -15.1	Fulfulde (NC)	N	56	55	52	56
AGP	Sénégal-Mandenka	AFR-O	12.7, -12.3	Mandenka (NC)	S	199	197	195	196
VC	Sénégal-Sérère	AFR-O	14.1, -16.4	Sérère (NC)	S	47	47	48	47
VC	Tchad-ArabeBaggara	AFR-C	13.2, 18.0	Arabe (AA)	N	49	51	51	49
VC	Tchad-Dangaléat	AFR-C	12.2, 18.5	Dangaléat (AA)	S	49	49	49	50
VC	Tchad-Daza	AFR-C	18.2, 20.6	Daza (NS)	SN	39	40	41	40
VC	Tchad-Maba	AFR-C	13.9, 20.8	Maba (NS)	S	41	42	42	38
EC	Ethiopie-Amhara-(Keketeya)	AFR-E	11.2, 39.9	Amharique (AA)	S	48	82	66	73
EC	Ethiopie-Oromo	AFR-E	11.3, 39.6	Oromo (AA)	S	21	31	23	33
JC	Syrie-Alaouite	ASI-O	33.5, 36.3	Arabe (AA)	S	26	20	24	26
JC	Syrie-Druze	ASI-O	33.5, 36.3	Arabe (AA)	S	83	84	83	85
JC	Syrie-ChrétienMaronite	ASI-O	33.5, 36.3	Arabe (AA)	S	24	24	8*	23
JC	Syrie-ChrétienOrthodoxe	ASI-O	33.5, 36.3	Arabe (AA)	S	97	95	92	98
JC	Syrie-Sunnite	ASI-O	33.5, 36.3	Arabe (AA)	S	76	79	78	79
VC	Slovaquie-(Galanta)	EUR-C	48.2, 17.7	Slovaque (IE)	S	33	34	32	33
VC	Slovaquie-(Namestovo)	EUR-C	49.4, 19.5	Slovaque (IE)	S	33	36	36	36
VC	Slovaquie-(NovaBana)	EUR-C	48.4, 18.6	Slovaque (IE)	S	26	27	27	26
VC	Slovaquie-(Skalica)	EUR-C	48.8, 17.2	Slovaque (IE)	S	29	33	26	32
VC	Slovaquie-(StaraLubovna)	EUR-C	49.3, 20.7	Slovaque (IE)	S	25	30	31	31

TABLE 4.1 – Tableau récapitulatif des populations échantillonnées. Les collecteurs sont Hacene Brouk (HB), Jean-Michel Dugoujon (JMD), Viktor Černý (VC), Jacques Chiaroni (JC), Eric Crubézy (EC) et le laboratoire d'Anthropologie, génétique et peuplements de Genève (AGP). Voir l'introduction pour les affiliations. La colonne « Région » donne les régions géographiques telles que définies par l'ONU (voir Figure 4.1) tout en adoptant la nomenclature (francisée) de [Nunes et al., 2014] : Afrique du nord (AFR-N), Afrique de l'ouest (AFR-O), Afrique centrale (AFR-C), Afrique de l'est (AFR-E), Asie de l'ouest (ASI-O), Europe centrale (EUR-C). La localisation correspond à la latitude et la longitude des lieux d'échantillonnage, en unités décimales. Les styles de vie correspondent à : sédentaires (S), semi-nomades (SN), nomades (N). Les parenthèses pour les langues indiquent la famille linguistique : afro-asiatique (AA), nilo-saharien (NS), niger-congo (NC) et indo-européen (IE). Les quatre colonnes « Tailles d'échantillons » donnent le nombre d'individus disponibles pour les analyses (individus possédant un génotype non ambigu, hors-réplicats), les populations ne possédant pas une taille d'échantillon suffisante (marquée par une *) ont été retirées de l'analyse.

2.2 Présentation des populations

La Figure 4.3 illustre les lieux d'échantillonnage des différentes populations africaines de l'étude, ainsi que les familles linguistiques auxquelles appartiennent les langues parlées par les populations étudiées.

Linguistique

La Figure 4.2 donne les répartitions géographiques des locuteurs des quatre familles linguistiques parlées en Afrique (afro-asiatique, nilo-saharienne, niger-congo, khoisane, en plus des langues indo-européennes et austronésienne).

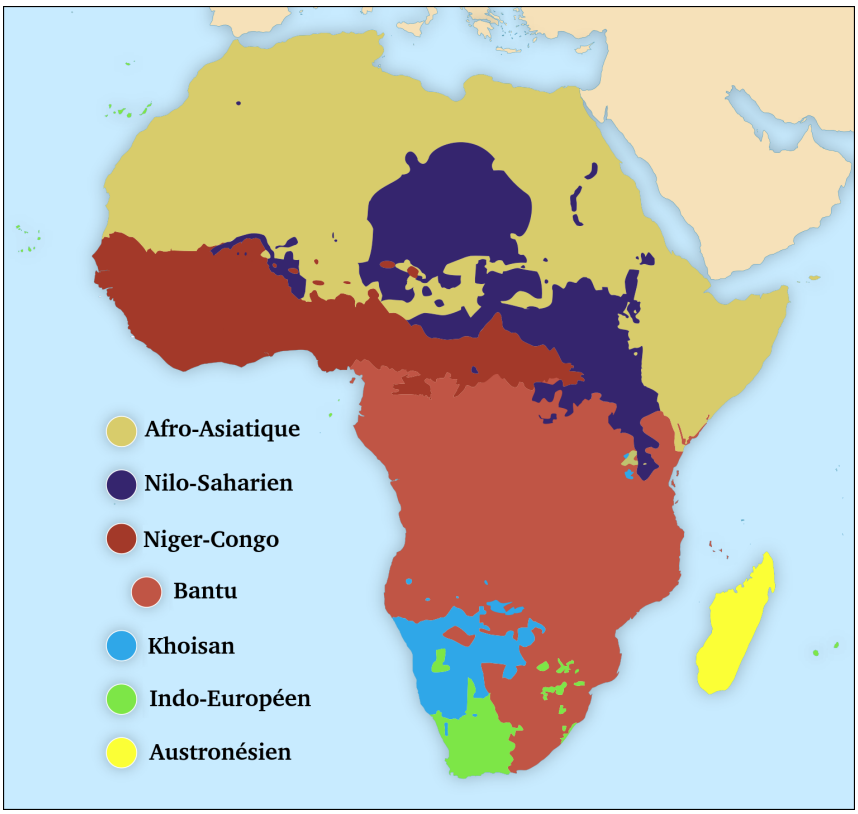


FIGURE 4.2 – Carte représentant les distributions géographiques des grandes familles linguistiques parlées en Afrique. Adapté de Wikimedia-Commons, licence CC-BY-SA 4.0

Les populations africaines de l'étude parlent des langues rattachées à trois grandes familles [Eberhard et al., 2019] :

— La famille **afro-asiatique**, présente dans toute la moitié nord de l'Afrique et regroupant 377 langues au sein de sept sous-familles : les langues berbères (27 langues), sémitiques (79 langues, incluant les langues arabes), cuchitiques (45 langues), tchadiques (193 langues), omotiques (31 langues), égyptienne (uniquement représentée par le copte) et l'ongota (une langue éthiopienne non rattachée aux autres familles) ;

— La famille **niger-congo**, regroupant 1'542 langues, principalement dans la sous-famille atlantique-congo (1'444 langues), mais aussi la sous-famille mandé (74 langues), ainsi que la sous-famille kordofanienne (23 langues) et le *mbre* (langue non rattachée aux précédentes sous-familles et parlée en Côte d'Ivoire) ;

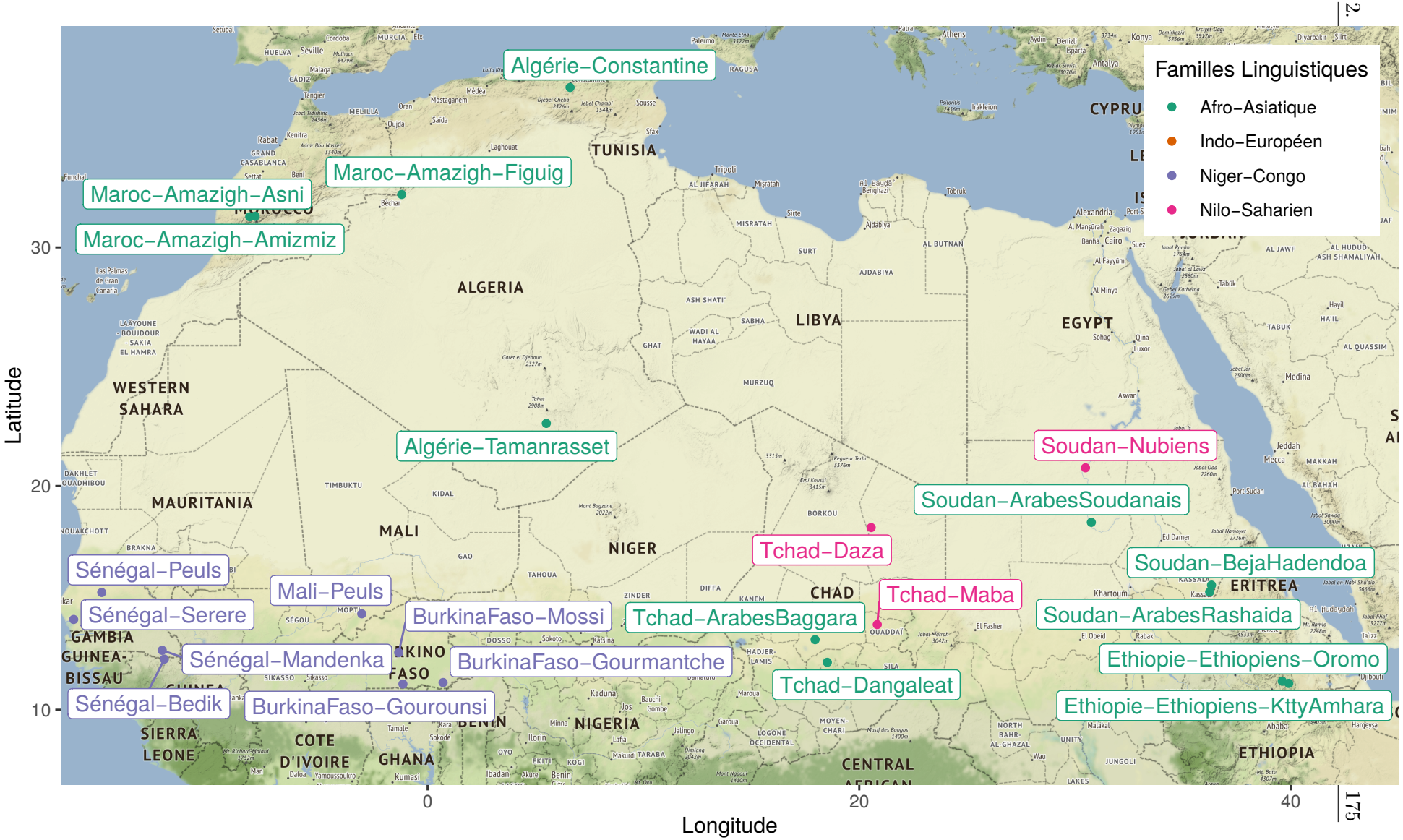


FIGURE 4.3 – Carte des lieux d'échantillonnage des populations africaines de l'étude, ainsi que des familles linguistiques auxquelles leurs langues sont rattachées. Fond de carte : Natural Earth.

— La famille **nilo-saharienne**, parlée principalement dans les régions centre et est de l'Afrique et regroupant 206 langues au sein de quatre sous-familles : les langues kuliak (3 langues), sahariennes (10 langues), songhai (9 langues) et celles regroupées dans la sous-famille « noyau et satellites » (184 langues).

La dernière famille de langues parlées en Afrique, la famille des langues **khoisanes**, n'est pas représentée dans cette étude. C'est une famille regroupant les langues dites « à clicks », qui est parlée principalement dans le sud de l'Afrique, ainsi que par les Sandawe dans l'est de l'Afrique [Tishkoff et al., 2009], bien que leur affiliation aux Khoisans du point de vue linguistique soit encore débattue [Güldemann and Stoneking, 2008].
Des langues de la famille **austronésienne**, d'origine sud-est-asiatique, sont aussi parlées à Madagascar et des langues **indo-européennes** peuvent être retrouvées, témoins du passé colonial de certains états européens.

Nous allons maintenant présenter les différentes populations étudiées dans ce chapitre. Les informations des sections suivantes proviennent principalement des travaux de Master de Élodie Chevallier et Maeva Pasquier qui ont, elles aussi, travaillé sur ces populations [Chevallier, 2015, Pasquier, 2016].

Afrique du nord

Les populations d'Afrique du nord représentent 15 populations provenant de trois pays différents.

Trois populations sont marocaines et appartiennent à l'ethnie Imazighen (pluriel d'Amazigh), plus connue sous l'exonyme [5] « berbères ». Ce sont des populations traditionnellement nomades, parlant des langues faisant partie de la famille afro-asiatique, le *tamazight*. Ces trois populations ont été échantillonnées en trois localités : Amizmiz et Asni dans le Haut-Atlas et Figuig, proche de la frontière avec l'Algérie.

Huit populations sont algériennes. Deux de ces populations sont aussi des Imazighen, les Mozabites de Ghardaia et les Kabyles de Tizi Ouzou. Trois populations arabes ont été échantillonnées dans les villes d'Annaba, de Constantine et de Tebessa. Elles parlent des langues *arabes*, de la famille afro-asiatique.
Les populations algériennes comprennent aussi des Tamasheq (aussi connus par l'exonyme « Touaregs »), provenant de la ville de Tamanrasset dans le sud de l'Algérie, en bordure du Sahara. Les Tamasheq parlent le *tamasheq*, une langue berbère de la famille afro-asiatique.

Les quatre dernières populations nord-africaines sont des populations soudanaises dont deux sont des populations nomades ou semi-nomades.
Les Nubiens sont une population d'effectif réduit au Soudan (autour de 200'000 individus [Fogel, 2003]), principalement retrouvée au nord du Soudan et au sud de l'Égypte. Cette population s'est installée il y a 3'000 ans dans la vallée du Nil et son histoire est très liée à celle de l'Égypte (les « pharaons noirs » de l'Égypte entre -1'000 et -655 étaient des Nubiens). Les Nubiens parlent une langue de la famille nilo-saharienne, le *nubien*.
Les Arabes soudanais sont une population arrivée au Soudan au 12ème siècle depuis la

5. Un exonyme est un nom donné à un groupe de personnes par un autre groupe de personnes. Le choix a été fait dans cette étude d'utiliser, autant que possible, les noms par lesquels ces populations se désignent elles-mêmes (autonymes).

péninsule arabique. Ils ont été échantillonnés dans la ville de Karima, au nord du Soudan et parlent une langue *arabe* faisant partie de la famille afro-asiatique.
Les Arabes Rashaayda sont des Bédouins, arrivés vers la fin du 19ème siècle depuis l'Arabie Saoudite. C'est un peuple de pasteurs nomades, parlant une langue *arabe* (famille afro-asiatique).
Finalement, la dernière population appartient au groupe des Beja, les Beja Hadendoa. Les Beja sont un ensemble de populations semi-nomades vivant majoritairement au Soudan et en Érythrée et parlant le *beja* (ou *bedawiyet*), une langue de la famille afro-asiatique.

Afrique de l'ouest

L'Afrique de l'ouest est une région représentée dans cette étude par 11 populations provenant de quatre pays différents.

Au Mali, les populations étudiées sont : les Dogon, sédentaires pratiquant l'agriculture et la métallurgie et parlant des langues *dogon* rattachées à la famille niger-congo (sous-famille atlantique-congo) ; les Bambara, appartenant au peuple mandingue, qui parlent le *bambara*, une langue de la famille niger-congo (sous-famille mandé) ; les Tamasheq, aussi échantillonnés au Mali (voir la population algérienne de Tamanrasset) ; les Peuls, éleveurs nomades et parlant le *fulfulde*, une langue niger-congo (sous-famille atlantique-congo) échantillonnés au Mali [Ba et al., 2015], mais aussi au Sénégal.

Au Sénégal, en plus des Peuls, trois populations locutrices niger-congo ont été échantillonnées : les Mandenkalu, grand groupe d'Afrique de l'ouest et étudié en détails dans le Chapitre 2, parlant le *mandenka* (sous-famille mande) ; les Sérère, agriculteurs et pêcheurs parlant le *sérère* (sous-famille atlantique-congo) ; et les Bédik, très petite population (2'500 individus recensés en 2015) parlant le *bédik* (sous-famille atlantique-congo).

Finalement, trois populations burkinabé ont été échantillonnées, toutes aussi locutrices d'une langue niger-congo : les Gourounsi, agriculteurs sédentaires parlant le *gourounsi* (sous-famille atlantique-congo, aussi connu sous le nom *lyélé*) ; les Mossis (langage : *mossi*, sous famille atlantique-congo), ethnie majoritaire du Burkina Faso et agriculteurs sédentaires ; les Gourmantché (langue : *gourmantché*, sous-famille atlantique-congo), actuellement agriculteurs sédentaires.

Afrique centrale

La région d'Afrique centrale comprend quatre populations originaires du Tchad : les Dangaléat, agriculteurs parlant une langue (le *dangaléat*) de la famille afro-asiatique ; les Mabas, éleveurs et locuteurs d'une langue nilo-saharienne (langue *maba*, aussi appelée *waddayen*) ; les Daza, éleveurs semi-nomades et parlant le *daza*, une langue nilo-saharienne ; les Arabes Baggara, pasteurs nomades parlant l'arabe tchadien, une langue afro-asiatique.

Afrique de l'est

L'Afrique de l'est est représentée par deux populations, les Oromo et les Amhara. Les Oromo constituent la population majoritaire d'Ethiopie et sont soit des agriculteurs

sédentaires soit des pasteurs nomades parlant l'*oromo*, une macro-langue [6] de la famille afro-asiatique regroupant plusieurs dialectes. Les Amhara, la deuxième population (en nombre d'individus) d'Éthiopie, sont majoritairement des agriculteurs sédentaires et parlent l'*amharique* (famille afro-asiatique).

Populations non africaines

Des populations provenant de trois pays hors d'Afrique (Syrie, Slovaquie et Pakistan) ont été utilisées, afin de servir de populations de références pour ces régions géographiques.
Les cinq populations syriennes ont été définies principalement sur la base de leurs religions : Chrétiens Maronites, Chrétiens Orthodoxes, Sunnites, Alaouites et Druzes. Toutes parlent une langue arabe (afro-asiatique).
Les populations slovaques ont été nommées en fonction de la localité d'échantillonnage (Stara Lubovna, Nova Bana, Namestovo, Skalica et Galanta), et toutes parlent le *slovaque*, une langue de la famille indo-européenne.
Cinq populations pakistanaises (Asie du sud) ont aussi été échantillonnées (Brahui, Burushaski, Hazara, Parsi et Sindhi) par Qasim Mehdi (Centre for Human Genetics, Sindh Institute of Urology and Transplantation, Karachi, Pakistan). Ces populations ont été séquencées en même temps que les autres populations de l'étude, mais les génotypes obtenus n'ont pas été utilisés car ils étaient incertains à cause d'une importante dégradation de l'ADN des échantillons (communication personnelle de Lydie Brunet [7], ayant réalisé les librairies de séquençage ADN).

2.3 Génotypage par séquençage ADN

La préparation de la librairie de séquençage a été réalisée par Lydie Brunet en suivant la méthode décrite par [Galan et al., 2010]. Après marquage par des tags oligonucléotidiques et ajout des adaptateurs pour l'émulsion-PCR, tous les amplicons d'un même locus ont été multiplexés et séquencés en 3 séries [8] (pour chaque locus) par un pyro-séquençeur 454 GS-FLX, en utilisant les réactifs « Lib-L Titanium Series ».
Les exons 2 de quatre gènes de classe II (HLA-DRB1, -DQA1, -DQB1, -DPB1) ont été ainsi séquencés. La Table 4.2 donne le nombre d'individus séquencés pour chacun des quatre loci (incluant de 306 à 504 individus répliqués afin de s'assurer de la reproductibilité de la méthode), en plus des 12 contrôles H_20 pour chaque locus :

Locus	**DRB1-Exon2**	**DQA1-Exon2**	**DQB1-Exon2**	**DPB1-Exon2**
Nb. individus	3'458	3'446	3'457	3'458
Dont réplicats	306	317	504	314

TABLE 4.2 – Nombre d'individus séquencés à chaque locus et nombre de réplicats (inclus dans le nombre total d'individus). Les échantillons des populations pakistanaises (voir page 178) sont inclus.

6. Ensemble de langues regroupant ici l'oromo Borana-Arsi-Guji, l'oromo de l'est, l'orma et l'oromo du centre-ouest.
7. À l'époque collaboratrice au laboratoire d'Anthropologie, génétique et peuplements de l'Université de Genève, Suisse.
8. *runs*

Le séquençage a été réalisé par Beckman Coulter Genomics (*Genomic Services, Danvers, Massachuetts*).

2.4 Traitement des lectures de séquençage

Les lectures de séquençage ont premièrement été filtrées sur la base de leur PhredScore. Le PhredScore est relié à la probabilité d'une erreur de lecture sur une base par la formule suivante :

$$P = 10^{\frac{-Q}{10}}$$

où P est la valeur de PhredScore et Q la probabilité d'une erreur de lecture du nucléotide.

Le filtre sur le PhredScore a été réalisé à l'aide du logiciel Mothur [Schloss et al., 2009] et de la commande `trim.seqs`, en considérant comme valeur seuil un PhredScore moyen (sur l'ensemble de la lecture) de 30, soit une probabilité d'erreur de 10^{-3} par nucléotide.

Les résultats de séquençage ont ensuite étés traités à l'aide du logiciel `MADaM` (voir Chapitre 3), avec les paramètres suivants : seuil T_1 (nombre minimum de séquences par individu) de 50, une séquence de référence par locus (fournies par M. Galan, disponibles en annexe S-31) et une *eValue* minimale de 1e-10 pour le BLAST [Altschul et al., 1990], une taille minimale (respectivement maximale) des séquences de 200pb (respectivement 400 pb).

Les résultats de séquençage pour DQB1-Exon2 contenaient aussi des lectures de DQB2-Exon2, amplifié par les amorces PCR à cause de la grande similarité entre ces gènes (gènes paralogues [Lenormand et al., 2012]). Le filtre markovien implémenté dans `MADaM` a été utilisé afin de séparer les lectures correspondant à des séquences DQB1-Exon2 des lectures correspondant aux séquences de DQB2-Exon2. Ce filtre consistant en un classificateur à apprentissage supervisé, il a été entraîné sur cinq séquences de DQB2-Exon2 (les seules séquences disponibles au moment des analyses) et 931 séquences de DQB1-Exon2.
De même, les amplifications PCR pour DRB1-Exon2 ayant aussi amplifié les exons 2 de HLA-DRB3, -DRB4, -DRB5, -DRB6 et -DRB7 (gènes paralogues issus de duplications [Satta et al., 1996b]), le filtre markovien a aussi été utilisé, entrainé sur 2'038 séquences de DRB1-Exon2 et DRB3-Exon2, 50 séquences DRB4-Exon2, 43 séquences DRB5-Exon2, trois séquences DRB6-Exon2 et deux séquences DRB7-Exon2 [9].

Trois valeurs de perplexité ont été utilisées pour l'étape de t-SNE [van der Maaten and Hinton, 2008] : 30, 40 et 50 et, afin de circonvenir aux risques de minimum local, cinq t-SNE ont été réalisées à chaque fois.

Les séquences nucléotidiques des vrais variants identifiés par `MADaM` sont données en annexe S-41.

2.5 Nomenclature des allèles

Du fait que seuls les exons 2 de chacun des quatre gènes ont été séquencés, l'assignation d'une séquence spécifique d'un exon 2 à un allèle nominal du troisième champ

9. Téléchargées le 15/02/2019 depuis la base de données IPD-IMGT/HLA, v3.35.

(tel que défini par la nomenclature MHC/HLA [Marsh et al., 2010]) n'est que rarement possible. L'assignation à un allèle défini au premier champ est toutefois possible pour l'ensemble des séquences de DRB1-Exon2, DQA1-Exon2 et DQB1-Exon2, mais pas pour DPB1-Exon2. L'assignation à un allèle (unique) du second champ est beaucoup plus rare et varie selon les loci. En effet, pour DRB1-Exon2 et DPB1-Exon2 respectivement 11 et 12 séquences peuvent être attribuées à un allèle unique au second champ, contre une seule séquence pour DQA1-Exon2 et aucune séquence de DQB1-Exon2.
L'annexe S-44 donne, pour chaque séquence d'exon 2 de cette étude, les allèles nominaux possibles (jusqu'au 3ème champ). Certaines séquences, notamment celles de DQB1-Exon2, correspondent à un grand nombre d'allèles nominaux (la séquence DQB1*2902 est partagée par 85 allèles définis au troisième champ).

Une nomenclature spécifique à cette étude a été adoptée afin de distinguer les séquences des exons 2 des allèles HLA nominaux rapportés officiellement dans la banque de données de reference IMGT/HLA [Robinson et al., 2015].
Ainsi, lorsqu'il est question d'allèles HLA nominaux, leur nom est toujours précédé du préfixe **HLA-** (exemple : HLA-DRB1*13:04), tandis que les noms des séquences d'exons 2 sont définies sans ce préfixe (exemple : DRB1*3155).

2.6 Extraction des codons du site de reconnaissance de l'antigène

Localisés, pour les gènes de classe II, sur les exons 2, les codons ARS [10] des gènes HLA sont les codons qui codent pour le site de reconnaissance de l'antigène [Bjorkman et al., 1987a, Bjorkman et al., 1987b], liant les peptides antigéniques pour les présenter aux cellules T. Ces codons présentent une diversité bien plus importante que les autres codons de l'exon 2 [Parham, 1988].

Afin d'étudier ces codons de manière séparée, pour chaque locus trois types de séquences (appelés ici « jeux nucléotidiques ») ont été considérées : 1) soit la séquence de l'exon 2 au complet, 2) soit uniquement les codons codant pour le site de reconnaissance de l'antigène, 3) soit uniquement les codons ne codant pas pour le site de reconnaissance de l'antigène (non-ARS). Les positions de ces codons ARS correspondent à ceux indiqués dans la Figure 1.6 (page 17), rappelés dans le Tableau 4.3.

	Région	**Positions des codons ARS**
Classe II	Exon 2α	7, 9, 11, 22, 24, 31, 32, 43, 52, 53, 54 58, 59, 62, 65, 66, 69, 72, 73, 76
	Exon 2β	9, 11, 13, 14, 15, 26, 27, 28,30, 37, 38, 47, 56 57, 61, 67, 71, 74, 78, 79, 82, 85, 86, 89, 90

TABLE 4.3 – Positions des codons ARS pour les exons 2 des loci de classe II. Les positions proviennent de [Reche and Reinherz, 2003] (voir la Figure 1.6).

2.7 Analyses de génétique des populations

Le test d'équilibre de Hardy-Weinberg (incluant le coefficient de consanguinité, calculé par Gene[Rate] comme paramètre du modèle alternatif à l'équilibre de Hardy-

10. *Antigen Recognition Site* : Site de Reconnaissance de l'Antigène.

Weinberg), l'estimation des fréquences alléliques, l'estimation de l'hétérozygotie et le test d'Ewens-Watterson-Slatkin, ainsi que les estimations de déséquilibres de liaison (globaux et haplotypiques) ont été réalisés avec Gene[Rate] [Nunes et al., 2014] (`hla-net.eu`, voir page 37).

Le logiciel Arlequin [Excoffier and Lischer, 2010] a été utilisé pour les calculs de diversité moléculaire (π et S, page 41), les tests de sélection (test du D de Tajima, page 43) et les calculs des distances génétiques entre populations (coefficient de coancestralité de Reynolds Θ_w, page 46).

Les calculs de richesse allélique ont été réalisés à l'aide de la formule détaillée dans le Chapitre 1 (Introduction) (page 42) et implémentée dans un script Python3.

Analyse Factorielle des Correspondances

Une Analyse Factorielle des Correspondances [Benzécri, 1973, Greenacre, 1984] (AFC, voir page 48) a été réalisée sur les fréquences alléliques estimées pour chaque population. L'algorithme d'estimation des fréquences alléliques implémenté dans Gene[Rate], utilisant l'algorithme d'*Expectation Maximization* (voir page 38), considère comme pré-condition l'équilibre de Hardy-Weinberg. Ainsi, toute distribution de fréquences alléliques, estimée pour une population qui ne serait pas à l'équilibre de Hardy-Weinberg, n'est pas représentative de la population puisque la pré-condition n'est pas respectée. Seules les fréquences alléliques des 31 populations à l'équilibre de Hardy-Weinberg et présentant une taille d'échantillon d'au moins 20 individus ont été utilisées pour réaliser l'AFC.

L'analyse factorielle des correspondances a été réalisée à l'aide du logiciel R [R Core Team, 2020] et de la librairie `ade4` [Chessel et al., 2004, Dray et al., 2007, Dray and Dufour, 2007, Bougeard and Dray, 2018].

AMOVA

Afin d'étudier différents facteurs pouvant avoir un rôle sur la distribution de la variance moléculaire parmi les populations africaines, une analyse de variance moléculaire (AMOVA, page 47) a été réalisée [Excoffier et al., 1992].
Les effets de quatre facteurs ont été testés dans cette analyse :

- l'effet de la structuration géographique des populations, en utilisant les régions géographiques où vivent ces populations (voir Figure 4.1) ;
- la structuration liée à la linguistique, en utilisant les familles linguistiques des langues parlées par les populations (Table 4.1) ;
- L'effet du style de vie (nomade, semi-nomade ou sédentaire) (Table 4.1) ;
- La possible pression de sélection liée à la malaria, utilisant les données de prévalence de *Plasmodium falciparum* en l'an 2000 (voir page 182 pour l'obtention de ces données), et un seuil minimal de $pfpr2000$ [11] de 5% pour considérer une population comme exposée au parasite.

L'AMOVA a été réalisée avec le logiciel R [R Core Team, 2020] et la librairie `ade4` [Chessel et al., 2004, Dray et al., 2007, Dray and Dufour, 2007,

11. Prévalence de *Plasmodium falciparum* en l'an 2000.

Bougeard and Dray, 2018]. À des fins de comparaison entre les loci, seules les populations ayant des génotypes exploitables pour les quatre loci ont été utilisées : les populations d'Imazighen d'Asni, de Peuls du Mali, de Bédik du Sénégal et de Chrétiens Maronites de Syrie ne possédant pas de génotypes pour DQB1-Exon2, elles n'ont pas été intégrées à l'étude d'AMOVA.

Une analyse d'AMOVA a aussi été réalisée de la même manière en considérant l'ensemble des populations comme faisant partie du même groupe afin de calculer un ϕ_{st} global comme indice de fixation.

Test de Mantel

Afin d'étudier la relation entre les distances géographiques et génétiques entre populations, un test de Mantel a été réalisé [Mantel, 1967] (voir page 48).
Les distances géographiques ont été calculées à l'aide de la librairie `geosphere` [Hijmans, 2019a] pour R et des localisations d'échantillonnage des populations (voir Table 4.1).
Les relations génétiques entre les populations ont été estimées à l'aide de la distance de Reynolds Θ_w [Reynolds et al., 1983] calculée avec Arlequin [Excoffier and Lischer, 2010]. Afin de tenir compte de la fonction immunitaire des molécules codées par les loci étudiés, les Θ_w ont été calculés en considérant 1) soit l'exon 2 au complet, 2) soit uniquement les codons codant pour le site de reconnaissance de l'antigène (ARS, [Reche and Reinherz, 2003]) et 3) soit uniquement les codons ne codant pas pour le site de reconnaissance de l'antigène (non-ARS).
Le test de Mantel a été réalisé avec la librairie `ade4` [Chessel et al., 2004, Dray et al., 2007, Dray and Dufour, 2007, Bougeard and Dray, 2018] pour R.

Étude de la corrélation entre les fréquences alléliques HLA et la prévalence d'un pathogène

Les exons 2 des gènes de classe II séquencés dans cette étude codent pour les domaines $\alpha 1$ (pour HLA-DQA1) ou $\beta 1$ (pour HLA-DRB1, -DQB1 et -DPB1) des molécules de classe II. Ces deux domaines forment la région de liaison au peptide sur les molécules HLA de classe II et sont donc potentiellement soumis à une sélection liée aux pathogènes de l'environnement. Cette étude porte sur l'analyse de la relation entre les fréquences des allèles et la prévalence de la malaria due au pathogène *Plasmodium falciparum*.

Les données de prévalence de la malaria liée à *Plasmodium falciparum* ont été téléchargées depuis le site du Malaria Atlas Project (`https://map.ox.ac.uk/explorer/#/`) [Bhatt et al., 2015].
L'échantillonnage des populations ayant été réalisé avant 2000, les données de prévalence de l'an 2000 (les plus anciennes disponibles) ont été utilisées. Les données se présentaient sous la forme de rasters géo-référencés et les valeurs de prévalence (appelée ci-après $pfpr2000$, pour *Plasmodium falciparum prevalence 2000*) ont été extraites à l'aide de la librairie `raster` [Hijmans, 2019b] pour R.

Comme expliqué page 181, l'algorithme d'estimation des fréquences alléliques implémenté dans Gene[Rate] nécessite comme pré-condition l'équilibre de Hardy-Weinberg

pour que les fréquences alléliques calculées soient représentatives des populations et comparables entre elles.
Les populations retenues pour l'analyse sont les populations africaines à l'équilibre de Hardy-Weinberg et ayant une taille minimale de 20 individus par échantillon. Cela représente 20 populations. La différence entre le nombre de variables explicatives (103 au total) et le nombre d'observations (20) est trop importante pour réaliser des modélisations en modèles linéaires (telles que réalisées dans [Sanchez-Mazas et al., 2017]) puisque cela risque de conduire à un sur-paramétrage [12] des modèles [13]. Le choix a été fait de tester uniquement les corrélations entre les fréquences alléliques et la prévalence du *P. falciparum*.

Afin de tenir compte d'une structure géographique dans les distributions de fréquences alléliques des populations, une correction de ces fréquences alléliques en fonction de la géographie a été appliquée. Pour ce faire, des modèles linéaires ont été réalisés pour expliquer la fréquence de chaque allèle dans les populations de l'étude comme une fonction de la distance géographique entre le lieu d'échantillonnage et la distance à Addis-Abeba [14]. De chacun de ces modèles ont été extraits les résidus (la part de chaque fréquence allélique non expliquée par la distance géographique) et c'est sur ces résidus que la suite des analyses a été réalisée.

Afin de différencier ces nouvelles variables [15] des fréquences alléliques, la nomenclature `locus.identifiant` (ex : DPB1.66) sera utilisée à la place du nom de l'allèle (`locus*identifiant`, ex : DPB1*66).

Pour chacune de ces variables, la corrélation entre la variable et la $pfpr2000$ a été calculée en utilisant trois indices de corrélation : le coefficient de Pearson, le ρ de Spearman ou le τ de Kendall (similaire au protocole utilisé dans [Sanchez-Mazas et al., 2017]).

12. *Overfitting*
13. Une règle empirique en modélisation étant de ne pas avoir plus d'une variable explicative pour 10 observations.
14. Addis-Abeba, capitale de l'Éthiopie, a été retenue comme point de référence car l'Afrique de l'est serait la zone d'origine des humains anatomiquement modernes dans le modèle de « l'Origine Africaine Récente » [Stringer and Andrews, 1988, Prugnolle et al., 2005a].
15. Qui ne sont plus des fréquences alléliques car corrigées pour la géographie.

3 Résultats

3.1 Traitement des résultats de séquençage

Génotypage automatisé des individus

Le filtre markovien pour DQB1-Exon2 a rejeté 130'419 lectures (soit 7.8% des lectures totales) identifiées comme séquences de DQB2-Exon2. Le filtre markovien appliqué aux lectures de DRB1-Exon2 a rejeté quant à lui 183'221 lectures (soit 11.4% des lectures totales). Ces dernières ont été identifiées comme étant des séquences de loci non ciblés par l'étude (exons 2 de HLA-DRB4, -DRB5, -DRB6 et -DRB7). Les séquences correspondant à DRB3-Exon2, indiscernables des séquences de DRB1-Exon2 par le filtre markovien, ont été retirées au cas par cas, manuellement, à l'aide d'un BLAST réalisé en local sur les données des exons 2 de HLA-DRB1 et -DRB3. Pour DQA1-Exon2 et DPB1-Exon2, le filtre markovien n'a pas été appliqué puisqu'aucun autre locus n'a été co-amplifié.

La Table 4.4 résume le nombre de lectures totales pour chaque locus (toutes séries confondues), le nombre de lectures assignées à des individus (sur la base des tags oligonucléotidiques), le nombre de variants (séquences uniques au sein d'un individu), le nombre d'allèles détectés par `MADaM` et le nombre d'individus possédant plus de 50 séquences (conformément au seuil T_1 fixé à 50) et pour lesquels un génotype a pu être déterminé.

Locus	**DRB1-Exon2**	**DQA1-Exon2**	**DQB1-Exon2**	**DPB1-Exon2**
Lectures totales	1'605'818	1'813'556	1'664'894	1'536'091
Lectures rejetées	712'393	238'750	628'441	489'565
(Filtre markovien)	183'221	0	130'419	0
(Autres filtres)	529'172	238'750	498'022	489'565
Lectures assignées	893'425	1'574'806	1'036'453	1'046'526
Nb. variants	318'065	551'581	336'172	383'339
Nb. allèles	75	61	27	52
Nb. ind. gen.	3'010	3'186	2'769	2'453

TABLE 4.4 – Pour chaque locus de l'étude, Lectures totales : nombre de lectures totales ; Lectures rejetées : nombre de lectures rejetées par les différents filtres (PhredScore, BLAST, taille, filtre markovien) ; (Filtre markovien) : nombre de séquences spécifiquement rejetées par le filtre markovien de `MADaM` ; (Autres filtres) : nombre de séquences rejetées par les autres filtres ; Lectures assignées : nombre de lectures assignées aux individus après les différents filtres ; Nb. variants : nombre de variants (séquences uniques au sein d'un individu) ; Nb. allèles : nombre d'allèles différents identifiés ; Nb. ind. gen. : nombre d'individus pour lesquels plus de 50 séquences étaient disponibles (seuil T_1, voir page 179) et un génotype identifié.

Reproductibilité de la méthode

La Table 4.5 donne le nombre de paires de réplicats pour lesquelles des résultats ont pu être obtenus, ainsi que le nombre de paires de réplicats concordant et ne concordant pas. Pour ces derniers, la raison de la discordance a été investiguée et ils ont été classés en trois catégories : 1) « Pré-informatique » regroupe l'ensemble des erreurs liées aux étapes avant le traitement bio-informatique, incluant entre autres les erreurs liées à la PCR lors

de la préparation de la librairie (*e.g.* vrais allèles non amplifiés), 2) « MADaM » regroupe les erreurs liées au traitement informatique (détection à tort d'un artefact comme vrai variant ou non détection d'un vrai variant) et 3) « Inconnue » regroupe les autres erreurs pour lesquelles la cause n'a pas pu être déterminée avec certitude.
Les erreurs classées dans « Pré-informatique » et « Inconnue » correspondent à l'ensemble des erreurs pour lesquelles MADaM n'aurait jamais pu identifier le ou les mêmes vrais variants chez les deux individus de la paire de réplicats.

Locus	**Paires de réplicats valides**	**Source des erreurs**		
		Pré-informatique	MADaM	Inconnue
DRB1-Exon2	255/267 (95.5%)	4	5	3
DQA1-Exon2	256/281 (91.1%)	2	20	3
DQB1-Exon2	347/386 (83.9%)	32	6	1
DPB1-Exon2	263/267 (98.5%)	2	2	0

TABLE 4.5 – Résultats des analyses des réplicats à chacun des quatre loci de l'étude. Paires de réplicats valides : nombre de paires de réplicats exploitables (pour lesquels des génotypes ont pu être obtenus pour les deux réplicats du même individu) présentant le même génotype (concordance) ainsi que le nombre de paires de réplicats exploitables total (incluant les discordances de génotypes) ; Source des erreurs : 1) « Pré-informatique » regroupe l'ensemble des erreurs liées aux étapes avant le traitement bio-informatique incluant entre autres les erreurs liées à la PCR lors de la préparation de la librairie (*e.g.* vrais allèles non amplifiés), 2) « MADaM » regroupe les erreurs liées au traitement informatique (détection à tort d'un artefact comme vrai variant et non détection d'un vrai variant) et 3) « Inconnue » regroupe les autres erreurs pour lesquelles la cause n'a pas pu être déterminée avec certitude.

Les 20 erreurs imputables à MADaM pour DQA1-Exon2 sont principalement dues à un allèle (DQA1*7) sous-amplifiant fortement (jusqu'à un facteur 10 par rapport au premier allèle d'un individu) et qui n'a pas été détecté comme vrai variant pour au moins un des deux individus de ces réplicats (voir le Chapitre 3, page 161 pour la discussion de ce problème). La cause de ce problème étant connue, l'ensemble des génotypes individuels ont été vérifiés manuellement afin de s'assurer de ne pas rejeter à tort cette séquence chez les individus qui en sont porteurs.

L'erreur majoritaire pour DQB1-Exon2 lors du traitement informatique est liée à la présence résiduelle de variants provenant de DQB2-Exon2 qui n'ont pas étés filtrés lors du filtre markovien. La cause de cette faiblesse du filtre est probablement due au petit nombre de séquences DQB2-Exon2 utilisées (et disponibles) pour entraîner le filtre.
Pour éviter de considérer (à tort) des artefacts comme des vrais variants, seuls les allèles présents chez au moins deux individus (hors réplicats) ont été conservés. Les allèles singletons ont été considérés comme des faux positifs (artefacts détectés à tort comme vrais variants).

Tailles d'échantillons

La taille d'un échantillon de population est un facteur influençant la détection d'allèles rares, ces derniers risquant de ne pas être échantillonnés si la taille d'échantillon est trop petite.

Le choix d'une taille minimale d'échantillon doit alors être un compromis entre la fréquence (dans la population) des allèles que l'on veut détecter et ce qui est réalisable en termes d'échantillonnage et de séquençage.
La Table 4.1 (page 173) donne pour chaque locus le nombre d'individus (taille d'échantillon) pour lesquels des typages ont pu être obtenus à chaque locus pour cette étude. Les tailles d'échantillons de populations (tous loci confondus) sont en moyenne de 49 individus, mais varient fortement (de 2 à 199) puisque l'écart-type est de 38. Si on utilise, par exemple, un seuil (arbitraire) de 100 individus minimum, sur les 42 populations de l'étude, seules trois seraient retenues. Basé sur les tailles d'échantillons disponibles et le polymorphisme de ces loci, une taille minimale de 20 individus a finalement été retenue, ce qui permet de détecter les allèles de fréquences supérieures ou égales à $1/2N = 1/40$, soit 2.5%.
Cela conduit à l'éviction de six populations de l'étude (aucun locus ne présentant plus de 20 individus pour chacune de ces populations) :

- Algérie-(ElOued) (2 individus) ;
- Algérie-(Oran) (6 individus) ;
- Algérie-(Tebessa) (14-18 individus) ;
- Algérie-(TiziOuzou) (15-18 individus) ;
- Mali-Bambara (9-12 individus) ;
- Mali-Tamasheq (1-3 individus).

De plus, quatre populations ne présentent pas assez d'individus testés au locus DQB1-Exon2 :

- Sénégal-Bédik (3 individus) ;
- Syrie-ChrétiensMaronites (8 individus) ;
- Mali-Peuls (18 individus) ;
- Maroc-Amazigh-(Asni) (aucun individu).

Ce locus ne sera donc pas traité pour ces quatre populations.

Pour les autres populations, les tailles d'échantillons de chaque locus sont toutes supérieures ou égales à 20 et ces populations sont donc retenues pour la suite des analyses.

3.2 Tests d'équilibre de Hardy-Weinberg & coefficients de consanguinité

Parmi les populations avec plus de 20 individus et afin de pouvoir calculer les fréquences alléliques et haplotypiques, l'équilibre de Hardy-Weinberg (HW) a été testé à chacun des loci pour chacune des populations. Ce test d'équilibre de Hardy-Weinberg réalisé avec Gene[Rate] [Nunes et al., 2014] se base sur une comparaison de deux modèles, le premier ayant comme hypothèse l'équilibre de Hardy-Weinberg et le second une déviation de cet équilibre (due, par exemple, à la consanguinité). Les coefficients de consanguinité estimés selon le second modèle ont été analysés et intégrés aux critères de décision.

Quatre loci étant testés pour chaque population, une correction de Holm-Bonferroni pour tests multiples est appliquée [Holm, 1979]. Le seuil du risque pour une population de rejeter à tort l'hypothèse de l'équilibre de Hardy-Weinberg devient $\alpha' = \frac{0.05}{4}$.

La Figure 4.4 montre, pour chaque population ayant au moins un locus rejetant l'équilibre de Hardy-Weinberg, la pValeur du test de Hardy-Weinberg (après correction) à ce locus. Dans toutes les autres situations, l'hypothèse d'équilibre de Hardy-Weinberg est conservée.

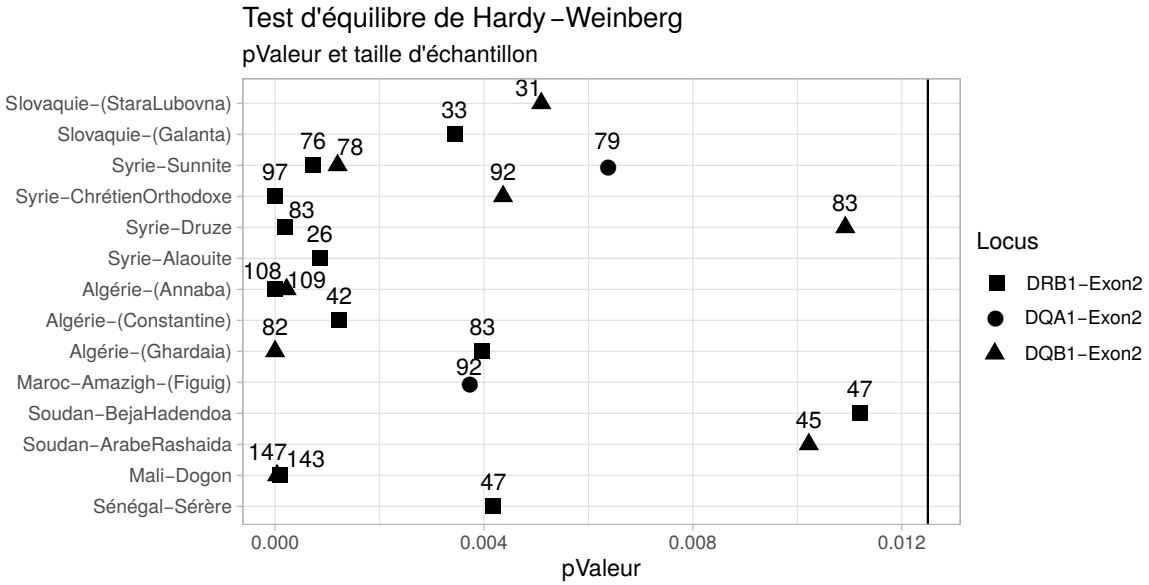


FIGURE 4.4 – Résultats significatifs du test d'équilibre de Hardy-Weinberg : pValeurs du test du Hardy-Weinberg pour chaque locus rejetant l'équilibre de Hardy-Weinberg au seuil $\alpha' = 0.0125$. Les nombres correspondent aux tailles d'échantillons de ces populations pour ces loci.

Aucune population ne rejette l'équilibre de Hardy-Weinberg au locus DPB1-Exon2, mais ce n'est pas le cas aux autres loci : le nombre de populations rejetant l'équilibre de Hardy-Weinberg aux autres loci est de 11 pour DRB1-Exon2, 2 pour DQA1-Exon2 et 8 pour DQB1-Exon2.

La Table 4.6 donne les coefficients de consanguinité moyens estimés ainsi que le nombre de populations rejetant l'équilibre de Hardy-Weinberg à chaque locus. Les valeurs numériques exactes sont disponibles en annexe S-46.

Locus	**DRB1-Exon2**	**DQA1-Exon2**	**DQB1-Exon2**	**DPB1-Exon2**
Nb. pop. testées	36	36	32	36
Rejets HWe	11	2	8	0
$\overline{CC}$	0.097 ± 0.075	0.043 ± 0.048	0.101 ± 0.074	0.031 ± 0.042

TABLE 4.6 – Résumé, pour chacun des quatre loci de l'étude, de : NB. pop. testées : nombre de populations pour lesquelles l'équilibre de Hardy-Weinberg a été testé ; Rejets HWe : nombre de populations avec plus de 20 individus rejetant l'hypothèse d'équilibre de Hardy-Weinberg au seuil $\alpha' = 0.0125$; $\overline{CC}$: Coefficient de consanguinité moyen ($\pm$ écart-type) observé à ce locus pour l'ensemble des populations.

On observe les coefficients de consanguinité moyens les plus élevés pour DQB1-Exon2 et DRB1-Exon2, mais aussi une forte dispersion de ces coefficients.

Les critères de décision retenus pour exclure une population de la suite de l'étude sur la base du test d'équilibre de Hardy-Weinberg et des coefficients de consanguinité sont :

1. trois loci en déséquilibre (DPB1-Exon2 ne rejetant jamais l'hypothèse d'équilibre de Hardy-Weinberg), **ou**
2. deux loci en déséquilibre, si, pour l'un des loci, le coefficient de consaguinité observé fait partie des 25% de coefficients les plus élevés observés à ce locus. Les seuils, pour qu'un coefficient de consanguinité observé soit parmi les 25% les plus élevés sont : DQA1-Exon2 : 0.073 ; DQB1-Exon2 : 0.149 ; DRB1-Exon2 : 0.155.

Ces deux critères amènent l'exclusion de cinq populations :

— Selon le **Critère 1 :**
  — Syrie-Sunnites, loci en déséquilibre de Hardy-Weinberg : DRB1-Exon2 (HW pValeur<0.001), DQA1-Exon2 (HW pValeur=0.006) et DQB1-Exon2 (HW pValeur=0.001) ;
— Selon le **Critère 2 :**
  — Algérie-(Annaba) (DRB1-Exon2 : HW pValeur<0.001 & CC=0.193, DQB1-Exon2 : HW pValeur=0.001 & CC=0.160) ;
  — Algérie-(Ghardaia) (DQB1-Exon2 : HW pValeur<0.001 & CC=0.299)
  — Mali-Dogon (DQB1-Exon2 : HW pValeur<0.001 & CC=0.153) ;
  — Syrie-ChrétiensOrthodoxes (DRB1-Exon2 : pValeur<0.001 & CC=0.170).

La Table 4.7 donne pour chaque locus le nombre de populations initialement disponibles pour l'étude, ainsi que le nombre de populations ayant une taille d'échantillons d'au moins 20 individus et, parmi ces dernières, le nombre de populations considérées à l'équilibre de Hardy-Weinberg.

Locus	**DRB1-Exon2**	**DQA1-Exon2**	**DQB1-Exon2**	**DPB1-Exon2**
Nb pop. totales	42	42	41	42
Nb pop $\geq$ 20 ind.	36	36	32	36
Pop éq. HW	31	31	27	31

TABLE 4.7 – Récapitulatif, pour chaque locus, du nombre de populations initialement disponibles pour l'étude, du nombre de populations pour lesquelles au moins 20 individus ont donné des génotypes exploitables et, parmi ces populations, du nombre de populations considérées à l'équilibre de Hardy-Weinberg.

3.3 Hétérozygotie

La Figure 4.5 montre les distributions des hétérozygoties estimées (H) à chaque locus en fonction de l'origine géographique, africaine ou non africaine (Europe et Syrie), des populations.

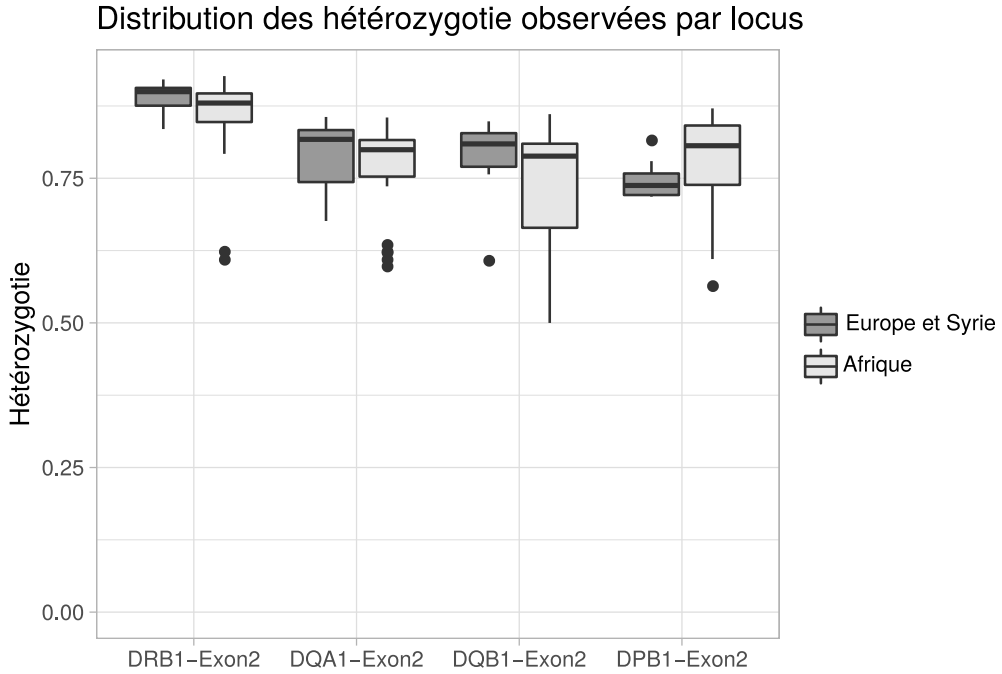


FIGURE 4.5 – Distribution des hétérozygoties estimées à chacun des loci, selon l'origine géographique (Afrique, ou Europe et Syrie) des populations.

Après correction pour tests multiples (méthode *fdr*, [Benjamini and Hochberg, 1995]), il n'y a pas de différences significatives d'hétérozygotie entre les populations africaines et celles d'Europe et de Syrie (pValeurs du test de Kruskall-Wallis, corrigées, pour DRB1-Exon2, DQA1-Exon2, DQB1-Exon2 et DPB1-Exon2 de respectivement 0.588, 0.588, 0.447 et 0.096). Il n'est toutefois pas possible de tester cette différence par région géographique à cause d'un trop petit nombre de populations pour certaines régions (par exemple il n'y a que 2 populations en Afrique de l'est). On observe une différence significative entre les loci (test de Kruskal-Wallis, pValeur = 3.771e-11), DRB1-Exon2 présentant une hétérozygotie supérieure (voir Table 4.8 et Figure 4.5).

Locus	**DRB1-Exon2**	**DQA1-Exon2**	**DQB1-Exon2**	**DPB1-Exon2**
Nb. pops.	31	31	27	31
$H \pm \sigma^2$	0.864 ± 0.069	0.773 ± 0.074	0.752 ± 0.098	0.774 ± 0.075

TABLE 4.8 – Moyenne et écart-type de l'hétérozygotie estimée ($H \pm \sigma^2$) et nombre de populations (Nb. pops.) pour chacun des quatre loci de l'étude.

La Figure 4.6 montre, pour chaque population et à chaque locus, les hétérozygoties estimées.

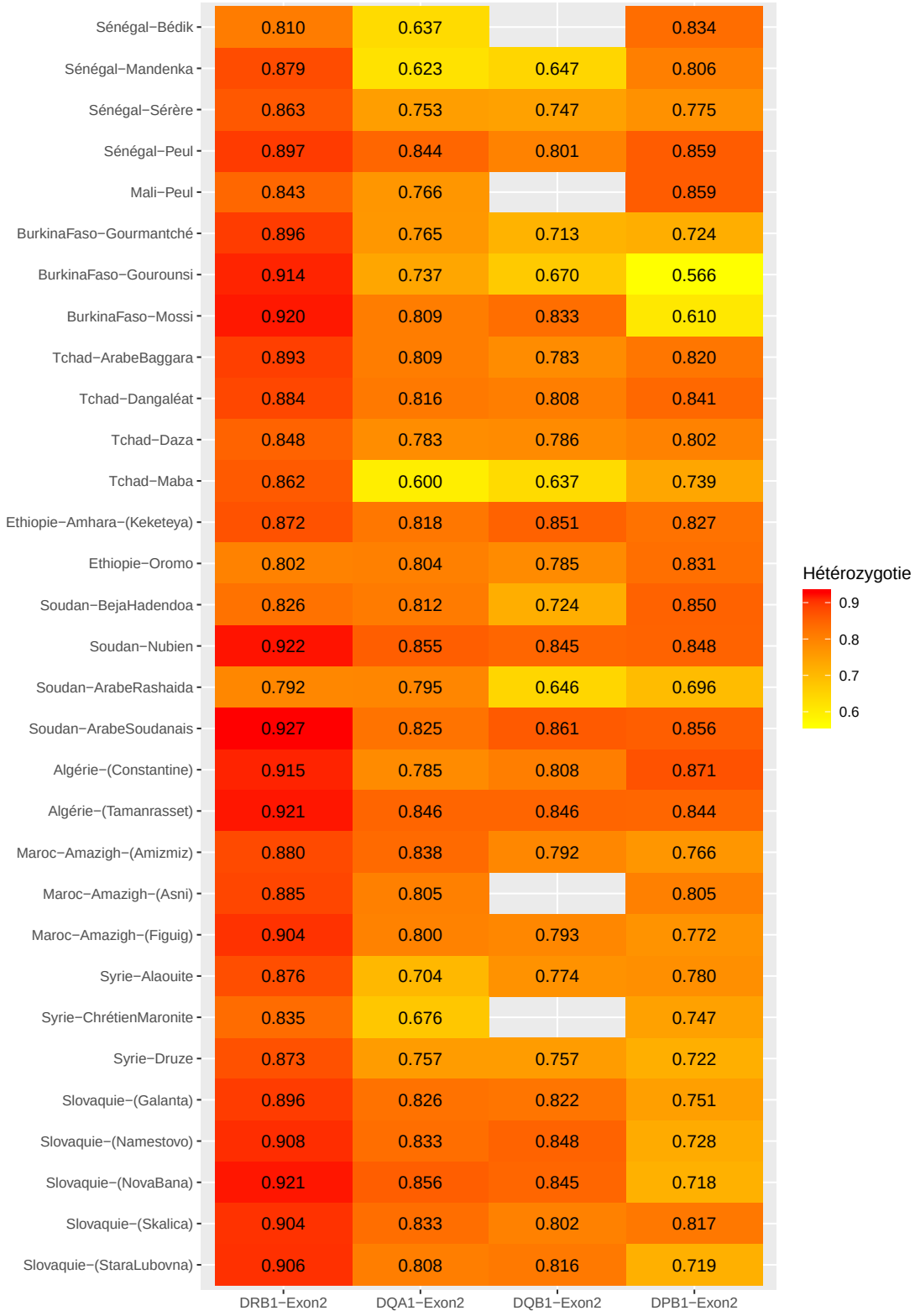


FIGURE 4.6 – Carte de chaleur des hétérozygoties estimées pour chacune des populations à chacun des loci. L'échelle de couleur indique le taux d'hétérozygotie et les valeurs numériques correspondent à l'hétérozygotie estimée. Les cases blanches correspondent aux quatre populations pour lesquelles les tailles d'échantillons au locus DQB1-Exon2 sont inférieures à 20 (voir page 185).

Quelques populations montrent une hétérozygotie particulièrement réduite (≤ 0.700, en jaune dans la Figure 4.6) à certains loci (majoritairement DQA1-Exon2 et DQB1-Exon2) :

— Les Mandenkalu du Sénégal, aux loci DQA1-Exon2 ($H = 0.623$, n=197) et DQB1-Exon2 ($H = 0.647$, n=195) ;
— Les Bedik du Sénégal, au locus DQA1-Exon2 ($H = 0.637$, n=39) ;
— Les Maba du Tchad, aux loci DQA1-Exon2 ($H = 0.600$, n=42) et DQB1-Exon2 ($H = 0.637$, n=42) ;
— Les Mossi du Burkina Faso, au locus DPB1-Exon2 ($H = 0.610$, n=34) ;
— Les Gourounsi du Burkina Faso, aux loci DQB1-Exon 2 ($H = 0.670$, n=33) et DPB1-Exon2 ($H = 0.566$, n=33) ;
— Les Arabes Rashaida du Soudan, aux loci DQB1-Exon2 ($H = 0.646$, n=45) et DPB1-Exon2 ($H = 0.696$, n=48) ;
— Les Chrétiens Maronites de Syrie, au locus DQA1-Exon2 ($H = 0.676$, n=24).

3.4 Richesse allélique

Plus la taille d'un échantillon est élevée, plus il y a de chances d'observer un allèle peu fréquent et donc plus le nombre d'allèles observés sera élevé. Il n'est donc pas judicieux de comparer le nombre d'allèles observés entre deux populations, à un même locus, si ces deux populations n'ont pas la même taille d'échantillon. Pour cela, il est préférable d'utiliser la richesse allélique, correspondant au nombre d'allèles attendus dans un sous-échantillon de taille plus petite, estimée selon la méthode de El Mousadik et Petit en 1996 [El Mousadik and Petit, 1996] reposant sur le principe de raréfaction défini par Hurlbert en 1971 [Hurlbert, 1971]. Ainsi la richesse allélique d'un échantillon de population ne peut jamais être supérieure au nombre d'allèles observés dans l'échantillon, elle peut être égale (si la taille du sous-échantillon est égale à la taille d'échantillon initiale) ou inférieure (si le sous-échantillon est de taille inférieure à l'échantillon).

La Figure 4.7 montre les distributions, à chaque locus, du nombre d'allèles détectés et de la richesse allélique calculée en considérant une taille d'échantillon correspondant à la plus petite taille d'échantillon observée à ce locus.

Les tailles d'échantillons les plus petites observées à chaque locus sont :

— DRB1-Exon2 : 21 individus (Oromo d'Éthiopie) ;
— DQA1-Exon2 : 24 individus (Chrétiens Maronites de Syrie) ;
— DQB1-Exon2 : 23 individus (Oromo d'Éthiopie) ;
— DPB1-Exon2 : 23 individus (Chrétiens Maronites de Syrie).

Les tests de corrélation (estimateur de Pearson) ne montrent pas de corrélation entre la taille d'échantillon et le nombre d'allèles détectés (pValeur=0.1413) ou la richesse allélique (pValeur=0.3573).

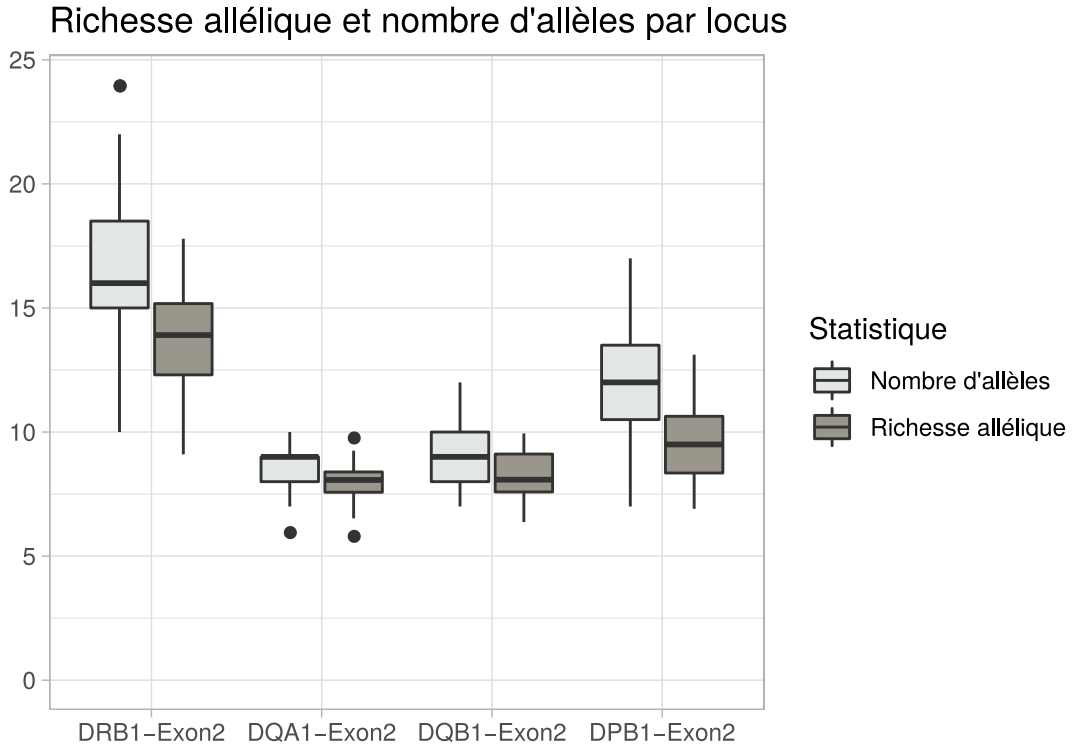


FIGURE 4.7 – Pour chaque locus, boîtes à moustaches représentant la distribution du nombre d'allèles détectés (en gris clair) et de la richesse allélique (gris foncé) calculée en considérant une taille d'échantillon correspondant à la plus petite taille d'échantillon observée à chaque locus DRB1-Exon2 : 21 (Ethiopie-Oromo), DQA1-Exon2 : 24 (Syrie-ChrétiensMaronites), DQB1-Exon2 : 23 (Ethiopie-Oromo), DPB1-Exon2 : 23 (Syrie-ChrétiensMaronites).

La Table 4.9 donne, pour chaque paire de loci, la pValeur (après correction pour tests multiples par méthode *fdr*) du test de Kruskal-Wallis comparant les richesse alléliques des deux loci.

Locus 1	**Locus 2**	**pValeur**
DRB1-Exon2	DQA1-Exon2	$1.44 \cdot 10^{-10}$
DRB1-Exon2	DQB1-Exon2	$4.24 \cdot 10^{-10}$
DRB1-Exon2	DPB1-Exon2	$1.63 \cdot 10^{-8}$
DQA1-Exon2	DQB1-Exon2	0.3376
DQA1-Exon2	DPB1-Exon2	$5.20 \cdot 10^{-5}$
DQB1-Exon2	DPB1-Exon2	$9.38 \cdot 10^{-4}$

TABLE 4.9 – Tableau récapitulant les différents tests de Kruskal-Wallis comparant les richesse alléliques entre chaque paire de loci. La colonne pValeur indique la pValeur du test de Kruskal-Wallis, après correction pour tests multiples par la méthode *fdr*.

Concernant les différences inter-locus, seuls les loci DQA1-Exon2 et DQB1-Exon2 ne sont pas significativement différents.

Ainsi, DRB1-Exon2 apparaît comme le locus le plus diversifié en termes de richesse

allélique (13.71 ± 2.25), suivi de DPB1-Exon2 (9.58 ± 1.56), puis de DQA1-Exon2 et DQB1-Exon2, qui présentent des richesses similaires (7.98 ± 0.84 et 8.25 ± 0.99).

3.5 Fréquences Alléliques

DRB1-Exon2

La Figure 4.8 montre les distributions de fréquences alléliques pour DRB1-Exon2 dans toutes les populations testées et à l'équilibre de Hardy-Weinberg (n=31). Les allèles n'excédant jamais 10% de fréquence dans l'ensemble des populations ont été regroupés dans la catégorie « Autres (<10%) ». Les valeurs numériques sont données en annexe S-43 et les correspondances entre les séquences d'exons 2 et les allèles nominaux HLA en annexe S-44.

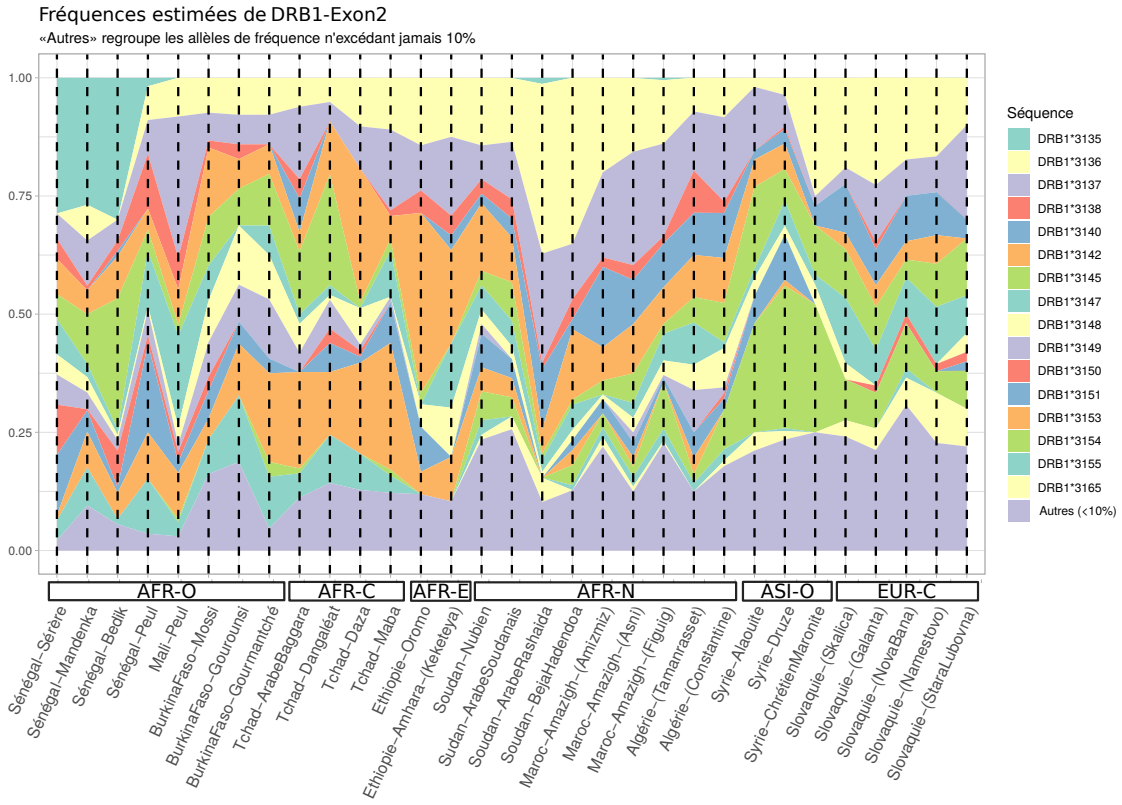


FIGURE 4.8 – Distribution des fréquences alléliques au locus DRB1-Exon2. Ne sont représentés que les allèles dont la fréquence observée dépasse 10% pour au moins une population, « Autres (<10%) » regroupant alors tous les allèles dont la fréquence n'excède jamais 10% dans ces populations. Les bandeaux au desssus des noms de populations indiquent la région géographique où vivent ces populations : AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

DRB1*3135 (correspondant à HLA-DRB1*13:04 en termes d'allèle HLA nominal) est observé à des fréquences élevées dans trois populations du Sénégal : les Bédik (30%), les Mandenkalu (27%) et les Sérère (29%). En dehors de ces trois populations, cette séquence n'est retrouvée qu'à une très basse fréquence (≤ 2%) chez les Imazighen de Figuig, les

Arabes Rashaida et les Peuls du Sénégal.

Les populations du Tchad sont caractérisées par une fréquence plus élevée de DRB1*3153 (correspondant essentiellement à HLA-DRB1*08) : les Arabes Baggara (20%), les Dangaléat (13%), les Maba (26%) et les Daza (19%).

Les Arabes Rashaida et les Beja Hadendoa montrent les plus hautes fréquences de DRB1*3136 (respectivement 36% et 35%) qui correspond essentiellement à HLA-DRB1*07. Cette séquence est aussi retrouvée à des fréquences élevées en Afrique du nord dans les populations Imazighen (de 13 à 20%), ainsi qu'en Europe centrale (jusqu'à 23% pour les Slovaques de Galanta). En Asie de l'ouest, seuls les Chrétiens Maronites de Syrie ont des fréquences élevées (25%) de cette séquence tandis que les Alaouites et les Druzes montrent des fréquences nettement plus basses (2 et 4%), ces trois populations ayant aussi les fréquences les plus élevées de DRB1*3154 (de 23 à 30%, correspondant à HLA-DRB1*11).

Les Peuls du Sénégal et du Mali montrent des différences notables de fréquences alléliques : DRB1*3151 (HLA-DRB1*10) a une fréquence élevée chez les Peuls du Sénégal (19%, allèle le plus fréquent, contre 2% chez les Peuls du Mali), alors que les Peuls du Mali ont, comme allèles fréquents, DRB1*3148 (19% contre 11% pour les Peuls du Sénégal), correspondant à HLA-DRB1*13, et DRB1*3137 (correspondant à HLA-DRB1*03), la séquence DRB1-Exon2 la plus fréquente avec 30% chez les Peuls du Mali, contre 7% chez les Peuls du Sénégal.

Les deux populations d'Afrique de l'est (Oromo et Amhara) sont caractérisées par une fréquence élevée de DRB1*3142 (HLA-DRB1*13) : 38% chez les Oromo et 20% chez les Amhara. Cette séquence est aussi retrouvée à une fréquence élevée chez les Daza du Tchad (28%).

Finalement, DRB1-Exon2 se démarque aussi des autres loci de l'étude par la part importante des séquences de faibles fréquences (catégorisées en «Autres (<10%)» sur la Figure 4.8), en lien avec le grand nombre de séquences différentes observées pour ce locus (voir Figure 4.7) et la présence de beaucoup de séquences à des fréquences intermédiaires comme le montre la Figure 4.12 illustrant les distributions de fréquences alléliques observées à chacun des quatre loci de l'étude, toutes populations confondues.

DQA1-Exon2

La Figure 4.9 montre les distributions de fréquences alléliques pour DQA1-Exon2 dans toutes les populations testées et à l'équilibre de Hardy-Weinberg (n=31). Les allèles n'excédant jamais 10% de fréquence dans l'ensemble des populations ont été regroupés dans la catégorie « Autres (<10%) ». Les valeurs numériques sont données en annexe S-43 et les correspondances entre les séquences d'exons 2 et les allèles nominaux HLA en annexe S-44.

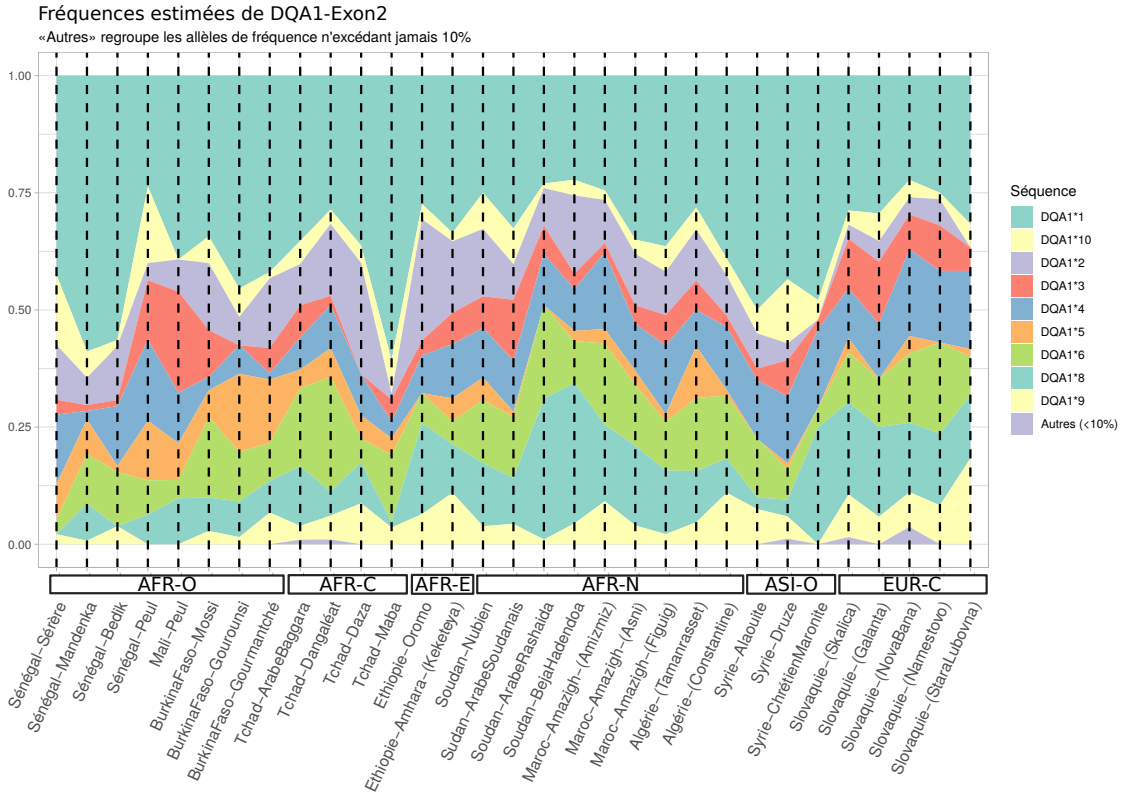


FIGURE 4.9 – Distribution des fréquences alléliques au locus DQA1-Exon2. Ne sont représentés que les allèles dont la fréquence observée dépasse 10% pour au moins une population, « Autres (<10%) » regroupant alors tous les allèles dont la fréquence n'excède jamais 10% dans ces populations. Les bandeaux au desssus des noms de populations indiquent la région géographique où vivent ces populations : AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

Les distributions de fréquences alléliques pour DQA1-Exon2 montrent que les celles-ci sont peu associées à la géographie. Quelques allèles montrent toutefois des variations notables entre plusieurs régions ou populations. DQA1*1 (HLA-DQA1*05) est très fréquent chez les Bédik du Sénégal (56%), les Mandenkalu du Sénégal (59%), les Maba du Tchad (61%), les Alaouites de Syrie (50%), les Druze de Syrie (44%) et les chrétiens Maronites de Syrie (48%). DQA1*5 (HLA-DQA1*04) est retrouvé fréquement en Afrique de l'ouest chez les Gourounsi (17%) et Gourmantché (14%) du Burkina Faso, les Peuls du Sénégal (13%), ainsi qu'en Afrique du nord chez les Tamasheq de Tamanrasset (11%). DQA1*8 (HLA-DQA1*02), quant à lui, est retrouvé à des fréquences élevées chez deux populations nomades du Soudan, les Beja Hadendoa (30%) et les Arabes Rashaida (30%). Les deux

populations Peuls (du Sénégal et du Mali) sont les seules à n'avoir aucun allèle « Autres (<10%) ».

DQB1-Exon2

La Figure 4.10 montre les distributions de fréquences alléliques pour DQB1-Exon2 dans toutes les populations testées et à l'équilibre de Hardy-Weinberg (n=27). Les allèles n'excédant jamais 10% de fréquence dans l'ensemble des populations ont été regroupés dans la catégorie « Autres (<10%) ». Les valeurs numériques sont données en annexe S-43 et les correspondances entre les séquences d'exons 2 et les allèles nominaux HLA en annexe S-44.

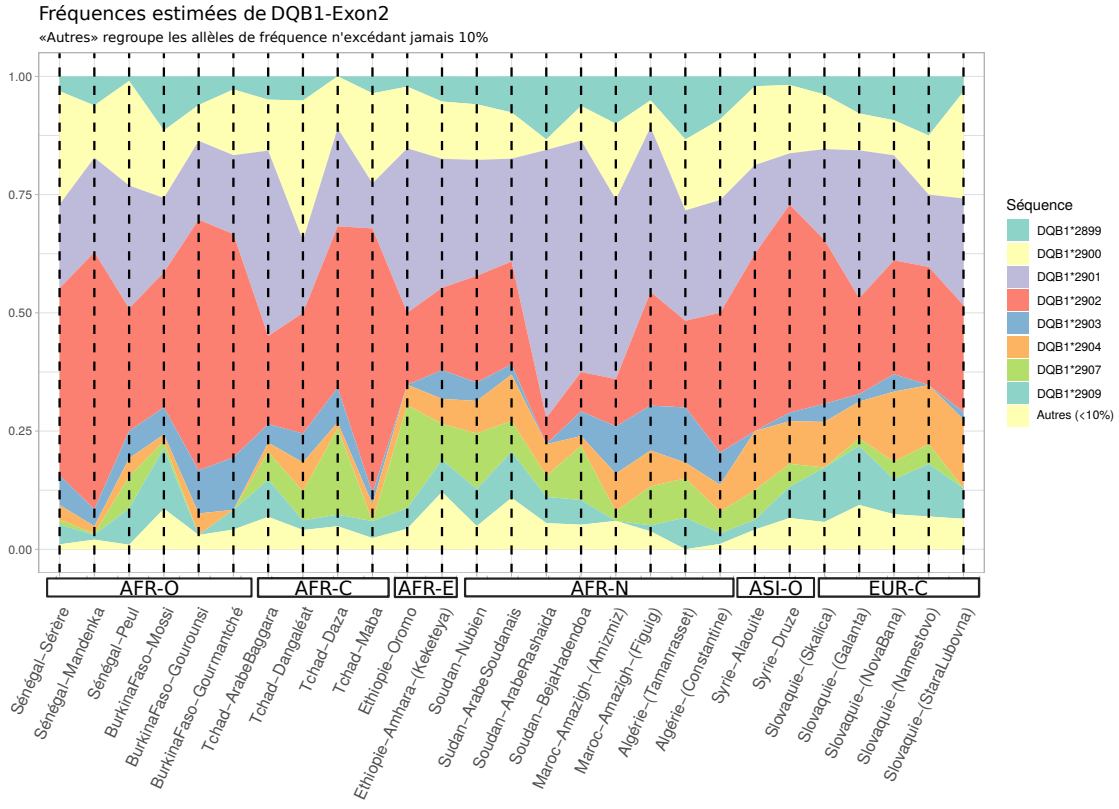


FIGURE 4.10 – Distribution des fréquences alléliques au locus DQB1-Exon2. Ne sont représentés que les allèles dont la fréquence observée dépasse 10% pour au moins une population, « Autres (<10%) » regroupant alors tous les allèles dont la fréquence n'excède jamais 10% dans ces populations. Les bandeaux au desssus des noms de populations indiquent la région géographique où vivent ces populations : AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

À l'instar de DQA1-Exon2, les distributions de fréquences alléliques pour DQB1-Exon2 montrent peu de correspondance avec la géographie, à l'exception de DQB1*2901 et DQB1*2902.

DQB1*2902 (en rouge sur la Figure 4.10 et pouvant correspondre à HLA-DQB1*03) est très fréquent dans les populations d'Afrique de l'ouest (fréquence moyenne de $41 \pm 12\%$) à l'exception des Peuls du Sénégal (26%) et des Mossi du Burkina Faso (29%).

Cet allèle est retrouvé aussi à une fréquence élevée en Afrique centrale (fréquence moyenne $34 \pm 16\%$), principalement chez les Maba du Tchad (56%).
Trois populations nomades, les Imazighen d'Amizmiz, les Beja Hadendoa et les Arabes Rashaida montrent au contraire une fréquence réduite de DQB1*2902 (respectivement 10%, 8% et 6%), mais une fréquence plus élevée de DQB1*2901 (respectivement 38%, 49% et 57%), ce dernier pouvant correspondre à HLA-DQB1*02.
Ces résultats sont à nuancer puisque ces deux séquences DQB1*2901 et DQB*2902 peuvent correspondre à un très grand nombre d'allèles définis au second champ (53 allèles différents pour DQB1*2901 et 73 pour DQB1*2902). Il pourrait donc s'agir non pas d'un seul allèle fréquent à chaque fois, mais de l'addition d'un grand nombre d'allèles différents et partageant tous un même exon 2.

DPB1-Exon2

La Figure 4.11 montre les distributions de fréquences alléliques pour DPB1-Exon2 dans toutes les populations testées et à l'équilibre de Hardy-Weinberg (n=31). Les allèles n'excédant jamais 10% de fréquence dans l'ensemble des populations ont été regroupés dans la catégorie « Autres (<10%) ». Les valeurs numériques sont données en annexe S-43 et les correspondances entre les séquences d'exons 2 et les allèles nominaux HLA en annexe S-44.

Les distributions de fréquences alléliques pour DPB1-Exon2 montrent davantage de différences entre les régions géographiques que celles des autres loci.

DPB1*64 (pouvant correspondre à HLA-DPB1*17:01:01, 131:01 ou 460:01) et DPB1*66 (pouvant correspondre à HLA-DPB1*01:01:01, 162:01:02 ou 733:01) sont fréquents en Afrique de l'ouest, avec des fréquences alléliques moyennes de $20 \pm 12\%$ pour DPB1*64 et $33 \pm 22\%$ pour DPB1*66. La deuxième région où ces allèles sont les plus fréquents est l'Afrique centrale avec respectivement $9 \pm 2\%$ et $10 \pm 13\%$, suivie de l'Afrique du nord avec $8 \pm 4\%$ et $6 \pm 7\%$. DPB1*64 est surtout fréquent chez les populations du Sénégal (fréquences de 24 à 35%), tandis que DPB1*66 est surtout fréquent dans les populations du Burkina Faso (fréquences de 49 à 64%). On retrouve ce dernier allèle fréquent aussi chez les Dangaléat du Tchad (31%), les populations nomades et semi-nomades d'Afrique du nord (Tamasheq de Tamanrasset et Imazighen d'Amizmiz, Asni et Figuig).

DPB1*76 (pouvant correspondre à HLA-DPB1*10:01:01, *650:01 et *673:01) n'est retrouvé à des fréquences assez élevées que chez les Peuls du Mali (23%) et du Sénégal (13%), ainsi que chez les Imazighen d'Asni (8%) et de Figuig (7%). Il n'est détecté que chez les Touaregs de Tamanrasset (3%), les Imazighen d'Amizmiz (3%) et les Arabes Rashaida (2%).

DPB1*71 semble montrer aussi un cline géographique important, avec des fréquences basses en Afrique de l'ouest et qui augmentent en suivant un axe Afrique centrale → Afrique de l'est → Afrique du nord → Asie de l'ouest → Europe centrale, pour atteindre $43 \pm 7\%$.

Finalement, notons les fréquences élevées des séquences catégorisées sous « Autres (<10%) » chez les Oromo d'Ethiopie (26%) et les Arabes Soudanais (22%). Cette catégorie regroupe toutes les séquences dont la fréquence ne dépasse jamais 10%, correspondant à

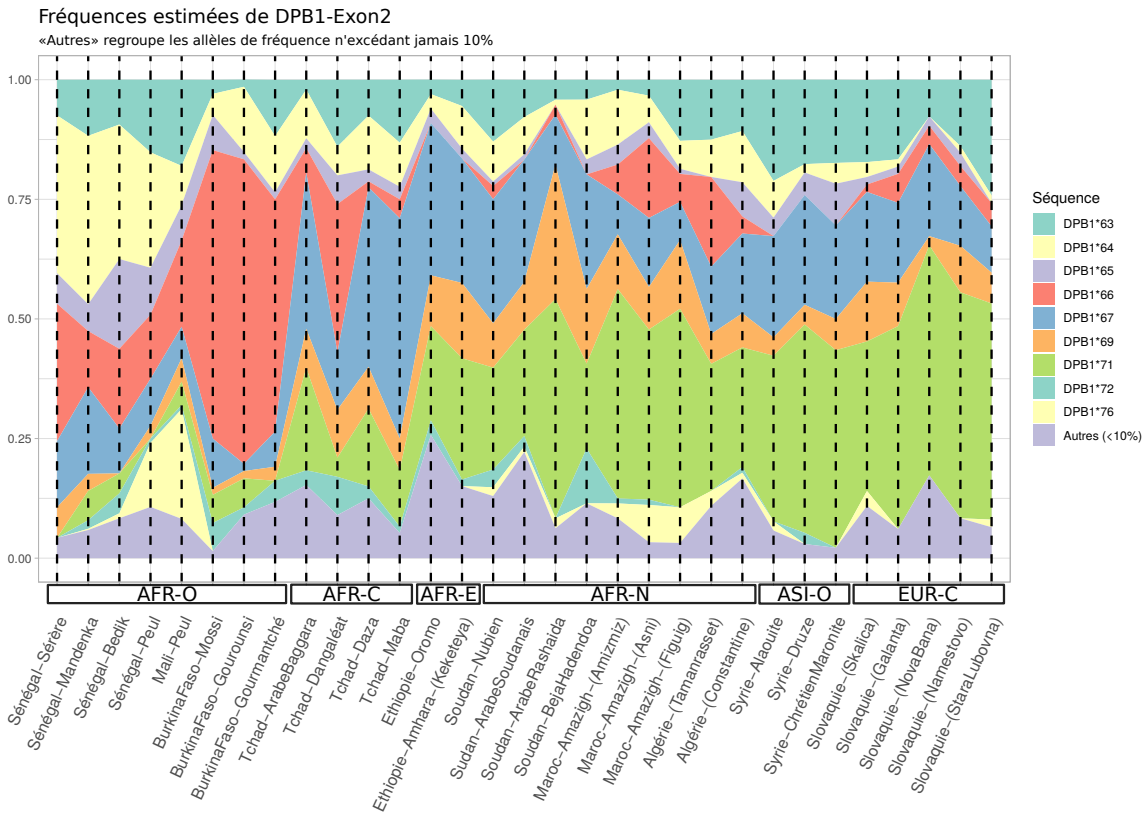


FIGURE 4.11 – Distribution des fréquences alléliques au locus DPB1-Exon2. Ne sont représentés que les allèles dont la fréquence observée dépasse 10% pour au moins une population, « Autres (<10%) » regroupant alors tous les allèles dont la fréquence n'excède jamais 10% dans ces populations. Les bandeaux au desssus des noms de populations indiquent la région géographique où vivent ces populations : AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

des allèles peu fréquemment observés dans cette étude. Pour les Oromo d'Éthiopie, ces « Autres » correspondent à 6 séquences différentes dont DPB1*68 (HLA-DPB1*15:01:01) et DPB1*77 (HLA-DPB1*11:01:01), toutes deux à 8% de fréquence. Au contraire, chez les Arabes Soudanais, cette catégorie regroupe 9 séquences différentes, dont aucune ne dépasse les 5% (DPB1*68 et DPB1*77 étant les plus fréquentes avec 4% de fréquence).

La Figure 4.12 montre, pour chacun des quatre loci de l'étude, les distributions de l'ensemble des fréquences alléliques.

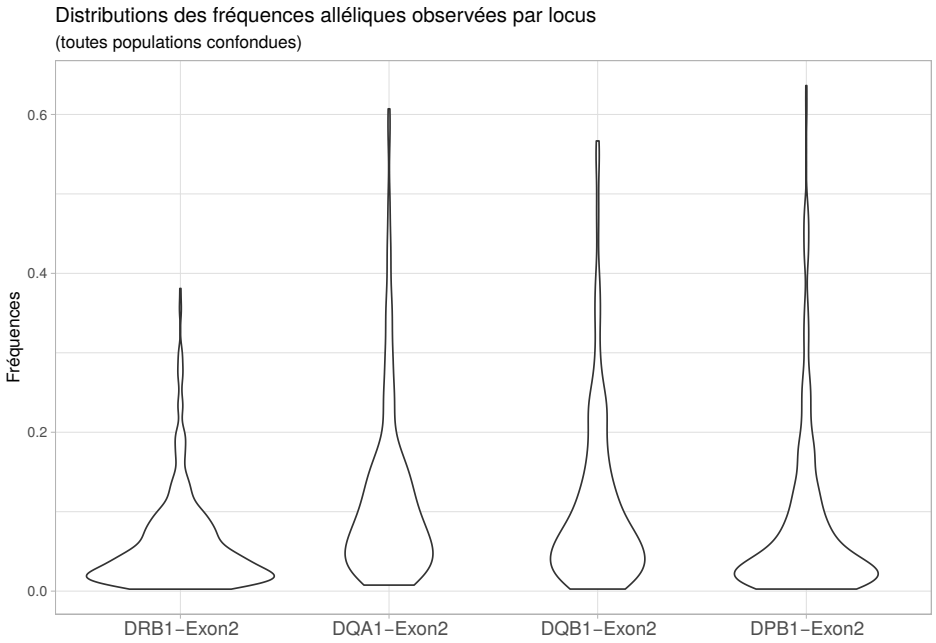


FIGURE 4.12 – Graphiques en violons représentant la distribution des fréquences alléliques estimées à chaque locus de l'étude, toutes populations confondues.

DRB1-Exon2 ne montre aucune fréquence supérieure à 38% (correspondant à DRB1*3142 chez les Oromo d'Ethiopie) et 84.5% des fréquences alléliques observées sont inférieures ou égales à 10%, contre respectivement 57.8%, 64.7% et 74.0% pour DQA1-Exon2, DQB1-Exon2 et DPB1-Exon2.

3.6 Déséquilibres de liaison

Déséquilibres de liaison globaux

La Figure 4.13 montre le résultat des tests non-paramétriques de déséquilibre de liaison global entre chacune des trois paires consécutives de loci. La valeur numérique correspond au quantile du test non-paramétrique de déséquilibre de liaison global [16], les cases en orange à un déséquilibre de liaison global significatif entre les deux loci et les cases vertes une absence de significativité. La liste complète des tests de déséquilibre de liaison global pour toutes les paires de loci est donnée en annexe S-45.

16. Pour rappel, il doit être supérieur ou égal à 95 pour que le déséquilibre de liaison global soit significatif. Voir le Chapitre 1.

FIGURE 4.13 – Résultats des tests non-paramétriques de déséquilibre de liaison global entre chacune des trois paires de loci consécutives. La valeur indiquée correspond au quantile du test non-paramétrique de déséquilibre de liaison global et la couleur indique la significativité (en orange) ou non (en vert) du déséquilibre de liaison global (les cases blanches correspondent à une absence de calcul par manque de données à l'un des loci). Les populations sont regroupées dans les facettes par régions géographiques, AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; EUR-C : Europe centrale ; ASI-O : Asie de l'ouest.

La Figure 4.13 indique une absence de déséquilibre de liaison global entre DQB1-Exon2 et DPB1-Exon2 pour la plupart des populations, à l'exception des Peuls et Mandenkalu du Sénégal, des Maba et Dangalét du Tchad, des Beja Hadendoa et Arabes Rashaida du Soudan et des Imazighen de Figuig. On observe toujours un déséquilibre de liaison global entre DRB1-Exon2∼DQA1-Exon2 et DQA1-Exon2∼DQB1-Exon2, sauf pour :

- Les Gourmantché du Burkina Faso (DRB1-Exon2∼DQA1-Exon2, quantile=93) ;
- Les Maba du Tchad (DRB1-Exon2∼DQA1-Exon2, quantile=34) ;
- Les Tamasheq de Tamanrasset (DRB1-Exon2∼DQA1-Exon2, quantile=1) ;
- Les Alaouites de Syrie (DRB1-Exon2∼DQA1-Exon2 et DQA1-Exon2∼DQB1-Exon2, quantiles=76) ;
- Les Oromo d'Éthiopie (DRB1-Exon2∼DQA1-Exon2 et DQA1-Exon2∼DQB1-Exon2, quantiles de 93 et 94 respectivement).

La Table 4.10 donne, pour chaque paire de loci, la proportion de populations présentant un déséquilibre de liaison global significatif (selon le test non-paramétrique implémenté dans Gene[Rate]).

	DQA1-Exon2	**DQB1-Exon2**	**DPB1-Exon2**
DRB1-Exon2	26/31	25/27	8/31
DQA1-Exon2	—	25/27	7/31
DQB1-Exon2	—	—	7/27

TABLE 4.10 – Pour chaque paires de loci, rapport du nombre de populations présentant un déséquilibre de liaison global significatif (quantile du test non-paramétrique de déséquilibre de liaison global égal à 95 ou plus) et du nombre de populations testées.

On observe un fort déséquilibre de liaison entre tous les loci à l'exception de DPB1-Exon2, pour lequel seules sept populations montrent un déséquilibre de liaison global significatif entre DQB1-Exon et DPB1-Exon2.
Ces résultats indiquent une distance génétique entre DQB1-Exon2 et DPB1-Exon plus importante qu'attendue selon la distance physique (autour de 415'000 pb [The MHC sequencing consortium, 1999]), cohérent avec la présence d'un point chaud de recombinaison méïotique proche du gène TAP2, entre HLA-DQB1 et HLA-DPB1 [Martin et al., 1995].

Du point de vue des populations, les Dangaléat du Tchad, les Peuls (du Sénégal et du Mali), les Imazighen de Figuig, les Mandenkalu du Sénégal et les Beja Hadendoa du Soudan présentent un déséquilibre de liaison global aux trois paires de loci. Au contraire, les Oromo d' Éthiopie et les Alaouite de Syrie ne présentent aucun déséquilibre de liaison global.

Déséquilibres de liaison haplotypiques

Le test de déséquilibre de liaison par haplotype implémenté dans Gene[Rate] repose sur l'utilisation des résidus standardisés et leur comparaison à une distribution de χ^2 à un degré de liberté (voir Chapitre 1 page 39).

Selon ce test, un haplotype est en déséquilibre de liaison positif et significatif si

les résidus standardisés sont supérieurs ou égaux à 1.96 [17]. Toutefois, afin d'éviter une sur-détection de déséquilibre de liaison pour des haplotypes de très faibles fréquences (typiquement moins de 5% [18]) un second seuil de fréquence observée minimale de 5% est utilisé ici pour considérer un haplotype en déséquilibre de liaison significatif et positif (en plus de la valeur-seuil imposée par les résidus standardisés).

La Table 4.11 donne les haplotypes en déséquilibre de liaison positif et significatif les plus fréquents (fréquence observée moyenne d'au moins 10%) par région géographique. La liste complète des 394 haplotypes en déséquilibre de liaison positif pour chaque population est donnée en annexe S-45.

La Table 4.11 ne montre qu'un seul haplotype pour l'Afrique de l'ouest qui ait une fréquence moyenne d'au moins 10%. Il s'agit de DQA1*1∼DQB1*2902, avec une fréquence moyenne de $24 \pm 16\%$. Cet haplotype est aussi retrouvé comme haplotype fréquent dans chacune des autres régions de l'étude. En termes d'allèles nominaux, DQA1*1 correspond à l'exon 2 de HLA-DQA1*05:01 et HLA-DQB1*2902 correspond à l'exon 2 de, notamment, HLA-DQB1*03:01 ou HLA-DQB1*03:19. Une étude réalisée sur les mêmes individus de la population Mandenka du Sénégal séquencés à haute résolution[Goeury et al., 2018a] (voir Chapitre 2) a identifié, dans cette population, cet haplotype comme étant HLA-DQA1*05:01:01∼HLA-DQB1*03:19. Il n'est toutefois pas possible de déterminer s'il s'agit des mêmes allèles pour les autres populations.
Parmi les autres haplotypes présentés dans cette table, DRB1*3142∼DQA1*2 est retrouvé en Afrique centrale ($12 \pm 10\%$) et en Afrique de l'est ($20 \pm 7\%$). DRB1*3142 correspond à l'exon 2 de HLA-DRB1*13:02 et DQA1*2 à l'exon 2 de HLA-DQA1*01:02.
DRB1*3136∼DQA1*8 est, quant à lui, retrouvé en Afrique de l'est ($14 \pm 4\%$), en Afrique du nord ($15 \pm 9\%$) et en Europe centrale ($16 \pm 4\%$). DRB1*3136 et DQA1*8 peuvent correspondrent aux exons 2 de HLA-DRB1*07:01 et HLA-DQA1*02:01, respectivement.
L'haplotype DRB1*3137∼DQA1*1 est retrouvé en Afrique de l'est ($13 \pm 4\%$) et en Afrique du nord ($14 \pm 7\%$), DRB1*3137 correspondant à l'exon 2 de HLA-DRB1*03:01.
Deux haplotypes de cette table sont retrouvés en Afrique du nord et en Europe centrale, il s'agit de DQA1*8∼DQB1*2901 (fréquence en Afrique du nord de $14 \pm 9\%$ et en Europe centrale de $12 \pm 5\%$) et DRB1*3136∼DQB1*2901 (fréquence en Afrique du nord de $12 \pm 8\%$ et en Europe centrale de $11 \pm 5\%$). En termes d'allèle nominal, DQB1*2901 correspond à l'exon 2 de HLA-DQB1*02:01 ou de HLA-DQB1*02:02.

3.7 Diversité moléculaire

La Table 4.12 donne les indices de diversité moléculaire (diversité moléculaire π moyenne par site nucléotidique et nombre de sites polymorphiques S moyen par site nucléotidique), ainsi que les valeurs D du test de Tajima pour chacun des quatre loci de l'étude et différents jeux de nucléotides : ensemble des nucléotides de l'exon 2, uniquement nucléotides codant pour le site de reconnaissance de l'antigène (ARS) et uniquement nucléotides ne codant pas pour le site de reconnaissance de l'antigène (non-ARS).

Que ce soit pour la diversité nucléotidique ou pour le nombre de sites polymorphiques, les mêmes tendances sont observées :

17. Seuil correspondant à un risque α de 5% pour une distribution de loi normale centrée réduite.
18. Pour une taille d'échantillon de 20 individus diploïdes (taille minimale d'échantillon dans cette étude), cela correspond à 2 copies de l'haplotype.

Haplotype	**Fréquence moyenne**	**Région**
DQA1*1∼DQB1*2902	0.239 ± 0.159	AFR-O
DQA1*1∼DQB1*2902	0.283 ± 0.096	AFR-C
DRB1*3142∼DQA1*2	0.123 ± 0.1	AFR-C
DRB1*3153∼DQA1*1	0.105 ± 0.066	AFR-C
DRB1*3142∼DQA1*2	0.197 ± 0.067	AFR-E
DQA1*1∼DQB1*2902	0.169 ± 0.029	AFR-E
DRB1*3142∼DQB1*2907	0.146 ± 0.072	AFR-E
DRB1*3136∼DQA1*8	0.142 ± 0.035	AFR-E
DRB1*3137∼DQA1*1	0.134 ± 0.038	AFR-E
DRB1*3136∼DQA1*8	0.151 ± 0.093	AFR-N
DQA1*8∼DQB1*2901	0.142 ± 0.088	AFR-N
DRB1*3137∼DQA1*1	0.141 ± 0.071	AFR-N
DRB1*3137∼DQB1*2901	0.135 ± 0.075	AFR-N
DQA1*1∼DQB1*2902	0.129 ± 0.081	AFR-N
DRB1*3136∼DQB1*2901	0.115 ± 0.084	AFR-N
DRB1*3154∼DQA1*1	0.189 ± 0.139	ASI-O
DQA1*1∼DQB1*2902	0.154 ± 0.196	ASI-O
DRB1*3154∼DQB1*2902	0.115 ± 0.146	ASI-O
DQA1*1∼DQB1*2902	0.212 ± 0.037	EUR-C
DRB1*3136∼DQA1*8	0.156 ± 0.036	EUR-C
DQA1*8∼DQB1*2901	0.115 ± 0.054	EUR-C
DRB1*3136∼DQB1*2901	0.11 ± 0.051	EUR-C

TABLE 4.11 – Liste des haplotypes en déséquilibre de liaison positif, de fréquence observée moyenne d'au minimum 10%. La fréquence moyenne correspond à la fréquence observée moyenne pour chaque région (ainsi que l'écart-type). Pour chaque région les haplotypes sont ordonnés par fréquences décroissantes. AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; EUR-C : Europe centrale ; ASI-O : Asie de l'ouest.

Locus	Région génique	$\pi.n \pm \sigma$	$S.n \pm \sigma$	$\overline{D} \pm \sigma$
DRB1-Exon2	Exon 2	0.076 ± 0.009	0.228 ± 0.011	2.164 ± 0.664
DQA1-Exon2		0.088 ± 0.007	0.202 ± 0.003	3.256 ± 0.556
DQB1-Exon2		0.076 ± 0.006	0.204 ± 0.008	2.746 ± 0.520
DPB1-Exon2		0.029 ± 0.005	0.088 ± 0.007	1.989 ± 0.872
DRB1-Exon2	ARS	0.160 ± 0.020	0.451 ± 0.022	2.425 ± 0.781
DQA1-Exon2		0.100 ± 0.008	0.248 ± 0.008	2.933 ± 0.482
DQB1-Exon2		0.155 ± 0.013	0.400 ± 0.011	2.913 ± 0.517
DPB1-Exon2		0.040 ± 0.008	0.107 ± 0.00	2.228 ± 1.030
DRB1-Exon2	non-ARS	0.042 ± 0.005	0.138 ± 0.008	1.560 ± 0.514
DQA1-Exon2		0.083 ± 0.007	0.185 ± 0.004	3.072 ± 0.548
DQB1-Exon2		0.041 ± 0.004	0.120 ± 0.008	2.129 ± 0.479
DPB1-Exon2		0.025 ± 0.004	0.081 ± 0.01	1.521 ± 0.686

TABLE 4.12 – Diversité moléculaire π moyenne par site nucléotidique ($\pi.n$), nombre moyen de sites polymorphiques S par site nucléotidique ($S.n$) et valeur moyenne du D du test de Tajima, selon les différents jeux de nucléotides (colonne « Région génique ») : l'ensemble des nucléotides de l'exon 2 (Exon 2), uniquement les nucléotides codant pour le site de reconnaissance de l'antigène (ARS), uniquement les nucléotides ne codant pas pour le site de reconnaissance de l'antigène (non-ARS). Toutes les valeurs sont données avec un écart-type.

DPB1-Exon2 est moins diversifié que les trois autres loci en termes de diversité moléculaire ($\pi.n$) ou de nombre moyen de sites polymorphiques ($S.n$) par site nucléotidique, quelle que soit la région génique considérée (exon 2 complet, ARS et non-ARS). Du point de vue de la diversité moléculaire des exons 2 complets ($\pi.n$), DQA1-Exon2, DQB1-Exon2 et DRB1-Exon2 semblent similaires, mais pour les codons codant pour l'ARS, DQB1-Exon2 et DRB1-Exon2 se démarquent par une plus grande diversité moléculaire $\pi.n$ et un plus grand nombre de site polymorphiques ($S.n$), notamment pour DRB1-Exon2 avec un $S.n = 0.451 \pm 0.022$ pour les codons ARS. Au contraire, pour les codons non-ARS, DQA1-Exon2 montre une plus grande diversité que les trois autres loci ($\pi.n = 0.083 \pm 0.007$ et $S.n = 0.185 \pm 0.004$).

En conclusion, DPB1-Exon2 est moins diversifié sur le plan moléculaire que les trois autres loci, DQA1-Exon2 montre une diversité élevée pour les codons non-ARS et DQB1-Exon2 et DRB1-Exon2 montrent une très grande diversité pour les codons ARS.

3.8 Tests de neutralité sélective

Test de Ewens-Watterson-Slatkin

La Figure 4.14 donne, pour chaque population, les intervalles de pValeurs du test de Ewens-Watterson-Slatkin, calculées par *bootstrap*. Le test est bilatéral : au seuil $\alpha = 0.05$, si une des bornes de l'intervalle de pValeurs est inférieure à 0.025 (ligne pointillée à gauche), il s'agit d'un rejet de l'hypothèse nulle de neutralité en faveur d'un excès d'hétérozygotes, tandis que si l'une des bornes de l'intervalle de pValeurs est supérieure à

0.975 (ligne pointillée à droite) il s'agit d'un rejet de l'hypothèse de neutralité en faveur d'un excès d'homozygotes. Les résultats pour l'ensemble des populations sont donnés en annexe S-47.

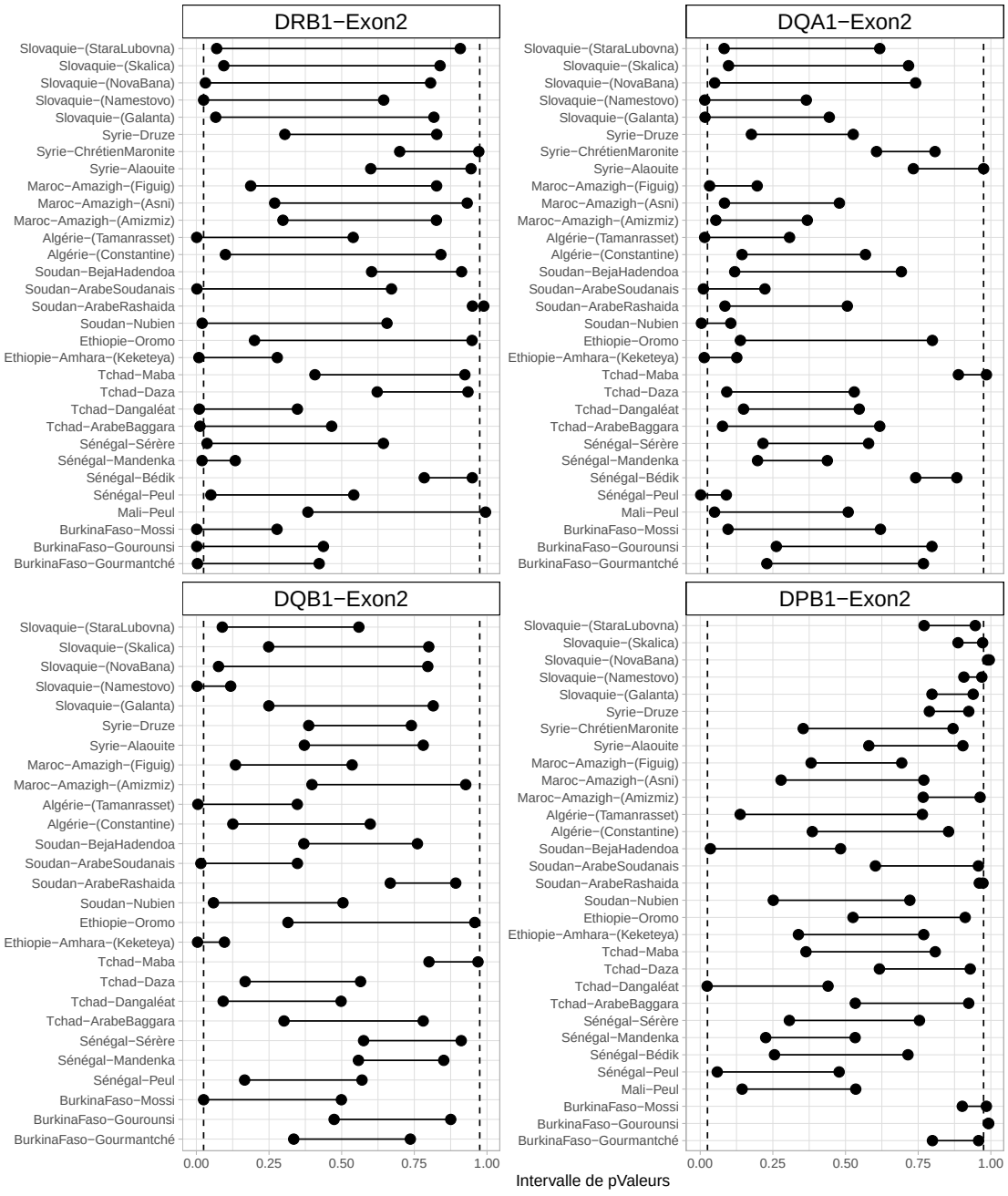


FIGURE 4.14 – Pour chaque population et à chaque locus de l'étude, résultats du test d'Ewens-Watterson-Slatkin. Les segments indiquent l'intervalle de pValeurs obtenues par *bootstrap*. Les lignes en pointillés indiquent les seuils de significativité pour un risque $\alpha = 0.05$. Le test étant bilatéral, l'hypothèse de neutralité est rejetée si l'intervalle de pValeurs possède une de ses bornes à moins de 0.025 (hypothèse de neutralité rejetée en faveur d'un excès d'hétérozygotes) ou à plus de 0.975 (hypothèse de neutralité rejetée en faveur d'un excès d'homozygotes).

La Table 4.13 donne les proportions de rejets du test d'Ewens-Watterson-Slatkin (EWS) à chacun des quatre loci.

	Rejets (hétérozygotes)	**Neutralité**	**Rejets (homozygotes)**
DRB1-Exon2	11	18	2
DQA1-Exon2	7	22	2
DQB1-Exon2	5	22	0
DPB1-Exon2	1	27	3

TABLE 4.13 – Pour chaque locus de l'étude, résultats du test d'Ewens-Watterson-Slatkin. Rejets (hétérozygotes) : nombre de populations rejetant la neutralité en faveur d'un excès d'hétérozygotes ; Neutralité : nombre de populations ne rejetant pas la neutralité ; Rejets (homozygotes) : nombre de populations rejetant la neutralité en faveur d'un excès d'homozygotes.

Le locus DRB1-Exon2 montre le plus de rejets de la neutralité en faveur d'un excès d'hétérozygotes (35% des populations) tandis que DPB1-Exon2 montre le moins de rejets de la neutralité en faveur d'un excès d'hétérozygotes (3% des populations). Les rejets de la neutralité sont majoritairement en faveur d'un excès d'hétérozygotes, tandis que les rejets de la neutralité en faveur d'un excès d'homozygotes concernent :

- Au locus **DPB1-Exon2**, les Slovaques de Nova Bana ($H = 0.718$), dû à la fréquence élevée de DPB1*71 (48%), ainsi que les Mossi ($H = 0.610$) et Gourounsi ($H = 0.566$) du Burkina Faso, dû aux fréquences élevées de DPB1*66 (respectivement 60 et 64%) ;
- Au locus **DQA1-Exon2**, les Alaouites de Syrie ($H = 0.704$) et les Maba du Tchad (H=0.600), dû à la fréquence de DQA1*1 (respectivement de 50 et 61%) ;
- Au locus **DRB1-Exon2**, les Arabes Rashaida du Soudan ($H = 0.792$), dû aux fréquences de DRB1*3136, DRB1*3137 et DRB1*3140 de respectivement 30, 25 et 14%, et les Peuls du Mali ($H = 0.843$), dû aux fréquences pour DRB1*3137 et DRB1*3147 de respectivement 30 et 19%).

Test du D de Tajima

La Table 4.14 donne pour chaque population les valeurs D de Tajima à chaque locus, selon les jeux de nucléotides considérés : l'ensemble des nucléotides de l'exon 2 (Exon 2), uniquement les nucléotides codant pour le site de reconnaissance de l'antigène (ARS), uniquement les nucléotides ne codant pas pour le site de reconnaissance de l'antigène (non-ARS). Les D significativement différents de 0 (après correction *fdr* [Benjamini and Hochberg, 1995] pour tests multiples) sont indiqués par des '*', '**' et '***'.

		DRB1-Exon2			DQA1-Exon2			DQB1-Exon2			DPB1-Exon2		
Région	**Population**	**Exon 2**	**ARS**	**non-ARS**	**Exon 2**	**ARS**	**non-ARS**	**Exon 2**	**ARS**	**non-ARS**	**Exon 2**	**ARS**	**non-ARS**
AFR-O	Sénégal-Sérère	2.24*	2.62*	1.46	3.49***	3.44**	2.95*	3.07**	3.15**	2.51*	2.11*	2.95*	1.3
AFR-O	Sénégal-Mandenka	2.02	2.42*	1.19	2.99*	2.78*	2.72*	3.09*	3.14*	2.44*	3.5**	4.28**	2.31*
AFR-O	Sénégal-Bedik	0.95	1.39	0.25	2.72*	2.74*	2.44*	–	–	–	3.1**	3.03*	2.54*
AFR-O	Sénégal-Peul	2.65*	2.6*	2.36*	3.71***	3.24**	3.57**	3.59**	3.7**	2.89*	2.36*	3.38**	1.47
AFR-O	Mali-Peul	1.82	2.05	1.28	3.5***	3.28**	3.47**	–	–	–	2.62*	3.21*	1.88
AFR-O	BurkinaFaso-Mossi	1.66	1.89	1.15	2.75*	2.34*	2.71*	2.63*	2.71*	2.14*	0.9	0.49	1.01
AFR-O	BurkinaFaso-Gourounsi	1.64	1.95	1.06	2.4*	2.38*	2.17*	1.66	1.75	1.31	0.7	0.69	0.59
AFR-O	BurkinaFaso-Gourmantché	1.13	1.28	0.78	2.57*	2.37*	2.43*	2.51*	2.56*	2.1	1.64	1.59	1.41
AFR-C	Tchad-ArabeBaggara	1.96	2.21*	1.38	3.54***	3.14**	3.18**	3.14**	3.49**	2.26*	2.82*	2.89*	2.29*
AFR-C	Tchad-Dangaléat	1.66	1.7	1.39	3.09**	2.57*	2.87*	3.22**	3.33**	2.58*	2.86*	2.57*	2.53*
AFR-C	Tchad-Daza	1.24	1.61	0.61	3.17**	2.78*	3.29**	2.93*	3.13**	2.22*	1.58	2.44*	0.91
AFR-C	Tchad-Maba	1.96	2.2*	1.43	1.99	1.64	1.98	1.74	1.87	1.31	1.67	1.82	1.29
AFR-E	Ethiopie-Oromo	1.76	2.14*	1.1	3.17**	2.85*	3.04**	3.14**	3.14***	2.7*	2.23*	2.24*	1.9
AFR-E	Ethiopie-Amhara-(Keketeya)	2.74*	2.94*	2.03	4.08***	3.52**	3.92***	3.58**	3.69**	2.87*	3.09*	3.2**	2.46*
AFR-N	Soudan-Nubien	2.63*	2.92*	1.93	3.64***	3.17**	3.31**	3**	3.39**	2.09	2.63*	2.78*	2.08
AFR-N	Sudan-ArabeSoudanais	2.57*	3.03*	1.73	3.58***	3.04**	3.3**	2.67*	3.06**	1.81	2.07	2.82*	1.39
AFR-N	Soudan-ArabeRashaida	3.75**	4.33***	2.45*	3.55**	3.42**	3.25**	2.58*	2.65*	2.09	3.05*	2.85*	2.59*
AFR-N	Soudan-BejaHadendoa	3.16**	3.82**	1.94	3.57***	3.35**	3.11**	3.01**	3.08**	2.47*	3.2**	3.34**	2.58*
AFR-N	Maroc-Amazigh-(Amizmiz)	3.11**	3.77**	1.94	3.75***	3.33**	3.59**	2.71*	3.01**	2	2.19*	3.03*	1.4
AFR-N	Maroc-Amazigh-(Asni)	2.79*	3.19**	1.95	3.52**	3.35**	3.46**	–	–	–	2.29*	2.64*	1.72
AFR-N	Maroc-Amazigh-(Figuig)	2.84*	3.02*	2.21*	4.27***	3.75**	4.06***	3.5**	3.58**	2.82*	2.88*	3.05*	2.2
AFR-N	Algérie-(Tamanrasset)	2.02*	2.36*	1.37	3.12**	2.79*	3**	3.12**	3.15**	2.65*	1.94	2.31*	1.47
AFR-N	Algérie-(Constantine)	2.11*	2.19*	1.73	3.36***	2.88*	3.3**	2.99**	3.31**	2.17*	1.98	2.82*	1.23
ASI-O	Syrie-Alaouite	0.91	0.75	1	2.33*	1.93	2.35*	2.37*	2.69*	1.63	1.44	1.41	1.18
ASI-O	Syrie-Druze	1.86	1.56	1.99	3.91***	3.21**	3.86**	2.19*	2.5*	1.46	0.86	0.73	0.78
ASI-O	Syrie-ChrétienMaronite	2.03*	2.08*	1.74	2.04*	2.22*	1.75	–	–	–	0.75	0.93	
EUR-C	Slovaquie-(Skalica)	2.42*	2.67*	1.81	3.48***	3.19**	3.09**	2.06*	2.17*	1.64	1.34	1.86	0.84
EUR-C	Slovaquie-(Galanta)	2.71**	2.98**	2.03	3.4**	3.18**	3.4**	2.33*	2.72*	1.55	1.15	0.95	1.08
EUR-C	Slovaquie-(NovaBana)	2.39*	2.68*	1.74	3.33***	3.01**	3.21**	2.03*	2.21*	1.51	0.16	0.12	0.16
EUR-C	Slovaquie-(Namestovo)	2.6*	2.88*	1.92	3.5**	3.04**	3.2**	2.64*	2.63*	2.23*	1.59	1.59	1.31
EUR-C	Slovaquie-(StaraLubovna)	1.78	1.92	1.39	3.37**	3**	3.26**	2.61*	2.82*	1.98	0.94	1.06	0.71
	Total	17/31	21/31	3/31	30/31	29/31	29/31	25/27	25/27	14/27	15/31	19/31	7/31

TABLE 4.14 – Valeurs du D de Tajima pour chaque population, pour les quatre loci de l'étude selon le jeu nucléotidique considéré : l'ensemble des nucléotides de l'exon 2 (Exon 2), uniquement les nucléotides codant pour le site de reconnaissance de l'antigène (ARS), uniquement les nucléotides ne codant pas pour le site de reconnaissance de l'antigène (non-ARS). Significativité : '*' : pValeur<0.05 ; '**' : pValeur < 0.01 ; '***' : pValeur < 0.001. AFR-O : Afrique de l'ouest, AFR-C : Afrique centrale, AFR-E : Afrique de l'est, AFR-N : Afrique du nord, ASI-O : Asie de l'ouest, EUR-c : Europe centrale.

La Figure 4.15 illustre les résultats du test de neutralité sélective de Tajima pour chacun des quatre loci de l'étude selon les jeux nucléotidiques considérés (ensemble des nucléotides de l'exon 2, uniquement les codons codant pour le site de reconnaissance de l'antigène ou uniquement les codons ne codant pas pour le site de reconnaissance de l'antigène).

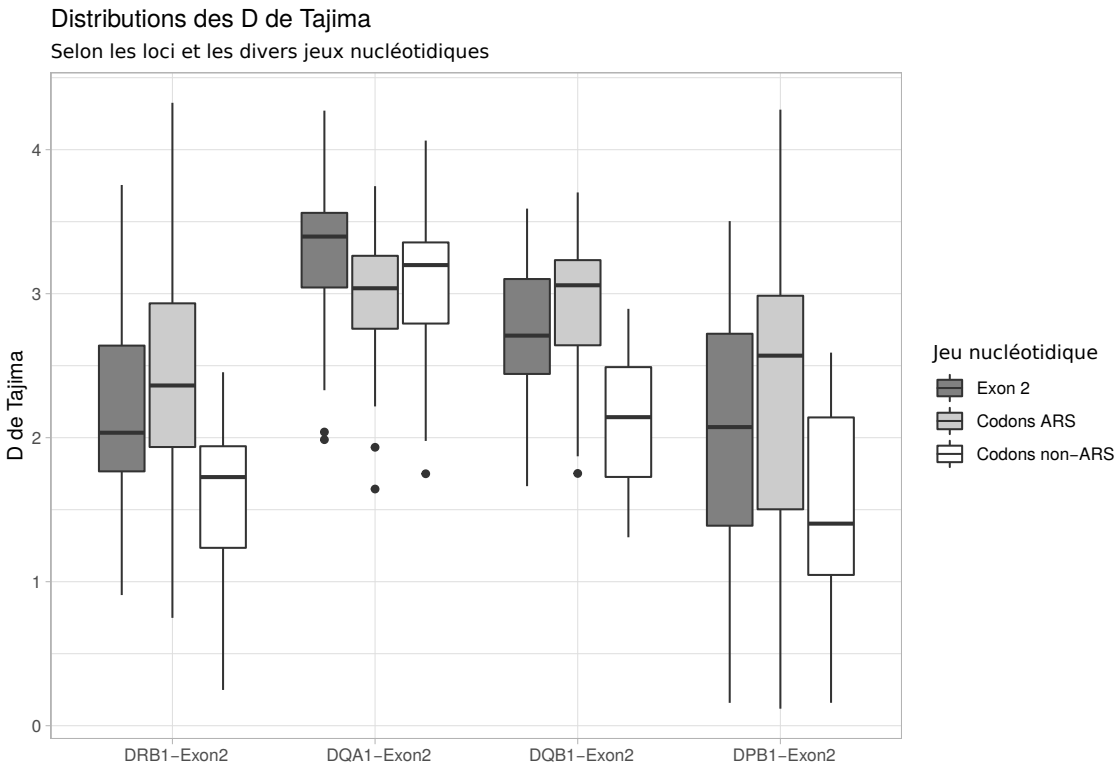


FIGURE 4.15 – Distribution des D de Tajima pour chacun des quatre loci de l'étude, selon les jeux nucléotidiques considérés : l'ensemble des nucléotides de l'exon 2 (Exon2), uniquement les codons codant pour le site de reconnaissance de l'antigène (Codons ARS), uniquement les codons ne codant pas pour le site de reconnaissance de l'antigène (Codons non-ARS).

Aucune valeur D de Tajima n'est négative. Parmi les valeurs significativement supérieures à zéro, les résultats suivants sont observés :

— **DRB1-Exon2** montre le plus de différences entre les codons ARS (68% des populations ont un D significatif) et les codons non-ARS, où seuls les Imazighen de Figuig, les Arabes Rashaida et les Peuls du Sénégal ont un D significativement différent de 0 ;

— **DQA1-Exon2** montre le plus grand nombre de populations avec un D de Tajima significatif pour les trois régions géniques. Seuls les Maba du Tchad ne montrent pas de D significatif (aux trois régions géniques) et les Chrétiens Maronites de Syrie ne montrent un D significatif que pour l'exon 2 complet (les D calculés pour les codons ARS ou non-ARS ne sont pas significatifs). C'est le locus pour lequel le plus de populations montrent des D de Tajima significatifs (96% des populations ont un $D_{Exon\ 2}$ significatif), même pour les codons non-ARS (94% des populations avec un $D_{non-ARS}$ significatif) ;

— **DQB1-Exon2** montre des résultats similaires à DQA1-Exon2 en termes de nombre de populations pour lesquels le D de Tajima est significatif pour l'exon 2 au complet ou les codons ARS, puisque seuls les Maba du Tchad ($D_{Exon2} = 1.74$ et $D_{ARS} = 1.87$) et les Gourounsi du Burkina Faso ($D_{Exon2} = 1.66$ et $D_{ARS} = 1.75$) n'ont pas un D significatif. A ce locus, 93% des populations ont un D significatif aux codons ARS contre 52% aux codons non-ARS. (Les résultats pour les Imazighen d'Asni, les Peuls du Mali, les Bédik du Sénégal et les Chrétiens Maronites de Syrie ne sont pas disponibles, faute d'un nombre suffisant de génotypes exploitables pour ces populations au locus DQB1-Exon2) ;
— **DPB1-Exon2** montre lui aussi une différence entre les codons ARS et les codons non-ARS, puisque sur les 31 populations, 19 montrent un D significatif aux codons ARS contre 7 au codons non-ARS.

3.9 Relations entre populations

Analyse Factorielle des Correspondances

Les Figures 4.16 à 4.19 représentent les deux premières composantes de la même analyse factorielle des correspondances (AFC), chacune illustrant les contributions des allèles les plus fréquents (critère : fréquence minimum de 10% dans une population) à chaque locus (de haut en bas : DRB1-Exon2, DQA1-Exon2, DQB1-Exon2, DPB1-Exon2) par des lignes pointillées. La représentation des composantes 1 et 3 et les contributions de l'ensemble des allèles sont données en annexe S-48.

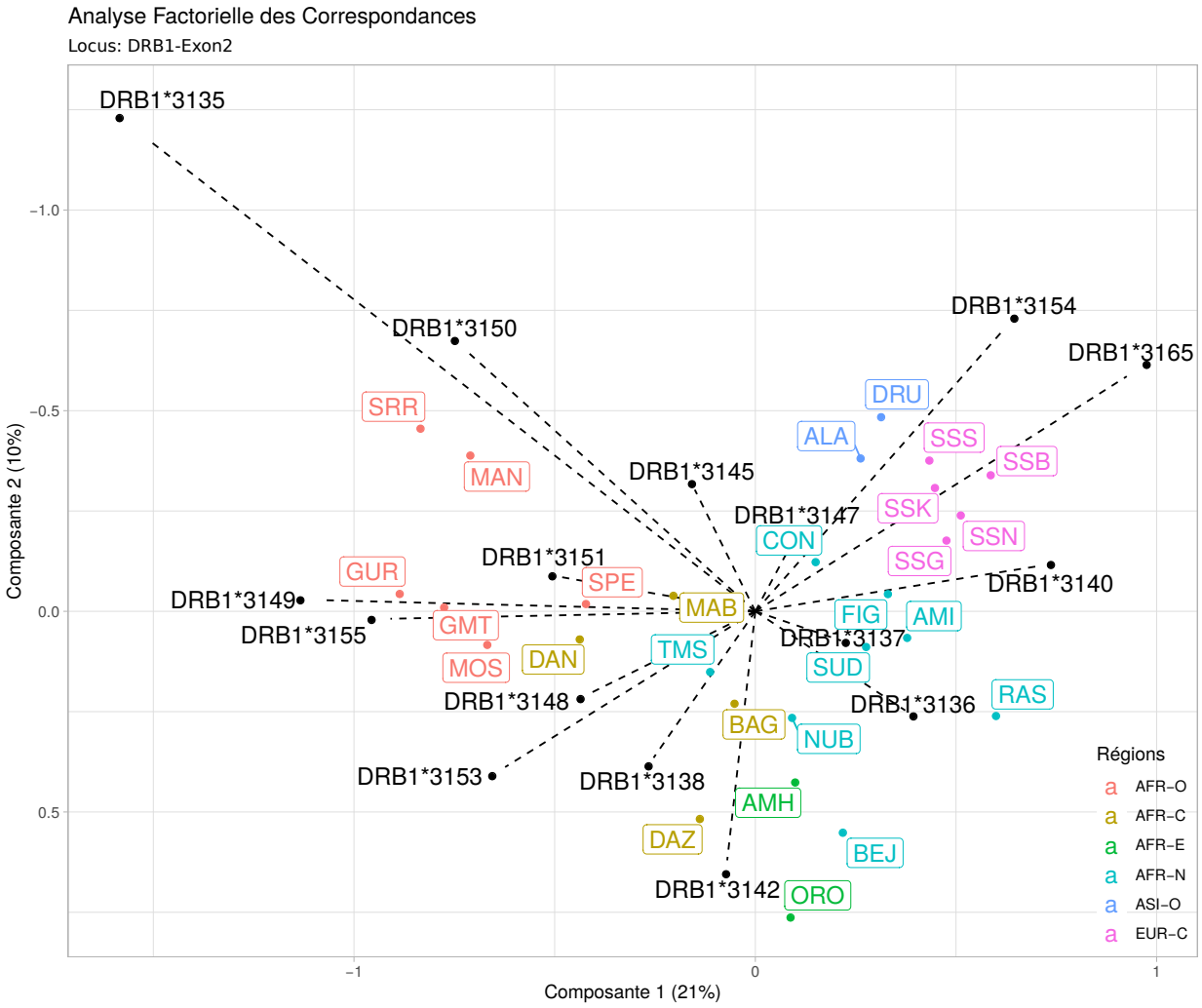


FIGURE 4.16 – Représentation graphique de l'analyse factorielle des correspondances (AFC) conduite sur les fréquences alléliques de chaque population. La composante 2 a été inversée afin de mieux représenter l'association avec la géographie. Les lignes pointillées représentent la position des allèles DRB1-Exon2 les plus fréquents (fréquence d'au moins 10% dans une population) dans l'AFC et les acronymes à trois lettres correspondent aux populations suivantes : SRR : Sénégal-Sérère ; MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ; SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

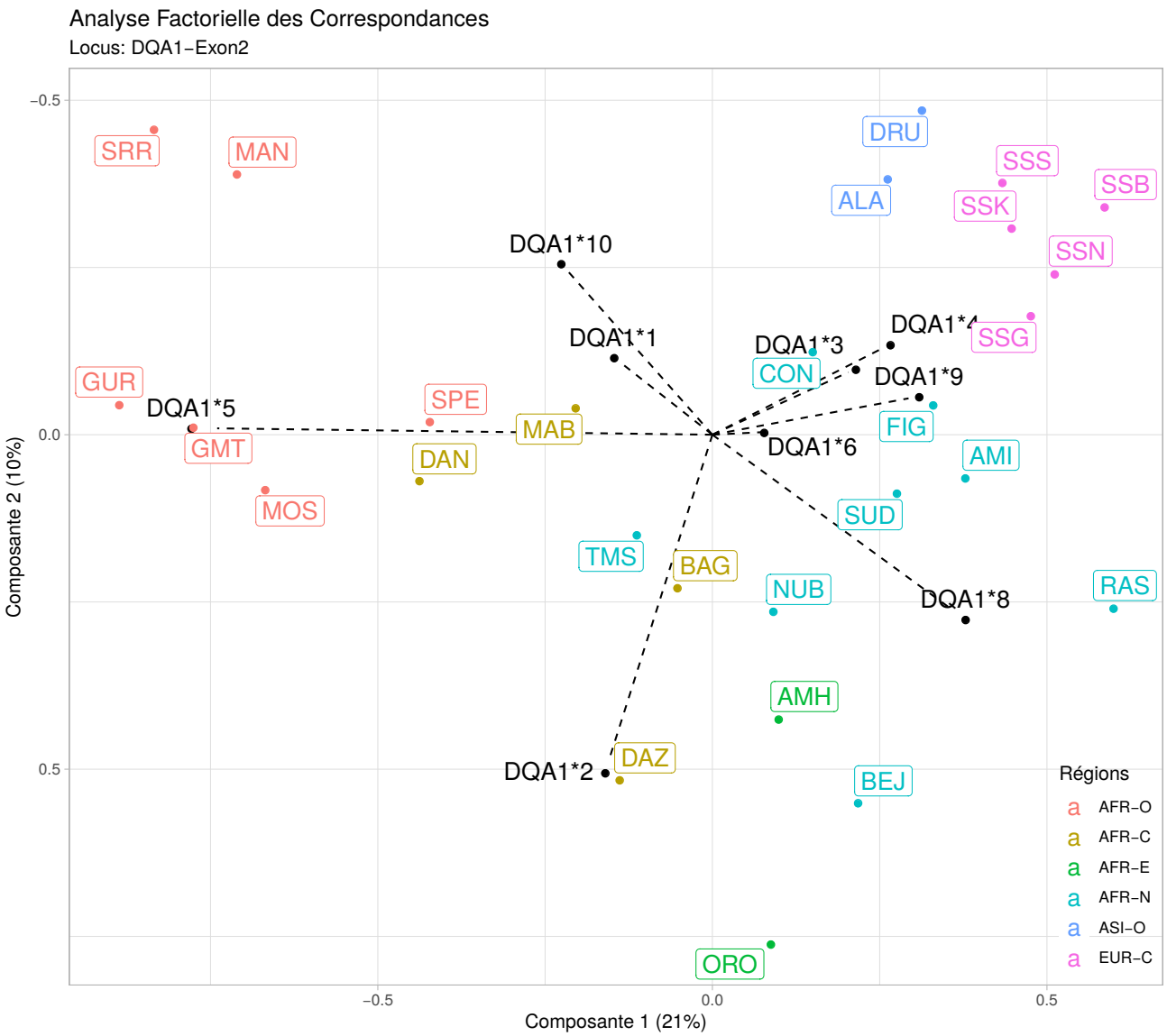


FIGURE 4.17 – Représentation graphique de l'analyse factorielle des correspondances (AFC) conduite sur les fréquences alléliques de chaque population. La composante 2 a été inversée afin de mieux représenter l'association avec la géographie. Les lignes pointillées représentent la position des allèles DQA1-Exon2 les plus fréquents (fréquence d'au moins 10% dans une population) dans l'AFC et les acronymes à trois lettres correspondent aux populations suivantes : SRR : Sénégal-Sérère ; MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ; SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

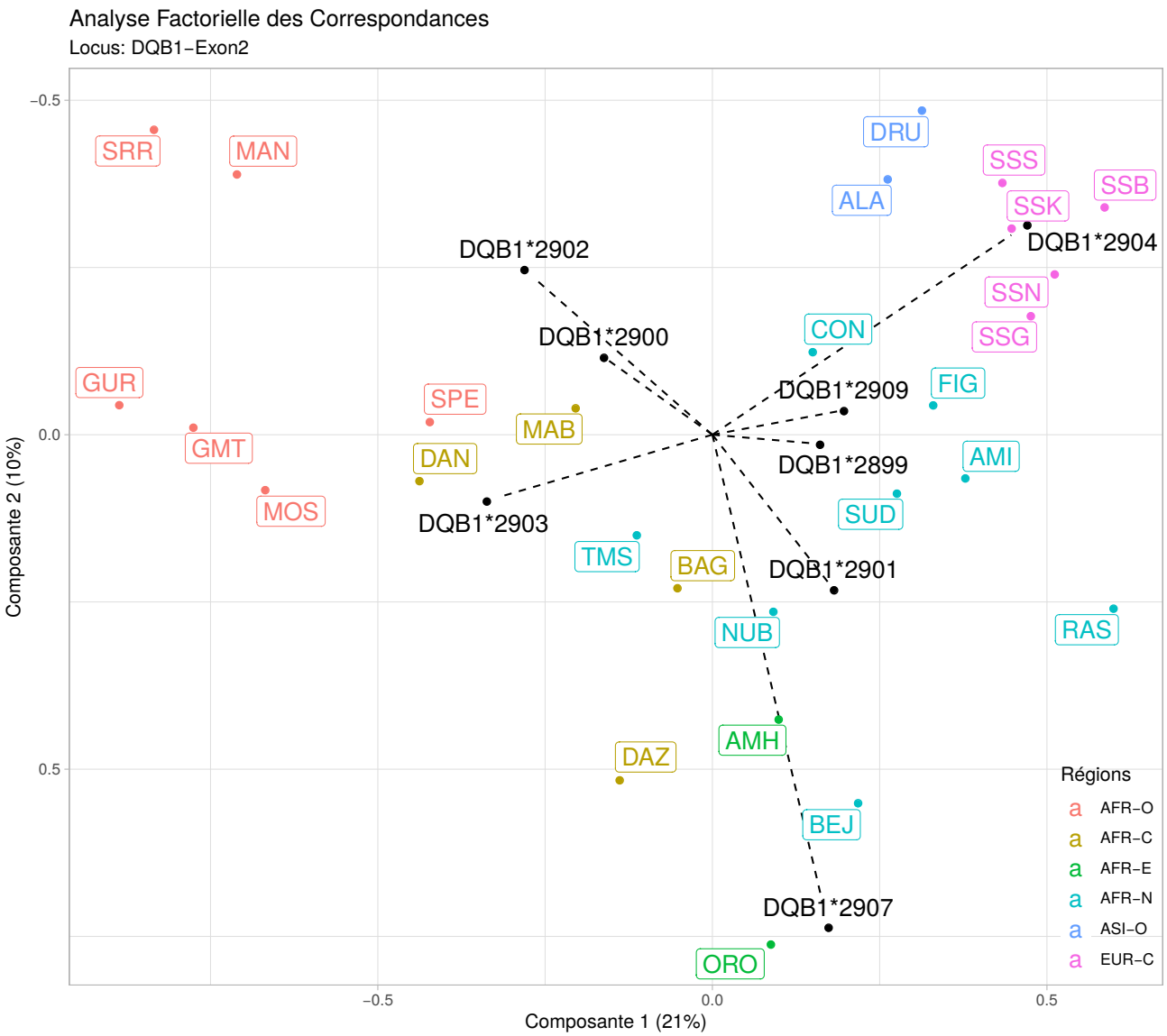


FIGURE 4.18 – Représentation graphique de l'analyse factorielle des correspondances (AFC) conduite sur les fréquences alléliques de chaque population. La composante 2 a été inversée afin de mieux représenter l'association avec la géographie. Les lignes pointillées représentent la position des allèles DQB1-Exon2 les plus fréquents (fréquence d'au moins 10% dans une population) dans l'AFC et les acronymes à trois lettres correspondent aux populations suivantes : SRR : Sénégal-Sérère ; MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ; SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

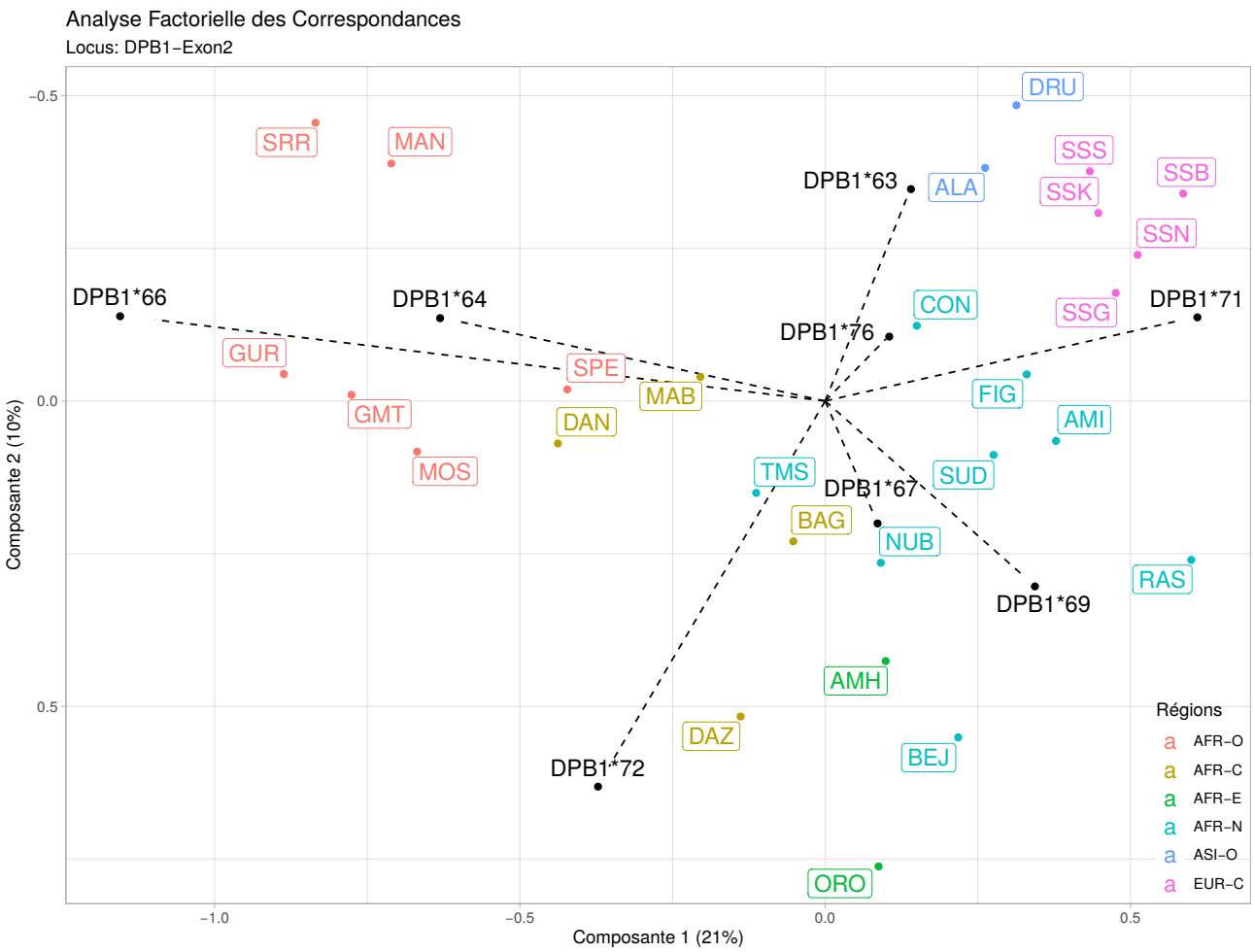


FIGURE 4.19 – Représentation graphique de l'analyse factorielle des correspondances (AFC) conduite sur les fréquences alléliques de chaque population. La composante 2 a été inversée afin de mieux représenter l'association avec la géographie. Les lignes pointillées représentent la position des allèles DPB1-Exon2 les plus fréquents (fréquence d'au moins 10% dans une population) dans l'AFC et les acronymes à trois lettres correspondent aux populations suivantes : SRR : Sénégal-Sérère ; MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ; SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

Les deux premières composantes de l'AFC montrent une structuration géographique des populations : la première composante sépare, à gauche, les populations d'Afrique de l'ouest des autres populations, tandis que la seconde composante sépare les populations d'Asie de l'ouest et d'Europe (en haut à droite), les populations d'Afrique du nord (au milieu à droite) et les populations d'Afrique de l'est (en bas à droite).

Les populations d'Afrique de l'ouest se séparent en deux groupes qui correspondent à leur origine géographique. On retrouve, d'un côté, les populations burkinabé, associées à l'allèle DPB1*66 (correspondant aux allèles nominaux HLA-DPB1*01:01:01, *162:01:02 et *733:01), et de l'autre côté, les populations sénégalaises (exception faite des Peuls du Sénégal), associées aux allèles DRB1*3135 (seules populations où cet allèle est retrouvé, correspondant à HLA-DRB1*13:04) et DPB1*64 (correspondant aux allèles nominaux HLA-DPB1*17:01:01, *131:01 et *460:01).
Les populations d'Afrique centrale sont séparées entre les Dangaléat et les Maba, proches des populations d'Afrique de l'ouest, les Arabes Baggara, proches des Tamasheq de Tamanrasset (la plus à gauche de toutes les populations d'Afrique du nord) et les Daza, au milieu en haut, plus proches des populations d'Afrique de l'est.
Pour les populations d'Afrique du nord, les Beja Hadendoa sont proches des populations d'Afrique de l'est (Oromo et Amhara), tandis que les populations du Maghreb (Algérie et Maroc dans cette étude) sont proches des populations d'Asie de l'ouest et d'Europe. Les Arabes Rashaida sont éloignés des autres populations d'Afrique du nord en se projetant le plus à droite sur la première composante.

Analyses d'échelonnement multi-dimensionnel

Les six figures suivantes (Figures 4.20 à 4.27) montrent les résultats des analyses d'échelonnement multi-dimensionnel (MDS [19]) basées sur les Θ_w entre populations, pour les quatre loci de l'étude.
La Figure 4.20 représente le résultat de la MDS basée sur les Θ_w calculés entre les 31 populations pour lesquelles le profil génétique DRB1-Exon2 était disponible et la Figure 4.21 représente la proportion de paires de populations présentant un Θ_w significativement différent entre chaque région de l'étude.

Trois axes sont nécessaires pour cette MDS afin d'obtenir une valeur de stress acceptable (qui est de 0.071 [20], la plus élevée des quatre loci). La même MDS sur seulement deux axes présente une valeur de stress de 0.123.
L'axe 1 de la MDS (MDS1) sépare les populations d'Afrique de l'ouest des populations d'Afrique du nord, les quatre populations d'Afrique centrale étant proches des populations burkinabé (Gourmantché, Mossi et Gourounsi) et les populations d'Afrique de l'est (Amhara et Oromo) étant localisées entre les populations Imazighen d'Asni et Figuig et les populations Daza et Maba d'Afrique centrale. L'axe 1 (MDS 1) sépare aussi à l'extrême droite les populations Beja Hadendoa, Arabes Rashaida et, dans une moindre mesure, les Imazighen d'Amizmiz.
L'axe 2 (MDS2) sépare les populations syriennes des autres populations (en haut de la Figure) et sépare aussi les deux populations Peuls du Mali (en bas) et du Sénégal (au milieu).

19. *Multi Dimensional Scaling.*
20. Rappel : Le stress d'une MDS est une mesure de la fiabilité de cette dernière. Une règle empirique pour interpréter ce stress est la suivante : stress < 0.05 : excellente représentation, < 0.1 : acceptable, < 0.2 : faible et 0.4 : très mauvaise. Voir Chapitre 1 pour le détail du calcul du stress.

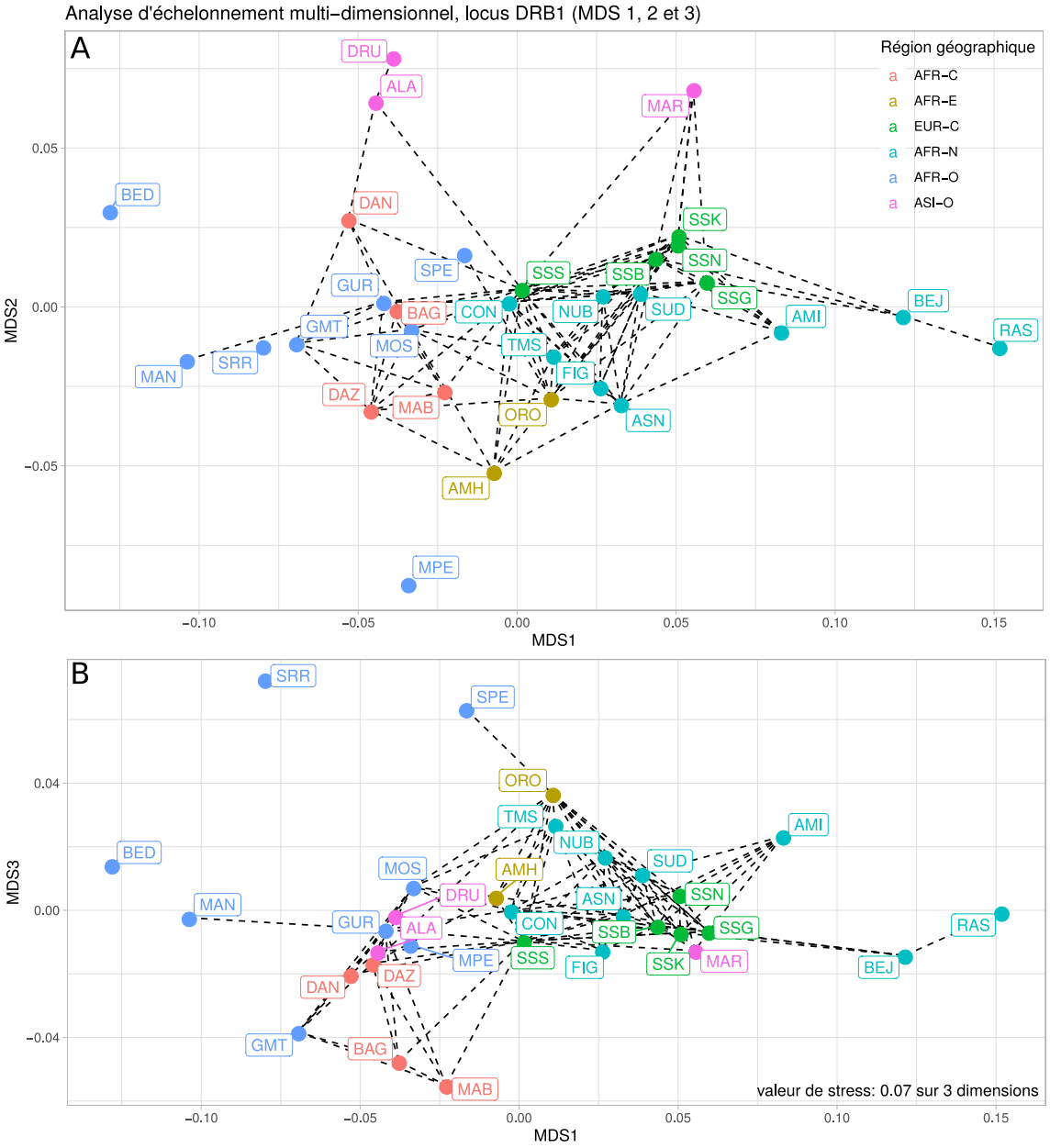


FIGURE 4.20 – Représentation graphique des axes 1 et 2 (Figure A) et des axes 1 et 3 (Figure B) de l'analyse d'échelonnement multi-dimensionnel (MDS) basée sur les distances de Reynolds Θ_w entre les populations, pour les séquences d'exons 2 de HLA-DRB1. Les lignes en pointillés correspondent à des différences (Θ_w) non significativement différentes de zéro. Les noms courts des populations correspondent à : SRR : Sénégal-Sérère ;MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ;SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

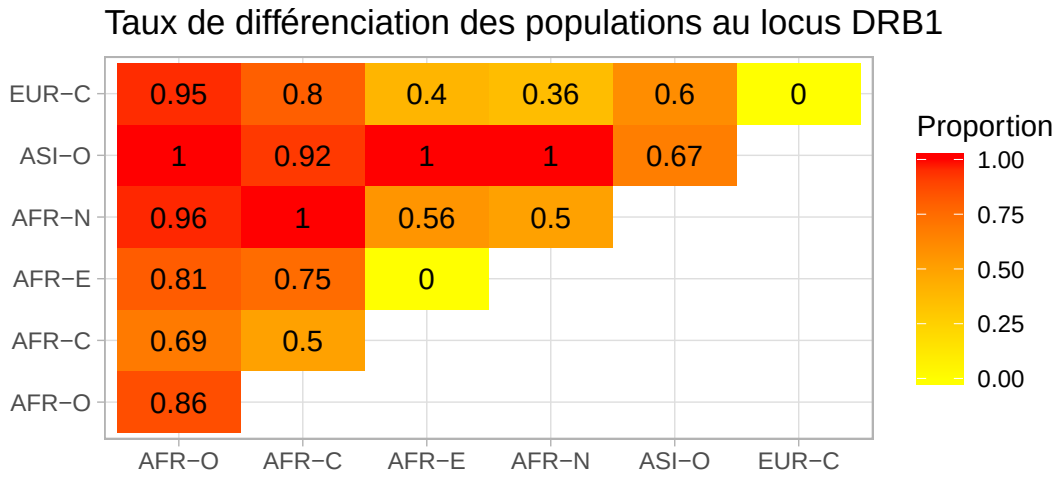


FIGURE 4.21 – Proportions de paires de populations présentant un Θ_w significativement différent de zéro dans et entre chaque région géographique de l'étude pour le locus DRB1-Exon2. AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

L'axe 3 (MDS3) différencie principalement les populations d'Afrique de l'ouest, avec les Sérère et Peuls du Sénégal en haut en les Gourmantché du Burkina Faso en bas.

Beaucoup de Θ_w ne sont pas significativement différents de zéro, ce qui est représenté par les lignes en pointillés. Seuls les Bédik et Sérère du Sénégal ainsi que les Peuls du Mali montrent des distances significativement supérieures à zéro avec toutes les autres populations.
La Figure 4.21 montre les proportions de paires de populations présentant un Θ_w significativement différent selon les régions d'où proviennent les populations. L'Afrique de l'ouest montre le plus de paires de populations significativement différentes à l'intérieur de la région (en bas à gauche sur la diagonale), tandis que l'Afrique de l'est et l'Europe centrale ne présentent aucune différence entre leurs populations à l'intérieur des régions. L'Asie de l'ouest et l'Afrique de l'ouest sont les régions présentant le plus de différences avec les autres régions (minimum de 60% de paires différentes entre l'Asie de l'ouest et l'Europe centrale). De manière générale, pour les populations africaines, la différenciation entre les populations (c'est-à-dire la proportion de paires significativement différentes) montre une composante géographique selon un axe Afrique de l'ouest $\iff$ Afrique centrale $\iff$ Afrique de l'est $\iff$ Afrique du nord, les populations de chaque région étant en moyenne plus proches des régions voisines (sur cet axe) que des régions plus éloignées.

Pour ce locus, l'indice de fixation Φ_{ST} est de 0.052.

La Figure 4.22 représente le résultat de la MDS basée sur les Θ_w calculés entre les 31 populations pour lesquelles le profil génétique DQA1-Exon2 était disponible et la Figure 4.23 représente la proportion de paires de populations présentant un Θ_w significativement différent entre chaque région de l'étude.

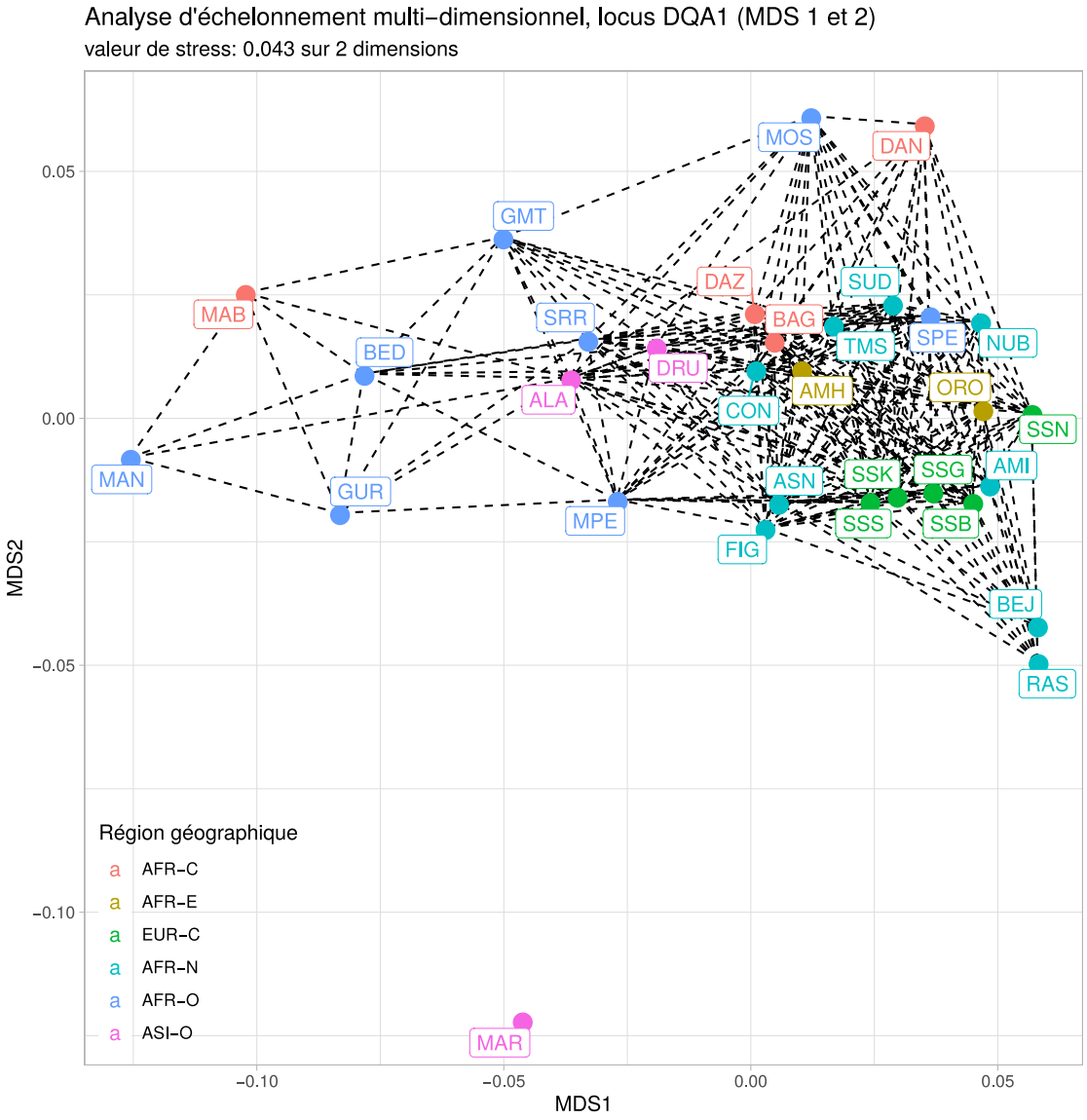


FIGURE 4.22 – Représentation graphique des axes 1 et 2 de l'analyse d'échelonnement multi-dimensionnel (MDS) basée sur les distances de Reynolds Θ_w entre les populations, pour les séquences d'exons 2 de HLA-DQA1. Les lignes en pointillés correspondent à des différences (Θ_w) non significativement différentes de zéro. Les noms courts des populations correspondent à : SRR : Sénégal-Sérère ;MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ;SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

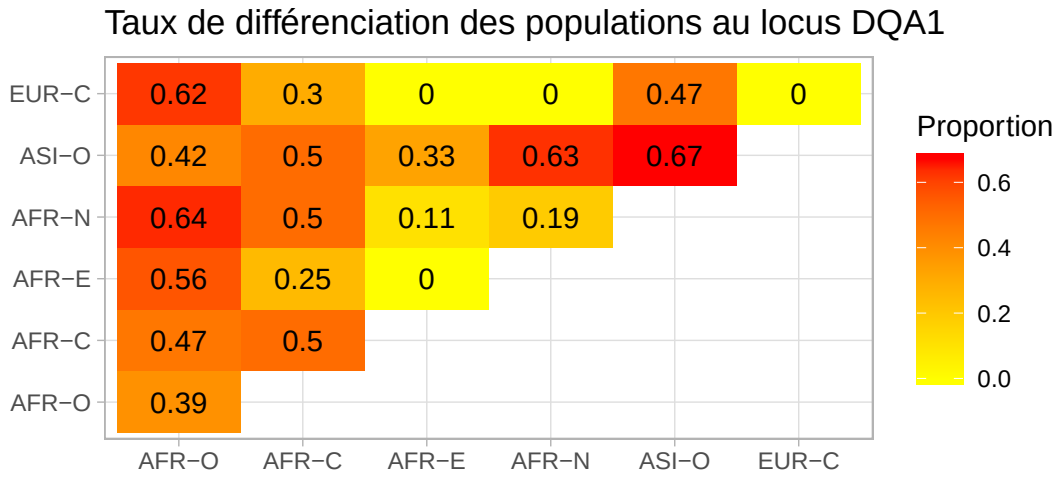


FIGURE 4.23 – Proportions de paires de populations présentant un Θ_w significativement différent de zéro dans et entre chaque région géographique de l'étude pour le locus DQA1-Exon2. AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

Seuls deux axes sont nécessaires pour cette MDS afin d'obtenir une valeur de stress correspondant à une excellente représentation (0.043). Toutefois, peu de structure est observée (beaucoup de différences non significatives).
La MDS1 sépare à gauche les populations d'Afrique de l'ouest des autres populations, à l'exception des Mossi du Burkina Faso et des Peuls du Sénégal (à droite sur la MDS1) ainsi que les Maba du Tchad (à gauche).
La MDS2 sépare principalement les Chrétiens Maronites de Syrie des autres populations (la cause probable étant la fréquence élevée des trois allèles dans cette population : DQA1*1, DQA1*8 et DQA1*4 [21] respectivement de 50, 22 et 17%) et, dans une moindre mesure, cette dimension sépare les Beja Hadendoa et les Arabes Rashaida des autres populations.
Seuls les Chrétiens Maronites de Syrie montrent des distances de Reynolds significativement différentes de zéro avec l'ensemble des populations.

Pour ce locus, l'indice de fixation Φ_{ST} est de 0.036.

La Figure 4.23, est l'analogue de la Figure 4.21 pour DQA1-Exon2 et montre aussi une structure géographique mais moins marquée. Sur l'axe Afrique de l'ouest $\iff$ Afrique centrale $\iff$ Afrique de l'est $\iff$ Afrique du nord les populations de régions voisines sont plus similaires que les populations de régions plus éloignées, mais cet effet est moins visible que dans la Figure 4.21. Les populations d'Afrique de l'est et d'Europe centrale ne montrent pas de différences à l'intérieur de ces régions.

La Figure 4.24 représente le résultat de la MDS basée sur les Θ_w calculés entre les 27 populations pour lesquelles le profil génétique DQB1-Exon2 était disponible et la Figure 4.25 représente la proportion de paires de populations présentant un Θ_w significativement différent entre chaque région de l'étude.

21. Ces trois séquences pouvant correspondre respectivement aux exons 2 des allèles nominaux HLA-DQA1*05:01, DQA1*02:01 et DQA1*03:01.

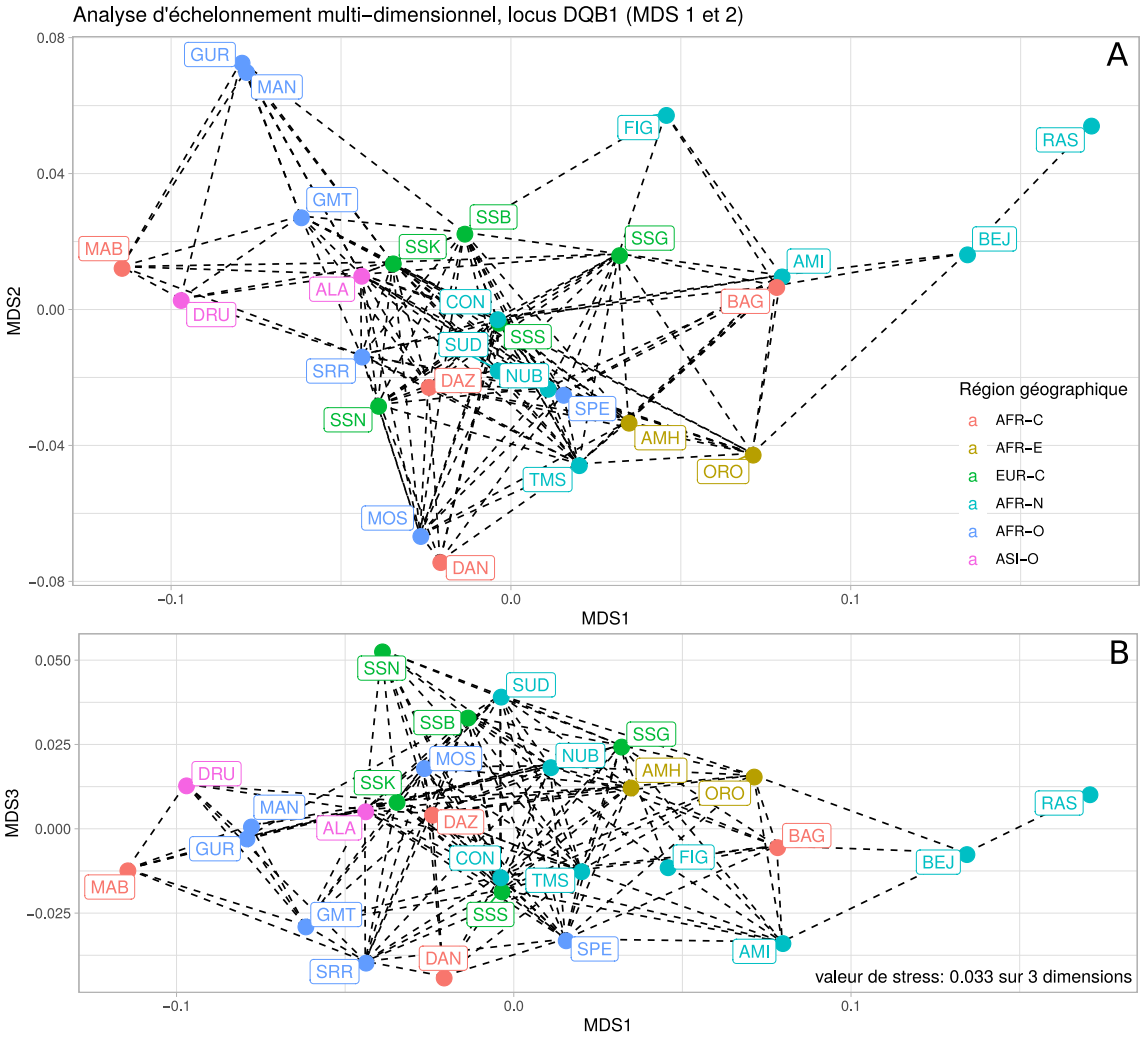


FIGURE 4.24 – Représentation graphique des axes 1 et 2 (Figure A) et des axes 1 et 3 (Figure B) de l'analyse d'échelonnement multi-dimensionnel (MDS) basée sur les distances de Reynolds Θ_w entre les populations, pour les séquences d'exons 2 de HLA-DQB1. Les lignes en pointillés correspondent à des différences (Θ_w) non significativement différentes de zéro. Les noms courts des populations correspondent à : SRR : Sénégal-Sérère ;MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ;SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

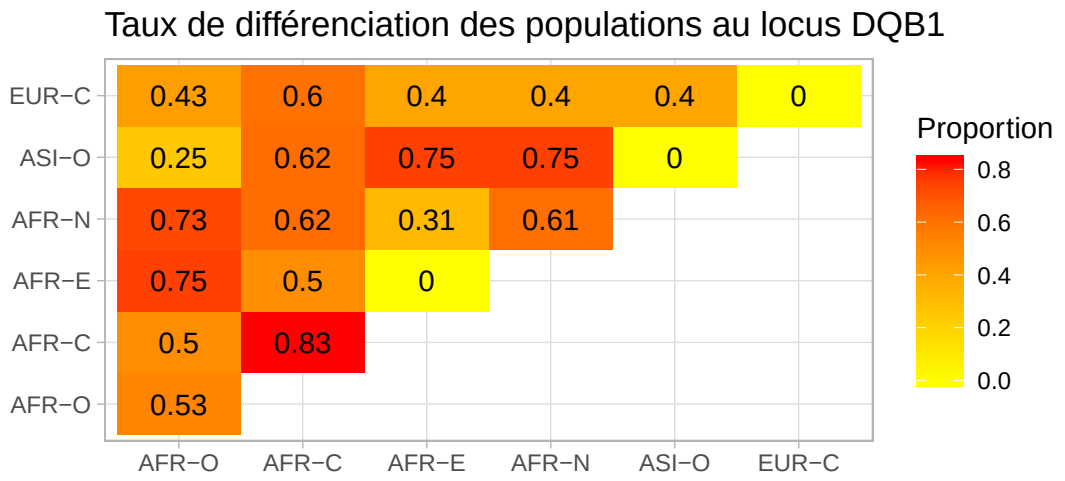


FIGURE 4.25 – Proportions de paires de populations présentant un Θ_w significativement différent de zéro dans et entre chaque région géographique de l'étude pour le locus DQB1-Exon2. AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

Trois axes sont nécessaires pour cette MDS afin d'obtenir une valeur de stress excellente (0.033). La même MDS sur seulement deux axes présente une valeur de stress de 0.081. Aucune structuration géographique n'est observable, à priori, dans cette analyse et aucune population ne montre de distance significativement différente de zéro avec toutes les autres populations.
Les Beja Hadendoa, les Arabes Rashaida et les Imazhigen de Figuig s'éloignent des autres populations (Figure 4.24, en haut à droite). Les Gourounsi du Burkina Faso et les Mandenkalu du Sénégal sont très proches (en haut à gauche sur la Figure 4.24.A) et éloignés des autres populations sur la MDS2, bien que les distances avec les populations les plus proches (sur la MDS) ne soient pas significativement différentes de zéro.

Pour ce locus, l'indice de fixation Φ_{st} est de 0.044.

La Figure 4.25 est l'analogue de la Figure 4.21 pour DQB1-Exon2 mais ne montre, quant à elle, aucune structure géographique, principalement car les populations sont plus dissimilaires à l'intérieur des régions géographiques qu'entre les régions voisines.

La Figure 4.26 représente le résultat de la MDS basée sur les Θ_w calculés entre les 31 populations pour lesquelles le profil génétique DPB1-Exon2 était disponible et la Figure 4.27 représente la proportion de paires de populations présentant un Θ_w significativement différent entre chaque région de l'étude.

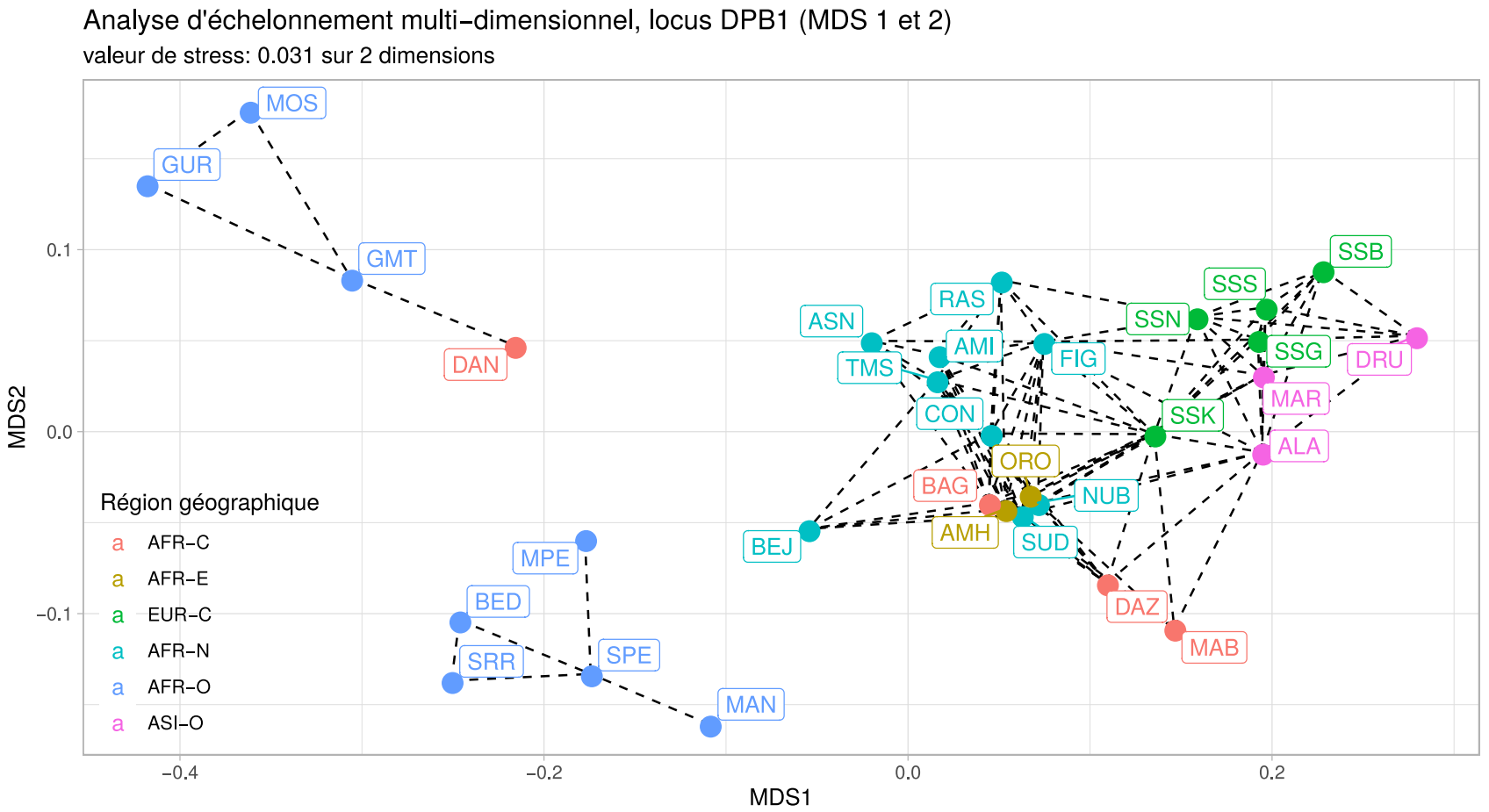


FIGURE 4.26 – Représentation graphique des axes 1 et 2 de l'analyse d'échelonnement multi-dimensionnel (MDS) basée sur les distances de Reynolds Θ_w entre les populations, pour les séquences d'exons 2 de HLA-DPB1. Les lignes en pointillés correspondent à des différences (Θ_w) non significativement différentes de zéro. Les noms courts des populations correspondent à : SRR : Sénégal-Sérère ;MAN : Sénégal-Mandenka ; BED : Sénégal-Bedik ; SPE : Sénégal-Peul ; MOS : BurkinaFaso-Mossi ; GUR : BurkinaFaso-Gourounsi ; GMT : BurkinaFaso-Gourmantché ; BAG : Tchad-ArabeBaggara ; DAN : Tchad-Dangaléat ; DAZ : Tchad-Daza ; MAB : Tchad-Maba ; ORO : Ethiopie-Oromo ; AMH : Ethiopie-Amhara-(Keketeya) ; NUB : Soudan-Nubien ; SUD : Sudan-ArabeSoudanais ; RAS : Soudan-ArabeRashaida ; BEJ : Soudan-BejaHadendoa ; AMI : Maroc-Amazigh-(Amizmiz) ; FIG : Maroc-Amazigh-(Figuig) ; TMS : Algérie-(Tamanrasset) ; CON : Algérie-(Constantine) ; ALA : Syrie-Alaouite ; DRU : Syrie-Druze ; SSK : Slovaquie-(Skalica) ; SSG : Slovaquie-(Galanta) ; SSB : Slovaquie-(NovaBana) ; SSN : Slovaquie-(Namestovo) ;SSS : Slovaquie-(StaraLubovna). AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

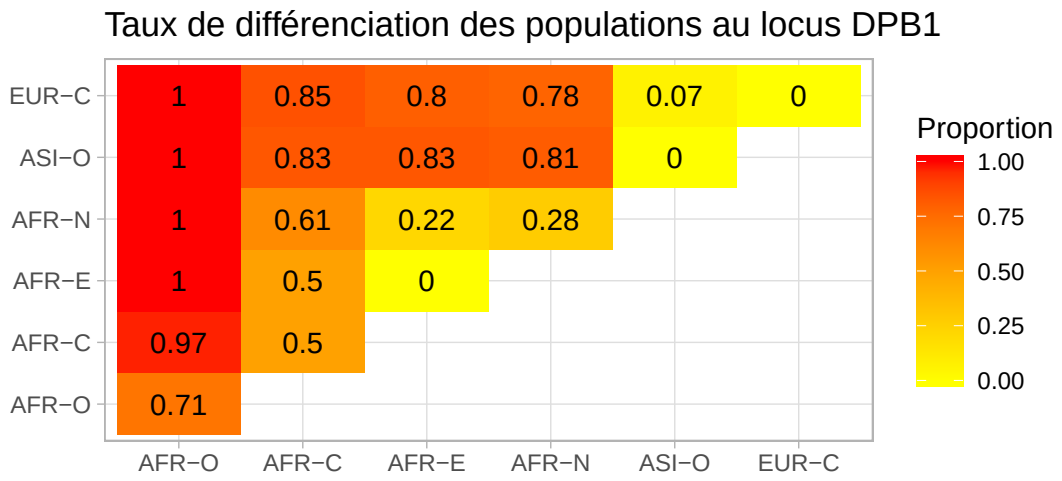


FIGURE 4.27 – Proportions de paires de populations présentant un Θ_w significativement différent de zéro dans et entre chaque région géographique de l'étude pour le locus DPB1-Exon2. AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

Cette MDS présente un valeur de stress excellente (0.031) sur uniquement deux dimensions. Il s'agit du locus différenciant le plus les populations avec un indice de fixation Φ_{st} de 0.108, ce qui se traduit sur la Figure 4.26 par des axes de la MDS très étendus (MDS1 va de -0.418 à 0.279 et la MDS2 va de -0.160 à 0.177).

Trois groupes sont observables sur cette figure :
— En haut à gauche, un groupe composé des trois populations burkinabé (Mossi, Gourounsi et Gourmantché) et des Dangaléat du Tchad ;
— En bas à gauche, un groupe composé des populations sénégalaises (Bédik, Peuls, Sérères et Mandenka) et des Peuls du Mali ;
— À droite, un groupe composé des 22 autres populations et présentant une certaine structuration géographique, la MDS1 séparant (de gauche à droite) les populations d'Afrique du nord des populations syriennes et slovaques tandis que la MDS2 sépare principalement les Daza et Maba du Tchad des autres populations. Les populations éthiopiennes (Amhara et Oromo) ainsi que les Nubiens du Soudan, Arabes Soudanais et Arabes Baggara du Tchad sont toutes très proches.

La Figure 4.27 donne les proportions de paires de populations significativement différentes entre les régions géographiques. Si les populations d'Afrique de l'est, Asie de l'ouest et d'Europe centrale ne montrent aucune différence à l'intérieur de ces régions, l'Afrique de l'ouest montre le plus de différences avec les autres régions en plus d'une forte différenciation à l'intérieur de l'Afrique de l'ouest, puisque les Dangaléat du Tchad et Gourmantché du Burkina Faso sont la seule paire de populations à montrer une différence non significativement différente de zéro (Figure 4.26). À l'instar de la Figure 4.21, on observe aussi une forte structuration géographique selon l'axe Afrique de l'ouest $\iff$ Afrique centrale $\iff$ Afrique de l'est $\iff$ Afrique du nord.

La Figure 4.28 montre, pour chaque locus, la proportions de paires de populations présentant un Θ_w significativement différent selon la région géographique d'où provient

l'une des populations de la paire. Seules les paires de populations composées de populations provenant de deux régions différentes ont été utilisées ici, les paires de populations d'une même région géographique n'ont pas été considérées dans cette analyse.

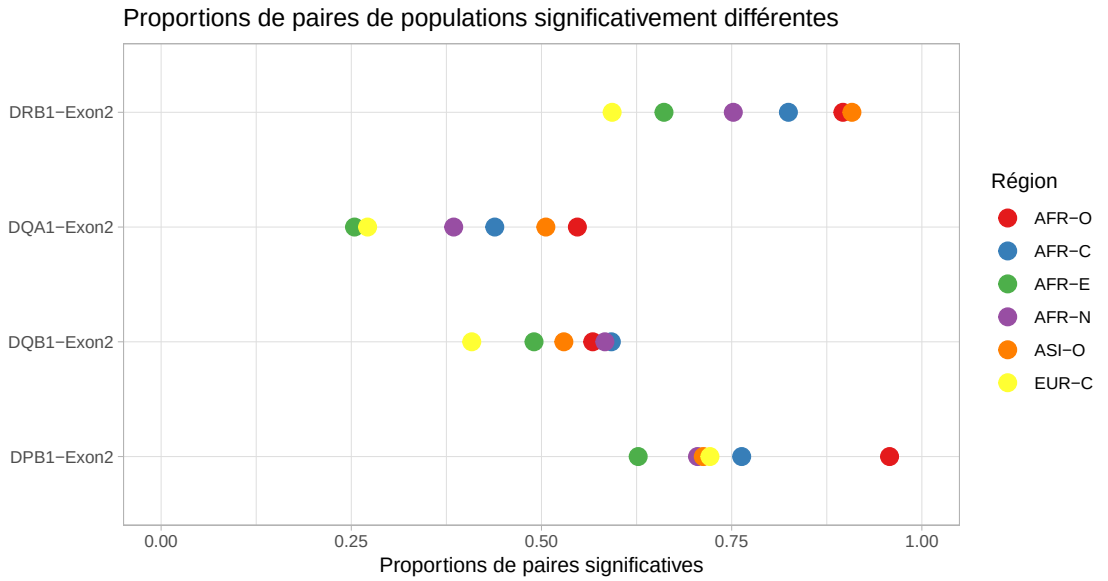


FIGURE 4.28 – Proportions de paires de populations présentant un Θ_w significativement différent selon l'origine d'une des populations de la paire et le locus considéré. Les paires correspondant à des populations d'une même région géographique n'ont pas été utilisées. AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est ; AFR-N : Afrique du nord ; ASI-O : Asie de l'ouest ; EUR-C : Europe centrale.

Les résultats de la Figure 4.28 apportent une précision numérique à ceux des Figures 4.20 à 4.27. On observe une différence entre les loci et entre les régions, les loci DRB1-Exon2 et DPB1-Exon2 montrant les plus grandes proportions de paires de populations significativement différentes (DQA1-Exon2 et DQB1-Exon2 montrant le moins de différences). Concernant les régions, l'Afrique de l'ouest est toujours la région montrant le plus de différences avec les autres régions, l'Afrique de l'est et l'Europe centrale étant les régions montrant le moins de différences (à l'exception de DPB1-Exon2 où l'Europe centrale montre des valeurs similaires à l'Afrique du nord et l'Asie de l'ouest).

3.10 Analyse de Variance Moléculaire

La Figure 4.29 montre le résultat de l'analyse de variance moléculaire (AMOVA) des quatre loci de l'étude dans les populations africaines et la Table 4.15 donne les différents groupes auxquels sont rattachées les populations.

Population	Région géographique	Famille linguistique	Modes de vie	Exposition au *P. falciparum*
Sénégal-Peul	AFR-O	NC	n	Oui
Sénégal-Mandenka	AFR-O	NC	s	Oui
Sénégal-Sérère	AFR-O	NC	s	Oui
BurkinaFaso-Gourmantché	AFR-O	NC	s	Oui
BurkinaFaso-Gourounsi	AFR-O	NC	s	Oui
BurkinaFaso-Mossi	AFR-O	NC	s	Oui
Tchad-ArabeBaggara	AFR-C	AA	n	Oui
Tchad-Dangaléat	AFR-C	AA	s	Oui
Tchad-Daza	AFR-C	NS	sn	Non
Tchad-Maba	AFR-C	NS	s	Oui
Ethiopie-Amhara-(Keketeya)	AFR-E	AA	s	Non
Ethiopie-Oromo	AFR-E	AA	s	Non
Soudan-BejaHadendoa	AFR-N	AA	n	Oui
Soudan-Nubien	AFR-N	NC	s	Non
Soudan-ArabeRashaida	AFR-N	AA	sn	Oui
Sudan-ArabeSoudanais	AFR-N	AA	s	Non
Algérie-(Constantine)	AFR-N	AA	s	Non
Algérie-(Tamanrasset)	AFR-N	AA	n	Non
Maroc-Amazigh-(Amizmiz)	AFR-N	AA	sn	Non
Maroc-Amazigh-(Figuig)	AFR-N	AA	sn	Non

TABLE 4.15 – Table résumant, pour chacune des catégories testées pour l'AMOVA, les groupes auxquel appartiennent les populations. Pour les régions géographiques les acronymes sont les suivants : AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est et AFR-N : Afrique du nord. Les familles linguistiques sont NC : niger-congo ; AA : afro-asiatique et NS : nilo-saharien. Les modes de vie sont s : sédentaires ; sn : semi-nomades et n : nomades. La colonne « Exposition au *P. falciparum* » précise si la prévalence du *Plasmodium falciparum*, en l'an 2000, sur le lieu d'échantillonnage était supérieure (oui) ou inférieure (non) à 5% (voir pages 182 et 226).

Pour les **régions géographiques**, DPB1-Exon2 montre le plus de variance à chaque stratification ($\Phi_{ST} = 0.102$, $\Phi_{SC} = 0.059$ et $\Phi_{CT} = 0.046$). DRB1-Exon2 montre plus de variance entre les groupes ($\Phi_{CT} = 0.035$) qu'entre les populations dans les groupes ($\Phi_{SC} = 0.029$). DQA1-Exon2 ne montre aucune structure liée aux groupes (Φ_{CT} non significativement différent de zéro).

Concernant les **familles linguistiques**, DPB1-Exon2 montre le plus de variance à chaque stratification ($\Phi_{ST} = 0.109$, $\Phi_{ST} = 0.063$ et $\Phi_{CT} = 0.049$). DQA1-Exon2 ne montre pas de structuration liée aux groupes (Φ_{CT} non significativement différent de zéro). DRB1-Exon2 et DQB1-Exon2 montrent leurs plus grandes variances entre les populations (les Φ_{ST} sont respectivement de 0.060 et 0.048), suivies de la variance

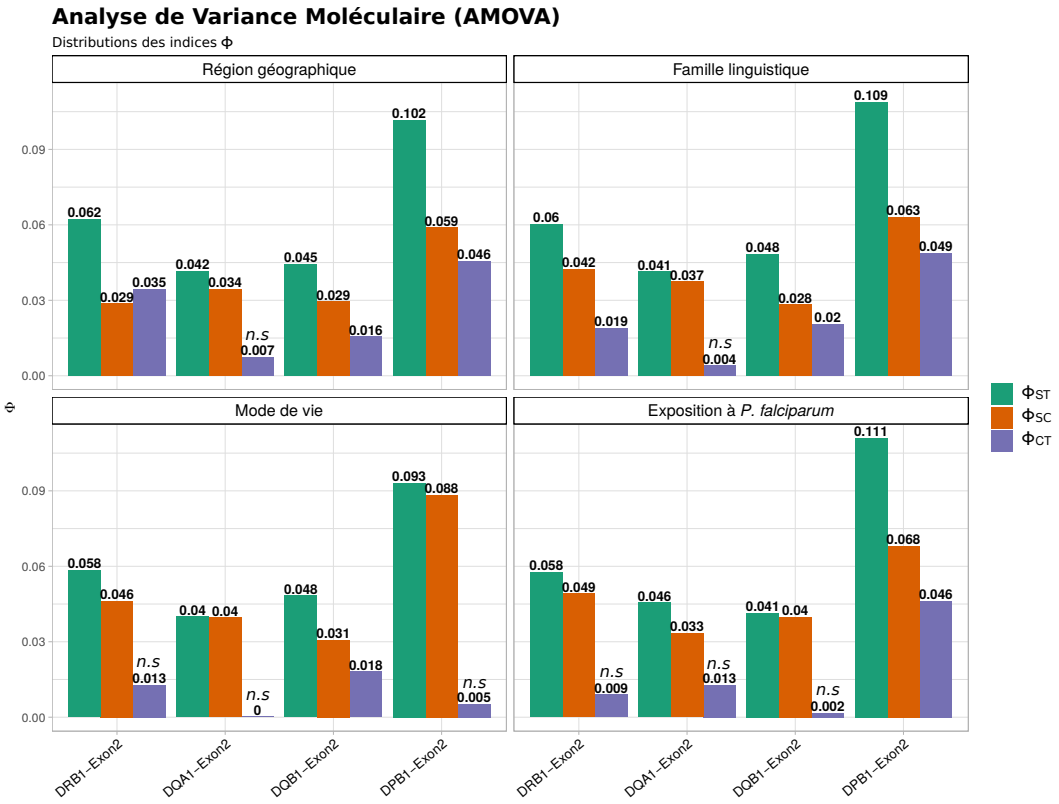


FIGURE 4.29 – Distributions des valeurs de Φ_{ST}, Φ_{SC} et Φ_{CT} selon les différents critères de catégorisation : la région géographique (quatre groupes), la famille linguistique (trois groupes), le mode de vie (trois groupes) ou l'exposition à la malaria (deux groupes) et pour chacun des quatre loci de l'étude. Se référer à la Table 4.15 pour le détail des différentes catégories. Les Φ non significatifs sont signalés par "*n.s*".

entre les populations dans les groupes (Φ_{SC} respectivement de 0.042 et 0.028) puis de la variance entre les groupes (Φ_{CT} respectivement de 0.019 et 0.020).

Pour les **modes de vie**, DRB1-Exon2, DQA1-Exon2 et DPB1-Exon2 montrent une importance variance entre les populations et entre les populations dans les groupes (Φ_{ST} de respectivement 0.058, 0.040 et 0.093 et Φ_{SC} de respectivement 0.046, 0.040 et 0.088) mais aucune structuration liée aux groupes (Φ_{CT} non significativement différent de zéro). Seul DQB1-Exon2 un Φ_{CT} significatif (0.018), mais inférieur au Φ_{SC} (0.031).

Finalement, pour **l'exposition au *P. falciparum***, DPB1-Exon2 montre encore le plus de variance à chaque stratification ($\Phi_{ST} = 0.111$, $\Phi_{SC} = 0.068$ et $\Phi_{CT} = 0.046$) et est le seul locus à montrer un Φ_{CT} significatif (les Φ_{CT} pour les trois autres loci n'étant pas significativement différents de zéro). Les quatre loci, DRB1-Exon2, DQA1-Exon2, DQB1-Exon2 et DPB1-Exon2 ont toujours une plus importante variabilité entre les populations (Φ_{ST} respectivement de 0.058, 0.046, 0.041 et 0.111) qu'entre les populations dans les groupes (Φ_{SC} respectivement de 0.049, 0.033, 0.040 et 0.068).

En conclusion, DPB1-Exon2 montre toujours la variance la plus élevée à chaque stratification à l'exception du mode de vie où ce locus ne montre pas de variance liée aux

groupes (DQB1-Exon2 étant le seul locus à montrer un Φ_{CT} significativement différent de zéro pour cette catégorie). DPB1-Exon2 est le seul locus à montrer un Φ_{CT} significatif concernant l'exposition au *P. falciparum*. La variance la plus élevée est toujours entre les populations (DQA1-Exon2 montrant toutefois des Φ_{ST} et Φ_{SC} similaires pour la catégorie des modes de vie) et la variance la plus faible est toujours entre les groupes, à l'exception de DRB1-Exon2 pour les régions géographiques, seul locus à montrer un $\Phi_{CT} > \Phi_{SC}$.

3.11 Test de Mantel

La Table 4.16 présente le résultat des tests de Mantel réalisés pour évaluer la corrélation de Pearson entre les distances de Reynolds et les distances géographiques entre les populations. Cette corrélation a été testée avec des distances calculées pour chacun des trois jeux nucléotidiques des quatre loci (exon 2, uniquement les codons ARS et uniquement les codons non-ARS). Les pValeurs (significativité des coefficients) ont été estimées par 9'999 permutations.

Locus	**Jeu nucléotidique**	**Corrélation (r) observée**	**pValeur associée**
DRB1-Exon2	Exon 2	-0.053	0.792
DQA1-Exon2		0.046	0.235
DQB1-Exon2		0.027	0.349
DPB1-Exon2		-0.093	0.926
DRB1-Exon2	ARS	0.033	0.304
DQA1-Exon2		0.053	0.207
DQB1-Exon2		0.024	0.365
DPB1-Exon2		-0.084	0.909
DRB1-Exon2	non-ARS	-0.038	0.708
DQA1-Exon2		0.012	0.424
DQB1-Exon2		0.075	0.161
DPB1-Exon2		-0.039	0.738

TABLE 4.16 – Résultats du test de Mantel testant la corrélation (r de Pearson) entre les distances de Reynolds et les distances géographiques des populations africaines. Pour chacun des quatre loci de l'étude, trois jeux nucléotidique ont été testés : l'exon 2 complet (Exon2), uniquement les codons codant pour le site de reconnaissance de l'antigène (ARS) et uniquement les codons ne codant pas pour le site de reconnaissance de l'antigène (non-ARS). Les pValeurs associées ont été obtenues avec 9'999 permutations.

Aucune valeur de corrélation n'est significative au seuil de 5%. Il n'y a donc pas de corrélation entre les distances génétiques et les distances géographiques pour les populations africaines de l'étude, quel que soit le locus ou le jeu nucléotidique considéré.

3.12 Association des fréquences alléliques avec la prévalence de la malaria

La Figure 4.30 est une carte représentant la distribution géographique de la prévalence de *P. falciparum* aux localisations dans lesquelles vivent les 20 populations de l'étude.

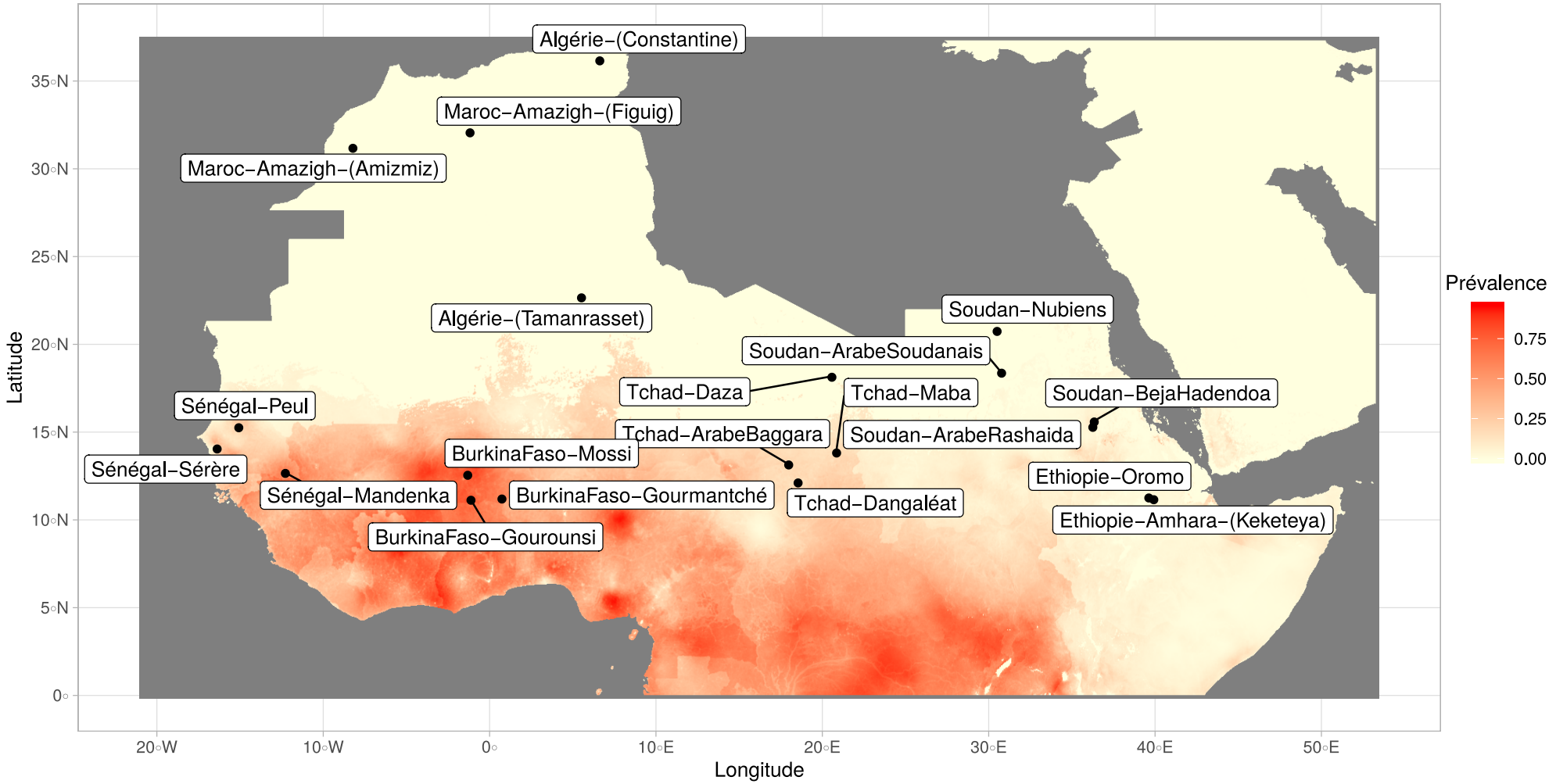


FIGURE 4.30 – Distribution de la prévalence de *Plasmodium falciparum* en l'an 2000 dans les régions de l'étude, selon les données du Malaria-Atlas-Project [Bhatt et al., 2015]. Les données pour la Tunisie, la Lybie et l'Égypte ne sont pas disponibles.

La Figure 4.30 montre une forte composante régionale dans la prévalence de *P. falciparum* en Afrique, l'Afrique de l'ouest étant la région la plus touchée, suivie de l'Afrique centrale et, dans une moindre mesure, de l'Afrique de l'est (cette région étant aussi touchée par le *Plasmodium vivax* [Bhatt et al., 2015], non étudié ici). En Afrique centrale, seuls les Daza du Tchad ne vivent pas dans une zone touchée par *P. falciparum*. Deux populations d'Afrique du nord (Soudan) sont aussi touchées : les Arabes Rashaida et les Beja Hadendoa. Pour les populations nomades, la valeur de $pfpr2000$ correspond à celle du lieu d'échantillonnage et peut ne pas représenter l'exposition réelle de ces populations au parasite.

La Figure 4.31 montre les différents coefficients de corrélation estimés pour les variables pour lesquelles au moins un des coefficients est supérieur ou égal à 0.3 (corrélation positive et significative entre la $pfpr2000$ et la fréquence de l'allèle).

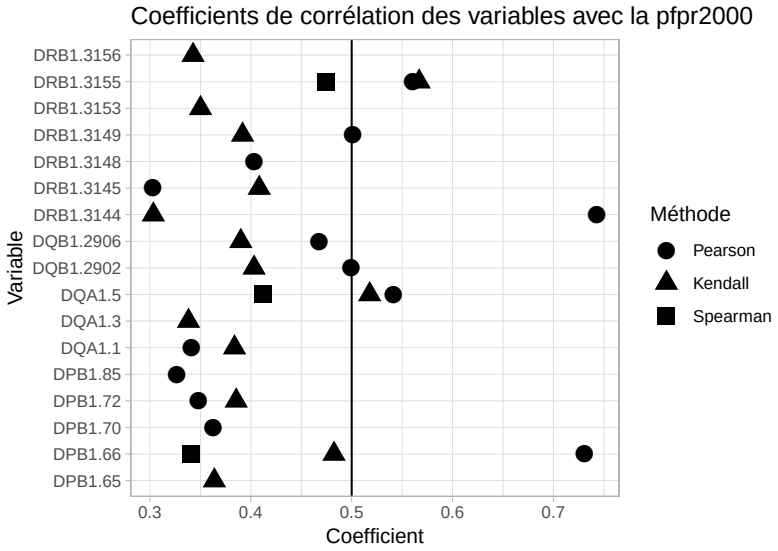


FIGURE 4.31 – Distribution des coefficients de corrélation (ronds : r de Pearson ; triangles : τ de Kendall ; carrés : ρ de Spearman) entre les variables (fréquences corrigées pour la géographie, voir page 182) et la prévalence de *P. falciparum* en l'an 2000. Par souci de clarté, seules les variables avec au moins un coefficient supérieur à 0.3 sont représentées. La barre verticale à $x = 0.5$ représente le seuil choisi pour considérer la corrélation comme significative (similaire au protocole décrit par [Sanchez-Mazas et al., 2017]).

Suivant le protocole décrit par Sanchez-Mazas *et al.* 2017, seules les variables présentant au moins un coefficient de corrélation supérieur à 0.5 seront conservées pour la suite des analyses [Sanchez-Mazas et al., 2017], à l'exception de DQB1.2902 dont le r de Pearson est inférieur à 0.5 mais très proche (0.499). Seul DRB1.3155 montre un τ de Kendall supérieur à 0.5, et plusieurs variables (DRB1.3144, DRB1.3149, DQB1.2902) ne montrent qu'un seul coefficient supérieur à 0.5. Deux variables, DRB1.3144 et DPB1.66 montrent des valeurs élevées de corrélation (coefficient de Spearman, $\rho > 0.75$).

La Figure 4.32 montre la relation entre les fréquences (non corrigées pour la géographie) des séquences retenues à l'étape précédente et la prévalence de *P. falciparum*.

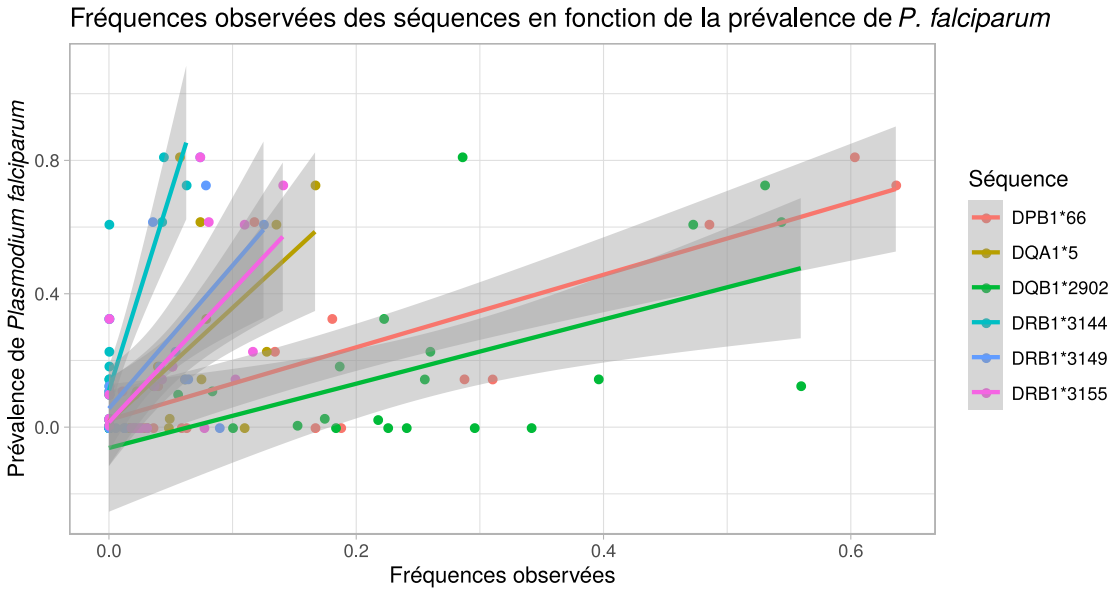


FIGURE 4.32 – Relation, pour les allèles les plus fortement associés (présentant au moins un des coefficients de corrélation supérieur ou égal à 0.5), entre la prévalence de *P. falciparum* et les fréquences des allèles.

Les six séquences retenues comme ayant une fréquence (corrigée pour la géographie) corrélée avec la prévalence du *P. falciparum* sont DRB1*3144, DRB1*3149, DRB1*3155, DQA1*5, DQB1*2902 et DPB1*66. Les quatre premières (DRB1*3144, DRB1*3149, DRB1*3155 et DQA1*5) sont toutes à des fréquences inférieures à 0.2, au contraire des deux dernières (DQB1*2902 et DPB1*66) qui ont des fréquences plus élevées (supérieures à 50%), notamment chez les Gourounsi et Mossi du Burkina Faso, les Mandenkalu du Sénégal ainsi que chez les Maba du Tchad. La Table 4.17 donne, pour chacune des populations, la prévalence de *P. falciparum* ainsi que les fréquences observées des six allèles identifiés comme associés à la malaria.

Population	**Région**	**Dist. à Addis A. (m)**	$pfpr$ 2000	**DRB1*3144** HLA-DRB1*08	**DRB1*3149** HLA-DRB1*03	**DRB1*3155** HLA-DRB1*11	**DQA1*5** HLA-DQA1*04	**DQB1*2902** HLA-DQB1*03	**DPB1*66** HLA-DPB1*01
Sénégal-Sérère	AFR-O	6'027'952	0.15	0.00	0.06	0.04	0.07	0.40	0.29
Sénégal-Mandenka	AFR-O	5'587'076	0.62	0.04	0.04	0.08	0.07	0.54	0.12
Sénégal-Peul	AFR-O	5'886'811	0.23	0.00	0.05	0.12	0.13	0.26	0.13
BurkinaFaso-Mossi	AFR-O	4'396'430	0.81	0.04	0.07	0.07	0.06	0.29	0.60
BurkinaFaso-Gourounsi	AFR-O	4'374'710	0.73	0.06	0.08	0.14	0.17	0.53	0.64
BurkinaFaso-Gourmantché	AFR-O	4'171'516	0.61	0.00	0.12	0.11	0.14	0.47	0.48
Tchad-ArabeBaggara	AFR-C	2'315'664	0.18	0.00	0.04	0.05	0.04	0.19	0.05
Tchad-Dangaléat	AFR-C	2'238'961	0.15	0.00	0.06	0.10	0.06	0.26	0.31
Tchad-Daza	AFR-C	2'210'506	0.00	0.00	0.01	0.08	0.00	0.34	0.01
Tchad-Maba	AFR-C	2'024'661	0.13	0.00	0.00	0.04	0.04	0.56	0.04
Ethiopie-Oromo	AFR-E	268'134	0.01	0.00	0.00	0.00	0.00	0.15	0.00
Ethiopie-Amhara-(Keketeya)	AFR-E	271'572	0.03	0.00	0.00	0.00	0.05	0.17	0.00
Soudan-Nubien	AFR-N	1'574'299	0.00	0.00	0.02	0.03	0.05	0.23	0.03
Soudan-ArabeSoudanais	AFR-N	1'349'372	0.02	0.00	0.00	0.00	0.00	0.22	0.00
Soudan-ArabeRashaida	AFR-N	747'471	0.10	0.00	0.00	0.00	0.00	0.06	0.02
Soudan-BejaHadendoa	AFR-N	775'721	0.11	0.00	0.00	0.01	0.02	0.08	0.00
Maroc-Amazigh-(Amizmiz)	AFR-N	5'437'489	0.00	0.00	0.00	0.02	0.03	0.10	0.06
Maroc-Amazigh-(Figuig)	AFR-N	4'841'325	0.00	0.00	0.01	0.03	0.02	0.24	0.06
Algérie-(Tamanrasset)	AFR-N	3'854'895	0.00	0.00	0.09	0.02	0.11	0.18	0.19
Algérie-(Constantine)	AFR-N	4'428'760	0.00	0.00	0.01	0.02	0.00	0.30	0.04

TABLE 4.17 – Pour chaque population, détail de la région géographique ; Dist. Addis A. (m) : distance (en mètres) à Addis-Abeba ; $pfpr2000$: prévalence du *P. falciparum* dans la zone d'échantillonnage ; colonnes suivantes : fréquences des allèles identifiés comme associés à la malaria. AFR-O : Afrique de l'ouest ; AFR-C : Afrique centrale ; AFR-E : Afrique de l'est et AFR-N : Afrique du nord.

Au vu de l'importante association entre DPB1-Exon2 et $pfpr2000$, des modèles linéaires étudiant la relation entre la $pfpr2000$, l'indice D de Tajima, le Θ_π et le Θ_S à chaque population ont été menés pour ce locus. Les résultats des modèles pour lesquels les coefficients sont significatifs et les résidus normalement distribués (test de Shapiro, seuil $\alpha = 0.05$) sont présentés dans les Équations 4.1 à 4.3 .

Codons ARS de DPB1-Exon2 :

$$Tajima.D^{ARS} = -1.56 \cdot pfpr2000 + 2.90 \tag{4.1}$$

ÉQUATION 4.1 – Résultat du modèle linéaire étudiant la relation entre la prévalence du *Plasmodium falciparum* en l'an 2000 ($pfpr2000$) et l'indice du D de Tajima pour les codons ARS de DPB1-Exon2 chez les 20 populations étudiées dans cette analyse.

$$\Theta_\pi^{ARS} = -1.03 \cdot pfpr2000 + 3.41 \tag{4.2}$$

ÉQUATION 4.2 – Résultat du modèle linéaire étudiant la relation entre la prévalence du *Plasmodium falciparum* en l'an 2000 ($pfpr2000$) et l'indice du Θ_π pour les codons ARS de DPB1-Exon2 chez les 20 populations étudiées dans cette analyse.

Codons non-ARS de DPB1-Exon2 :

$$\Theta_\pi^{non-ARS} = -1.41 \cdot pfpr2000 + 5.14 \tag{4.3}$$

ÉQUATION 4.3 – Résultats des modèles linéaires étudiant la relation entre la prévalence du *Plasmodium falciparum* en l'an 2000 ($pfpr2000$) et l'indice du Θ_π pour les codons non-ARS de DPB1-Exon2 chez les 20 populations étudiées dans cette analyse.

Les Équations 4.1 à 4.3 montrent que, pour DPB1-Exon2, la pfpr2000 est négativement associée avec le D de Tajima pour les codons ARS, ainsi qu'avec le Θ_π. Cela indique que, pour DPB1-Exon2, une prévalence plus élevée de *P. falciparum* est associée à un D de Tajima réduit aux codons ARS, dû à une plus faible diversité nucléotidique des codons ARS.
Pour les codons non-ARS, seul le Θ_π montre une association significative avec la $pfpr2000$.

4 Discussion

4.1 Résumé des résultats obtenus

Cette étude portait sur les résultats de séquençage des exons 2 de quatre loci HLA de classe II (HLA-DRB1, -DQA1, -DQB1 et -DPB1) pour 2'061 individus appartenant à 26 populations d'Afrique de l'ouest, centrale, de l'est et du nord ainsi que 10 populations d'Asie de l'ouest et d'Europe centrale, à titre de comparaison.

Les estimations d'hétérozygotie ont montré une hétérozygotie plus élevée pour DRB1-Exon2 que pour les autres loci, et quelques populations montrent des hétérozygoties particulièrement réduites aux loci DQA1-Exon2 (Mandenka et Bédik du Sénégal, Maba du Tchad et Chrétiens Maronites de Syrie), DQB1-Exon2 (Mandenka du Sénégal, Gourounsi du Burkina Faso, Maba du Tchad et Arabes Rashaida du Soudan) et DPB1-Exon2 (Mossi et Gourounsi du Burkina Faso et Arabes Rashaida du Soudan).

Les calculs de richesse allélique ont montré une richesse allélique plus élevée pour DRB1-Exon2 suivi de DPB1-Exon2, DQA1-Exon2 et DQB1-Exon2 montrant des valeurs de richesse allélique plus réduites (et similaires entre les deux loci).

L'analyse des distributions de fréquences alléliques montre d'importantes différences entre les loci. DQA1-Exon2 et DQB1-Exon2 montrent peu de différences liées à la géographie, à l'exception de DQB1*2902 fréquent surtout en Afrique de l'ouest et DQB1*2901 fréquent chez les Imazighen d'Amizmiz (Maroc), Beja Hadendoa et Arabes Rashaida (Soudan).
DRB1-Exon2 montre davantage d'effets liés à la géographie, avec la séquence DRB1*3135 très fréquente dans les populations sénégalaises (à l'exception des Peuls) et retrouvée à l'état de traces chez les Imazighen de Figuig (Maroc), Arabes Rashaida (Soudan) et Peuls du Sénégal.
DPB1-Exon2, quant à lui, montre le plus d'effets liés à la géographie avec notamment la séquence DPB1*64 très fréquente dans les populations sénégalaises et DPB1*66 très fréquente dans les populations burkinabé.

L'étude du déséquilibre de liaison global a mis en évidence des déséquilibres surtout entre les loci DRB1-Exon2∼DQA1-Exon2∼DQB1-Exon2, mais peu de déséquilibre incluant DPB1-Exon2. L'analyse des déséquilibres de liaison haplotypiques a montré plusieurs haplotypes fréquents (fréquence moyenne d'au moins 10%) dans les régions étudiées, à l'exception de l'Afrique de l'ouest où seul l'haplotype DQA1*1∼DQB1*2902 dépasse les 10% de fréquence moyenne (pour la région).

Les analyses de diversité moléculaire ont mis en évidence la plus faible diversité au locus DPB1-Exon2 (comparé aux trois autres loci), DQA1-Exon2 montrant une diversité élevée pour les codons non-ARS, DQB1-Exon2 et DRB1-Exon2 montrant une très grande diversité pour les codons ARS.

Le test du D de Tajima a montré des valeurs de D (significativement différentes de zéro) plus élevées pour les codons ARS (indiquant une sélection balancée) de tous les loci sauf DQA1-Exon2, où l'ensemble des codons ARS et non-ARS montrent des valeurs de D similaires et les plus élevées des quatre loci.

L'analyse factorielle des correspondances a mis en évidence une structuration géographique des populations, surtout due à des allèles DPB1-Exon2 et DRB1-Exon2 (qui apparaissent comme variables explicatives importantes).

L'analyse d'échelonnement multidimensionnel montre des résultats similaires, avec surtout DPB1-Exon2 (et dans une moindre mesure DRB1-Exon2) montrant une structuration géographique des populations, les deux autres loci DQA1-Exon2 et DQB1-Exon2 montrant le moins de différences entre les populations.

L'analyse de variance moléculaire montre que seul DRB1-Exon2, lorsque l'on considère les groupes de populations selon la géographie, a un Φ_{CT} supérieur au Φ_{SC}. DPB1-Exon2 est le locus à montrer les Φ_{ST} les plus importants dans les quatre catégories. Ce locus montre aussi les Φ_{CT} les plus importants pour toutes les catégories, à l'exception du mode de vie, sans qu'il soit possible de parler d'un effet lié aux groupes, les Φ_{SC} étant à chaque fois supérieurs. De manière générale, la variance la plus élevée est toujours entre les populations (Φ_{ST}) et la plus faible est toujours entre les groupes (Φ_{CT}).

Le test de Mantel n'a pas mis en évidence de corrélation significative entre les distances génétiques et géographiques pour les populations africaines.

Finalement, l'analyse des liens entre fréquences alléliques et prévalence de la malaria ($pfpr2000$) a permis d'identifier six séquences comme présentant une corrélation significative, quatre dont les fréquences n'excèdent jamais 20% (DRB1*3144, *3149, *3155 et DQA1*5) et deux dont les fréquences dépassent 20% dans plusieurs populations (DQB1*2902 et DPB1*66).

4.2 Forces évolutives agissant sur chacun des gènes

Les analyses menées sur les quatre loci chez les mêmes individus (mêmes échantillons de populations) et concernant 1) la diversité moléculaire (π, hétérozygotie, richesse allélique, distributions de fréquences alléliques), 2) les résultats des tests de neutralité (test d'Ewens-Watterson-Slatkin et test du D de Tajima) et 3) les liens entre les loci (tests de déséquilibres de liaison) suggèrent que les quatre loci sont soumis à des mécanismes évolutifs différents.

DRB1-Exon2

DRB1-Exon2 montre l'hétérozygotie la plus élevée des quatre loci (0.864 ± 0.069, Table 4.8) et, à l'échelle des populations, le plus de rejets de la neutralité dus à un excès d'hétérozygotes (Figure 4.14).
La diversité moléculaire (nombre des sites polymorphiques et diversité nucléotidique, voir Table 4.12) élevée de ce locus n'est cependant observable qu'aux codons ARS considérés séparément, puisque ces mêmes mesures sur l'exon 2 complet ou sur les codons non-ARS ne montrent pas une diversité plus élevée qu'aux autres loci. C'est aussi le locus qui montre le moins de valeurs de D de Tajima significatives aux codons non-ARS, puisque seulement trois populations ont un D de Tajima significatif et positif à ces codons (Peuls du Sénégal, Imazighen de Figuig et Arabes Rashaida du Soudan).

HLA-DRB1-Exon2 semble donc soumis à une sélection balancée de type avantage de l'hétérozygote, visible par l'importante diversité moléculaire et les excès d'hétérozygotes dans 10 des 31 populations de l'étude. La grande différence dans le nombre de D de Tajima significatifs entre les codons ARS et non-ARS suggère que ce sont principalement les codons ARS qui évoluent par sélection balancée.
Les deux seules populations montrant un excès d'homozygotes à ce locus sont les Arabes Rashaida du Soudan et les Peuls du Mali. Dans les conditions d'excès d'hétérozygotes observées à ce locus, ce signal est donc certainement une signature démographique particulièrement forte pour ces deux populations que le mode de vie nomade a pu isoler génétiquement, d'où une perte de diversité par dérive génétique, en accord avec des résultats obtenus par une autre étude aux loci HLA-A et -B [Sanchez-Mazas et al., 2017].

DQA1-Exon2

DQA1-Exon2 montre une diversité moléculaire intermédiaire, mais se démarque par des valeurs de D de Tajima très élevées ($D_{Exon2} = 3.256 \pm 0.556$), même pour les codons non-ARS ($D_{non-ARS} = 3.072 \pm 0.548$, Table 4.12 et Figure 4.15). Ces derniers montrent un plus grand nombre de sites polymorphiques (0.185 ± 0.004 sites polymorphiques par positions) et un $\pi.n$ [22] plus élevé (0.083 ± 0.007) que les codons non-ARS des autres loci (Table 4.12).
En 2016, Lindo *et al.* ont suggéré un changement de régime de sélection pour HLA-DQA1, qui serait passé d'une ancienne sélection positive à une récente sélection négative pour des populations natives nord-américaines [Lindo et al., 2016]. Il est possible d'émettre une hypothèse similaire, de changement de régime de sélection au locus DQA1-Exon2, qui serait compatible avec les observations faites sur les populations africaines de notre étude. En effet, les valeurs élevées du D de Tajima associées au petit nombre d'allèles identifiés (richesse allélique de 7.98 ± 0.84, Figure 4.7) soutiennent l'hypothèse d'une ancienne sélection balancée, puis d'un régime de sélection directionnelle en faveur d'un ou de quelques-uns des allèle(s), les autres allèles ayant alors été éliminés.
Une autre hypothèse pour expliquer les valeurs élevées du D de Tajima de DQA1-Exon2 et la diversité moléculaire observée serait de considérer que la sélection n'agit pas seulement sur les codons ARS, comme cela semble être le cas pour d'autres loci tels que HLA-DRB1 et -DQB1 [Goeury et al., 2018a], mais que des codons non-ARS pourraient aussi être impliqués indirectement dans la liaison au peptide [Valdes et al., 1999], en étant proches du peptide lié par la molécule [Reche and Reinherz, 2003].

DQB1-Exon2

DQB1-Exon2, à l'instar de DQA1-Exon2, montre des motifs intermédiaires entre DRB1-Exon2 et DPB1-Exon2 (voir page 235). Ses codons ARS sont fortement diversifiés (diversité moléculaire moyenne par nucléotide $\pi.n$ de 0.155 ± 0.013 et nombre de sites polymorphiques moyen par nucléotide $S.n$ de 0.400 ± 0.011, Table 4.12) à l'opposé de ses codons non-ARS pour lesquels les valeurs sont intermédiaires (par rapport aux trois autres loci) sur ces mêmes statistiques ($\pi.n$ de 0.041 ± 0.004 et $S.n$ de 0.120 ± 0.008).

Le test d'Ewens-Watterson-Slatkin (Figure 4.14) montre quatre populations avec un excès significatif d'hétérozygotes (Tamasheq de Tamanrasset, Arabes Soudanais, Amhara

22. Diversité moléculaire moyenne par site.

d'Éthiopie et les Slovaques de Namestovo) et aucune avec des excès d'homozygotes, tandis que le test de Tajima montre une différence importante des valeurs D significatives entre les codons ARS et non-ARS (92% des populations montrent des valeurs de D positives et significatives pour les codons ARS de DQB1-Exon2, contre 52% pour les codons non-ARS, Table 4.14), suggérant une différence de pression de sélection entre les codons ARS et non-ARS. La seconde interprétation d'un D de Tajima positif et significatif est une contraction démographique, mais un tel signal devrait alors s'observer sur l'ensemble du locus, c'est-à-dire aussi les codons non-ARS. Cette seconde hypothèse ne nous semble donc pas soutenue.

Ce locus semble toutefois impliqué dans la résistance au pathogène *P. falciparum* de la malaria puisque la séquence DQB1*2902 montre une forte corrélation de sa fréquence allélique avec la prévalence du parasite (Figures 4.31 et 4.32, Table 4.17).

Mais ce locus est aussi en fort déséquilibre de liaison avec DQA1-Exon2 et peut donc subir en partie les pressions de sélection agissant sur DQA1-Exon2. Les quatre populations montrant un excès significatif d'hétérozygotes au locus DQB1-Exon2 montrent aussi un excès significatif d'hétérozygotes au locus DQA1-Exon2 et un déséquilibre de liaison global significatif entre ces deux loci (Figure 4.13).
Ces résultats suggèrent une sélection balancée associative [Ohta and Kimura, 1970, Slatkin, 1995, Sanchez-Mazas, 2007], où la diversité génétique s'accumule par déséquilibre de liaison dans le voisinage d'un locus sous sélection balancée.

DPB1-Exon2

DPB1-Exon2 montre les valeurs de diversité moléculaire les plus basses des quatre loci (pour l'exon 2 complet, $\pi.n = 0.029 \pm 0.005$ et $S.n = 0.088 \pm 0.007$, Table 4.12), associées toutefois à une richesse allélique intermédiaire entre DRB1-Exon2 et DQB1-Exon2 (richesse allélique de 9.58 ± 1.56, Figure 4.7) et les plus faibles valeurs de D de Tajima aux codons ARS (Tables 4.12 et 4.14). Ces résultats suggèrent une pression de sélection balancée plus faible que pour les trois autres loci.
Toutefois, ces pressions sélectives ne semblent pas similaires entre toutes les populations. Deux populations burkinabé, les Mossi et les Gourounsi, ont une fréquence élevée de DPB1*66 et cela se voit dans le résultat du test d'Ewens-Watterson-Slatkin (Figure 4.14) donnant un excès d'homozygotes. Cet allèle a été suggéré dans notre étude comme ayant un rôle de protection vis-à-vis de l'infection à *P. falciparum* (voir page 226).

DPB1-Exon2 montre peu de déséquilibre de liaison avec les autres loci étudiés. Aucune population non-africaine ne montre de déséquilibre entre DPB1-Exon2 et les autres loci (Figure 4.13). Pour les populations africaines, ce dernier n'est en déséquilibre avec les autres loci que dans des populations nomades ou semi-nomades (Peuls, Imazighen et Beja Hadendoa) ainsi que chez les Dangaléat du Tchad et les Mandenkalu du Sénégal.
La cause du déséquilibre de liaison pour les populations (semi-)nomades peut-être attribuée à de la dérive génétique due à leur isolement, de la même façon que la présence de ce déséquilibre de liaison chez les Dangaléat du Tchad pourrait être due à la petite taille de la population (60'000 individus en 2005 [Eberhard et al., 2019]), générant une dérive génétique rapide.
Pour les Mandenkalu du Sénégal, l'hypothèse d'une dérive génétique rapide est peu envisageable, cette population ayant été déjà étudiée pour d'autres marqueurs génétiques (immunoglobulines [Dard et al., 1996, Dard et al., 1997], HLA

[Tiercy et al., 1992, Tiercy et al., 1992], ADN mitochondrial [Graven et al., 1995], RFLP [Poloni et al., 1995], α et β-globulines [Martinson et al., 1995, Currat et al., 2002] et N-acetyltransferase 2 [Sabbagh et al., 2008]) montrant tous un haut niveau de diversité génétique résultant d'une probable expansion démographique [Excoffier and Schneider, 1999].
Le déséquilibre de liaison chez les Mandenkalu serait alors dû à un effet de sélection de HLA-DPB1*17:01 et HLA-DPB1*131:01 (correspondant à l'unique séquence DPB1*64, car partageant le même exon 2), bien que le pathogène responsable de cette sélection reste à définir [Goeury et al., 2018a].

L'analyse d'échelonnement multi-dimensionnel (MDS) basée sur les Θ_w de DPB1-Exon2 (Figure 4.26) montre une structuration des populations en trois groupes : un groupe comprenant les populations sénégalaises et maliennes (ci-après "groupe 1"), un groupe formé par les populations burkinabé et les Dangaléat du Tchad (ci-après "groupe 2") et un groupe formé par les autres populations (populations non-africaines de Syrie et de Slovaquie incluses, ci-après "groupe 3").
Les populations appartenant aux groupes 1 et 2, à gauche sur l'axe 1 de la MDS (MDS1), sont toutes des populations exposées à *P. falciparum*, la prévalence moyenne de *P. falciparum* (en l'an 2000) de ces populations étant de 0.45 ± 0.27 contre 0.04 ± 0.06 pour les populations du groupe 3, à droite sur la MDS1. Les populations des groupes 1 et 2 semblent donc se différencier génétiquement de celles du groupe 3 par l'effet d'une pression de sélection engendrée par l'exposition à *P. falciparum*. Les populations des groupes 1 et 2 se différencient aussi entre elles par les fréquences de DPB1*64 (plus fréquent dans les populations du groupe 1) et DPB1*66 (plus fréquent dans les populations du groupe 2) tel qu'illustré par la Figure 4.11 (voir aussi l'annexe S-43 pour les fréquences alléliques de chaque population). Cela suggère une différence en termes de pressions évolutives dues à la malaria d'une part, et d'histoire démographique, d'autre part (visible par la séparation des deux premiers groupes sur la MDS2).

DPB1-Exon2 montre donc une structure qui, bien que présentant une forte composante géographique (avec les populations d'Afrique de l'ouest d'un côté et les autres populations de l'autre), semble façonnée par une pression de sélection exercée (au moins en partie) par *P. falciparum* sur plusieurs de ces populations.

Conclusions

Les quatre loci étudiés montrent des profils génétiques et moléculaires différents.
DRB1-Exon2 est le locus le plus diversifié (en termes de diversité moléculaire et de richesse allélique) et évoluerait par une pression de sélection balancée de type avantage de l'hétérozygote agissant principalement sur les codons ARS.
DQA1-Exon2 montre des valeurs de D de Tajima significatives et positives très élevées avec un petit nombre d'allèles (en comparaison avec DRB1-Exon2 et DPB1-Exon2), deux hypothèses pouvant expliquer cela. La première implique un changement de régime de sélection pour les populations de l'étude, passant d'une ancienne sélection balancée à une sélection directionnelle. La deuxième hypothèse concerne le fait que les codons non-ARS seraient, eux aussi, la cible de la sélection car étant impliqués indirectement dans la liaison au peptide.
DQB1-Exon2 semble impliqué dans la résistance à *P. falciparum*, notamment la séquence DQB1*2902, tout en montrant un fort déséquilibre de liaison global avec DQA1-Exon2,

expliquant les similarités entre ces deux loci en termes de richesse allélique et diversité moléculaire.
Finalement, DPB1-Exon2 montre les plus faibles valeurs de diversité moléculaire et serait sous sélection positive pour les populations exposées à *P. falciparum*.

4.3 Forces évolutives agissant sur les populations

Populations d'Afrique du nord

Les Tamasheq de Tamanrasset et les Imazighen (d'Amizmiz, Asni et Figuig), bien que non exposés à la malaria ($pfpr2000 < 0.01$), possèdent la séquence DPB1*66 à une fréquence relativement élevée (Tamasheq de Tamanrasset : 19%, Imazighen d'Asni : 17%, Imazighen d'Amizmiz : 6% et Imazighen de Figuig : 6%, voir Figure 4.11). Ils partagent aussi un allèle, DPB1*76, avec les Peuls (fréquences alléliques de 3% pour les Tamasheq de Tamanrasset, 8% pour les Imazighen d'Asni, 7% pour ceux de Figuig, 3% pour ceux d'Amizmiz, 23% pour les Peuls du Mali et 13% pour ceux du Sénégal).
Les Tamasheq sont une population nomade en voie de sédentarisation, ayant largement participé au commerce trans-saharien dans le passé [Mayor, 2011]. Des analyses basées sur l'ADN mitochondrial, sur des loci HLA de classe I et à l'échelle du génome ont suggéré un flux génique entre les Peuls et une (ou des) population(s) nord-africaine(s) sur la base 1) de la présence d'une composante nord-africaine (23%) dans les génomes Peuls [Triska et al., 2015], 2) de la présence d'haplogroupes mitochondriaux partagés (U5b1b1b, H1cb) [Kulichová et al., 2017] et 3) d'une similarité génétique au locus HLA-A [Sanchez-Mazas et al., 2017]. De plus, une étude menée en 2010 par Harich *et al.* sur les haplogroupes mitochondriaux de 81 Marocains suggère que la plupart des séquences fréquentes en Afrique sub-saharienne et retrouvées dans les génomes d'Africains du nord proviennent d'un ancien commerce trans-saharien d'esclaves [Harich et al., 2010].
La présence au sein des populations Tamasheq de Tamanrasset, Imazighen et Peuls des séquences DPB1*66 et DPB1*76, fréquentes en Afrique de l'ouest, peut donc résulter de flux géniques entre populations à travers le Sahara.

Les Imazighen d'Amizmiz, Arabes Rashaida et Beja Hadendoa montrent des fréquences très élevées de DQB1*2901 (respectivement 38, 57 et 49%, Figure 4.11) et un fort déséquilibre de liaison (toutes les paires de loci étant en déséquilibre de liaison global, voir Figure 4.13). Les Arabes Rashaida montrent aussi un excès d'homozygotes au locus DRB1-Exon2 (Figure 4.14) et une fréquence élevée de DRB1*3136 (FA [23]=36%, aussi observée chez les Beja Hadendoa avec FA=35%), ainsi qu'un D de Tajima significatif et positif pour les codons non-ARS de DRB1-Exon2. Ce résultat est surprenant connaissant les conditions d'hétérozygotie habituellement observées à ce locus. Les Beja Hadendoa et les Arabes Rashaida montrent enfin une position extrême sur l'axe 1 de la MDS basée sur les Θ_w de DRB1-Exon2 (Figure 4.20). À l'instar des Peuls, il est vraisemblable que le mode de vie nomade de ces populations, caractérisé par un isolement et un petit effectif, ait entraîné une dérive génétique rapide expliquant ces observations.

23. Fréquence allélique

Peuls

Un des résultats notables de cette étude concerne les Peuls du Sénégal et du Mali. Si ces populations ont en commun une fréquence élevée de DQA1*1, elles montrent par contre des signaux différents pour DRB1-Exon2 : les Peuls du Mali ont un excès d'homozygotes à ce locus (Figure 4.14), dû à la haute fréquence (Figure 4.8) de DRB1*3137 (30%) et DRB1*3147 (19%), tandis que les Peuls du Sénégal ont un D de Tajima significatif et positif pour les codons non-ARS de ce même locus. DRB1-Exon2 évolue sous une sélection balancée de type avantage de l'hétérozygote ayant pour cible les codons ARS (voir page 233). Dans ces conditions d'hétérozygotie élevée observées au locus DRB1-Exon2, l'excès d'homozygotes pour les Peuls du Mali, ainsi que le D de Tajima significatif et positif aux codons non-ARS (non soumis à la sélection et donc plus enclins à révéler des signaux démographiques) des Peuls du Sénégal peuvent être interprétés comme une signature démographique. Ces populations étant nomades, leur mode de vie a pu les isoler génétiquement et causer une dérive génétique rapide, voire une contraction démographique (d'après les valeurs du D de Tajima aux codons non-ARS).

En 2011, une étude menée par Černý *et al.* avait étudié la diversité du segment hyper-variable I (HVS-I) de l'ADN mitochondrial et de huit marqueurs STR du chromosome Y sur 432 individus Peuls provenant de cinq pays africains, ainsi que de 470 individus provenant de populations sédentaires voisines [Černý et al., 2011]. Cette étude avait mis en évidence une plus grande homogénéité des populations Peuls (en termes de F_{ST} et R_{ST} [24]), tout en révélant des traces de contraction démographique au niveau de l'ADNmt. Dans notre étude, les résultats obtenus aux loci DQA1-Exon2 et DPB1-Exon2 sont similaires à ceux de Černý *et al.* puisque les deux populations Peuls ne sont pas différenciées, les Θ_w (distances de Reynolds) estimées pour ces deux loci n'étant pas significativement différents de 0. Toutefois, les résultats pour DRB1-Exon2 sont différents, puisque le Θ_w entre les deux populations Peuls est significatif, indiquant une différence dans les fréquences alléliques de ces deux populations au locus DRB1-Exon2. La différence entre les résultats de notre étude et celle de Černý *et al.* peut s'expliquer par la localité d'échantillonnage des populations. En effet, les populations étudiées par [Černý et al., 2011] ont été majoritairement échantillonnées autour du lac Tchad, alors que celles de notre étude proviennent de localités géographiquement plus éloignées les unes des autres (voir la Figure 4.3). Il est possible que l'homogénéité des populations Peuls observée par Černý *et al.* soit due à la proximité géographique entre ces populations. De plus, la différence de fréquences alléliques entre les Peuls du Mali et du Sénégal, au locus DRB1-Exon2, pourrait s'expliquer par l'effet de la dérive génétique rapide, causant une différenciation plus importante de ces populations à ce locus.

Il a été rapporté que les Peuls du Burkina Faso seraient moins affectés par la malaria que les autres populations burkinabé [Modiano et al., 1996], cette résistance étant liée à un haut niveau d'IFN-γ [25] et une plus faible activité (liée à un déficit fonctionnel) des cellules T régulatrices [Torcia et al., 2008, Boström et al., 2012] impliquant une réponse plus importante du système immunitaire en cas d'infection par *P. falciparum*.
Une étude basée sur les gènes HLA de classe I [Sanchez-Mazas et al., 2017] a identifié trois allèles possiblement protecteurs vis-à-vis de la malaria

24. Distances génétiques estimées à partir des données (respectivement) mitochondriales et du chromosome Y.

25. L'interféron γ est une molécule de la classe des cytokines, impliquée dans la réponse inflammatoire en cas d'infection.

(HLA-A*74:01, HLA-B*53:01 et HLA-B*78:01) et indiqué que les Peuls, bien qu'exposés à la malaria, ne montraient pas de fréquences de ces allèles (notamment HLA-B*53) aussi élevées que dans les populations exposées de manière similaire, ce qui suggère que les Peuls posséderaient un mécanisme de défense contre ce pathogène différent des populations voisines.
La présente étude, basée sur les gènes de classe II, montre des résultats différents : les deux populations Peuls ont une fréquence de DPB1*66 (18% pour les Peuls du Mali et 13% pour les Peuls du Sénégal) proche de celle prédite par les modèles (24% pour les Peuls du Mali et 18% pour les Peuls du Sénégal, Figure 4.32) et les Peuls du Mali montrent une fréquence de DQB1*2902 (22%) plus basse qu'attendue selon le modèle (33%), sans être non plus extrême. Il semble donc que les loci de classe II participent, chez les Peuls, à la défense contre le pathogène *P. falciparum*.

Une étude des gènes HLA de classe II, menée en 2009 sur un échantillon de 43 Peuls du Burkina Faso non-apparentés, a identifié deux allèles HLA fréquents chez les Peuls, HLA-DRB1*04 (13%) et HLA-DQB1*02 (36%) [Lulli et al., 2009]. Ces deux allèles ont été considérés, dans une méta-étude menée en 2008, comme étant associés à un plus haut risque de maladies auto-immunes [Fernando et al., 2008]. Les auteurs de la première étude ont émis l'hypothèse que ces deux allèles pourraient être impliqués dans la plus faible susceptibilité à la malaria et la réaction immunitaire plus importante en cas d'infection (par ce pathogène) chez les Peuls.
HLA-DRB*04 peut correspondre dans notre étude à DRB1*3138 (fréquence allélique de 12% pour les Peuls du Sénégal et 7% pour les Peuls du Mali) et HLA-DQB1*02 peut correspondre à DQB1*2901 (26% chez les Peuls du Sénégal, les données de ce locus n'étant pas disponibles pour les Peuls du Mali). Bien que ces deux allèles montrent des fréquences similaires à celles de l'étude de [Lulli et al., 2009], ils n'ont pas été identifiés dans notre étude comme ayant des fréquences corrélées à la prévalence de la malaria (Figure 4.32). Si ces allèles sont associés à un rôle protecteur vis-à-vis de la malaria spécifiquement chez les Peuls, il est alors cohérent que notre étude basée sur un grand nombre de populations (22 populations d'Afrique) n'ait pas pu les identifier.

Notre étude met en évidence un possible rôle protecteur des molécules HLA de classe II contre *P. falciparum* chez les Peuls du Mali et du Sénégal, tout en soulignant que certains des allèles protecteurs, suggérés par d'autres études, ne sont pas identifiés dans notre travail.

Afrique de l'ouest

Les populations sénégalaises, à l'exception des Peuls, sont les seules populations à montrer une fréquence élevée de DRB1*3135 (Mandenka : 27%, Bédik : 30% et Sérère 29%, voir Figure 4.8), cet allèle correspondant à l'unique allèle HLA-DRB1*13:04 en termes d'exon 2. Cet allèle a aussi été identifié à des fréquences élevées (40%) en Gambie [Hill et al., 1992a] (grâce à l'étude d'un échantillon incluant des populations variées : Mandenka, Wolof, Jola et Fulani).
L'étude menée sur les Mandenkalu du Sénégal (Chapitre 2 et publication associée [Goeury et al., 2018a]) a mis en évidence que cet allèle serait issu d'une conversion allélique entre deux allèles HLA-DRB1 : DRB1*11:02:01 comme receveur, l'allèle donneur n'ayant pas pu être identifié parmi une liste de 195 donneurs potentiels (voir l'annexe S-24 pour la liste des donneurs potentiels et le Chapitre 2, page 89, pour le détail de cette

conversion allélique).
Sa forte fréquence dans ces populations suggère une récente sélection directionnelle [Hill et al., 1992a]. Chez les Mandenkalu du Sénégal (Chapitre 2), cet allèle est en déséquilibre de liaison avec HLA-DQA1*05:01 et HLA-DQB1*03:19, le premier possédant un exon 2 (la région de liaison au peptide) identique à DQA1*01:01, rapporté en 1994 comme étant plus fréquent chez les individus immunisés contre *Onchocerca volvulus* [Meyer et al., 1994]. *O. volvulus*, transmis par des diptères du genre *Simulium*, est le parasite responsable de l'onchocercose (aussi appelé cécité des rivières) et est une maladie montrant une prévalence élevée en Afrique de l'ouest [O'Hanlon et al., 2016]. Il est donc possible que DRB1*3135 (HLA-DRB1*13:04) ait atteint des fréquences élevées dans les populations sénégalaises de l'étude, soit par déséquilibre de liaison avec des allèles sous sélection directionnelle (HLA-DQA1*05:01 et HLA-DQB1*03:19) pour la malaria et/ou l'onchocercose, soit en évoluant lui-même par sélection directionnelle.

La représentation graphique MDS, basée sur les Θ_w calculés pour DPB1-Exon2 (Figure 4.26), met en évidence deux groupes distincts pour les populations d'Afrique de l'ouest (à gauche sur l'axe 1) : un groupe composé des populations sénégalaises et maliennes (uniquement représentées ici par les Peuls du Mali), en bas (axe 2), et un groupe composé des populations burkinabé, en haut. Outre le fait que ces deux groupes de populations soient géographiquement séparés, il existe aussi une différence linguistique. Les populations du premier groupe (Sénégal et Mali) parlent des langues de la famille niger-congo, sous-famille atlantique-congo, branche atlantique (à l'exception des Mandenkalu qui parlent une langue de la sous-famille mandé). Les populations du second groupe (Burkina-Faso) parlent aussi toutes des langues de la famille niger-congo et sous-famille atlantique-congo, mais de la branche volta-congo [Eberhard et al., 2019]. De plus, DPB1-Exon2 montre les valeurs les plus élevées de différences inter-groupes Φ_{CT} (0.049) dans l'AMOVA basée sur les familles linguistiques (Figure 4.29). Il est donc probable que cette différence entre les deux groupes observée sur la MDS provient, entre autres, d'une histoire différente, reflétée par la différence linguistique et expliquant les différences génétiques.

Les huit populations ouest-africaines étudiées ici se différencient également des autres populations au locus DPB1-Exon2 par les fréquences élevées de DPB1*64 et DPB1*66 (Table 4.17 et Figure 4.11). Ces deux allèles distinguent aussi les populations burkinabé (où DPB1*66 est majoritaire) des populations sénégalaises et maliennes (dans lesquelles DPB1*64 est majoritaire).
La fréquence DPB1*66 semble fortement associée à la prévalence de la malaria (voir Figures 4.31 et 4.32). Sa fréquence élevée peut s'expliquer, dans les populations d'Afrique de l'ouest étudiées ici, par son possible rôle protecteur vis-à-vis de la malaria, ce qui expliquerait sa fréquence plus élevée dans les populations burkinabé où la prévalence de la malaria est plus importante (la $pfpr2000$ des lieux d'échantillonnage est de 0.33 ± 0.25 pour le Sénégal, contre 0.72 ± 0.10 pour le Burkina Faso).
La séquence DPB1*64 peut, quant à elle, correspondre à trois allèles HLA connus : HLA-DPB1*17:01:01, HLA-DPB1*131:01 et HLA-DPB1*460:01. Parmi ces trois allèles, on sait que pour les Mandenkalu, DPB1*64 correspond aux deux premiers, avec des fréquences similaires (22 et 20%, voir le Chapitre 2). Il serait donc intéressant de savoir si, pour les autres populations du Sénégal et du Mali, cette séquence correspond aussi à deux allèles différents. Cette information permettrait alors de mieux caractériser les flux géniques entre ces populations.

Afrique centrale et de l'est

Dangaléat

Les Dangaléat du Tchad se retrouvent proches des populations du Burkina Faso sur la MDS basée sur les Θ_w de DPB1-Exon2 (Figure 4.26), le Θ_w entre les Dangaléat et les Gourmantché n'étant pas significativement différent de zéro, ceci indique des fréquences alléliques similaires à ce locus.
Les Dangaléat sont une population exposée à la malaria et possèdent l'allèle DPB1*66 à une fréquence élevée (31%) mais montrent toutefois un excès d'hétérozygotes à ce locus (Figure 4.14), contrairement aux observations faites pour les Mossi et Gourounsi du Burkina Faso (où la fréquence élevée de DPB1*66 induit un excès d'homozygotes).

L'hypothèse qui pourrait expliquer cette fréquence est de considérer une sélection de type directionnelle pour DPB1*66, en rapport à un rôle protecteur vis-à-vis de la malaria. L'équation 4.1 montre que la valeur du D de Tajima des codons ARS de DPB1-Exon2 est négativement associée avec la prévalence de *P. falciparum*. Ces résultats suggèrent qu'au moins un allèle DPB1-Exon2 (probablement DPB1*66) soit sous sélection directionnelle positive, lié à un rôle protecteur de la malaria. Cette hypothèse expliquerait alors la fréquence élevée de cet allèle chez les populations burkinabé, fortement exposées à la malaria ($pfpr2000$ allant de 0.61 à 0.81), et sa fréquence plus basse chez les Dangaléat, population plus faiblement exposée ($pfpr2000 = 0.15$). Les Sérère du Sénégal ont une exposition au parasite proche des Dangaléat ($pfpr2000 = 0.15$) et une fréquence de DPB1*66 similaire (29% chez les Sérère, 31% chez les Dangaléat), soutenant l'hypothèse d'une sélection positive dont l'intensité serait proportionnelle à la pression exercée par la malaria.

Il reste toutefois à expliquer l'excès d'hétérozygotes à ce locus ($H = 0.841$). Deux hypothèses peuvent expliquer ce résultat. La première est de considérer que cette sélection positive exercée par la malaria agit sur une ancienne sélection balancée, probablement de type avantage des hétérozygotes. En effet, la valeur du D de Tajima pour l'exon 2 complet est supérieure ($D = 2.86$, Table 4.14) à celle estimée pour les codons ARS ($D = 2.57$). Une étude de 2011, étudiant les profils moléculaires des exons 2 de HLA-DPB1 pour 4'911 individus, avait déjà observé des valeurs de D de Tajima significativement supérieures à 0 pour certaines populations [Buhler and Sanchez-Mazas, 2011]. Ces résultats avaient alors été interprétés comme le signe d'une ancienne sélection balancée pour ce locus. Dans notre cas, une ancienne sélection balancée de type avantage des hétérozygotes expliquerait l'hétérozygotie importante et la valeur du D de Tajima sur l'exon 2 au complet. La plus faible valeur de D de Tajima pour les codons ARS s'expliquerait, quant à elle, par la pression de sélection positive exercée par la malaria sur ce locus et dont les codons ARS seraient la cible.
La seconde hypothèse est de considérer un effet lié à (au moins) un évènement de fusion de populations. En effet, les Dangaléat montrent un fort déséquilibre de liaison global (toutes les paires de loci étant en déséquilibre, Figure 4.13) qui peut être interprété, avec l'excès d'hétérozygotes au locus DPB1-Exon2, comme le signal d'une fusion de populations génétiquement différentes (tel qu'observé chez les Cham du Vietnam, voir Chapitre 2). Une étude de 2009, menée par Poloni *et al.* sur de l'ADN mitochondrial de deux populations Nyangatom et Daasanach du sud de l'Éthiopie, avait elle aussi mis en évidence une importante structuration génétique associée à une diversité élevée [Poloni et al., 2009]. Les auteurs avaient interprété ce signal comme le résultat de processus de migrations et

d'absorptions de groupes de populations, entremêlés d'épisodes d'isolation. Ainsi, il est possible qu'une succession d'évènements de fusions de populations et d'isolements culturels explique, pour les Dangaléat du Tchad, les résultats observés aux loci HLA de classe II.

Maba

Les Maba sont des locuteurs nilo-sahariens (langue *maban*, sous-famille maban, famille satellite) entourés de populations locutrices afro-asiatiques (parlant notamment l'arabe) et les résultats obtenus pour les loci DQA1-Exon2 et DQB1-Exon2 ont conduit à formuler l'hypothèse d'un isolement des Maba, causé par une importante différence culturelle vis-à-vis de leurs voisins géographiques [Pasquier, 2016].
Les MDS basées sur les loci DQA1-Exon2, DQB1-Exon2 et DPB1-Exon2 (Figures 4.22, 4.24 et 4.26) montrent toutes une différence significative entre les Maba et les autres populations d'Afrique centrale.
La fréquence très élevée de DQA1*1 chez les Maba (61%, Figure 4.9) cause un faible taux d'hétérozygotes ($H = 0.600$, Figure 4.6) visible par un excès d'homozygotes à ce locus (Figure 4.14) dans cette population. De même, DQB1-Exon2 montre une faible hétérozygotie ($H = 0.637$), causée par la fréquence de DQB1*2902 (FA=56%).
Pour cette population, les trois loci DQA1-Exon2, DQB1-Exon et DPB1-Exon2 sont tous les trois en déséquilibre de liaison global (Figure 4.13). Seuls 23% des populations étudiées ici montrent un déséquilibre de liaison entre DQA1-Exon2 et DPB1-Exon (Table 4.10), résultat explicable par la présence d'un point chaud de recombinaison meïotique proche du gène TAP2, entre HLA-DQB1 et HLA-DPB1 [Martin et al., 1995]. Ainsi, l'important déséquilibre de liaison entre ces trois loci pourrait ici s'expliquer par un effet de la dérive génétique rapide, qui peut causer un déséquilibre de liaison plus important [Vangenot et al., 2020].
Les résultats obtenus ici soutiennent l'hypothèse d'un isolement culturel des Maba, causant une dérive génétique plus rapide et expliquant la diversité plus réduite de cette population, l'important déséquilibre de liaison et les distances génétiques significatives avec les populations voisines.

Toutefois, à l'instar des autres populations tchadiennes de l'étude, les Maba montrent aussi une fréquence élevée de DRB1*3153 [26] (27%). De plus, le locus DRB1-Exon2 est le seul à ne pas être en déséquilibre de liaison global (avec l'un des trois autre loci) pour cette population.
Une étude parue en 2018 et basée sur les génomes de 751 individus habitant la bande du Sahel (Tchad, Burkina Faso, Mali, Sud-Soudan et Soudan) a mis en évidence, chez les populations Daza, Kanembu et Tubu, un évènement de mélange [27] entre des Africains de l'est et du centre-ouest, il y a 950 ans (intervalle de confiance à 95% allant de 210 à 1'740 ans) [Shriner and Rotimi, 2018]. La Figure 4.8, montrant les distributions de fréquences alléliques de DRB1-Exon2 met en évidence une séquence, DRB1*3155, présente dans les populations d'Afrique de l'ouest ($7.5 \pm 4.5\%$) et d'Afrique centrale ($6.7 \pm 2.9\%$) et absente des autres régions (<1%) à l'exception de l'Afrique du nord ($1.6 \pm 1.1\%$). Si l'on considère des évènements de mélanges entre des populations d'Afrique de l'ouest et de l'est, alors DRB1*3155 pourrait être un signal de flux génique entre ces populations. Cette hypothèse pourrait alors expliquer le motif particulier observé au locus DRB1-Exon2.

26. Pouvant correspondre à HLA-DRB1*08:04 en termes d'exon 2.
27. *Admixture* en anglais.

Afrique de l'est

L'Afrique de l'est est représentée par deux populations éthiopiennes, les Oromo et les Amhara.

En 1998, une étude de Fort *et al.*, menée sur 83 Oromo et 98 Amhara génotypés pour trois loci HLA de classe II (HLA-DRB1, -DQA1 et -DQB1) n'avait pas identifié de différences significatives dans les fréquences alléliques de ces deux populations [Fort et al., 1998]. Notre étude, intégrant en plus les données de DPB1-Exon2, confirme ces résultats par l'absence de significativité des Θ_w entre ces deux populations pour l'ensemble des quatre loci (Figures 4.20 à 4.27).

En 2017, une étude menée par Sanchez-Mazas *et al* et portant sur le séquençage de trois loci de classe I (HLA-A, -B et C) pour 484 individus provenant de 11 populations avait mis en évidence un chevauchement des populations d'Afrique centrale, de l'est et du nord dans des analyses d'échelonnement multi-dimensionnel [Sanchez-Mazas et al., 2017], expliqué par des flux géniques entre populations dans ces régions d'Afrique. Les résultats de notre étude, obtenus sur quatre loci de classe II, soutiennent aussi cette hypothèse, visible sur les MDS (4.20 à 4.26) par le grand nombre de paires de populations qui ne sont pas significativement différentes.

Il apparaît alors qu'au contraire de l'Afrique de l'ouest, où les populations montrent une forte structuration génétique, l'Afrique centrale et l'Afrique de l'est soient des régions où les populations sont moins différenciées entre elles, indiquant des échanges plus importants entre ces populations.

4.4 Impact du pathogène *P. falciparum*

L'étude des relations entre les fréquences alléliques observées pour les loci de classe II et la prévalence de la malaria au Sahel (due à *P. falciparum*) suggère une association entre la malaria et les allèles DPB1*66 et DQB1*2902, mais aussi DRB1*3144, DRB1*3149, DRB1*3155 et DQA1*5.

Si plusieurs études ont établi des liens entre des allèles de classe I et la malaria ([Sanchez-Mazas et al., 2017, Hill et al., 1992b]) chez les humains mais aussi chez les bonobos [de Groot et al., 2018], peu d'études ont établi un lien entre la malaria et des allèles HLA de classe II, bien que HLA-DPB1 ait été plusieurs fois étudié [Hill et al., 1991, Stephens et al., 1995, May et al., 1999].

Cette dernière étude [May et al., 1999] portait sur la comparaison des profils génétiques de deux groupes de patients gabonais, infectés par la malaria et montrant soit peu (ou pas) de symptômes, soit des symptômes sévères de la maladie. Cette étude a alors mis en évidence une fréquence plus élevée de HLA-DPB1*01:01 [28] et de HLA-DQA1*04:01 [29] dans le groupe de patients montrant peu de symptômes, comparé au groupe montrant des symptômes sévères de la malaria. Les auteurs ont alors suggéré un rôle protecteur de ces deux allèles vis-à-vis de la malaria. Les résultats de notre étude suggèrent, eux aussi, un rôle protecteur de ces deux allèles vis-à-vis de la malaria.

Hill *et al.* ont mis en évidence, en 1991, une corrélation entre la fréquence de l'haplotype DQB1*05:01∼DRB1*13:02 et la sévérité des symptômes de la malaria dans une cohorte d'enfants en Gambie [Hill et al., 1991]. Les enfants porteurs de cet haplotype

28. Pouvant correspondre à DPB1*66 au niveau de l'exon 2.
29. Correspondant à la séquence DQA1*5.

étaient significativement plus fréquents dans le groupe présentant des symptômes modérés (ne nécessitant pas d'hospitalisation) que dans le groupe présentant des symptômes sévères (taux d'hémoglobine inférieur à $5g.dl^{-1}$ et pronostic vital engagé en l'absence de traitements médicaux).
Ces résultats sont différents de ceux obtenus dans notre étude, notamment pour DQB1*05:01∼DRB1*13:02, où aucun des deux allèles n'est retrouvé significativement corrélé avec la prévalence de la malaria, sans pour autant que les conclusions de Hill *et al.* et les nôtres s'opposent.
Premièrement, il y a une différence de conception expérimentale entre les deux études. Hill *et al.* ont comparé des individus infectés mais ne présentant pas la même réponse immunitaire (pas les mêmes symptômes) tandis que notre étude repose sur une corrélation entre les fréquences alléliques observées et la prévalence de la malaria due à *P. falciparum*. Ainsi, notre étude diffère de celle de Hill *et al.* en étant une étude populationnelle, ne tenant pas compte de la manifestation des symptômes de la malaria, alors que celle de Hill *et al.* est une étude menée sur une cohorte de patients.
Ensuite, Hill *et al.* ont d'abord mis en évidence un lien entre la sévérité des symptômes et la présence de la spécificité sérologique DR13. Ce n'est que dans un deuxième temps que des séquençages ADN ont identifié l'allèle HLA-DRB1*13:02. L'allèle identifié dans notre étude, HLA-DRB1*13:04 appartient aussi à la famille DR13 et il est possible que ces deux allèles soient interchangeables en termes d'antigènes présentés.
Finalement, ce ne sont pas les mêmes populations qui ont été analysées dans les deux études, notre étude n'incluant pas d'échantillons de populations gambiennes. Or notre étude a montré que l'effet de la pression de sélection de la malaria touchait aussi plusieurs allèles (Figure 4.32), de manière différente selon les populations (voir page 238). Cet effet est particulièrement visible avec la différence de fréquence de DPB1*66 et de la prévalence du *P. falciparum* entre les populations sénégalaises et burkinabé (Table 4.17).

En conclusion, notre étude propose les séquences DRB1*3144, DRB1*3149, DRB1*3155, DQA1*5, DQB1*2902 et DPB1*66 comme des séquences protectrices vis-à-vis de la malaria. En termes d'allèles HLA nominaux ces séquences correspondent, respectivement, aux allèles du premier champ HLA-DRB1*08, -DRB1*03, -DRB1*11, -DQA1*04, -DQB1*03 et -DPB1*01.

5 Conclusion

La présente étude, portant sur l'analyse moléculaire des exons 2 de quatre loci HLA de classe II (HLA-DRB1, -DQA1, -DQB1 et -DPB1) chez 2'061 individus appartenant à 36 populations d'Afrique (ceinture du Sahel et Afrique du nord), d'Asie de l'ouest et d'Europe centrale, fournit de nouveaux indices quant aux forces évolutives agissant sur ces loci, mais aussi sur ces populations.

Concernant les forces évolutives agissant sur chacun des quatre loci de l'étude, d'importantes différences sont observées. DRB1-Exon2 montre une diversité élevée, qui serait le résultat d'une pression de sélection balancée de type avantage de l'hétérozygote agissant principalement sur les codons ARS (codant pour le site de reconnaissance de l'antigène). DQA1-Exon2 montre des signaux particuliers sur les codons non-ARS (ne codant pas pour le site de reconnaissance de l'antigène), impliquant soit un changement de régime de sélection, passant d'une ancienne sélection balancée à une sélection directionnelle, soit que les codons non-ARS de DQA1-Exon2 soient eux aussi impliqués dans la liaison au peptide antigénique. Le locus DQB1-Exon2, bien que fortement en déséquilibre de liaison global avec DQA1-Exon2 et donc co-évoluant avec ce dernier, montre des signaux de sélection liés à la malaria. Finalement, le locus DPB1-Exon2 montre lui aussi des signaux de sélection liés à la malaria, ainsi qu'une importante structure géographique.

À propos des populations étudiées, des signaux de dérive génétique rapide sont observés pour les Imazighen d'Amizmiz, les Arabes Rashaida, les Beja Hadendoa ainsi que pour les deux populations Peuls (du Sénégal et du Mali). Ces populations sont toutes caractérisés par un mode de vie (semi-)nomade et il est possible que ce mode de vie soit responsable de cette dérive génétique rapide.

Les allèles partagés, au locus DPB1-Exon2, entre les populations d'Afrique sub-saharienne et les populations d'Imazighen et de Tamasheq d'Afrique du nord, suggèrent l'existence de flux géniques au travers du Sahara. De la même façon, le partage d'allèles entre les populations d'Afrique de l'ouest et d'Afrique centrale, au locus DRB1-Exon2, suggère, quant à lui, des flux géniques entre ces deux régions d'Afrique sub-saharienne.

Le locus DRB1-Exon2 montre aussi une association avec la géographie, notamment au Sénégal (Afrique de l'ouest), avec l'allèle HLA-DRB1*13:04, probablement issu d'une conversion allélique et ayant atteint des fréquences élevées, soit par sélection positive, soit par balayage sélectif en étant en déséquilibre de liaison avec des allèles eux-mêmes possiblement sous sélection (notamment HLA-DQA1*05:01 et -DQB1*03:19, voir le Chapitre 2). Les populations d'Afrique de l'ouest montrent aussi une forte structuration, au locus DPB1-Exon2, en deux groupes distincts. Cette structuration, analogue à la structuration linguistique de ces populations, peut refléter des différences historiques, mais aussi une pression de sélection, due à la malaria, différente entre les deux groupes.

Finalement, l'étude de l'association entre les fréquences alléliques et la prévalence de la malaria liée à *P. falciparum* a identifié six séquences dont la fréquence était corrélée à la prévalence de la malaria. Ces séquences sont DRB1*3144, DRB1*3149, DRB1*3155, DQA1*5, DQB1*2902 et DPB1*66 et correspondent, respectivement, aux allèles du premier champ HLA-DRB1*08, DRB1*03, DRB1*11, DQA1*04, DQB1*03 et DPB1*01.

La principale limitation de cette étude réside dans la technologie disponible au moment des séquençages, qui a obligé à se limiter aux exons 2 de ces quatre gènes, afin de pouvoir étudier un grand nombre de populations et suffisamment d'individus par population. Cette limitation exclut, de fait, une part non négligeable de la variabilité de ces gènes, par exemple l'exon 3 de HLA-DQB1 (voir Chapitre 5) ou les régions introniques qui, potentiellement moins soumises à la sélection, peuvent fournir des informations sur les effets démographiques agissant sur les populations. Les nouvelles technologies de séquençage maintenant disponibles devraient permettre de dépasser cette limitation et c'est ce qui est actuellement en cours au Laboratoire d'anthropologie, génétique et peuplements avec le récent projet HLA-AFRICA visant à appliquer les dernières technologies de séquençage (séquençage de gènes complets) sur les gènes HLA d'un corpus de populations incluant celles de notre étude.

Chapitre 5

Analyses statistiques du contenu des bases de données IPD-IMGT/HLA

1 Introduction

Dans le chapitre 2, « Étude comparée des Mandenka du Sénégal et des Cham du Vietnam », trois techniques de typage HLA ont été utilisées et comparées (PCR-SSO, NGS-454 et NGS-MiSeq) pour les mêmes individus d'une même population (les Mandenkalu du Sénégal). Bien qu'ayant toutes pour but la caractérisation de l'information moléculaire des allèles HLA, ces techniques de typage différaient soit dans la façon de lire l'information moléculaire (indirectement par hybridation de l'ADN avec des sondes marquées, pour PCR-SSO, ou directement par séquençage de l'ADN, pour les NGS-454 et NGS-MiSeq) soit dans la région génique qu'elles étaient capables de typer : exons 2 (et exons 3 pour les gènes de classe I) pour PCR-SSO, exons 2 des gènes de classe II pour NGS-454 et gènes complets pour NGS-MiSeq.
Ces techniques de typage ont montré chacune des différences qui ne pouvaient être expliquées, au moins en partie, que par des différences d'ordre technologique (25 années séparant les premiers typages PCR-SSO des typages NGS-MiSeq).
L'origine de ces différences se trouverait alors dans les régions géniques ciblées spécifiquement par chaque technique. En effet, quand les mêmes régions étaient ciblées (PCR-SSO et NGS-454), alors de hauts taux de correspondances étaient observés entre les typages obtenus par les deux techniques. Au contraire, quand les régions ciblées étaient différentes (PCR-SSO et NGS-MiSeq pour les loci de classe II) alors les taux de correspondance étaient plus faibles, ou cela menait à un nombre d'ambiguïtés très variable (de 2.8 allèles par séquence pour HLA-DPB1 à 12.3 pour HLA-DQB1, par typage NGS-454).
Un des objectifs de ce chapitre est donc de trouver une méthode capable de quantifier l'information de chacune des différentes régions (introns et exons) des gènes HLA, afin de localiser les régions qui doivent être ciblées pour obtenir un typage avec peu (ou pas) d'ambiguïtés.
Le choix a été fait d'utiliser l'entropie de Shannon comme mesure de la quantité d'information contenue dans les différentes régions géniques ; cela a permis d'expliquer que cibler une seule ou les deux régions géniques (en l'occurrence les exons 2 et 3) apportait une quantité d'information différente, selon les gènes considérés.
La théorie de l'information de Shannon a été proposée par Claude Shannon en 1948 [Shannon, 1948] dans le but de fournir une base théorique dans le domaine du traitement du signal et des technologies de l'information. L'un des points-clefs de cette théorie est l'introduction de l'entropie comme mesure de la quantité d'information d'un signal

délivré par une source d'information. Les propriétés mathématiques de cette fonction d'entropie permettent aussi de calculer l'entropie de deux signaux (entropie conjointe) et l'information mutuelle à ces deux signaux (traduisant l'indépendance ou non des sources d'information).
L'entropie telle que définie par Shannon possède un usage répandu, entre autres, en biologie, que ce soit dans le domaine de l'écologie, en tant que métrique d'évaluation de la biodiversité d'un milieu [Joshi et al., 2006], ou en génétique, avec l'identification de gènes impliqués dans des maladies [Monaco et al., 2019] ou l'analyse de réseaux de gènes [Hausser and Strimmer, 2009].
L'entropie, en plus des métriques associées (entropie conjointe et information mutuelle), sera utilisée ici pour analyser le contenu des bases de données IPD-IMGT/HLA [1] ainsi que la répartition de l'information entre les différentes régions des loci HLA.

Dans le Chapitre 3 (page 125, à propos de l'algorithme `MADaM`), un filtre est utilisé pour séparer les lectures des différents gènes (et pseudo-gènes) co-amplifiés par une spécificité non totale des amorces PCR (c'est-à-dire les exons 2 de HLA-DQB1 et -DQB2 dans le cas des séquençages HLA-DQB1 et les exons 2 de HLA-DRB1, -DRB3, -DRB4, -DRB5, -DRB6 et -DRB7 pour les séquençages HLA-DRB1). Ce filtre est un classificateur reposant sur une décomposition en chaînes de Markov des séquences nucléotidiques (voir page 132 pour une description détaillée de son fonctionnement).
Les chaînes de Markov sont des processus stochastiques, aussi appelés processus de Markov, dans lesquels la distribution de probabilité de l'état du système au temps $T + 1$ ne dépend que de l'état du système au temps T (le système n'a pas de *mémoire*).
Ces systèmes possèdent un large champ d'application incluant, entre autres, la reconnaissance de la parole [Juang and Rabiner, 1991], l'indexation de pages web [Page et al., 1999], la simulation de la dérive génétique au sein d'une population [Watterson, 1996], les dynamiques populationnelles en écologie [Leslie, 1945, Cáceres and Cáceres-Saez, 2011], la prédiction de fonction génique [Do and Choi, 2006] ou l'alignement de séquences nucléotidiques [Durbin, 1998].
Cette décomposition en chaînes de Markov a été appliquée ici aux gènes HLA-A, -B, -C, -DRB1/3/4, -DQA1, -DQB1, -DPA1 et -DPB1 ainsi qu'à quatre gènes de classe I du chimpanzé, Patr-A, -Al, -B et -C, permettant de valider la méthode implémentée dans `MADaM` en confirmant que la décomposition en chaînes de Markov des séquences nucléotidiques des différentes régions des gènes HLA permet bien de les identifier de manière claire et précise.

Dans la première partie de cette étude exploratoire, la théorie de l'information sera utilisée pour décrire le contenu des bases de données *gen* et *nuc* d'IPD-IMGT/HLA. Dans un premier temps, le contenu de ces deux bases de données sera comparé afin de s'assurer qu'il n'y a pas de biais particulier de représentativité de la diversité HLA entre ces deux bases de données. En effet, ces deux bases de données diffèrent dans le nombre d'allèles représentés mais aussi dans la couverture de ces allèles. Il y a moins d'allèles représentés dans la base de données *gen*, mais les séquences disponibles couvrent tout le gène au contraire de la base de données *nuc* qui possède plus d'allèles mais avec des couvertures plus réduites, généralement juste l'exon 2, voire l'exon 3. Le but ici est de s'assurer que même si la base de données *gen* compte moins d'allèles, elle soit bien un échantillonnage aléatoire de la base de données *nuc*. La base de données *gen* est la plus intéréssante des

1. *Immuno Polymorphism Datababse - ImMunoGeneTics/Human Leukocyte Antigen*, base de données répertoriant l'ensemble des séquences HLA (complètes ou partielles) connues et servant ainsi de référence pour la nomenclature des noms des allèles HLA.

deux pour étudier la distribution de l'information tout le long des gènes, en incluant les régions non codantes (absentes de la base de données *nuc*).
Dans un second temps, l'entropie, l'information mutuelle absolue et relative ainsi que le gain relatif d'information seront utilisés pour explorer les contributions respectives des exons 2 et 3 à la discrimination des allèles de huit gènes : HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1. Les entropies calculées pour l'ensemble des régions disponibles pour ces huit gènes (régions non codantes incluses) seront aussi utilisées pour étudier la distribution de l'information en fonction de la taille de ces régions, de leur fonction (codante / non codante) et des gènes qui les portent.

Dans une seconde partie, les séquences des différentes régions (codantes et non codantes) de 10 gènes HLA d'humain (HLA-A, -B, -C, -DRB1/3/4, -DQA1, -DQB1, -DPA1 et -DPB1) et quatre gènes Patr de chimpanzé (Patr-A, -AL, -B et -C) seront décomposées en chaînes de Markov de premier ordre, ces dernières représentées à l'aide d'une méthode de visualisation et de réduction de dimensionalité afin de valider la méthode implémentée dans `MADaM`.
De plus, cette analyse cherche aussi à analyser les similaritées (ou différences) entre les différentes régions d'un même gène mais aussi entre les régions de différents gènes, pour apporter des indices supplémentaires permettant une meilleure compréhension des processus évolutifs du HLA (ou plus généralement du MHC).

2 Matériel et Méthodes

2.1 Provenance des données

L'ensemble des données HLA utilisées dans ce chapitre proviennent des bases de données d'IPD-IMGT/HLA [Robinson et al., 2015] (`https://www.ebi.ac.uk/ipd/imgt/hla/`). Deux bases de données différentes sont disponibles : la première, `gen`, ne contient que les séquences des allèles HLA entièrement séquencés (au minimum du premier au dernier exon) et la seconde, `nuc`, contient l'ensemble des séquences alléliques même partielles (majoritairement les exons 2 en plus des exons 3 pour les classe I).

Pour le calcul des entropies des régions géniques, les huit gènes HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1 provenant des deux bases de données ont été utilisés (ces séquences téléchargées le 13 avril 2017, version 3.28 de la base de données). Pour les décompositions en chaînes de Markov, seules les données de la base de données `gen` ont été utilisées (séquences téléchargées le 9 Avril 2018, version 3.31), les gènes concernés étant HLA-A, -B -C, -DRB1/3/4, -DQA1, -DQB1, -DPA1 et -DPB1. L'analyse des décompositions en chaînes de Markov inclut aussi 311 séquences du MHC (en l'occurrence Patr-A, -AL, -B et -C) du chimpanzée (*Pan sp.*) provenant de la base de données GenBank du NCBI [2] et dont les numéros d'accession sont disponibles en annexe S-51.

Les données étant disponibles sous plusieurs formats (incluant Fasta, PIR et MSF), le choix a été fait d'utiliser le format « *Alignments* », ce dernier présentant l'avantage d'avoir les séquences alignées (permettant d'avoir un alignement similaire pour toutes les régions de tous les gènes et évitant le recours à un outil externe d'alignement de séquences), les positions numérotées (la position 1 correspondant au premier nucléotide de l'exon 1) et les régions géniques identifiées.

2.2 Calcul des entropies

Pour le calcul des entropies, seules les séquences (des différentes régions géniques) entièrement séquencées ont été utilisées (les positions non séquencées étant symbolisées par un ***** dans la base de données), afin d'éviter le problème d'une incertitude sur la séquence (si deux séquences diffèrent d'une seule position non séquencée sur une des deux séquences, il est impossible de savoir si ces séquences sont identiques ou différentes). En plus des différentes séquences de chaque région, les séquences des exons 2 et 3 ont été concaténées afin de pouvoir calculer l'entropie des exons 2 et 3 (considérées ensemble) et l'information mutuelle aux exons 2 et 3.

Sur la base des fréquences de chaque séquence unique (au sein d'une même région d'un même gène), quatre statistiques ont été calculées : l'indice de diversité de Shannon ou entropie (H, équation 5.1), l'information mutuelle à deux régions (MI, équation 5.2), l'information mutuelle relative (rMI, équation 5.3) et le gain relatif d'information (rIG, équation 5.4) [Shannon, 1948, Butte and Kohane, 1999, Hausser and Strimmer, 2009].

2. National Center for Biotechnology Information.

$$H_x = -\sum_{i=1}^{S} P_i \cdot log2(Pi) \tag{5.1}$$

ÉQUATION 5.1 – Indice de Shannon H (ou entropie de Shannon) d'une région x possédant S séquences différentes, Pi étant la fréquence de la séquence i et $log2$ le logarithme base 2 (ainsi l'unité est le bit).

$$MI_{x,y} = H_x + H_y - H_{x,y} \tag{5.2}$$

ÉQUATION 5.2 – Information mutuelle MI à deux régions x et y à partir de l'information de la région x (H_x), de la région y (H_y) et de l'information des deux régions considérées ensemble ($H_{x,y}$).

L'information mutuelle décrit la quantité d'information commune à deux régions, c'est-à-dire quelle quantité de l'information d'une seconde région est déjà contenue dans la première (et est donc redondante).

La Figure 5.1 illustre l'information relative de deux régions géniques différentes à deux loci différents. Pour le premier locus, l'information mutuelle (en bleu) est réduite, indiquant que les deux régions géniques fournissent une information différente et peu redondante (les deux régions possédant une information strictement complémentaire et non redondante si et seulement si $MI = 0$). Pour le second locus, bien que les deux régions géniques montrent une quantité d'information plus élevée, l'information mutuelle relative est elle aussi plus importante, indiquant qu'une importante fraction de l'information de la région 2 (la moins informative à ce locus) est déjà contenue dans la région 1 (toute l'information de la région 2 serait contenue dans la région 1 si et seulement si $MI = H_{region\ 2}$). Ainsi, dans le cas où l'information de la région 1 est disponible (pour les deux loci), la région 2 du second locus est moins informative que la région 2 du premier locus (malgré une valeur de H plus importante).

L'information mutuelle à deux régions dépendant directement de l'information de chacune des régions, il n'est pas possible de comparer les informations mutuelles absolues entre elles. À cette fin, il faut alors utiliser l'information mutuelle relative (equation 5.3), qui correspond à la proportion de l'information de la région la moins informative déjà contenue dans la région la plus informative (l'inverse étant impossible).

$$rMI_{x,y} = \frac{MI_{x,y}}{min(H_x, H_y)} \tag{5.3}$$

ÉQUATION 5.3 – Information mutuelle relative à deux régions, $MI_{x,y}$ étant l'information mutuelle de ces deux régions (voir équation 5.2) et $min(H_x, H_y)$ la plus petite valeur entre H_x et H_y.

Ainsi l'information mutuelle relative (rMI) est définie entre 0 (les deux régions apportent une information différente et non redondante) et 1 (l'information de la région la moins informative est entièrement contenue dans la région la plus informative).
L'information mutuelle relative étant définie entre 0 et 1, il est alors possible de définir

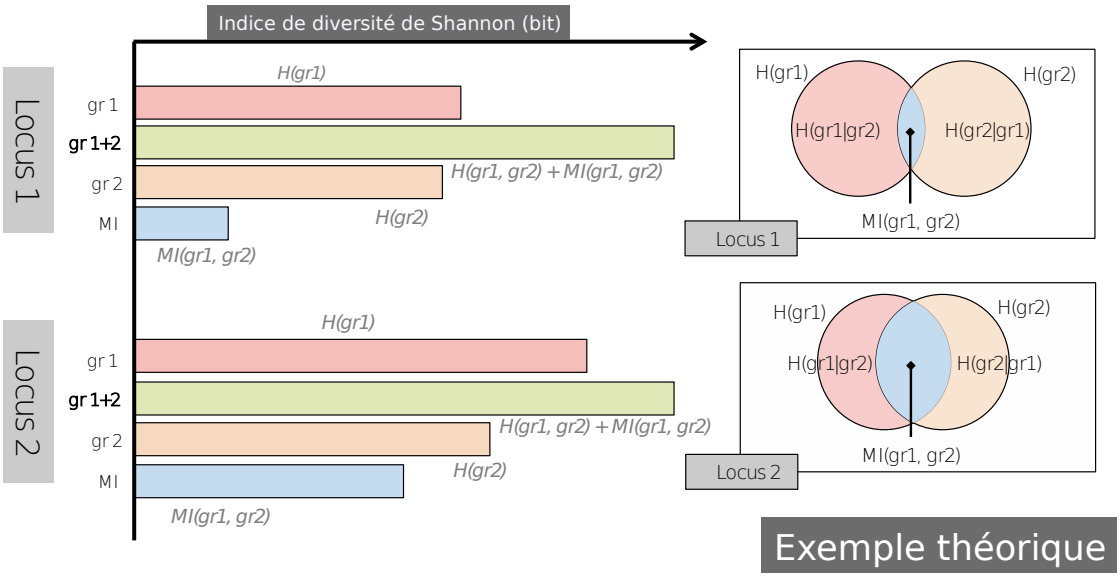


FIGURE 5.1 – Illustration de l'information mutuelle à l'aide de deux loci théoriques 1 et 2, chacun composé de deux régions géniques gr1 et gr2. L'entropie (équation 5.1) des gr1 est représentée par des barres rouges, celle des gr2 par des barres oranges, les barres en vert représentant l'information totale des deux régions pour chaque locus et les barres bleues représentant l'information mutuelle (équation 5.2). Les diagrammes de Venn, à droite, utilisent le même code couleur mais ne reprennent pas exactement les mêmes proportions que pour le graphique en barres, à gauche. Malgré une quantité d'information identique aux deux loci (barres vertes), le locus 1 possède des régions moins informatives, mais aussi une information mutuelle plus réduite tandis que le locus 2 montre des régions plus informatives assorties d'une information mutuelle plus élevée.

une autre mesure appelé gain relatif d'information (equation 5.4).

$$rIG_{x,y} = 1 - rMI_{x,y} \tag{5.4}$$

ÉQUATION 5.4 – Gain relatif d'information rIG à partir de l'information mutuelle relative rMI calculée avec l'équation 5.3.

Cette mesure représente la proportion d'information de la région la moins informative qui n'est pas contenue dans la région la plus informative et varie de 0 (si l'information de la région la moins informative est totalement redondante avec celle de la région la plus informative) à 1 (si l'information de la région la moins informative est totalement différente de celle de la région la plus informative).

2.3 Biais des bases de données

Les bases de données `gen` et `nuc` de IPD-IMGT/HLA ne présentent pas les mêmes données (voir sous-section 2.1). En effet, la base de données `gen` ne présentant que les allèles pour lesquels l'ensemble des régions sont entièrement séquencées, elle possède moins d'allèles que la base de données `nuc`, bien qu'étant entièrement comprise dans cette

dernière (du moins les exons). La question est alors de savoir si ce sous-échantillonnage a été réalisé aléatoirement ou si seuls certains allèles sélectionnés par avance ont été entièrement séquencés. De plus, afin de pouvoir utiliser l'une ou l'autre des bases de données, il est nécessaire de s'assurer que les éventuels biais d'échantillonnage soient eux aussi aléatoires (non dépendants de l'entropie des régions).
Pour cela, deux statistiques seront calculées en se basant sur les entropies précédemment calculées. La première statistique, appelée R1, consiste en un rapport entre l'entropie d'une région génique observée dans une des bases de données et l'entropie maximale que pourrait posséder cette région dans cette base de données (équation 5.5). L'entropie maximale est donnée par $H^{max} = log2(S)$ et correspond au cas où l'ensemble des séquences uniques possèdent la même fréquence (et sont donc équiprobables).

$$R1 = \frac{H_i}{H_i^{max}} \tag{5.5}$$

ÉQUATION 5.5 – Rapport R1, mettant en rapport l'entropie H_i observée à une région i avec l'entropie maximale théoriquement observable à cette région.

La seconde statistique, appelée R2 (équation 5.6), met en relation les rapports R1 (d'une même région) calculés sur les deux bases de données. Si ce rapport est proche de 1, alors les contenus de l'une ou l'autre des bases de données ne sont pas biaisés. Si l'entropie observée est N fois plus petite dans une des bases de données que dans une autre, mais que l'entropie maximale est aussi N fois plus petite, alors il s'agit d'un échantillonnage aléatoire.

$$R2 = \frac{R1_i^{gen}}{R1_i^{nuc}} \tag{5.6}$$

ÉQUATION 5.6 – Calcul du rapport R2, mettant en relation les rapports R1 calculés pour les bases de données `gen` et `nuc`.

2.4 Comparaison des régions géniques

Les données génétiques HLA provenant de IPD-IMGT/HLA (voir sous-section 2.1) ont été premièrement séparées par régions géniques (les différents UTR, introns et exons) et seules les séquences uniques ont été conservées. Ensuite, à l'instar des analyses d'entropie, les régions présentant des positions non séquencées (indiqués par des ***** dans les fichiers de la base de données) ont été retirées de l'étude. Ensuite, l'ensemble des *gaps* (représentés par des **-** dans la base de données) ont été supprimés (la décomposition en chaînes de Markov ne s'appliquant que sur les quatre nucléotides ATCG).

Les données génétiques de Patr n'étant pas disponibles sous la forme d'alignements (comme le sont les données HLA), elles ont d'abord été alignées sur des gènes HLA de référence complets (ne disposant pas de gènes Patr entièrement séquencés) à l'aide de MAFFT [Katoh, 2002]. Les gènes HLA de référence utilisés étaient HLA-A*01:01:01:01 pour Patr-A et Patr-Al, HLA-B*07:02:01:01 pour Patr-B et HLA-C*01:02:01:01 pour Patr-C. Les séquences des exons des gènes de référence ont ensuite été ajoutées et alignées (à l'aide de l'option `-add` de MAFFT), ce qui a permis de découper en régions

géniques les séquences Patr. Les données HLA retirées, seules les séquences ne comportant aucun nucléotide ambigu (selon la nomenclature IUPAC [IUPAC-IUB, 1970]), présentant moins de 40% de différences (mesurées par la distance de Hamming) avec les autres séquences analogues et possédant moins de 20% de *gaps* ont été conservées. Ces valeurs ont été choisies afin de retirer les séquences les plus divergentes et donc potentiellement mal assignées. Seules les séquences correspondant aux exons 1 à 5 et aux introns 1 à 4 ont été utilisées, les régions suivantes étant faiblement couvertes et difficiles à aligner.

Pour les séquences HLA, comme pour Patr, seules les séquences uniques ont été conservées. La Table 5.1 fournit le nombre de séquences par gène et par région ayant été retenues pour la suite des analyses.

La décomposition en chaînes de Markov de premier ordre [3] a été réalisée à l'aide d'un script Python3 et, pour chaque séquence, les fréquences des états de transition entre les quatre nucléotides ont été calculées. Ces chaînes de Markov permettent alors de décrire chaque séquence par 16 variables différentes (illustrées sur la Figure 5.2), représentant alors 16 dimensions qui ont été ensuites réduites à deux par la méthode t-SNE [van der Maaten and Hinton, 2008].

		$i+1$			
		A	**T**	**C**	**G**
i	**A**	0.09	0.16	0.38	0.37
	T	0.19	0.19	0.30	0.32
	C	0.20	0.13	0.31	0.36
	G	0.29	0.11	0.28	0.32

FIGURE 5.2 – Illustration de la décomposition en chaîne de Markov de la séquence de l'exon 2 de HLA-A*01:01:01. Les données sont représentées en lignes, et donnent les probabilités d'observer chacun des quatre nucléotides à la position $i+1$ uniquement en connaissant le nucléotide à la position i.

Afin de déterminer les hyper-paramètres [4] de l'algorithme t-SNE, une méthode dite de *GridSearch* a été appliquée. Trois hyper-paramètres ont été ainsi évalués : le facteur de perplexité (valeurs testées : 20, 30, 50, 60), le facteur d'exagération (valeurs testées : 2, 12, 22) et la réalisation ou non d'une analyse en composantes principales (ACP) préliminaire à la t-SNE. Ces valeurs ont été choisies afin d'explorer une gamme représentative de ces paramètres. Les autres hyper-paramètres ont été laissés par défaut : deux dimensions finales, 50 composantes principales conservées pour les ACP (lorsque réalisées), θ de 0.5, 1'000 itérations maximum, ACP centrées mais non réduites (lorsque réalisées), *momentum* de 0.5 (final de 0.8), η de 200.
Dans le but de réduire les effets stochastiques liés à l'initialisation pseudo-aléatoire de la t-SNE, chaque combinaison des hyper-paramètres testée par *GridSearch* a été répliquée cinq fois et seule la projection t-SNE présentant la plus petite valeur de divergence de Kullback-Leibler a été conservée. Dans un but de reproductibilité des résultats, l'état initial du générateur pseudo-aléatoire de R a été fixé à 42.

3. C'est-à-dire en ne considérant que la transition du nucléotide n-1 au nucléotide n. Transition est ici utilisé au sens mathématique et non biochimique.

4. Paramètres externes au modèle dont les valeurs ne peuvent pas être déduites des données.

	HLA										Patr			
	A	**B**	**C**	**DRB1**	**DRB3**	**DRB4**	**DQA1**	**DQB1**	**DPA1**	**DPB1**	**A**	**Al**	**B**	**C**
5-UTR	99	155	108	34	3	2	30	38	9	4	–	–	–	–
Exon 1	17	25	17	5	1	1	9	7	2	2	5	1	3	5
Intron 1	36	33	38	63	3	3	50	23	26	89	9	4	1	7
Exon 2	270	303	229	35	2	1	12	59	8	90	23	3	6	17
Intron 2	49	80	47	70	3	2	13	48	5	98	–	2	16	13
Exon 3	323	361	338	10	2	2	13	58	7	24	28	2	43	24
Intron 3	50	52	43	19	3	2	16	15	4	12	14	5	–	10
Exon 4	60	50	68	5	–	1	12	6	7	5	13	2	11	9
Intron 4	15	11	13	14	2	1	–	15	–	19	7	3	–	6
Exon 5	19	21	24	4	–	–	–	1	–	1	5	1	5	8
Intron 5	43	38	26	25	3	2	–	18	–	–	–	–	–	–
Exon 6	8	4	5	1	–	1	–	2	–	–	–	–	–	–
Intron 6	11	16	22	–	–	–	–	–	–	–	–	–	–	–
Exon 7	4	6	6	–	–	–	–	–	–	–	–	–	–	–
Intron 7	20	–	17	–	–	–	–	–	–	–	–	–	–	–
Exon 8	2	–	1	–	–	–	–	–	–	–	–	–	–	–
3-UTR	106	257	109	31	3	2	31	25	19	26	–	–	–	–

TABLE 5.1 – Récapitulatif du nombre total de séquences uniques utilisées pour réaliser les décompositions en chaînes de Markov et les réductions de dimensionalité par t-SNE. selon le locus et la région génique. Les « – » signifient qu'aucune donnée n'était disponible. Les en têtes de colonnes HLA et Patr représentent, respectivement, les séquences des MHC humain et du chimpanzé.

3 Résultats

3.1 Calcul des entropies

La Table 5.2 présente les indices de Shannon ainsi que le décompte des haplotypes et séquences présentes dans les deux bases de données pour les exons 2 et 3 des huit gènes HLA de l'étude.
Les valeurs d'entropie obtenues pour chacune des régions (en plus des exons 2 et 3) de chacun des gènes, pour les deux bases de données, sont disponibles en annexe S-52.
Dans le cas des allèles HLA, la différence l'entropie des différentes régions donne une information sur les régions à cibler afin d'obtenir un typage avec peu (voire pas) d'ambigüités.

		N.hap – N.seq		**Ind. Shannon**	
Locus	**Région**	*gen*	*nuc*	*gen*	*nuc*
A	Exon 2	193 - 576	1'497 - 3'912	5.82	7.46
	Exon 3	234 - 576	1'820 - 3'912	6.21	8.25
B	Exon 2	203 - 655	1'715 - 4'765	6.09	7.96
	Exon 3	255 - 655	2'193 - 4'764	6.65	8.89
C	Exon 2	226 - 705	1'236 - 3'510	5.90	7.17
	Exon 3	305 - 705	1'773 - 3'509	6.64	8.53
DRB1	Exon 2	28 - 43	1'950 - 2'008	4.63	10.89
	Exon 3	10 - 43	41 - 306	2.92	3.41
DQA1	Exon 2	11 - 54	20 - 75	2.94	3.47
	Exon 3	11 - 54	20 - 71	2.80	3.48
DQB1	Exon 2	56 - 129	855 - 1'065	4.99	9.18
	Exon 3	64 - 129	180 - 417	5.07	5.90
DPA1	Exon 2	5 - 15	27 - 37	2.01	4.44
	Exon 3	6 - 15	7 - 18	2.34	2.39
DPB1	Exon 2	77 - 166	679 - 806	5.40	9.05
	Exon 3	7 - 166	26 - 220	1.54	2.27

TABLE 5.2 – Statistiques descriptives des exons 2 et 3 des huit gènes HLA de l'étude, illustrant les différences entre les bases de données *gen* et *nuc*. N.hap et N.seq sont, respectivement, le nombre d'haplotypes (séquences uniques) et le nombre de de séquences totales disponibles, Ind. Shannon correspond à l'indice de Shannon (entropie) calculé pour cette région.

Cette Table met en évidence plusieurs différences entre les loci et les régions. Pour les loci de classe I, les exons 3 sont toujours plus informatifs (indice de Shannon plus élevé) que les exons 2. Les deux loci de classe II, HLA-DQA1 et -DQB1, montrent des valeurs semblables pour les entropies des exons 2 et 3 calculées à partir de la base de données `gen`, mais pas à partir de la base de données `nuc`, suggérant une différence d'échantillonnage entre les bases de données, visible ici par une plus faible représentativité de l'exon 3 de ces deux loci dans la base de données `nuc`. Cette différence de représentativité s'observe pour l'ensemble des loci de classe II, avec une importante différence du nombres de séquences et d'haplotypes entre bases de données. Par exemple, le locus HLA-DRB1 exon 2 comptabilise 1'950 haplotypes pour 2'008 séquences dans `nuc`, alors que le même locus ne comptabilise que 28 haplotypes et 43 séquences dans la base de données `gen`).
De manière générale, la différence d'entropie calculée pour les mêmes loci/régions entre

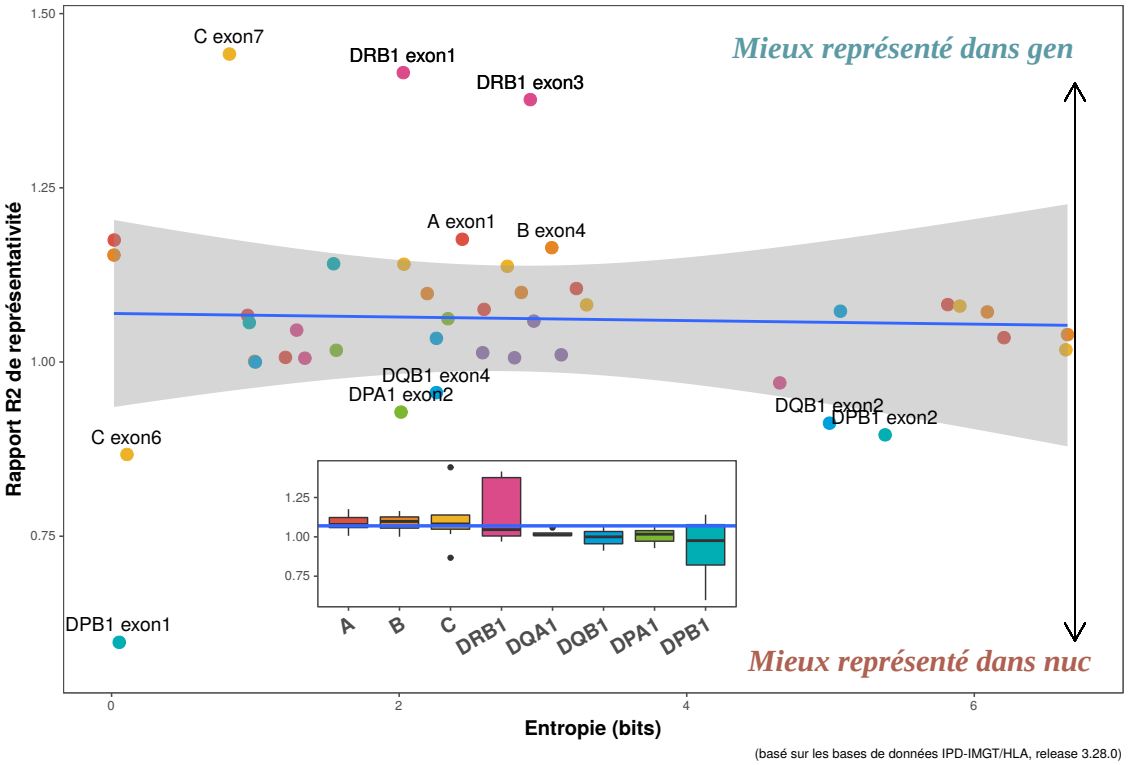


FIGURE 5.3 – Distribution des rapports R2 (correspondant aux rapports entre $R1_{gen}$ et $R1_{nuc}$) pour chacune des régions (exons) représentées dans les deux bases de données d'IPD/IMGT-HLA *gen* et *nuc*, en fonction de l'indice de Shannon (en bits) calculé d'après les séquences disponibles dans la base de données *gen*. La ligne bleue correspond à la régression linéaire entre le R2 et l'entropie, la zone grisée représentant l'intervalle de confiance à 95% de cette régression. Les étiquettes correspondent aux régions dont le R2 est hors de l'intervalle de confiance. Les boîtes à moustache (encart au milieu en bas) montrent la distribution de ces rapports R2 par gène. La couleur des points correspond aux différents gènes comme indiqué par les boîtes à moustaches.

`gen` et `nuc` varie selon les classes : pour les loci de classe I, le rapport H^{nuc}/H^{gen} moyen est de 1.83 ± 0.34, tandis que pour les classe II, ce rapport est de 1.99 ± 2.07. Les principales différences, observées pour les loci de classe II HLA-DRB1 exon 2 et HLA-DQB1 exon 2, indiquent que ce sont majoritairement les exons 2 qui ont été séquençés et non les gènes complets.

3.2 Biais des bases de données

La Figure 5.3 montre la distribution des rapports R2 pour chacune des régions géniques de chacun des loci étudiés, en fonction de l'entropie calculée à ces loci, d'après les séquences de la base de données `gen` (H_{gen}).

La ligne bleue horizontale correspond au résultat d'une régression entre les valeurs de R2 et de H_{gen}. L'ordonnée à l'origine de cette régression est de 1.07 et significative ($pValeur < 2e-16$) tandis que la pente n'est pas significativement différente de 0

($pValeur = 0.823$), indiquant qu'il n'y a pas d'effet significatif de l'entropie sur le rapport R2.

Le rapport R1 (equation 5.5) est calculé comme étant le rapport entre l'entropie observée à une région génique et l'entropie maximale atteignable à cette région. Le rapport R2, quant à lui, est le rapport des R1 calculés pour `gen` ($R1_{gen}$) et `nuc` ($R1_{nuc}$). De manière générale, pour l'ensemble des régions et des gènes, $R1_{gen}$ est supérieur de 0.03 ± 0.07 par rapport à $R1_{nuc}$, mais la différence n'est pas significative (test de Kruskal-Wallis, pValeur=0.1678).

Il est toutefois à noter que les régions ne possédant qu'un seul haplotype (*e.g.* HLA-DQB1 exon 5) ne sont pas représentées ici, puisque leur entropie maximale est de 0 ($log2(1)$), rendant le rapport R1 ($\frac{H}{H_{max}}$) incalculable.

Quelques régions se démarquent par un rapport $R2 >> 1$ (donc un $R1_{gen} >> R1_{nuc}$), indiquant qu'elles sont mieux représentées dans la base de données `gen` que `nuc` : HLA-B exon 4, HLA-C exon 7, HLA-DRB1 exon 1 et 3, HLA-DQB1 exon 6, HLA-DPA1 exon 2 et HLA-DPB1 exons 1 et 3. Deux autres régions se démarquent au contraire par un $R2 << 1$ (donc un $R1_{gen} << R1_{nuc}$), indiquant q'elles sont mieux représentées dans la base de données `nuc` que `gen`, HLA-C exon 6 et HLA-DPB1 exon 4.
Selon les boîtes à moustache en encadré de la Figure 5.3, HLA-DRB1 apparaît comme globalement mieux représenté dans `gen` (lié à ses exons 1 et 3) tandis que HLA-DPB1 semble mieux représenté dans `nuc` (lié à son exon 1 et, dans une moindre, à mesure son exon 2). Toutefois, aucun locus ni aucune région (*e.g.* les exons 2) ne semblent mieux représentés dans une base de données que dans une autre, indiquant qu'il n'y a pas de biais de représentativité entre les bases de données.

3.3 Distribution de l'entropie

La Figure 5.4 montre, pour chaque gène, la distribution de l'information par nucléotide (permet ainsi de comparer le contenu en information entre régions de différentes tailles) selon le type de région considérée (codante ou non codante).

Les huit gènes étudiés peuvent être classés en trois catégories selon l'information contenue dans leurs régions codantes ou non codantes. Premièrement, les gènes de classe I révèlent une information importante aussi bien à leurs régions codantes que non codantes. Ensuite, les trois gènes de classe II, HLA-DRB1, -DQA1 et -DQB1, montrent une différence entre les régions codantes, avec une information par nucléotide élevée et les régions non codantes montrant une information par nucléotide beaucoup plus faible. Finalement, les gènes HLA-DPA1 et -DPB1 montrent une information par nucléotide faible, que ce soit à leurs régions codantes ou non codantes.

Afin d'explorer comment l'information se distribue en fonction des régions et de leurs tailles, une analyse mettant en relation la taille et l'information des régions codantes a été réalisée par modèles linéaires.
Les modèles linéaires ont d'abord été réalisés en considérant séparement les régions codantes et non codantes, mais aussi les régions des gènes de classe I et celles des gènes de classe II. Les résultats de ces modèles sont représentés par les équations 5.7 à 5.10.

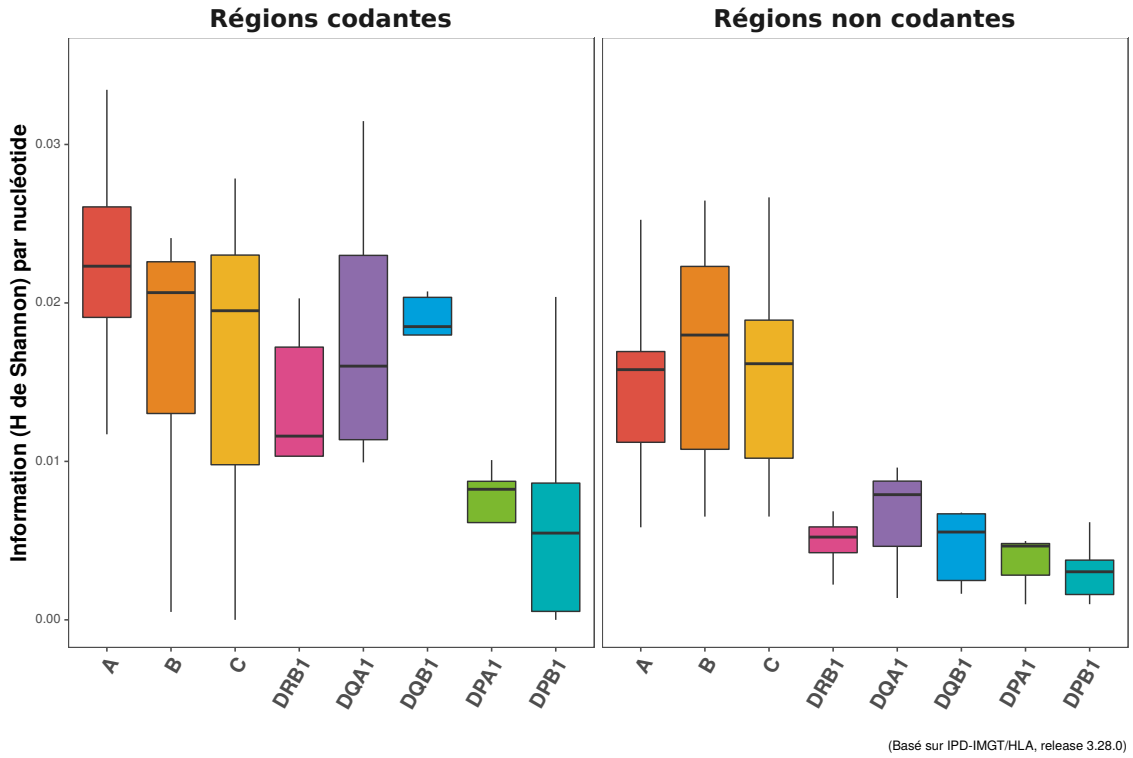


FIGURE 5.4 – Graphique en boîtes à moustaches illustrant, pour chacun des huit loci de l'étude (HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1), la quantité d'information (indice de Shannon) par nucléotide (en $bits \cdot nt^{-1}$) pour les régions codantes (exons, à gauche) et les régions non codantes (introns et UTR, à droite).

Classe I, régions codantes :

$$H = 1.84 \cdot 10^{-2} \cdot taille \tag{5.7}$$

ÉQUATION 5.7 – Résultat du modèle linéaire détaillant la relation entre l'entropie des régions géniques (`H`, en bits) et la taille de ces régions (en nucléotides) pour les régions codantes des gènes de classe I. Les modèles ont été réalisés à partir des données provenant de la base de données *gen* de IPD-IMGT/HLA v3.28.

Classe I, régions non codantes :

$$H = 2.38 + 2.47 \cdot 10^{-3} \cdot taille \tag{5.8}$$

ÉQUATION 5.8 – Résultat du modèle linéaire détaillant la relation entre l'entropie des régions géniques (`H`, en bits) et la taille de ces régions (en nucléotides) pour les régions non codantes des gènes de classe I. Les modèles ont été réalisés à partir des données provenant de la base de données *gen* de IPD-IMGT/HLA v3.28.

Classe II, régions codantes :

$$H = 1.25 \cdot 10^{-2} \cdot taille - 1.22 \cdot DP \tag{5.9}$$

ÉQUATION 5.9 – Résultat du modèle linéaire détaillant la relation entre l'entropie des régions géniques (`H`, en bits) et la taille de ces régions (en nucléotides) pour les régions codantes des gènes de classe II. `DP` représente une variable qualitative, prenant la valeur 1 (ou 0) si la région appartient (ou n'appartient) pas à HLA-DPA1/-DPB1. Les modèles ont été réalisés à partir des données provenant de la base de données *gen* de IPD-IMGT/HLA v3.28, en excluant l'intron 1 de HLA-DRB1 à cause d'une couverture incomplète.

Classe II, régions non codantes :

$$H = 3.05 + 6.87 \cdot 10^{-4} \cdot taille - 1.55 \cdot DP \tag{5.10}$$

ÉQUATION 5.10 – Résultat du modèle linéaire détaillant la relation entre l'entropie des régions géniques (`H`, en bits) et la taille de ces régions (en nucléotides) pour les régions non codantes des gènes de classe II. `DP` représente une variable qualitative, prenant la valeur 1 (ou 0) si la région appartient (ou n'appartient) pas à HLA-DPA1/-DPB1. Les modèles ont été réalisés à partir des données provenant de la base de données *gen* de IPD-IMGT/HLA v3.28, en excluant l'intron 1 de HLA-DRB1 à cause d'une couverture incomplète.

Les coefficients pour les régions codantes de loci de classe I et II étant du même ordre de grandeur, trois modèles imbriqués ont été testés. Le premier, `m1`, ne considère que la taille des séquences pour expliquer l'entropie de la région, le deuxième, `m2`, considère, en plus, les régions des gènes HLA-DP séparément (de manière similaire aux équations 5.9 et 5.10) et le troisième modèle, `m3`, considère en plus la classe (classe I ou II). Seul le premier modèle diffère des deux autres, après comparaisons deux à deux des modèles par ANOVA (m1-m2 : pValeur=0.0002 et m1-m3 : pValeur=0.0004). Il apparaît donc que pour les régions codantes, seule la taille et l'appartenance à HLA-DPA1 ou -DPB1 compte, la

classe n'étant pas un facteur significatif.
Les mêmes modèles ont été réalisés sur les régions non codantes, aboutissant à la même conclusion (pValeurs $< 10^{-6}$).

Il est donc possible de fusionner les équations 5.7 et 5.9, ainsi que 5.8 et 5.10 :

Régions codantes :

$$H = 1.55 \cdot 10^{-2} \cdot taille - 1.59 \cdot DP \tag{5.11}$$

ÉQUATION 5.11 – Modèles linéaires détaillant la relation entre l'entropie des régions géniques (`H`, en bits) et la taille de ces régions (en nucléotides) pour les régions codantes. Aucune différence significative n'étant observée entre les classes, les modèles regroupent ici les gènes de classe I et II. `DP` représente une variable qualitative, prenant la valeur 1 (ou 0) si la région appartient (ou n'appartient pas) à HLA-DPA1/-DPB1. Les modèles ont été réalisés à partir des données provenant de la base de données IPD-IMGT/HLA v3.28, en excluant l'intron 1 de HLA-DRB1 à cause d'une couverture incomplète.

Régions non codantes

$$H = 2.89 + 7.4 \cdot 10^{-4} \cdot taille - 1.48 \cdot DP \tag{5.12}$$

ÉQUATION 5.12 – Modèle linéaire détaillant la relation entre l'entropie des régions géniques (`H`, en bits) et la taille de ces régions (en nucléotides) pour les régions non codantes. Aucune différence significative n'étant observée entre les classes, les modèles regroupent ici les gènes de classe I et II. `DP` représente une variable qualitative, prenant la valeur 1 (ou 0) si la région appartient (ou n'appartient pas) à HLA-DPA1/-DPB1. Les modèles ont été réalisés à partir des données provenant de la base de données IPD-IMGT/HLA v3.28, en excluant l'intron 1 de HLA-DRB1 à cause d'une couverture incomplète.

La Figure 5.5 est une représentation graphique de ces modèles linéaires, illustrant la différence entre les régions codantes et non codantes, mais aussi entre les régions appartenant à HLA-DPA1/DPB1 ou non.

Les régions codantes accumulent toujours plus vite l'information avec la taille des séquences que les régions non codantes. Les régions codantes des loci de classe I sont celles qui accumulent le plus vite l'entropie avec la taille des séquences. Les régions non codantes de classe II sont, quant à elles, celles qui accumulent le moins vite l'information avec la taille des séquences, mais possèdent une valeur de base (l'ordonnée à l'origine) la plus importante (hormis pour les loci HLA-DPA1 et -DPB1).

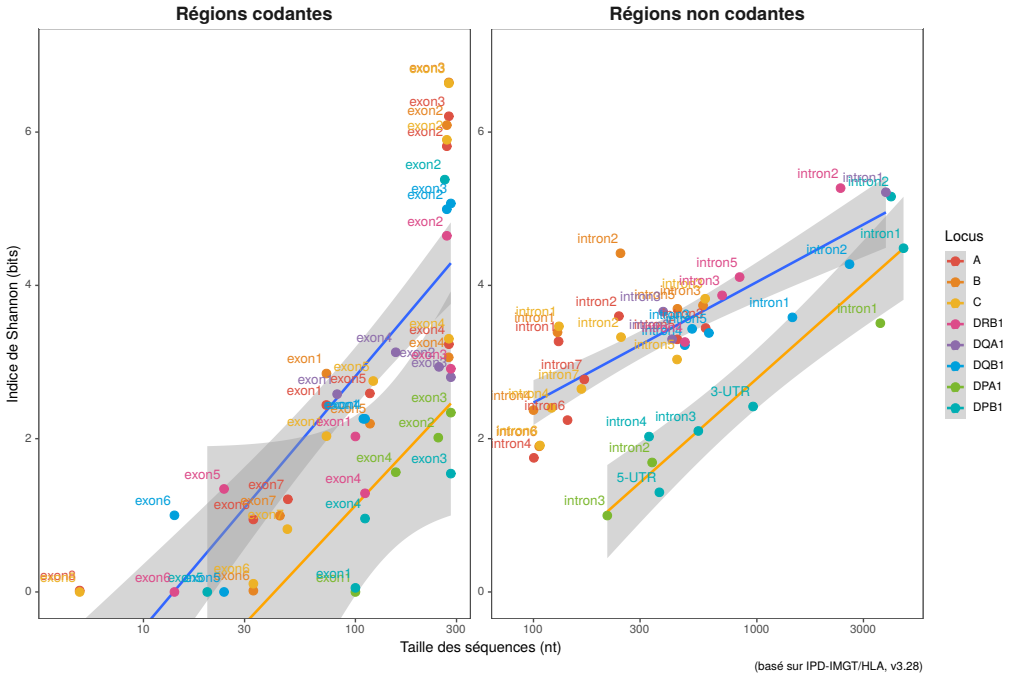


FIGURE 5.5 – Nuages de points illustrant la quantité d'information (indice de Shannon, en bits) des régions codantes (exons, à gauche) et non codantes (introns et UTR, à droite) des loci HLA de classe I et II en fonction de la taille (en nucléotides, nt) de ces régions. Attention à l'échelle logarithmique de l'axe des abscisses. Les différents gènes sont représentés par des couleurs spécifiques, les lignes continues (et les surfaces grisées) indiquent la relation linéaire (et l'intervalle de confiance à 95%) entre l'information et la taille de ces régions, la ligne bleue illustrant la relation observée pour les loci HLA-A, -B, -C, -DRB1, -DQA1 et -DQB1 tandis que la ligne orange illustre la relation pour HLA-DPA1 et -DPB1.

3.4 Information mutuelle et gain d'information

Les Figures 5.6 et 5.7 illustrent l'entropie (indice de diversité de Shannon) aux exons 2 et 3 (H2 et H3), l'entropie des exons 2 et 3 concaténés, ainsi que l'information mutuelle aux deux exons (equation 5.2), pour chacun des huit loci HLA de l'étude.

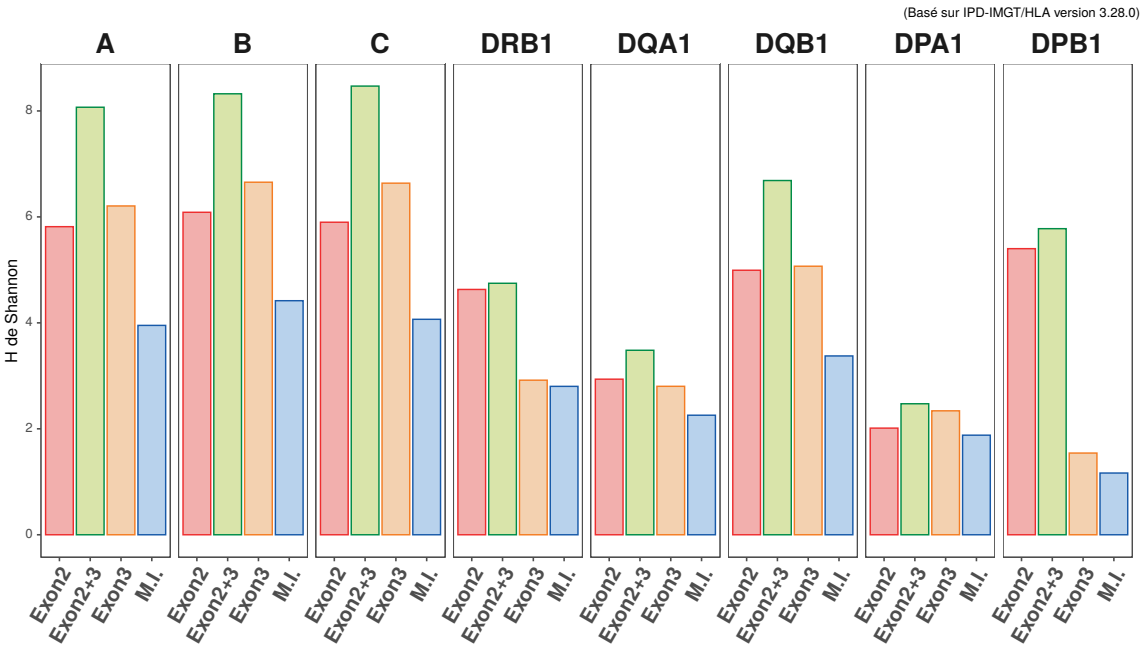


FIGURE 5.6 – Représentation graphique des indices de Shannon (entropie, ici en bits) des exons 2 (en rouge), 3 (orange), des exons 2 et 3 considérés conjointement (vert) et de l'information mutuelle de ces deux exons (bleu), pour les loci HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1.

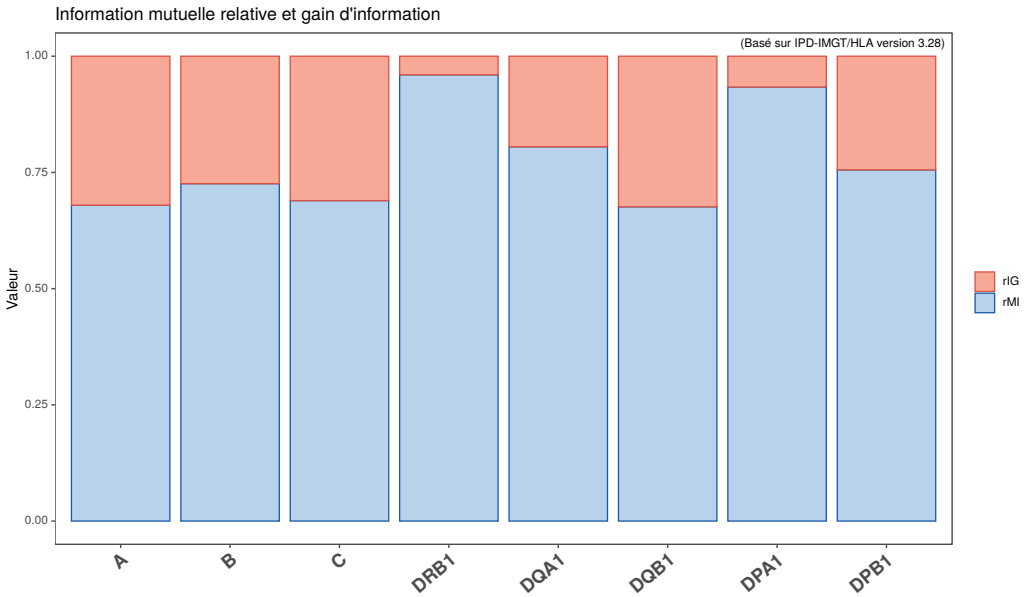


FIGURE 5.7 – Graphique en barres empilées illustrant la répartition de l'information mutuelle relative rMI (en bleu) et le gain d'information relatif rIG (en rouge) pour les exons 2 et 3 de chacun des huit loci HLA de l'étude (HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, -DPA1 et -DPB1).

La Table 5.3 donne les valeurs numériques de ces indices ainsi que l'information mutuelle relative (correspondant à l'information mutuelle standardisée de 0 à 1 pour pouvoir être comparée entre les loci) et le gain relatif d'information lorsque les deux exons sont considérés à la place d'un seul.

Locus	$N^A_{2,3}$	$N^S_{2,3}$	H_2	H_3	$H_{2,3}$	MI	rMI	rIG	H_{max}	$H_{2,3}/H_{max}$
A	414	576	5.82	6.21	8.07	3.95	0.68	0.32	9.17	0.88
B	516	655	6.09	6.65	8.32	4.42	0.73	0.27	9.36	0.89
C	583	705	5.90	6.64	8.47	4.07	0.69	0.31	9.46	0.90
DRB1	30	45	4.63	2.92	4.75	2.80	0.96	0.04	5.49	0.86
DQA1	25	54	2.94	2.80	3.48	2.26	0.81	0.19	5.75	0.61
DQB1	112	129	4.99	5.07	6.69	3.38	0.68	0.32	7.01	0.95
DPA1	9	15	2.01	2.34	2.47	1.88	0.93	0.07	3.91	0.63
DPB1	89	166	5.40	1.54	6.78	1.17	0.76	0.24	7.38	0.78

TABLE 5.3 – Pour chacun des huit gènes HLA de l'étude (HLA-A, -B, -C, -DPB1, -DQA1, -DQB1, -DPA1 et -DPB1) : $N^S_{2,3}$: nombre de séquences d'exons 2 et 3 combinés ; $N^A_{2,3}$: nombre d'allèles (séquences uniques) définis par les exons 2 et 3 combinés ; H_2, H_3, $H_{2,3}$: indices de diversité de Shannon (entropie) aux exons 2, aux exons 3 et aux deux exons considérés ensemble ; MI : information mutuelle entre ces deux exons ; rMI : information mutuelle relative ; rIG : gain d'information à considérer les deux exons ensemble ; H_{max} : entropie maximale atteignable (calculée avec $log_2(N^S_{2,3})$) ; $H_{2,3}/H_{max}$: rapport entre l'entropie des exons 2 et 3 et l'entropie maximale. Base de données *gen* de IPD-IMGT/HLA, version 3.28.

Les loci de classe I sont similaires entre eux et diffèrent des loci de classe II en montrant des indices H élevés à chacun des exons 2 et 3 (l'exon 3 étant supérieur), avec des valeurs de rMI les plus faibles des huit loci (en moyenne de 0.70) et donc un rIG important, indiquant que les deux exons possèdent une importante quantité d'information (indiqué par H) qui est peu redondante.

Pour les loci de classe II, les motifs observés sont différents. HLA-DRB1 possède un exon 2 très informatif, similaire aux loci de classe I, mais aussi une importante information mutuelle relative ($rMI = 0.96$) et un très faible gain d'information relatif ($rIG = 0.04$), indiquant que la majeure partie de l'information de l'exon 3 est déjà contenue dans l'exon 2 (son information est fortement redondante).
HLA-DPB1 possède les exons 2 et 3 les plus informatifs de tous les classe II ($H_{2,3} = 6.78$). Bien que son exon 3 soit le moins informatif des huit loci étudiés ($H_3 = 1.54$), son information est très peu redondante avec celle de l'exon 2 puisque l'information mutuelle relative est la seconde plus petite des loci de classe II ($rMI = 0.76$), derrière HLA-DQB1.
HLA-DQB1, quant à lui, est similaire aux classe I dans le sens où son exon 3 est sensiblement plus informatif que son exon 2 ($H_2 = 4.99$ et $H_3 = 5.07$) et une information mutuelle relative proche des loci de classe I ($rMI = 0.68$), indiquant que l'information est répartie de manière équilibrée entre les deux exons.
HLA-DQA1 montre des exons 2 et 3 moins informatifs que HLA-DQB1, avec tout de même un rIG de 0.19, indiquant un gain d'information substantiel lorsque l'on considère les deux exons au lieu d'un seul.
Les résultats pour HLA-DPA1 sont moins clairs et sont difficilement interprétables puisque ce locus possède peu d'allèles séquençés sur les deux exons (seules 15 séquences

sont disponibles dans la base de données *gen* d'IPD/IMGT-HLA pour ce locus). Ce problème touche plus généralement les loci de classe II, puisque, dans la version de la base de données utilisée (version 3.28), $99.98 \pm 0.02\%$ de tous les allèles de classe I connus ont leurs exons 2 et 3 entièrement séquencés, contre seulement $34.05 \pm 30.71\%$ des allèles de classe II. Cela indique que, pour les classe I, la grande majorité des séquences des exons 2 et 3 de tous les allèles connus était disponible, avec peu de différences entre les trois loci, tandis que, pour les gènes de classe II, beaucoup moins de séquences d'exons 2 et 3 étaient disponibles, avec de grandes différences entre les cinq gènes. Cela influence directement les entropies calculées, puisque l'entropie maximale qu'un locus puisse atteindre (en supposant que chaque séquence soit différente) est donnée par $log2(S)$, où S est le nombre de séquences différentes disponibles pour une région.

3.5 Comparaison des régions géniques

Après application de la méthode *GridSearch* pour explorer les résultats de 120 t-SNE différentes, les paramètres retenus, correspondant à la t-SNE montrant la plus petite divergence de Kullback-Leibler (KL), sont présentés dans la Table 5.4 :

Paramètre	**Valeur**
θ	0.5
Perplexité	60
KL final	0.421
η	200
Exagération	12
ACP préliminaire	Non

TABLE 5.4 – Paramètres retenus (pour la meilleure t-SNE) par la méthode *GridSearch* parmi les différentes t-SNE réalisées. La meilleure t-SNE a été sélectionnée sur la base de la plus faible valeur de divergence de KL. Les paramètres en gras correspondent à ceux explorés par *GridSearch*.

La t-SNE a été réalisée sur l'ensemble des données disponibles (gènes HLA de classe I et classe II, et gènes Patr) et les Figures 5.8, 5.9 et 5.10 sont issus de cette même t-SNE mais tous les loci ne sont pas simultanément représentés.
Un des principaux avantages d'utiliser une t-SNE, dans le cas présent, est la conservation de la proximité entre les points : deux points qui sont proches dans l'espace en hautes dimensions (les 16 dimensions produites par la décomposition en chaînes de Markov) restent proches dans l'espace en faibles dimensions (les deux dimensions de la projection t-SNE).

La Figure 5.8 montre le résultat, pour les gènes de classe I de l'humain (HLA-A, HLA-B, HLA-C) et du chimpanzé (Patr-A, Patr-AL, Patr-B et Patr-C), de la décomposition en chaînes de Markov et de la projection par t-SNE.

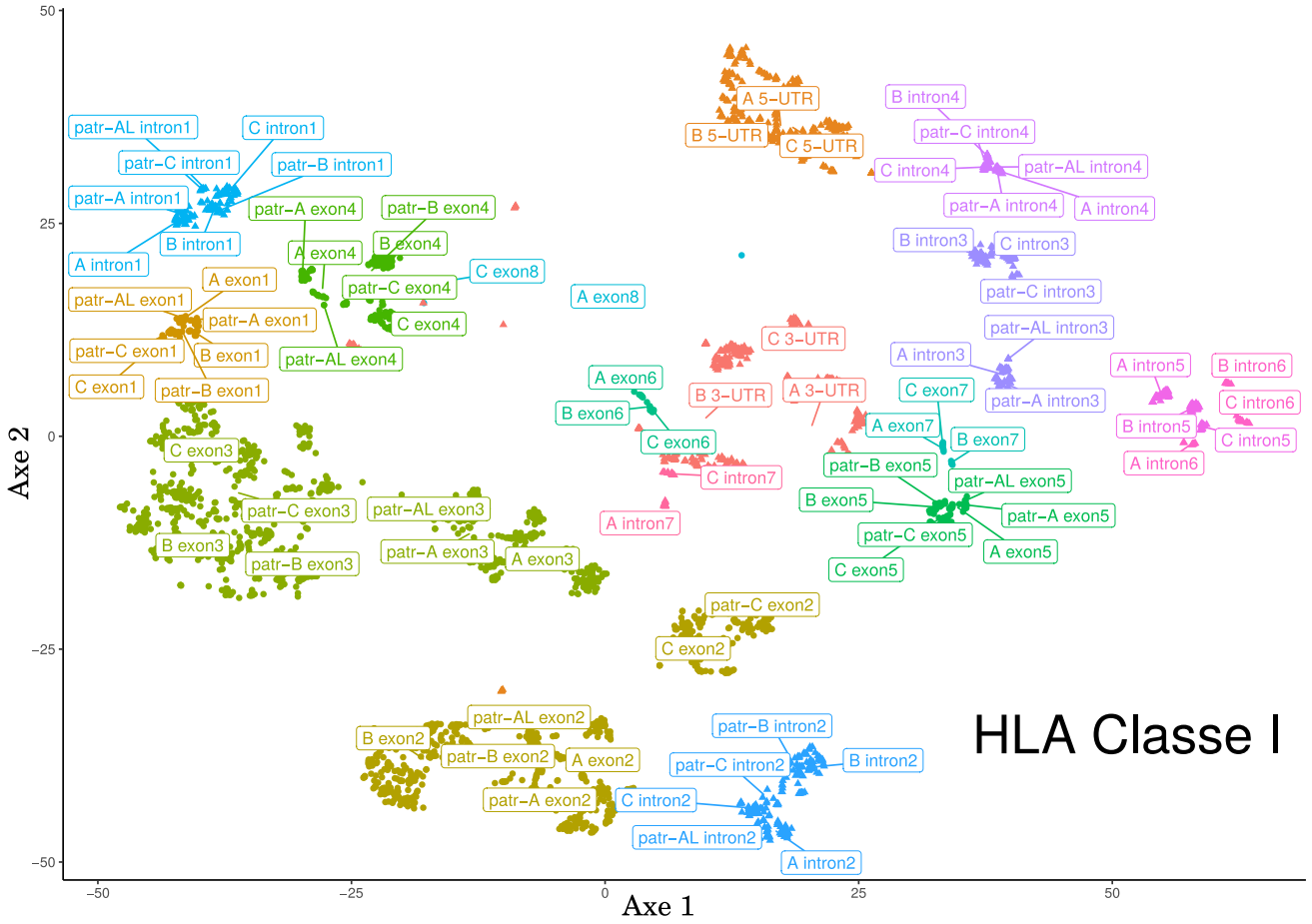


FIGURE 5.8 – Représentation graphique de la projection t-SNE des chaînes de Markov d'ordre 1 issues des séquences des différentes régions géniques pour chacun des loci de classe I de l'humain (HLA-A, -B et -C) et du chimpanzé (Patr-A, -AL, -B, -C). Les couleurs correspondent aux différentes régions géniques, les ronds aux régions codantes (exons) et les triangles aux régions non codantes (introns et UTR). Les étiquettes montrent la position des centroïdes pour chacune des régions géniques de chacun des loci.

La figure révèle que les mêmes régions géniques de différents gènes se regroupent entre elles, alors que les différentes régions d'un même gène ne se regroupent pas. Il existe donc une similarité plus grande entre les mêmes régions de différents gènes plutôt qu'entre les différentes régions d'un même gène. Les séquences correspondant aux loci Patr sont, elles aussi, groupées par régions avec les séquences HLA, illustrant une similarité inter-spécifique de ces régions.
Deux exceptions sont à noter, tout d'abord les séquences des exons 3 de HLA-A, Patr-A et Patr-AL sont séparées des séquences des exons 3 des autres loci, et les séquences des exons 2 des loci HLA-C et Patr-C sont elles aussi séparées de celles des exons 2 des autres loci.
Une proximité entre les séquences des exons 2 et des introns 2 est à noter, de même qu'entre les séquences des introns 1 et des exons 2, alors que cette proximité n'est retrouvée pour aucune des autres régions (par exemple les séquences des exons 4 et des introns 4 sont éloignées, de même que les séquences des exons 3 et introns 3).

La Figure 5.9 montre le résultat, pour les gènes de classe II (HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1 et HLA-DPB1), de la décomposition en chaînes de Markov et de la projection par t-SNE.

Les projections t-SNE des séquences des classe II diffèrent en plusieurs points de la projection des séquences de classe I. En premier lieu, une limitation est due au nombre de données disponibles, puisque nous disposions seulement de 1'319 séquences pour l'ensemble des 5 gènes de classe II, contre 3'966 pour les gènes de classe I. Ce plus faible nombre de données explique que les résultats soient plus clairsemés pour les loci de classe II que pour les loci de classe I.
Un motif particulier est observé concernant les exons 2. Les exons 2 des gènes codant pour les chaînes β (HLA-DRB1, -DQB1 et -DPB1) sont groupés ensemble (en bas à droite) et éloignés des autres régions tandis que les exons 2 des gènes codant pour les chaînes α sont plus disséminés et isolés les uns des autres. Concernant les exons 4, ceux des gènes codant pour les chaînes β sont regroupés ensemble et ceux des gènes codant pour les chaînes α sont eux aussi groupés de leur côté. Pour les exons 3, à part pour HLA-DQA1 exon 3, les séquences sont toutes regroupées.
Un autre résultat intéressant concerne les exons 2 de HLA-DPB1, dont un agrandissement est visible dans l'encadré en bas à gauche du graphique, où les séquences se projettent en deux groupes d'allèles distincts (liste disponible en annexe S-54).

La Figure 5.10 montre le résultat, pour les gènes HLA-DRB1, HLA-DRB3 et HLA-DRB4, de la décomposition en chaînes de Markov et de la projection par t-SNE.

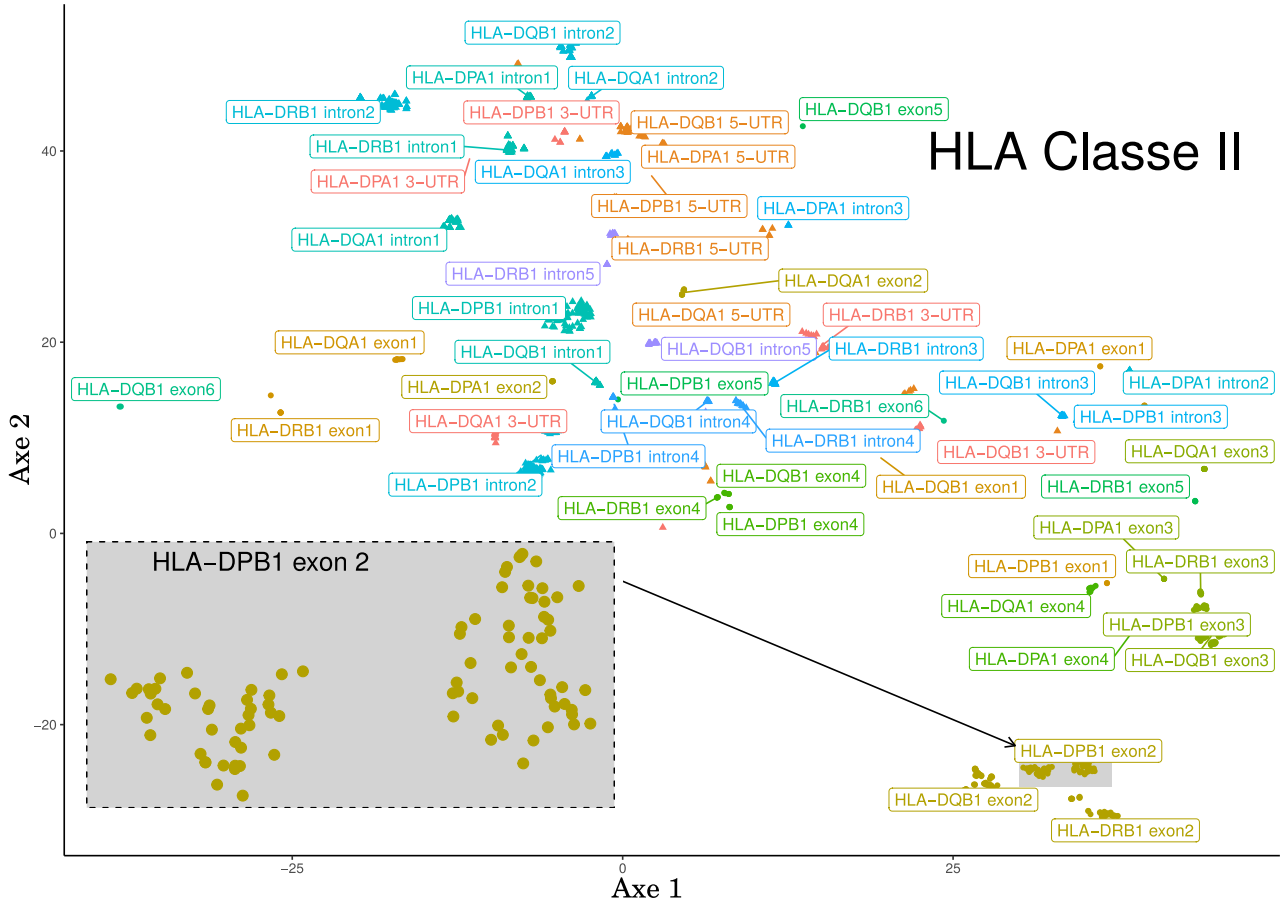


FIGURE 5.9 – Représentation graphique de la projection t-SNE des chaînes de Markov d'ordre 1 issues des séquences des différentes régions géniques pour chacun des loci de classe II (HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1 et HLA-DPB1 ; HLA-DRB-3 et HLA-DRB4 ne sont pas représentés). Les couleurs correspondent aux différentes régions géniques, les ronds aux régions codantes (exons) et les triangles aux régions non codantes (introns et UTR). Les étiquettes montrent la position des centroïdes pour chacune des régions géniques de chacun des loci. L'encadré grisé en bas à gauche est un agrandissement de la zone correspondant aux séquences de l'exon2 de HLA-DPB1.

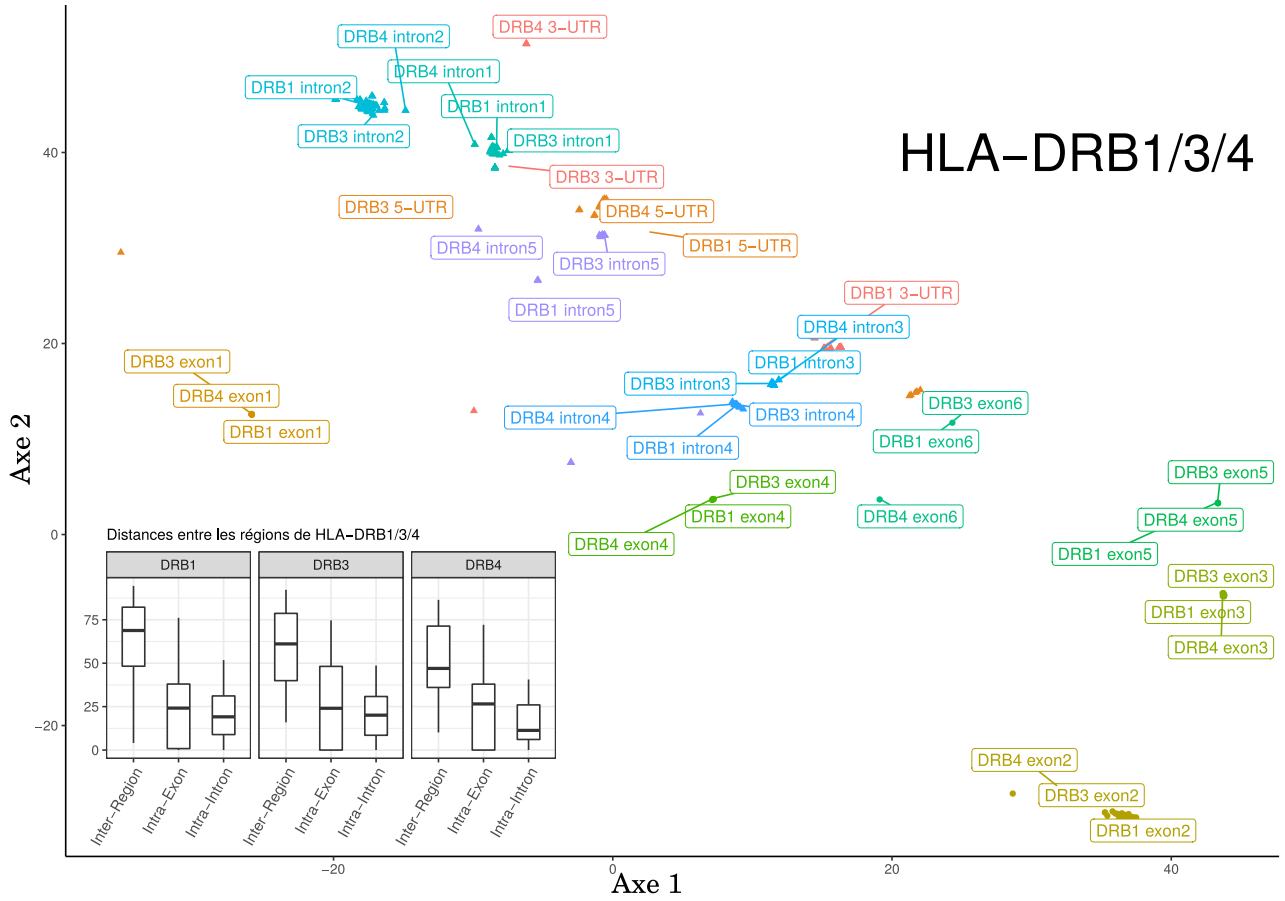


FIGURE 5.10 – Représentation graphique de la projection t-SNE des chaînes de Markov d'ordre 1 issues des séquences des différentes régions géniques pour chacun des loci HLA-DRB1/3/4. Les couleurs correspondent aux différentes régions géniques, les ronds aux régions codantes (exons) et les triangles aux régions non codantes (introns et UTR). Les étiquettes montrent la position des centroïdes pour chacune des régions géniques de chacun des loci. L'encadré grisé en bas à gauche correspond à la distribution des distances sur la tSNE entre les régions codantes (Intra-Exon), non-codantes (Intra-Introns), ou entre les régions codantes et non codantes (Inter-Region).

De manière similaire aux autres loci de classe I et II, les mêmes régions géniques des différents gènes se regroupent entre elles, alors que les différentes régions d'un même gène ne se regroupent pas.

Le résultat probablement le plus intéressant est visible dans l'encadré en bas à gauche. Ces boîtes à moustaches représentent les distributions des distances (sur la t-SNE) entre les séquences, soit entre les introns (Intra-Intron), soit entre les exons (Intra-Exon) soit entre les introns et les exons (Inter-Region). La projection t-SNE respectant la proximité entre les points, deux séquences similaires en hautes dimensions (les 16 dimensions données par la décomposition en chaînes de Markov) se retrouveront proches sur la projection en petites dimensions (les deux dimensions de la t-SNE). Il apparaît alors que les introns sont généralement similaires entre eux, que les exons sont aussi similaires entre eux, mais que les exons et les introns sont assez dissimilaires.

4 Discussion

Cette étude porte sur une description statistique du contenu des deux bases de données IPD-IMGT/HLA, `gen` et `nuc`. Premièrement, le contenu des bases de données a été évalué à l'aide d'une métrique dérivée de l'entropie afin de vérifier l'absence de biais dans le contenu de chacune d'entre elles. Ensuite, l'entropie ainsi que trois autres métriques reliées (entropie conjointe relative et absolue, gain relatif d'information) ont été utilisées afin de décrire la distribution de l'information le long de huit gènes HLA (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1 et HLA-DPB1) ainsi que la répartition de l'information entre leurs exons 2 et 3.
Pour finir, les huit gènes précédents ainsi que quatre gènes Patr du chimpanzé, homologues aux gènes HLA (Patr-A, Patr-AL, Patr-B et Patr-C), ainsi que les gènes HLA-DRB3 et HLA-DRB4 ont été étudiés à l'aide d'une décomposition en chaînes de Markov, suivie d'une réduction de dimensionalité à l'aide de l'algorithme t-SNE, en vue de valider la méthode au cœur du filtre markovien de `MADaM` (voir Chapitre 2, page 132), mais aussi d'étudier les similarités et différences entre les différentes régions de ces différents gènes et permettre une meillleure compréhension des processus évolutifs du MHC.

4.1 Entropies

L'analyse de la relation entre le rapport R2 et l'entropie des régions n'a pas montré de biais significatif dans le contenu de ces bases de données. Ainsi la base de données `gen` peut être considérée comme un sous-échantillonnage aléatoire de la base de données *nuc*. C'est-à-dire que la base de données *gen* présente, certes, moins d'allèles que *nuc*, mais ces allèles semblent avoir été échantillonnés aléatoirement parmi tout les allèles présents dans *nuc*. Cette conclusion est importante, puisque *gen* inclut aussi les données pour les régions non codantes (introns et UTR), permettant une comparaison entre les régions codantes et non codantes.

L'analyse de la distribution de l'entropie a montré que les régions codantes (des gènes classes I et II confondus) contiennent plus d'information que les régions non codantes à taille égale.
La Figure 5.5 montre que parmi les régions codantes, ce sont surtout les exons 2 et 3 (ainsi que l'exon 4 de HLA-C) qui possèdent une information importante. Ce résultat pourrait s'expliquer par le fait que les exons 2 (et 3 pour les gènes de classe I) sont soumis à une plus importante sélection, principalement sous la forme d'une sélection balancée qui va favoriser une plus grande diversité de ces régions (voir page 25). Toutefois, les autres exons ne participant pas directement à la présentation des antigènes montrent la même information par nucléotide que les exons 2 et 3, suggérant que si ces derniers sont plus informatifs, c'est uniquement lié à leur plus grande taille, comparé aux autres exons. Il reste alors à expliquer pourquoi les introns montrent une information par nucléotide bien plus faible que les exons (d'un facteur 10^{-2}, visible sur les équations de la Figure 5.12). Une étude de 1997 analysant les régions des introns 1 à 3 de HLA-A et HLA-B avait observé une importante conservation des séquences introniques au sein des différents lignages, suggérant une sélection négative des mutations ponctuelles sur ces introns, ces régions jouant probablement un rôle dans la conformation tridimensionnelle de l'ADN super-enroulé [Blasczyk et al., 2004]. Il a aussi été suggéré que les événements de conversion allélique entre les gènes de classe I étaient favorisés par la conservation d'une importante homologie entre les introns, facilitant ainsi l'appariement des régions variables

[Kourilsky, 1983]. Il est donc possible d'émettre une hypothèse similaire, expliquant la plus faible diversité des introns 1 à 7 de ces huit gènes HLA par une pression de sélection négative vis-à-vis des mutations ponctuelles sur les régions non-codantes.

4.2 Information mutuelle des exons 2 et 3

Les analyses de l'information mutuelle et du gain d'information ont révélé d'importantes différences entre les gènes de classe I et II, d'une part, et entre les gènes de classe II, d'autre part.

Les gènes de classe I semblent similaires quant à la répartition de leur information. Ils montrent une entropie des exons 2 élevée et une entropie des exons 3 plus élevée que celle des exons 2, une entropie conjointe très élevée associée à une information mutuelle relative la plus petite des huit loci, signifiant que l'information pouvant amener à l'identification précise d'un allèle se trouve partagée entre l'exon 2 et l'exon 3, tous deux étant nécessaires pour arriver à un typage avec peu d'ambiguïtés.

Les gènes de classe II montrent, au contraire, une plus importante hétérogénéité. Les trois gènes $II\beta$ possèdent une information des exons 2 et 3 plus importante que les deux gènes $II\alpha$ (Figure 5.6).
Ces trois gènes $II\beta$ divergent par contre quant à leur exon 3 et l'information mutuelle relative. HLA-DRB1 possède la majeure partie de son information dans l'exon 2, HLA-DQB1 possède un exon 3 aussi informatif que son exon 2 et l'exon 3 de HLA-DPB1, bien que peu informatif, montre peu de redondance de cette information avec son exon 2.
Pour les gènes de classe $II\alpha$, HLA-DQA1 montre un profil similaire à HLA-DQB1 puisque ses deux exons 2 et 3 sont similaires en termes d'information (bien que plus petite que l'information des exons 2 et 3 HLA-DQB1). HLA-DPA1 semble montrer lui aussi un exon 3 plus informatif que l'exon 2, mais le petit nombre de séquences disponibles rend incertain ce résultat.

En conclusion, à l'exception de HLA-DRB1 et HLA-DPA1 (ce dernier ne présentant pas assez de données pour assurer la fiabilité des résultats), les exons 3 des gènes de classe II possèdent aussi une information qui n'est pas redondante avec celle des exons 2. Ces résultats expliquent en partie les différences observées entre les diverses méthodes de typage des Mandenkalu pour les classe II, notamment pour HLA-DRB1 et HLA-DQB1, qui montrent de faibles taux de correspondances entre les typages PCR-SSO (ne ciblant à l'époque que l'exon 2) et NGS-MiSeq (couvrant entre autres les exons 2 et 3).

4.3 Décomposition en chaînes de Markov

A l'exception des régions pour lesquelles peu de séquences étaient disponibles, les séquences d'une région génique sont toujours regroupées ensemble, dans un groupe distinct et relativement compact, peu clairsemé. Ainsi les chaînes de Markov permettent d'isoler spécifiquement une région d'une autre, même entre les différents gènes. Cela justifie alors l'utilisation d'une décomposition en chaînes de Markov dans le filtre markovien de `MADaM`.

Les régions des gènes de classe I sont groupées par régions plutôt que par gènes et les séquences des régions Patr sont elles aussi regroupées avec les séquences HLA correspon-

dantes. Un modèle évolutif qui expliquerait ce résultat (schématisé dans la Figure 5.11) implique l'existence d'un gène MHC de classe I ancestral, au sein duquel les différentes régions (introns et exons) se seraient différenciées, acquérant leurs fonctionnalités. Ce gène se serait alors dupliqué, créant les loci A, B et C. Les différents introns et exons des gènes HLA et Patr étant groupés par région génique plutôt que par espèce, cela indique que cette duplication serait antérieure à la divergence entre les humains et les chimpanzés.

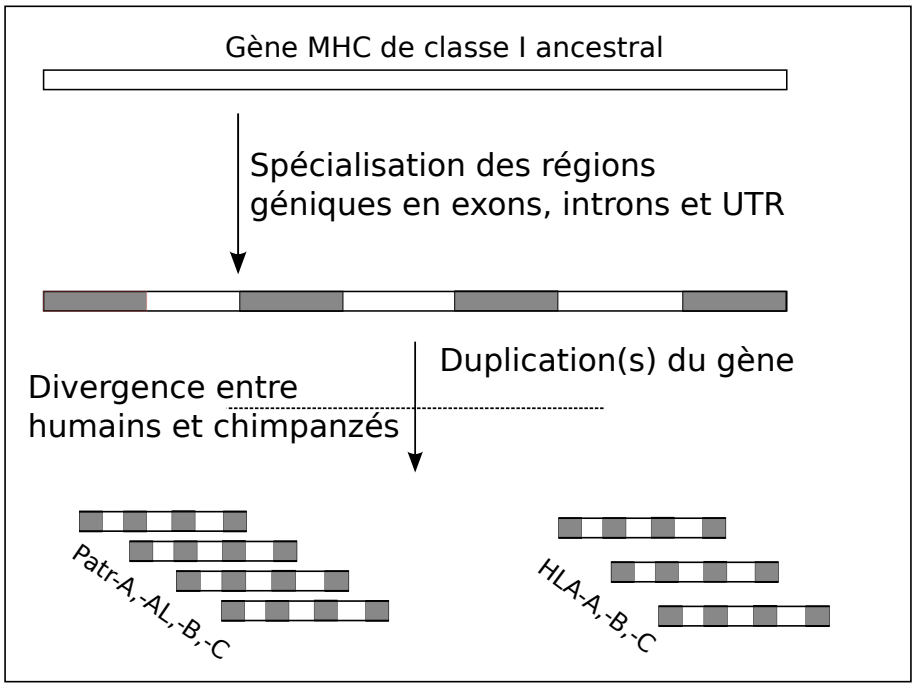


FIGURE 5.11 – Schéma illustrant l'hypothèse évolutive des gènes MHC de classe I proposée ici, dans laquelle les différentes régions (introns et exons) se seraient d'abord spécialisées dans un gène de classe I ancestral qui se serait ensuite dupliqué, le tout avant la divergence des humains et chimpanzés.

Concernant les régions de classe II, les exons 2 à 4 des gènes codant pour les chaînes β (HLA-DRB1, -DQB1 et -DPB1) sont groupés par régions homologues, alors que les régions des gènes codant pour les chaînes α ne le sont pas. Ces résultats rejoignent ceux obtenus par Takahashi en 2000, où les phylogénies des exons 2 à 5 des gènes $II\alpha$ et $II\beta$ n'étaient pas congruentes, suggérant une histoire évolutive différente de ces deux groupes de gènes [Takahashi, 2000].
Le fait que les régions des gènes $II\alpha$ ne soient pas groupées, à l'instar des régions des gènes $II\beta$, peut s'expliquer par le plus faible nombre de séquences disponibles pour ces régions. La Table 5.1 montre que pour les exons 2 à 4 des gènes HLA-DQA1, HLA-DQB1, HLA-DPA1 et HLA-DPB1 il y a en moyenne 4 fois plus de séquences pour les gènes β que pour les gènes α. Il est probable que le peu de séquences disponibles ne permettent pas de faire un regroupement de manière efficace, ce qui est observé étant alors non pas un effet de la sélection ou de l'évolution, mais simplement un artefact statistique dû à un manque de données. Ce manque de données pour HLA-DQA1 et HLA-DPA1 empêche alors de proposer une hypothèse évolutive (similaire à celle de la Figure 5.11) concernant les gènes de classe II.

Les exons 2 de HLA-DPB1, bien que regroupés ensemble (par rapport aux autres régions), présentent une sous-structure en deux groupes distincts non chevauchants, visibles sur la Figure 5.9.
Une récente étude menée par Vangenot *et al.*, comparant la diversité génétique du MHC des chimpanzés et des humains, a conclu que le locus DPB1 évoluait aussi bien par mutations ponctuelles que par recombinaison, au contraire des autres loci qui évoluent majoritairement par l'une ou l'autre des méthodes [Vangenot et al., 2020]. Ces résultats sont similaires à ceux obtenus précédemment par Otting *et al.*, dans une comparaison de la diversité du locus DPB1 des chimpanzés et macaques rhésus [Otting et al., 1998]. Il est possible que ces deux modes d'évolution pour HLA-DPB1 soient responsables de la structuration en deux groupes, soit par une évolution différente pour chacun des deux groupes (l'un regroupant des allèles évoluant majoritairement par mutation ponctuelle et l'autre par recombinaison), soit que la recombinaison ne se fasse pas aléatoirement entre les allèles, mais que les allèles d'un groupe échangent préferentiellement des séquences avec les allèles du même groupe.

Un résultat très intéressant est observé pour les loci HLA-DRB1/3/4, dont les exons sont similaires entre eux (proches sur la projection t-SNE), les introns aussi similaires entre eux, mais les introns et les exons très dissimilaires.
Une étude menée par Bergström *et al.* en 1998 a comparé les temps de divergence moyen, pour HLA-DRB1, entre les exons 2 d'une part et les introns 1 et 2 d'autre part [Bergström et al., 1998]. Cette étude avait alors estimé que le temps de divergence moyen entre les exons 2 était de 7.06 milions d'années, contre 250'000 ans pour les introns 1 et 2. La conclusion de cette étude est que les introns 1 et 2 ont une origine différente des exons 2. Une autre étude menée par Doxiadis *et al.* en 2008 est arrivée à la même conclusion en comparant les séquences des gènes DRB (introns 1 à 4 et exons 2) des humains, des chimpanzés et des macaques rhésus [Doxiadis et al., 2008b]. Dans cette étude, les phylogénies basées d'une part sur les exons 2 et d'autre part sur les introns 1 à 4 ne sont pas congruentes, la première montrant que les exons 2 des différentes espèces se mélangent tandis que la seconde phylogénie montre que les séquences introniques sont regroupées par espèces, indiquant un âge des introns bien plus récent.
Les résultats obtenus ici tendent à confirmer cette hypothèse, selon laquelle les introns et les exons de HLA-DRB1/3/4 auraient deux origines différentes, se traduisant par des distances plus importantes entre ces deux groupes qu'à l'intérieur de ces groupes.

5 Conclusion

Cette étude a analysé les bases de données IPD-IMGT/HLA à l'aide de la théorie de l'information, d'une part, et d'une décomposition en chaînes de Markov des séquences, d'autre part.

Dans une première partie, il a été mis en évidence que les régions codantes possèdent plus d'information (telle que définie par C. Shannon [Shannon, 1948]) que les régions non codantes, de même que les gènes HLA-DPA1 et HLA-DPB1 possèdent moins d'information (régions codantes et non codantes) que les autres loci HLA.
L'analyse de la distribution de cette information, ainsi que les analyses d'entropie conjointe, ont révélé une homogénéité des gènes de classe I entre eux, mais une forte hétérogénéité des gènes de classe II, où pour plusieurs d'entre eux (notamment HLA-DQB1) une partie non négligeable de l'information était localisée dans le troisième exon. Les méthodes de typage ne ciblant pour ces loci que l'exon 2 ne devraient alors pas chercher à assigner un allèle nominal, mais devraient considérer à la place une ambiguïté et donc l'ensemble des allèles possibles.

La deuxième partie de cette étude a apporté des indications quant à l'émergence des différents gènes de classe I chez l'humain et le chimpanzé, en supposant une origine antérieure à la séparation des deux lignées. Cette analyse a aussi mis en évidence, pour les gènes HLA-DRB1/3/4, la similarité des introns entre eux ainsi que des exons entre eux et la grande différence entre les introns et les exons, suggérant une origine différente des introns et des exons.

Comme précisé en introduction, cette étude ne représentait qu'une analyse exploratoire et plusieurs voies sont possibles pour poursuivre cette étude. Il serait intéressant d'étendre les analyses d'information mutuelle et de gain d'information en dehors des seuls exons 2 et 3, par exemple, il serait intéréssant de connaître l'information mutuelle entre l'intron 1 et l'exon 2 des différents gènes, ou entre l'intron 2 et l'exon 3.
Les analyses basées sur la décomposition en chaînes de Markov pourraient aussi intégrer des séquences MHC d'autres espèces dont la divergence des lignées vis-à-vis de la nôtre est plus ancienne que celle des chimpanzés (catarrhiniens par exemple), afin de tenter une datation relative des processus observés. De plus, il serait intéressant d'intégrer des séquences d'autres gènes HLA, tels que des gènes non-classiques (HLA-G, -E, ...), des pseudogènes (HLA-DRB9, -DQB3, -N) ou bien des gènes d'ancrage, (TRIM26, GNL1, ...) censés être plus conservés sur le plan évolutif [Amadou, 1999] afin d'étudier les relations entre les différentes régions géniques de gènes d'un même bloc évolutif (voir page 10).
Sur un plan technique, il serait aussi judicieux de réaliser des chaînes de Markov d'ordre supérieur (c'est-à-dire ne plus considérer les transitions d'un seul nucléotide à un autre, mais de deux ou trois nucléotides au nucléotide suivant), ce qui devrait permettre de mettre l'emphase sur les évènements de conversion et de transposition.

Chapitre 6

Discussion générale

1 Résumé du travail effectué

Ce travail de thèse est consacré à l'étude des facteurs évolutifs qui déterminent l'acquisition et le maintien de la diversité moléculaire HLA dans les populations humaines.

L'introduction générale (Chapitre 1) a présenté l'histoire évolutive du MHC, depuis son apparition dans un ancêtre commun aux Gnathostomes jusqu'à la structure actuelle du HLA. Nous avons alors vu que la région du MHC est caractérisée par un fort dynamisme, avec de nombreuses duplications et pertes de gènes (le modèle de « naissance et mort » des gènes) le long d'une charpente évolutivement stable (les gènes d'ancrage). Nous avons aussi présenté les différentes méthodes de génération du polymorphisme (la mutation ponctuelle et la recombinaison), ainsi que les processus qui vont influencer la diversité de ce polymorphisme dans les populations (sélection naturelle exercée par les pathogènes et effets démographiques).

Dans le Chapitre 2, nous avons présenté une analyse de typages à haute résolution de deux populations très différentes, permettant d'observer les effets, sur la diversité HLA, de deux de ces phénomènes : la sélection naturelle pour les Mandenkalu du Sénégal et une probable fusion de populations pour les Cham du Vietnam. Ce chapitre a aussi présenté une comparaison de trois méthodes de typage sur les mêmes individus Mandenka, illustrant l'apport de 25 ans de développement des méthodes de typage.

Le Chapitre 3 a présenté une méthode de traitement de lectures de séquençage, développée spécifiquement pour des séquençages de régions hautement polymorphiques, telles que le HLA et montrant d'excellents résultats avec des jeux de données d'espèces non modèles. Ce chapitre technique a aussi présenté les difficultés qui sont posées par les nouvelles méthodes de typage HLA et proposé des pistes pour résoudre ces problèmes.

La méthode développée au Chapitre 3 a servi pour traiter les résultats de séquençage de quatre exons 2 de classe II pour les mêmes individus (plus de 2'000) de 36 populations de la bande du Sahel (de l'Afrique de l'ouest à l'Afrique de l'est), d'Afrique du nord, mais aussi de Syrie (Asie de l'ouest) et de Slovaquie (Europe centrale) comme populations extra-africaines de référence (dans cette étude). Ces données, analysées dans le Chapitre 4, ont mis en évidence des structurations des populations, influencées par la géographie et le mode de vie mais aussi une importante pression de sélection engendrée par la malaria sur les populations exposées, bien que ces pressions de sélection s'exercent de manières différentes selon les populations.

Finalement, le Chapitre 5 présente deux analyses, non pas populationnelles, mais statistiques de la diversité HLA. Ces analyses ont permis de fournir une base théorique expliquant les différences observées entre les méthodes de typage dans le Chapitre 2, mais aussi

d'obtenir des indices supplémentaires sur les processus évolutifs à l'origine de la diversité du HLA.

Dans ce dernier chapitre nous allons discuter de manière plus générale les résultats de ce travail, selon trois axes de recherches. Premièrement, nous discuterons des mécanismes permettant la génération du polymorphisme HLA, qu'ils soient anciens (création des gènes HLA) ou plus récents (création des allèles HLA). La seconde partie concernera les processus qui ont déterminé la distribution de ce polymorphisme HLA au sein des populations. Finalement, la dernière partie discutera des mécanismes qui permettent à cette diversité de se maintenir, ou au contraire qui ont causé une perte de diversité dans certaines populations étudiées.

2 Axe 1 : Génération du polymorphisme

2.1 Génération ancienne

L'étude basée sur la décomposition en chaînes de Markov a mis en évidence que les différentes régions des gènes HLA de classe I se regroupaient par régions homologues et non par gènes (Figure 5.8). Ce résultat traduit une plus grande similarité entre les régions homologues des différents gènes qu'entre les régions d'un même gène. La présence des séquences de chimpanzé (Patr-A, -AL, -B et -C), dont la divergence avec les humains remonte de 6 à 12 millions d'années [Glazko, 2003, White et al., 2009, Wilkinson et al., 2011, Moorjani et al., 2016], et le fait que certains gènes soient encore plus anciens (l'apparition du gène Patr-AL est datée de 25 à 28 millions d'années [Piontkivska, 2003]) implique que cette similarité ne soit pas le résultat d'une homogénéisation par recombinaison récente (postérieure à la spéciation).

La datation des événements observés sur les Figures 5.8 à 5.10 est toutefois difficile. Une méthode serait d'utiliser des séquences d'espèces ayant divergé il y a plus longtemps, telles que des séquences de cercopithécidés ayant divergé il y a ∼25 millions d'années, ou de platyrrhiniens ayant divergé il y a ∼36 millions d'années [Glazko, 2003, Perelman et al., 2011, Doxiadis et al., 2012]. En l'absence d'une datation plus précise, nous ne pouvons qu'émettre des hypothèses sur l'ancienneté et les raisons de ces observations.
La première hypothèse est de considérer que ces événements sont relativement récents et que cette similarité est le fruit d'échanges de segments nucléotidiques fréquents entre ces régions homologues ayant homogénéisé les séquences chez un ancêtre commun aux humains et aux chimpanzés. Mais, à l'exception des exons 2 et 3 (pour lesquels les pressions de sélection semblent avoir causé une différenciation plus importante), nous n'observons pas de différences, au sein d'une même région, entre les séquences de chimpanzés et les séquences humaines. Or, si cette homogénéisation avait eu lieu chez un ancêtre commun, ces séquences auraient commencé à diverger après la spéciation, ce qui ne s'observe pas. L'utilisation de séquences de MHC d'espèces dont la divergence est antérieure à celle des humains et chimpanzés permettrait de tester cette hypothèse.
La deuxième hypothèse est de considérer que la décomposition en chaînes de Markov de premier ordre permet d'observer des processus beaucoup plus anciens. En effet, la décomposition de ces séquences en seulement 16 variables diminue le champ d'observation à des variations importantes dans les séquences, en réduisant les mutations ponctuelles à quelques pourcentages de différences entre deux variables. Considérer des chaînes de Markov d'ordres supérieurs fournirait plus de degrés de liberté pour décrire les différences entre les séquences (puisqu'elles ne seraient plus décrites par 16 variables mais 64 ou 256 pour des chaînes d'ordre 2 ou 3), permettant de mettre l'emphase sur des différences plus fines entre les séquences et donc des événements plus récents. Sous cette deuxième hypothèse et considérant la similarité entre les séquences des loci HLA-A, -B et -C, ce que nous observons alors est un phénomène bien plus ancien que la divergence entre les différents loci, c'est-à-dire l'apparition d'un gène ancestral à ces trois loci, pouvant remonter aussi loin que celle des différents blocs évolutifs (α et β) dans lesquels ils évoluent [Amadou, 1999].
Il faudrait alors continuer cette analyse en intégrant 1) des séquences d'espèces ayant divergé à une période bien plus ancienne que les humains et les chimpanzés , 2) des séquences d'autres gènes, soit des gènes non classiques (tels que les gènes d'ancrage ou des gènes MIC

présents depuis longtemps dans le MHC [Kulski et al., 1999, Kulski et al., 2002]), soit de gènes classiques dont l'origine est plus ancienne (tel que MHC-F [Piontkivska, 2003]), et 3) des chaînes de Markov d'ordres supérieurs afin de faire apparaître des événements plus récents.

Les résultats pour les gènes de classe II sont aussi intéressants. La Figure 5.9 montre que les exons 2 à 4 des gènes IIβ (HLA-DRB1, -DQB1 et -DPB1) sont groupés par régions homologues excluant les gènes IIα (HLA-DQA1 et -DPA1). Ce premier résultat est en accord avec l'étude faites par Hughes en 1990 qui a mis en évidence que pour la classe II, les gènes HLA-DRB1, -DQB1 et -DPB1 sont issus de duplications successives de gènes IIβ [Hughes and Nei, 1990], les phylogénies pour les gènes IIα et IIβ n'étant pas congruentes [Takahashi, 2000]. L'apparition des gènes HLA-DP ayant été datée de plus de 37 millions d'années [Klein et al., 1993a, Grahovac et al., 1993], ce regroupement des gènes IIβ suggère que les phénomènes observés dans cette étude (valable aussi pour les résultats observés aux gènes de classe I) soient plus anciens que cela. Une étude de Takahashi menée en 2000 et datant la divergence des différents groupes de gènes de classe II (DR, DQ, DP et DO) fait remonter la séparation des gènes DRB avec les gènes DQB/DPB à 200 millions d'années [Takahashi, 2000]. Les résultats de cette étude feraient alors remonter l'âge des processus observés pour les gènes de classe II à cette époque. Toutefois, les introns et certains exons (exons 1, 5 et 6) de ces mêmes gènes ne montrent pas ce regroupement en régions homologues. Cela concerne des régions où moins de séquences étaient disponibles, impliquant une possible limitation de la méthode en cas de données insuffisantes. Lorsque plus de données seront disponibles pour ces régions, il sera alors possible de réitérer cette analyse en incluant ces régions, afin de tester les hypothèses d'une similarité due à une origine commune ou due à une convergence évolutive.

Le second résultat intéressant concerne les gènes HLA-DRB. De manière similaire à ce qui a été observé pour les gènes de classe I, les différents gènes de la région DR se regroupent par régions homologues plutôt que par gènes (Figure 5.10). Nous savons que les gènes HLA-DRB descendent tous de plusieurs duplications à partir d'un gène MHC-DRB ancestral [Andersson et al., 1987], duplications datées de 36 à 58 millions d'années [Satta et al., 1996b]. Il semble qu'à l'instar des gènes de classe I, la décomposition en chaînes de Markov des séquences des gènes de classe II permet d'observer des phénomènes relativement anciens.

Toujours concernant les gènes HLA-DRB, l'encadré de la Figure 5.10 montre que les introns sont similaires eux, de même que les exons, tandis que les introns et les exons sont beaucoup moins similaires. Ceci indique que les introns et les exons ont des origines différentes, en accord avec l'étude réalisée par Bergström *et al.* en 1998 qui a montré que les introns et exons avaient des temps de divergence moyens très différents (250'000 ans pour les introns et 7.06 millions d'années pour les exons) [Bergström et al., 1998]. Si l'on considère les recherches menées par Hedrick en 1998, qui avançait que la recombinaison intra-génique permettait de générer rapidement de nouveaux variants ayant déjà passé le filtre de la sélection naturelle [Hedrick, 1998], nous pourrions dans notre cas considérer que les nouveaux allèles HLA-DRB aient été générés de la même façon, les exons anciens ayant passé le filtre de la sélection étant réutilisés dans un squelette d'introns plus récents : « *Old wine served in new skins* » [Doxiadis et al., 2008a].

2.2 Génération plus récente

La recombinaison intragénique permet aussi de générer du polymorphisme HLA sur une échelle de temps beaucoup plus récente que celle que nous venons de voir pour les gènes MHC-DRB, bien que l'importance de ce mécanisme par rapport à la mutation ponctuelle soit toujours discutée (voir Chapitre 1, page 23). Même si nous n'avons pas réalisé une étude approfondie sur l'ensemble des allèles HLA et leur origine possible, certains des résultats de ce travail méritent d'être discutés.

Tous les loci ne semblent pas concernés de la même façon par la recombinaison. Dans son travail de doctorat, Stéphane Buhler a mis en évidence que les loci HLA-B, HLA-DPB1 et HLA-DRB1 (dans cet ordre) montraient le plus d'événements de conversion génique [1] [Buhler, 2007]. Or, dans le Chapitre 2, les distributions de fréquences alléliques des Mandenkalu du Sénégal ont mis en évidence un allèle fréquent (le plus fréquent à ce locus) au locus HLA-B : HLA-B*35:01:01 (FA [2]=16%). En 1991, Allsopp a suggéré que cet allèle était le receveur dans une conversion allélique ayant créé HLA-B*53:01:01 (aussi présent chez les Mandenkalu, FA=6%) [Allsopp et al., 1991]. Aussi bien HLA-B*35:01 que B*53:01 ont été proposés comme des allèles apportant une résistance à la malaria, par des études *in-vitro* [Hill et al., 1992b], des études de cohortes de patients au Ghana [Yamazaki et al., 2011] et des études de prédiction de liaison de peptide et d'association entre la fréquence de l'allèle et la prévalence du parasite *P. falciparum* responsable de la malaria [Sanchez-Mazas et al., 2017]. Il est fort probable que HLA-B*35:01 (et B*53:01) assure ce rôle protecteur envers la malaria étant donné que les Mandenkalu sont dans une zone touchée par cette maladie (voir Figure 4.30) et que d'autres allèles HLA-B (HLA-B*15:03 et B*78:01, FA de 7 et 8%) de cette population ont aussi été rapportés comme protecteurs [Sanchez-Mazas et al., 2017].

Un autre allèle présent dans la population Mandenka, HLA-DRB1*13:04, proviendrait lui aussi d'une conversion allélique. Cet allèle, fréquent en Afrique de l'ouest [Tiercy et al., 1992, Hill et al., 1992a], a été proposé comme provenant d'une conversion allélique impliquant HLA-DRB1*11:02 aussi bien sur la base d'analyses sérologiques [Lee et al., 1990] que de RFLP [Hill et al., 1992a]. L'analyse que nous avons menée dans le Chapitre 2 a confirmé HLA-DRB1*11:02 comme receveur et a proposé une liste de 195 allèles comme donneurs potentiels. Au moment où cette analyse a été réalisée, la base de données IMGT/HLA recensait 1'913 allèles HLA-DRB1 [3], c'est-à-dire qu'un peu plus de 10% des allèles DRB1 sont des donneurs potentiels, simplement dans le cas de cette conversion allélique. Il est raisonnable de penser que ce fragment est régulièrement échangé entre les différents exons 2 de HLA-DRB1, expliquant son abondance et soulignant l'importance de la recombinaison intragénique pour la génération du polymorphisme au locus HLA-DRB1.
De plus, cet allèle n'est retrouvé qu'en Afrique de l'ouest. Il correspond à la séquence DRB1*3135 sur la Figure 4.8, qui est retrouvée uniquement dans les populations sénégalaises. L'étude menée par Hill *et al.* en 1992 a retrouvé cet allèle dans d'autres pays d'Afrique de l'ouest, tels que la Gambie et le Libéria, soulignant son absence dans les populations non-africaines [Hill et al., 1992a]. Une méta-étude à l'échelle mondiale menée par Solberg *et al.* en 2008 sur 497 populations n'a retrouvé cet allèle qu'en Afrique de

1. La conversion génique regroupant la conversion allélique, dans laquelle le fragment provient d'un autre allèle du même gène et la conversion ectopique, dans laquelle le fragment provient d'un gène paralogue.
2. Fréquence Allélique
3. IMGT/HLA v3.24

l'ouest [Solberg et al., 2008]. Ainsi son apparente absence dans d'autres populations que celles d'Afrique de l'ouest suggère une origine de cet allèle par une conversion allélique récente.

Comme nous venons de le voir, plusieurs exemples d'allèles issus de conversions alléliques sont retrouvés au sein d'une même population. Certains semblent même d'une origine récente, à l'instar des allèles de classe I identifiés dans les populations des Premières Nations comme provenant de conversions alléliques, et dont l'absence en dehors de ces populations, associée au peuplement récent des Amériques, indique une origine récente [Belich et al., 1992, Watkins et al., 1992, Parham et al., 1997].
En conclusion, nos résultats soulignent l'importance de la conversion allélique (et par extension de la recombinaison intragénique) comme facteur de génération du polymorphisme HLA, aussi bien pour les gènes de classe I que de classe II, ainsi que du caractère récent de certains de ces évènements de recombinaison.

Il nous faut alors discuter du deuxième mécanisme par lequel la diversité des gènes HLA est générée, à savoir la mutation ponctuelle. En 2003, Richman *et al.* pointe le fait que seule la mutation ponctuelle crée de nouveaux sites variables, la conversion allélique ne faisant que créer de nouvelles séquences à partir de ces sites [Richman et al., 2003a]. Ainsi, pour que la recombinaison soit efficace dans la génération de nouveaux variants, il faut au préalable que la mutation ponctuelle ait créé suffisamment de variabilité à l'intérieur des séquences.
Dans l'étude de Major *et al.* en 2013, qui réalise une comparaison entre des typages PCR-SSO et des typages basés sur des séquençages d'exomes par Illumina, les auteurs ont déterminé que les différences de typage les plus fréquentes étaient dues à des différences ponctuelles entre les allèles [Major et al., 2013]. Dans le Chapitre 2, les loci montrant le plus de différences de typages entre PCR-SSO et NGS-MiSeq (Table 2.6) sont HLA-DQB1 (18% de correspondances) et HLA-DPB1 (30% de correspondances). Pour le locus HLA-DQB1, la principale erreur est liée à l'allèle HLA-DQB1*03:19 (FA=44%), déterminé par PCR-SSO comme HLA-DQB1*03:01. La seule différence entre ces allèles est une différence dans l'exon 3 (codon 185, ACC/ATC), hors de portée des typages PCR-SSO à l'époque. De la même façon, pour HLA-DPB1, l'allèle identifié comme HLA-DPB1*17:01 est en réalité constitué de deux allèles, HLA-DPB1*17:01 et HLA-DPB1*131:01, dont les différences sont des mutations ponctuelles dans les exons 3 et 4.
L'analyse du profil de liaison de HLA-B*39:10, ne différant de B*39:01 que par un seul changement d'acide aminé dans le domaine protéique codé par l'exon 2, a mis en évidence que cette unique mutation changeait l'affinité de la poche B du site de liaison de l'antigène, passant d'une affinité pour les acides aminés basiques à une affinité pour les acides aminés non polaires, similaire à ce qui est observé pour HLA-B*53:01 [Yagüe et al., 1998]. Cette observation a permis de suggérer que cet allèle HLA-B*39:10 a acquis un rôle protecteur vis-à-vis de la malaria via cette mutation ponctuelle, illustrant alors l'importance de la mutation comme génération de nouvelle fonctionnalité.
L'étude de la répartition de l'information (Table 5.2 et Figure 5.6) dans les différentes régions géniques (Chapitre 5) a mis en évidence que les gènes de classe II, plus particulièrement HLA-DQA1, -DQB1 et -DPB1, possédaient une part non négligeable de leur information dans l'exon 3, cette information étant non redondante avec celle de l'exon 2 (Figure 5.7). Ce résultat permet d'expliquer l'importance des différences entre des méthodes de typage ciblant uniquement l'exon 2 (telle que la PCR-SSO pour les gènes

de classe II) et celles ciblant une partie bien plus grande du gène (telle que NGS-MiSeq, ciblant les gènes complets).
Il faut donc considérer que les études cherchant à comparer l'importance de la mutation ponctuelle par rapport à la recombinaison intragénique doivent tout d'abord s'assurer qu'elles ciblent une portion suffisamment importante des gènes étudiés, puisque de nombreuses mutations ponctuelles peuvent être localisées hors des exons 2 (et 3 pour les gènes de classe I) habituellement étudiés.

Répondre à la question de l'importance relative de la mutation ponctuelle ou de la recombinaison apparaît comme une tâche difficile. Les résultats de ces travaux ne permettent pas de déterminer laquelle est plus importante que l'autre, mais ils mettent en évidence que ces deux mécanismes sont importants. La recombinaison semble être un phénomène fréquent et rapide pour générer du polymorphisme, mais seule la mutation ponctuelle crée de nouveaux sites variables. Finalement, des études portant sur la comparaison entre la mutation ponctuelle et la conversion génique ont montré des différences entre les loci HLA. Les mutations ponctuelles semblent être la source principale de création de nouveaux variants pour HLA-A, -C, -DQA1 et -DQB1 tandis que la recombinaison serait prédominante pour HLA-B et -DRB1, HLA-DPB1, quant à lui, présentant un profil intermédiaire [Belich et al., 1992, Watkins et al., 1992, McAdam et al., 1994, Otting et al., 1998, de Groot et al., 2000, Buhler, 2007, Vangenot et al., 2020].

3 Axe 2 : Répartition du polymorphisme

Maintenant que nous avons vu comment était généré le polymorphisme HLA, la question suivante est de s'intéresser à la répartition de ce polymorphisme entre les populations. C'est-à-dire, de savoir quels sont les processus qui vont déterminer les jeux d'allèles qu'une population va posséder et comment ces derniers vont s'échanger d'une population à une autre.

Le premier exemple d'acquisition de polymorphisme présenté dans ce travail concerne la population Cham du Vietnam. L'une des hypothèses avancées pour expliquer la diversité observée (Figure 2.6), ou les motifs particuliers de déséquilibre de liaison haplotypique (Table 2.14 et Figure 2.14), est de considérer que les profils génétiques des Cham actuels sont le résultat d'un flux génique important, entre (au moins) une population austronésienne migrante et une population d'Asie de l'est continentale. Bien que l'origine exacte des deux populations soit encore à déterminer, ce premier exemple illustre comment une population peut acquérir un important polymorphisme (supérieur à celui d'une population africaine de grande taille, Figure 2.6) par un phénomène d'accumulation des diversités de plusieurs populations.

L'existence de flux géniques entre des populations sous-entend une possibilité de contact entre ces dernières, que ce soit par voie maritime pour les Cham ou par voie terrestre pour les populations africaines du Chapitre 4. Dans ce Chapitre 4, nous avons vu plusieurs exemples de flux géniques qui, selon leur magnitude, peuvent conduire d'un partage de quelques allèles à une homogénéisation des populations.

En Afrique de l'est, les Oromo et Amhara ne montrent aucune différence significative en termes de distance de Reynolds aux quatre loci étudiés (Figures 4.20 à 4.27), signifiant une similarité entre les profils génétiques des deux populations majoritaires d'Éthiopie (*c.f.* Figures 4.8 à 4.11). En 1998, une étude de Fort *et al.* menée sur ces deux populations typées pour trois loci de classe II (HLA-DRB1, -DQA1, -DQB1) avait déjà identifié une similarité entre ces populations, attribuée à un mélange [4] entre elles [Fort et al., 1998]. En 2001, deux études utilisant soit la distance de Reynolds au locus HLA-DRB1 entre 21 populations africaines [Renquin et al., 2001], soit la distance de Prevosti à ce même locus pour 22 populations de répartition mondiale (à l'exception des Amériques) [Sanchez-Mazas, 2001], avaient elles aussi mis en évidence une absence de différence significative entre ces deux populations d'Afrique de l'est. Il apparaît alors que la proximité géographique entre ces deux populations a permis un flux génique ayant homogénéisé leurs profils HLA aux loci de classe II, sans toutefois négliger la possibilité d'une convergence liée la sélection. En effet, ces deux populations vivant dans la même région géographique, il est possible de penser qu'elles sont exposées à une même pression en pathogènes qui aurait alors pu sélectionner des allèles similaires, mais 1) cette hypothèse n'exclut bien entendu pas celle d'une homogénéisation par flux géniques et 2) cette similarité s'observe sur l'ensemble des quatre loci, alors que HLA-DPB1, notamment, est censé être moins soumis à la sélection naturelle que les trois autres [Sanchez-Mazas, 2001] (voir aussi page 297 pour la discussion de la neutralité de HLA-DPB1).

Finalement, un flux génique est aussi probable, d'après nos résultats, entre les populations (semi-)nomades d'Afrique du nord (Tamasheq et Imazighen) et les Peuls

4. *admixture*

d'Afrique de l'ouest. Ce flux serait plus faible que pour les populations d'Éthiopie, puisqu'il n'a pas conduit à une homogénéisation des profils génétiques de ces populations, visible sur la Figure 4.26 par les distances de Reynolds significativement supérieures à zéro au locus DPB1-Exon2. Ceci indique une différence significative dans les distributions de fréquences alléliques de ces populations pour DPB1-Exon2, alors même que ce flux génique est principalement visible à ce locus. En effet, les résultats montrent que la séquence DPB1*76 est retrouvée (Figure 4.11) dans les populations Peuls du Sénégal (FA=13%) et du Mali (FA=23%), ainsi que dans les populations Imazighen d'Asni (FA=8%) et de Figuig (FA=7%) et, dans une moindre mesure, chez les Tamasheq de Tamanrasset (FA=3%). Une récente étude, menée par Sanchez-Mazas *et al.* en 2017, a mis en évidence des similarités entre les profils génétiques des Peuls et des populations d'Afrique du nord au locus HLA-A (fréquence élevée du lignage HLA-A*01), suggérant alors un échange génétique au travers du Sahara [Sanchez-Mazas et al., 2017]. La ville de Tamanrasset en Algérie, localité d'échantillonnage de la population Tamasheq, est située sur un des itinéraires des caravanes trans-sahariennes [Murdock, 1959], qui a alors pu être l'un des axes le long desquels s'est réalisé ce flux génique. Cette hypothèse est corroborée par une étude réalisée en 1978 sur les Tamasheq du Niger (où passe cet itinéraire trans-saharien) et ayant mis en évidence une cohabitation entre les Tamasheq et les Peuls de cette région [Bernus, 1993], cohabitation aussi signalée au Burkina Faso [Kuba et al., 2003].
Une étude de 2010, basée sur le séquençage d'ADN mitochondrial et de SNP sur le chromosome Y de 90 Tamasheq (répartis en trois échantillons de populations), a déterminé que 48% des haplotypes mitochondriaux étaient rattachés à des haplogroupes fréquents en Afrique sub-saharienne, concluant que les lignages d'Afrique de l'ouest et du centre étaient les plus fréquemment retrouvés chez ces individus [Pereira et al., 2010]. Les F_{ST} estimés à partir des SNP du chromosome Y placent les deux populations Tamasheq proches des populations d'Afrique du nord et la troisième proche des populations d'Afrique sub-saharienne. Ces résultats sont similaires à ceux de notre étude, soutenant l'hypothèse de flux géniques au travers du Sahara, entre les populations Tamasheq et les populations d'Afrique sub-saharienne.

Ces trois populations étant nomades (ou semi-nomades pour les Imazighen), il serait raisonnable de penser que leur mode de vie a pu influencer la répartition de ce polymorphisme, ce qui expliquerait la structure génétique observée. D'autres systèmes génétiques, tels que le gène NAT2 [5], ont déjà révélé un polymorphisme associé au mode de subsistance des populations. Une étude menée en 2011 par Sabbagh *et al.* sur 128 populations des cinq continents a mis en évidence une fréquence plus importante du phénotype d'acétylation rapide chez les populations de chasseurs-cueilleurs, tandis que le phénotype d'acétylation lente est retrouvé plus fréquemment dans les populations d'éleveurs et agriculteurs [Sabbagh et al., 2011]. Une autre étude réalisée par Patillon *et al.* en 2014 a identifié l'allèle `rs1799930-A` du gène NAT2 comme la cible d'une sélection positive lors de la transition du Néolithique [B. Patillon et al., 2014]. Toutefois, une étude de 2015 a identifié une plus grande proportion d'acétylateurs rapides chez les populations pastoralistes que chez les populations d'agriculteurs, mais ces effets se confondent aussi avec les climats dans lesquels évoluent ces populations, suggérant une possible double influence indépendante de ces deux facteurs [Podgorná et al., 2015].
L'étude menée au Chapitre 4 n'a pas identifié un rôle du mode de vie dans la distribution des fréquences alléliques HLA, à l'exception de DQB1-Exon2 qui montre un (faible) effet

5. Arylamine N-acetyltransferase, impliqué dans la détoxication de l'organisme.

du mode de vie dans l'AMOVA ($\Phi_{CT} = 0.018$, Figure 4.29), bien inférieur à la variance existant à l'intérieur des groupes ($\Phi_{SC} = 0.031$) et entre les populations ($\Phi_{ST} = 0.048$). Il ne semble donc pas que le mode de vie ait une influence détectable sur la distribution des polymorphismes des gènes HLA de classe II, du moins pour les populations africaines étudiées.

L'existence de ces flux géniques, ou plutôt leur absence, va aussi déterminer des différences entre les populations. La MDS basée sur des distances de Reynolds calculées pour le locus DPB1-Exon2 (Figure 4.26), ainsi que la proportion de paires de populations présentant des distances de Reynolds significatives pour DRB1-Exon2 (Figure 4.21), DPB1-Exon2 (Figure 4.27) et, dans une moindre mesure, DQA1-Exon2, Figure 4.23 montrent, pour l'Afrique, un modèle compatible avec l'isolement par la distance, les populations de régions proches possédant plus souvent des distances de Reynolds non significatives que les populations de régions éloignées. Cette structuration géographique est aussi visible sur l'AFC basée sur l'ensemble des loci (Figures 4.16 à 4.19).
Il apparaît alors que les populations d'Afrique centrale, de l'est et du nord forment un ensemble de populations présentant beaucoup de liens (peu de Θ_w significativement différents de zéro) tandis que les populations d'Afrique de l'ouest apparaissent comme un à deux groupes isolés (visibles sur la Figure 4.26) puisque seuls les Dangaléat du Tchad présentent un $\Theta_w \sim 0$ avec une population d'Afrique de l'ouest (les Gourmantché du Burkina Faso en l'occurrence).
L'analyse de la répartition des variants liés à la persistance de la lactase à l'âge adulte, un trait qui a été proposé comme lié au pastoralisme [Gerbault et al., 2009], montre lui aussi une différence de répartition entre l'est et l'ouest du Sahel. Les pastoralistes arabes de l'est du Sahel (représentés dans ce travail de doctorat par les Beja Hadendoa et les Arabes Rashaida) sont caractérisés par une fréquence plus importante du variant `G-13915` du gène *LCT*, identique à celui retrouvé dans les populations de la péninsule arabique et interprété comme une origine commune à ces populations [Enattah et al., 2008, Priehodová et al., 2014]. Au contraire, à l'ouest du Sahel, ce sont les variants `T-13910` et `A-22018`[6] qui ont été retrouvés à des fréquences plus élevées dans des populations Peuls [Coelho et al., 2005, Lokki et al., 2011], mais aussi chez les populations d'Arabes Baggara et de Mozabites d'Algérie [Ranciaro et al., 2014], suggérant des flux géniques entre ces populations.
Ces résultats sont similaires aux observations faites dans notre étude de différences significatives entre les populations de l'ouest et de l'est du Sahel.
Nos résultats pour les populations de l'est du Sahel sont similaires à ceux obtenus par Sanchez-Mazas *et al.* en 2017, qui a comparé les fréquences des allèles HLA-A et -B pour 40 populations d'Afrique et identifié un apparentement génétique important des populations d'Afrique centrale, de l'est et du nord [Sanchez-Mazas et al., 2017]. Cet apparentement a été interprété comme le résultat de migrations et d'échanges génétiques entre les populations sédentaires et nomades de ces régions. Plusieurs des populations de cette étude étant elles aussi des populations étudiées dans ce travail de doctorat, nous pouvons dire que les résultats pour les gènes de classe II soutiennent ceux basés sur les gènes de classe I.

Bien que la séparation entre les populations d'Afrique de l'ouest et les autres populations de la bande du Sahel, observée aux loci HLA de classe II, soit principalement due, selon nous, à l'effet de la sélection naturelle (discuté dans la section 4.1), un signal

6. Ces deux variants étant aussi retrouvés en Europe [Ranciaro et al., 2014], soutenant l'hypothèse d'une origine commune [Coelho et al., 2005].

lié à l'histoire de ces populations (et lié à la linguistique) est aussi envisageable.
L'origine de la plupart des allèles HLA de classe I retrouvés en Afrique est antérieure à la séparation des différents groupes linguistiques, comme l'a indiqué l'étude de Cao *et al.* en 2004, qui a souligné que l'ensemble des lignages alléliques de classe I étaient retrouvé dans toutes les populations sub-sahariennes étudiées [7], les mêmes allèles étant certaines fois retrouvés dans les mêmes haplotypes chez plusieurs populations d'origines différentes [Cao et al., 2004].
Linguistiquement, le partage de langues apparentées peut représenter une origine commune, comme c'est le cas avec les langues bantu en Afrique méridionale, dont la répartition suit le trajet de la migration bantu du 5ème millénaire avant notre ère [Li et al., 2014].
Dans ce travail, les populations d'Afrique de l'ouest sont toutes locutrices de langues de la famille niger-congo, seules représentantes (dans notre étude) de ce phylum linguistique. Nos résultats suggèrent que l'existence d'un groupe génétique indépendant (visible sur la Figure 4.26), regroupant l'ensemble des populations locutrices de langues niger-congo, représente une origine commune à ces populations [Černý et al., 2018]. La divergence entre les populations locutrices de langues niger-congo et nilo-sahariennes étant datée de 28'000 ans, selon certaines sources, on peut alors supposer que cette structure n'est pas plus ancienne [Shriner et al., 2015].

Ainsi, même si la distribution de la plupart des allèles ne semble pas associé à des différenciations linguistiques, il demeure que certains allèles spécifiques montrent des fréquences plus importantes dans certains groupes linguistiques que dans d'autres.
Par exemple les populations du Sénégal (à l'exception des Peuls) sont les seules populations à posséder la séquence DRB1*3135 à une fréquence élevée (minimum 25%, Figure 4.8), cette séquence étant absente des autres populations, à l'exception des Arabes Rashaida (FA<2%) et des Imazighen de Figuig (FA<1%). La Figure 4.16 montre que cette séquence est associée aux populations Sérère et Mandenka. Cette séquence correspond à l'unique allèle HLA-DRB1*13:04, déjà identifié chez les Mandenkalu au Chapitre 2 et retrouvé comme allèle HLA-DRB1 majoritaire en Gambie [Hill et al., 1992a] et plus généralement sur la côte de l'Afrique de l'ouest, du Sénégal au Sierra Leone [Solberg et al., 2008], c'est-à-dire la répartition de la branche atlantique (en plus de la sous-famille mandé) des langues niger-congo.
Un autre exemple concerne les séquences DPB1*64 et DPB1*66 [8]. La Figure 4.19 montre que ces deux séquences sont principalement associées aux populations d'Afrique de l'ouest et la distribution des fréquences alléliques en Figure 4.11 montre une différence entre les populations du Sénégal (locutrices de langues de la sous-famille mandé et de la branche atlantique des langues niger-congo), dans lesquelles DPB1*64 est prédominant, et celles du Burkina Faso (locutrices de langues de la branche volta-congo des langues niger-congo), dans lesquelles DPB1*66 est prédominant. Deux hypothèses ont été émises pour expliquer cette différence la première, (discutée plus en détail dans la section 4.1) étant celle d'une différence de pression de sélection due à la malaria (la fréquence de DPB1*66 étant fortement corrélée avec la prévalence de cette maladie) et la deuxième faisant appel à des différenciations linguistiques.
L'AMOVA (Figure 4.29), présentée au Chapitre 4, en se basant sur les différentes familles linguistiques, montre que c'est pour DPB1-Exon2 que la valeur de Φ_{CT} est maximale, indiquant que c'est ce locus qui montre le plus de structure liée à la famille linguistique.

7. Du Kénya (Nandi et Luo), du Mali (Dogon), d'Ouganda et de Zambie
8. Pouvant respectivement correspondre à HLA-DPB1*17:01/131:01 et HLA-DPB1*01:01.

Il serait alors intéressant de savoir si cette structure pour DPB1-Exon2 est toujours significative si l'on utilise les branches linguistiques (permettant dans ce cas de séparer les populations sénégalaises et burkinabé). Malheureusement, il n'est pas possible de réaliser cette analyse par manque de données suffisament représentatives (les 20 populations utilisées dans l'AMOVA sont représentées par 11 branches linguistiques différentes).
Cette séparation entre les populations sénégalaises et burkinabé est toutefois retrouvée dans l'étude de Černý *et al.* de 2018, où un arbre phylogénétique (reproduit dans la Figure 6.1), basé sur les fréquences des allèles HLA-A et -B (définis au premier champ), montre que les populations d'Afrique de l'ouest se séparent en deux groupes, l'un regroupant les Peuls et Mandenka et l'autre les Mossi, Gourounsi et Gourmantché [Černý et al., 2018].

De plus, l'étude de Triska *et al.* de 2015 conduite sur 2.2 millions de SNP pour 161 individus provenant de 13 populations sahéliennes a utilisé une analyse d'`admixture` [Alexander et al., 2009] pour évaluer les différentes composantes ancestrales de ces génomes [Triska et al., 2015]. Le meilleur résultat a été obtenu pour sept populations ancestrales (K=7) et montre deux composantes dans les génomes d'individus ouest-africains (à l'exception des Peuls), une première, fréquente dans les populations ouest-atlantiques (Mandenka, Gambiens et Mende) et une seconde, fréquente dans les populations du centre-ouest de l'Afrique (notamment Gourounsi, Gourmantché et Mossi). L'observation de ces deux composantes, peu retrouvées dans les génomes des autres populations (uniquement retrouvées de manière minoritaire chez les populations Daza et Kanembu du Tchad), va dans le sens d'une origine commune aux populations d'Afrique de l'ouest. Ces résultats sont similaires à ceux obtenus dans notre étude au locus DPB1-Exon2 avec la prédominance de la séquence DPB1*66 dans les populations burkinabé, alors que DPB1*64 est prédominante dans les populations sénégalaises. De même, la présence de HLA-DRB*13:04 (séquence DRB1*3135) dans les populations sénégalaises et son absence dans les populations burkinabé pourrait être expliquée par cette différence de composantes.
Il apparaît alors que les populations d'Afrique de l'ouest, déjà clairement séparées des autres populations d'Afrique sub-saharienne, sont composées de deux groupes ayant une histoire évolutive distincte.

En conclusion, nous venons de voir que les flux géniques entre les populations constituaient une part notable de la diversité des allèles HLA retrouvés dans les populations. Cet effet s'illustre de plusieurs manières et de plusieurs amplitudes, puisqu'elle peut concerner des flux géniques modestes, comme ceux observés pour les populations Imazighen, Tamasheq et Peuls, ou importants (par exemple pour les Cham) allant même jusqu'à homogénéiser des populations (les Oromo et Amhara d'Afrique de l'est). La distribution de ces flux géniques détermine aussi des régions qui vont être séparées et donc génétiquement plus distantes, telle l'Afrique de l'ouest, ou bien au contraire illustrer des mouvements complexes de populations pour les régions d'Afrique centrale, de l'est et du nord.

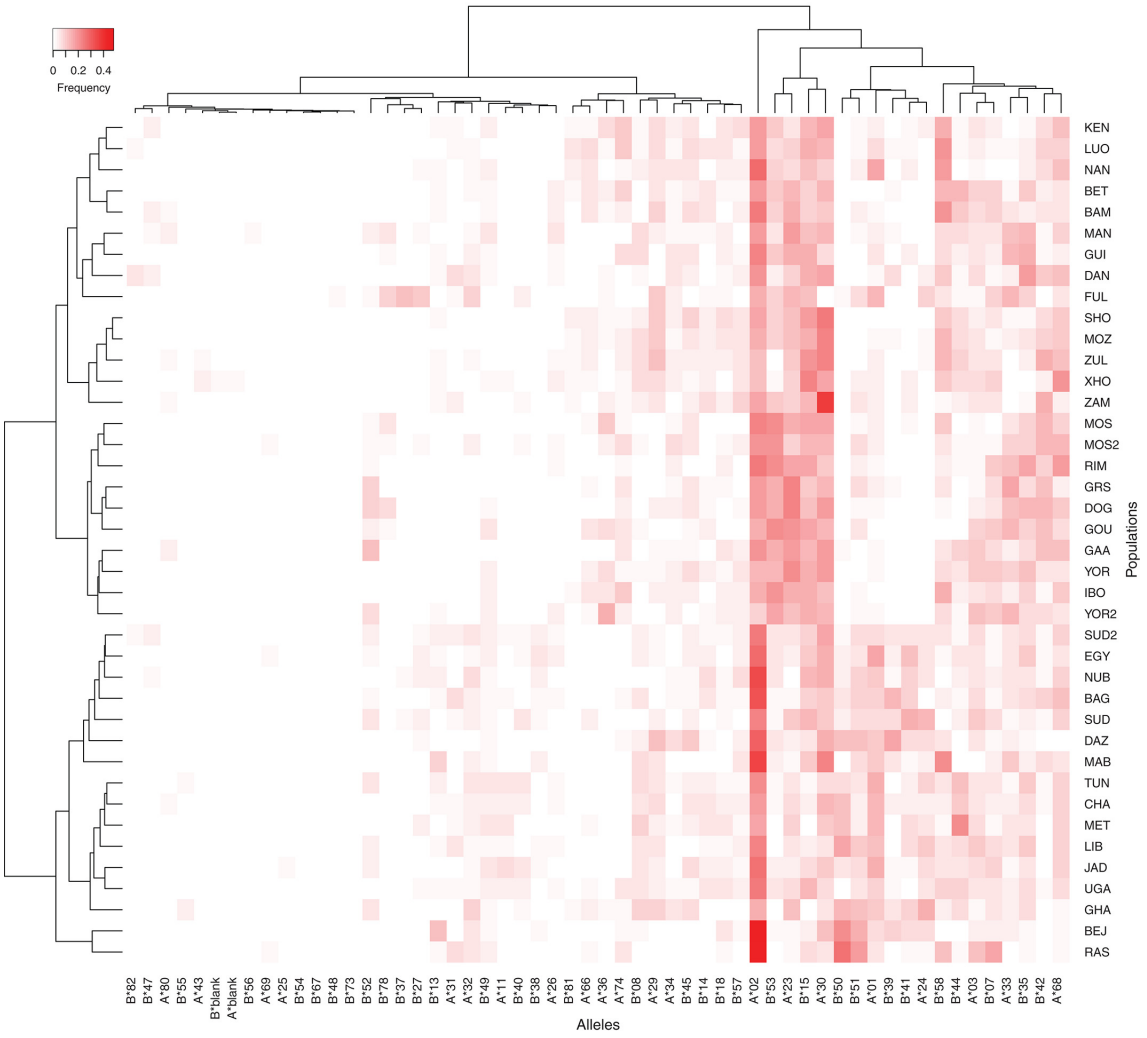


FIGURE 6.1 – Carte de chaleur des fréquences (résolution au premier champ) des allèles HLA-A et -B (axe des abscisse) pour quarante populations africaines. L'arbre du haut représente les relations entre ces allèles et l'arbre de gauche représente les relations entre les populations sur la base de leurs fréquences alléliques. Les deux arbres ont été construits à l'aide de la méthode de Ward [R Core Team, 2020]. Les noms courts des populations correspondent à : BAG, Arabe Baggara (Tchad) ; BAM, Bamileke (Cameroun) ; BEJ, Beja Hadendoa (Soudan) ; BET, Beti (Cameroun) ; CHA, Chaouya (Maroc) ; DAN, Dangaléat (Tchad) ; DAZ, Daza (Tchad) ; DOG, Dogons (Mali) ; EGY, Égyptiens (Égypte) ; FUL, Peuls (Burkina Faso) ; GAA, Ga Adangbe (Ghana) ; GHA, Ghannouch (Tunisie) ; GOU, Gourmantché (Burkina Faso) ; GRS, Gourounsi (Burkina Faso) ; GUI, Guiné Bissa (Guiné Bissau) ; JAD, El Jadida (Maroc) ; IBO, Ibo (Nigeria) ; KEN, Kényans (Kenya) ; LIB, Benghazi (Libye) ; LUO, Luo (Kenya) ; MAB, Maba (Tchad) ; MAN, Mandenka (Sénégal) ; MOS, Mossi (Burkina Faso) ; MOS2, Mossi (Burkina Faso) ; MOZ, Mozambique ; NAN, Nandi (Kenya) ; MET, Metalsa (Maroc) ; NUB, Nubiens (Soudan) ; RAS, Arabes Rashaida (Soudan) ; RIM, Rimaibé (Burkina Faso) ; SHO, Shona (Zimbabwe) ; SUD2, Soudanais (Soudan) ; SUD, Arabe soudanais (Soudan) ; TUN, Tunisiens du sud (Tunisie) ; UGA, Ugandiens (Uganda) ; XHO, Xhosa (Afrique du sud) ; YOR, Yoruba (Nigeria) ; YOR2, Yoruba (Nigeria) ; ZAM, Zambiens (Zambie) ; ZUL, Zulu (Afrique du sud). Figure provenant de [Černý et al., 2018], avec l'autorisation de l'auteure.

4 Axe 3 : Variations du polymorphisme

Nous avons présenté, dans la première partie de cette discussion, un aperçu des mécanismes qui permettent la génération *de novo* du polymorphisme HLA. Mais, que ce soit par mutation ponctuelle ou recombinaison intragénique, pour qu'un variant soit observable aujourd'hui, il a dû survivre aux deux forces évolutives majeures : la sélection naturelle et la dérive génétique. Dans cette dernière partie, nous allons voir ce que les résultats de ce travail de doctorat apportent de nouveau concernant ces deux forces.

4.1 La sélection naturelle

Le Chapitre 1 (Introduction générale) a présenté les trois types de sélection balancée décrites jusque là (Section 4.2). Il s'agit de l'avantage de l'hétérozygote, initialement proposé par Doherty et Zinkernagel en 1975 [Doherty and Zinkernagel, 1975], de la sélection fréquence-dépendante négative (ou avantage de l'allèle rare) [Slade and McCallum, 1992] et de la sélection fluctuant dans le temps et l'espace, conceptualisée par Hedrick en 2002 [Hedrick, 2002]. Néanmoins, la sélection balancée n'est pas le seul type de sélection observable sur les gènes HLA puisque plusieurs études ont déjà mis en évidence des signaux de sélection directionnelle, notamment pour HLA-B*53:01 en lien avec une protection vis-à-vis de la malaria en Afrique [Hill et al., 1991, Hill et al., 1992b, Sanchez-Mazas et al., 2017], des signaux de sélection directionnelle négative récente au locus HLA-DQA1 chez des Premières Nations de l'actuelle Colombie-Britannique [Lindo et al., 2016], ou bien au locus HLA-DRB1 d'une population Mongol pour laquelle l'allèle HLA-DRB1*12:02:01 semble avoir été sous sélection positive suite à la migration de cette population dans le sud de la Chine [Sun et al., 2015].

Nous allons à présent discuter des principaux résultats de ce travail concernant les différents signaux de sélection observés sur les loci HLA.

Cible de la sélection

Dans le Chapitre 5, la Table 5.3 montre que la majorité de l'entropie, c'est-à-dire de la variabilité, est localisée dans les exons 2 et 3. Ce résultat était attendu, sachant que ce sont les exons 2 (et exons 3 pour les gènes de classe I) qui codent pour les régions de liaison au peptide (voir Section 3.2), une plus haute diversité étant alors cohérente avec le mécanisme de sélection balancée généralement observé à ces loci [Bernatchez and Landry, 2003, Buhler and Sanchez-Mazas, 2011].
Au Chapitre 2, la Figure 2.10 détaillant les profils de diversité observés sur les différentes régions géniques de huit gènes HLA chez deux populations très différentes, montre que ce sont spécifiquement les codons ARS [9] qui sont caractérisés par une importante diversité. Ces résultats sont aussi visibles sur les Figures 2.4 et 2.5, représentant la diversité des nucléotides et acides aminés de ces mêmes gènes, où les codons ARS sont nettement plus diversifiés que les autres codons des exons 2 et 3. Les analyses du Chapitre 4, concernant les exons 2 de quatre loci de classe II, montrent des résultats similaires, notamment la Table 4.12 donnant pour chacun des quatre loci les valeurs de diversité nucléotidique, de fréquence des sites polymorphiques et des valeurs du D de Tajima selon si les codons codent ou non pour le site de reconnaissance de l'antigène. Les codons

9. Codant pour les acides aminés qui vont former la poche de liaison peptidique des molécules HLA.

ARS montrent toujours une diversité plus élevée (diversité nucléotidique et fréquence des sites polymorphiques) ainsi que des valeurs du D de Tajima plus élevés (à l'exception de DQA1-Exon2, discuté page 292).
Ces résultats sont en accord avec ceux obtenus jusqu'à présent, concernant la cible de la sélection sur HLA, notamment d'après Bitarello *et al.* en 2016 qui ont étudié le taux de substitutions synonymes et non-synonymes pour 3'821 allèles provenant des trois loci de classe I et qui ont mis en évidence que les codons ARS présentaient toujours un taux de substitutions non-synonymes plus élevé que les substitutions synonymes (alors que ce résultat n'est observé que pour 7% des codons non-ARS sur les mêmes exons) [Bitarello et al., 2016]. D'autres études, telles que celles menées sur des populations de chamois des Alpes ([Schaschl et al., 2004, Schaschl et al., 2005, Alvarez-Busto et al., 2007, Mona et al., 2008]), ont aussi mis en évidence une plus importante fréquence de sites polymorphiques ainsi que de mutations non-synonymes sur les codons ARS de rupi-DRB.

Il est donc apparent que les codons ARS sont la cible de la sélection balancée qui agit sur les loci HLA. Il reste maintenant à définir les modalités et preuves de cette pression de sélection, c'est-à-dire à identifier les similarités et différences entre loci et à déterminer l'origine de cette pression de sélection.

Evolution des loci de classe I

Plusieurs différences de pression de sélection s'observent sur les loci de classe I. La première est liée à la différence de pression de sélection sur les exons 2 et 3 des ces gènes. Les Figures 2.4 et 2.5 montrent une différence entre les exons 2 et 3 de HLA-A et -B, tout particulièrement visible sur la diversité des acides aminés du site de reconnaissance de l'antigène dans la population Cham (Figure 2.4, en bas à gauche). Il apparaît que pour HLA-A, les acides aminés de l'ARS sont nettement moins diversifiés sur la chaîne $\alpha 1$ par rapport à la chaîne $\alpha 2$, alors que le contraire est observé pour HLA-B. Une hypothèse pour expliquer cette différence entre HLA-A et -B, en termes de sélection, réside dans la poche qui va être décisive pour la liaison au peptide. En effet, le site de reconnaissance du peptide est composé de six poches (nommées de A à F), qui vont chacune accueillir un des neuf acides aminés du peptide nonamérique lié. Les poches B et F étant les plus décisives dans la liaison au peptide [Saper et al., 1991], cela a conduit à classer les allèles de classe I en *super-types* selon les propriétés physico-chimiques de ces deux poches [Sidney et al., 1996, Sidney et al., 2008]. Or, pour HLA-A, c'est la poche F (codée par l'exon 3) qui va être la plus décisive dans la liaison au peptide, tandis que pour HLA-B c'est la poche B (codée par l'exon 2) qui va être décisive [dos Santos Francisco et al., 2015]. Si l'on considère que c'est surtout l'exon 3 de HLA-A qui est sous sélection, alors l'exon 2 peut être plus enclin à afficher d'autres signaux, notamment démographiques. Ce résultat a été observé dans une étude sur des populations gitanes d'Europe, où HLA-A montrait, sur une MDS, les effets de la dérive génétique de ces populations (due à un ou plusieurs effets fondateurs) tandis que HLA-B et -DRB1 montraient l'origine commune de ces populations [Inotai et al., 2015]. Des résultats similaires ont été observés par Di *et al.* lors d'une étude du peuplement humain en Asie de l'est, où HLA-A montrait un fort effet de la démographie et peu d'effet de la sélection [Di et al., 2015].

L'analyse de la diversité en acides aminés des Mandenkalu (Figure 2.4) et des Cham (Figure 2.5) du Chapitre 2 montre aussi une plus faible diversité de la chaîne $\alpha 1$ pour

HLA-C comparée aux autres loci de classe I. Cette plus faible diversité pourrait s'expliquer par le rôle de ligand des KIR [10] plus important pour HLA-C comparé aux deux autres loci de classe I. En effet, si les KIR de lignage II ne se lient qu'aux molécules HLA-A et -B portant respectivement les épitopes A3/A11 et Bw4, l'ensemble des molécules HLA-C sont des ligands des KIR, selon deux épitopes mutuellement exclusif définis à la position 80 de la chaîne $\alpha 1$ [Winter and Long, 1997, Hilton et al., 2015]. L'étude menée par Bitarello *et al.* en 2016 sur les ratios dN/dS des gènes de classe I avait mis en évidence que les valeurs de dN et dS sont en moyenne deux fois plus petites pour HLA-C que pour HLA-A ou -B [Bitarello et al., 2016]. Les auteurs ont émis l'hypothèse que cette plus petite diversité pouvait être liée à la fonction de ligand KIR des molécules HLA-C et que l'évolution rapide des molécules KIR (observée en comparant les KIR des humains et chimpanzés [Khakoo et al., 2000, Vilches and Parham, 2002, Single et al., 2007]) exercerait une forte pression de sélection sur HLA-C, responsable des plus petits taux de substitution observés à ce locus.

Évolution des loci de classe II

De manière générale, à l'exception de DPB1-Exon2, les loci de classe II étudiés dans le Chapitre 4 montrent un indice de fixation Φ_{ST} peu important, allant de 0.036 pour DQA1-Exon2 à 0.052 pour DRB1-Exon2 (Section 3.9). Ces faibles valeurs de Φ_{ST} sont cohérentes avec le peu de structuration visible sur les MDS (Figures 4.20 4.22 et 4.24), notamment le nombre important de paires de populations pour lesquelles le Θ_w n'est pas significativement différent de zéro (visibles par les lignes en pointillés sur les Figures). L'un des signaux de la sélection balancée, observable lorsque l'on compare des populations, est la faible différenciation entre ces populations (selon une étude théorique [Muirhead, 2001], mais aussi sur des espèces non modèles telles que le rongeur *Spermophilus suslicus* [Biedrzycka and Radwan, 2008] ou la truite *Oncorhynchus mykiss* [Aguilar and Garza, 2006] mais aussi chez l'humain [Brandt et al., 2018]). Si l'on considère les valeurs du D de Tajima significativement supérieures à zéro aux codons ARS de ces loci pour 68 à 94% de ces populations (Table 4.14), alors la conclusion la plus probable est celle d'une forte sélection balancée agissant sur ces trois loci.

Toutefois, tous les loci ne semblent pas évoluer sous un unique régime de sélection. Le locus DRB1-Exon2 est caractérisé par une richesse allélique importante (Figure 4.7), le plus important nombre de rejets de la neutralité en faveur d'un excès d'hétérozygotes (test d'Ewen-Watterson-Slatkin, Table 4.13), en lien avec l'hétérozygotie la plus importante parmi celles des quatre loci ($H = 0.864 \pm 0.069$, Table 4.8). Ces résultats soutiennent un modèle de sélection balancée par avantage de l'hétérozygote, qui semble principalement agir sur les codons ARS (puisque 268% des populations ont des valeurs du D significativement supérieures à 0 pour les codons ARS de DRB1-Exon2 contre 10% pour les codons non-ARS, Table 4.14 et voir aussi la Figure 4.15).

Un autre résultat particulier est observé au locus DQA1-Exon2. Il s'agit du locus qui montre le plus de populations avec des valeurs de D de Tajima significativement supérieures à zéro (Figure 4.14), que ce soit pour l'exon 2 au complet (97% des populations), les codons ARS ou les codons non-ARS (94% des populations). Il s'agit aussi du locus présentant les valeurs de D de Tajima les plus élevées des quatre loci (pour

10. *Killer Immunoglobulin-like Receptor.*

l'exon 2 au complet, $\overline{D} \pm \sigma = 3.256 \pm 0.556$, Table 4.12). Ces observations suggèrent que l'effet de la sélection balancée s'exerce, soit de façon plus importante sur ce locus, soit depuis plus longtemps. La méta-analyse réalisée par Solberg *et al.* en 2008, sur 497 populations mondiales, a identifié le plus fort signal de sélection balancée pour HLA-DQA1 [Solberg et al., 2008] (selon l'écart normalisé d'homozygotie F_{nd} [Salamon et al., 1999]), cohérent avec les résultats de notre étude. DQB1-Exon2 montre un résultat similaire, avec les deuxièmes valeurs de D les plus élevées pour l'exon 2 au complet et les codons ARS (après DQA1-Exon2, Figure 4.15). Ce locus étant en fort déséquilibre de liaison global avec DQA1-Exon2 (dans 93% des populations, Figure 4.13 et Table 4.10), ces résultats pour HLA-DQB1 peuvent s'expliquer, soit par un phénomène de sélection balancée associative où la diversité génétique s'accumule par déséquilibre de liaison dans le voisinage d'un locus sous sélection balancée [Ohta and Kimura, 1970, Slatkin, 1995, Sanchez-Mazas, 2007], soit par une sélection balancée agissant sur les deux loci simultanément.
Néanmoins, les codons ARS de DQA1-Exon2 montrent un D de Tajima (Table 4.12 et Figure 4.15) légèrement inférieur (2.933 ± 0.482) à celui des codons non-ARS (3.072 ± 0.548) ou de l'exon 2 complet (3.256 ± 0.556), résultat qui n'est retrouvé pour aucun des trois autres loci. Ce résultat pourrait s'expliquer en considérant que HLA-DQA1 ait évolué sous une ancienne sélection balancée (expliquant, par sélection balancée associative, les valeurs du D de Tajima observées aux codons non-ARS) suivie d'un balayage sélectif récent dans ces populations et agissant sur les codons ARS (expliquant leur plus faible valeurs du D de Tajima). L'étude de Lindo *et al.*, menée en 2016 et comparant les génomes de 25 Premières Nations modernes et 25 Premières Nations datant d'avant la colonisation européenne, a déjà mis en évidence un changement de régime de sélection pour HLA-DQA1, alors passé d'une ancienne sélection directionnelle positive à une sélection négative à la suite de la colonisation européenne [Lindo et al., 2016]. Cette étude montre que la sélection qui s'applique sur les loci HLA n'est pas constante, soutenant l'hypothèse du changement de régime de sélection observé pour DQA1-Exon2 dans les populations étudiées au Chapitre 4.
Une autre hypothèse pour expliquer cette différence de pression de sélection entre les codons ARS et non-ARS serait de considérer que la sélection balancée n'agisse pas uniquement sur les codons ARS, mais aussi sur des codons voisins. Cette hypothèse rejoint celle émise par Valdes *et al.* en 1999, qui avait observé des variations importantes du polymorphisme en dehors des codons ARS, ou du moins, des positions des codons ARS inférées à partir de HLA-DR, car la structure cristallographique de HLA-DQ était inconnue à l'époque de la publication [Valdes et al., 1999]. Cette hypothèse semble toutefois peu probable, puisque les positions des codons ARS utilisées dans ce travail de doctorat proviennent de l'étude de Reche et Reinherz de 2003, où les codons ARS ont été identifiés par cristallographie et confirmés par analyse bio-informatique de la variabilité moléculaire [Reche and Reinherz, 2003].

Pour conclure sur les loci de classe II, nous venons de voir que ces derniers sont caractérisés par une importante sélection balancée visible, par exemple pour HLA-DRB1, sous la forme d'un avantage de l'hétérozygote. Toutefois, cette sélection balancée n'est pas une règle canonique, car le locus HLA-DQA1 semble être plus enclin à des variations de type de sélection, comme le suggèrent les résultats du Chapitre 4 où ce locus aurait évolué sous une ancienne sélection balancée, mais plus récemment sous un régime de sélection directionnelle.

Pressions de sélection due à la malaria

Un des résultats particulièrement intéressants pour les loci HLA de classe II est la présence d'un haplotype étendu chez les Mandenkalu, formé de HLA-DQA1*05:01∼ HLA-DQB1*03:19∼DRB1*13:04. Cet haplotype semble avoir été la cible d'un balayage sélectif lié à une fonction supposée de protection contre un ou plusieurs pathogènes, par exemple celui de l'onchocercose, par HLA-DQA1*05:01 et DQB1*03:19 [Meyer et al., 1994, Witter et al., 2007], cette hypothèse n'expliquant toutefois pas la présence de HLA-DRB1*13:04 dans cet haplotype. Deux hypothèses non mutuellement exclusives peuvent être émises pour expliquer la présence de cet allèle.

Dans la section 4.4, nous avons discuté de l'hypothèse d'apparition de HLA-DRB1*13:04 comme résultat d'une conversion allélique impliquant HLA-DRB1*11:02:01 comme donneur. La première hypothèse est alors de considérer que HLA-DRB1*11:02:01 était déjà présent dans un haplotype avec HLA-DQA1*05:01 et DQB1*03:19 et que DRB1*13:04 aurait alors été amené à une fréquence élevée par auto-stop génétique [11]. Les analyses de déséquilibre de liaison haplotypique du Chapitre 4 (Matériel supplémentaire S-45.2) ont identifié les séquences DRB1*3155, DQA1*1 et DQB1*2902 (pouvant correspondre respectivement à HLA-DRB1*11:02, DQA1*05:01 et DQB1*03:01/03:19) comme étant en déséquilibre de liaison (les trois paires de séquences) dans les populations Dangaléat du Tchad et Mossi du Burkina Faso, mais aussi (déséquilibre entre deux paires de loci sur les trois) les Gourounsi du Burkina Faso et les Peuls du Sénégal. Une étude menée en 2009 et comparant des populations Peuls, Mossi et Rimaibé du Burkina Faso a identifié l'haplotype HLA-DRB1*11∼DQB1*03 comme fréquent dans deux de ces populations (11% chez les Mossi et 21% chez les Rimaibé) [Lulli et al., 2009]. Le niveau de résolution disponible dans cette étude ne permet pas de déterminer s'il s'agit de HLA-DRB1*11:02:01 et -DQB1*03:19, mais elle met toutefois en évidence qu'un haplotype contenant ces deux lignages de classe II est retrouvé à des fréquences élevées dans des populations d'Afrique de l'ouest. Un haplotype à trois loci, composé des allèles HLA-DRB1*11:02∼DQA1*05:01∼DQB1*03:01 [12] a, quant à lui, été retrouvé à une fréquence de 3% dans une étude menée en 2001 au Cameroun [Pimtanothai et al., 2001]. Finalement, l'étude de Witter *et al.* en 2007, qui a rapporté pour la première fois HLA-DQB1*03:19, a mis en évidence que cet allèle était communément associé à l'allèle DRB1*11:02 [Witter et al., 2007]. Il semble ainsi que l'haplotype HLA-DQA1*05:01∼DQB1*03:19∼DRB1*11:02 ne soit pas rare en Afrique de l'ouest, ce qui soutient l'hypothèse que HLA-DRB1*13:04 ait été, dès son apparition, en déséquilibre de liaison avec HLA-DQA1*05:01 et DQB1*03:19.

La seconde hypothèse est de considérer que HLA-DRB1*13:04 possède une fonction protectrice vis-à-vis d'un pathogène présent en Afrique de l'ouest. Une récente étude (communication personnelle de A. Sanchez-Mazas, Mai 2020) a appliqué des méthodes de prédiction de liaisons peptidiques pour différents allèles HLA-DRB1 sur des peptides dérivés du protéome de *Plasmodium falciparum*, dont les résultats sont représentés sur la Figure 6.2.

11. *Hitchhiking*

12. Pour rappel, HLA-DQB1*03:19, ne diffère de -DQB1*03:01 que d'une seule substitution dans l'exon 3 et a été rapporté en 2007 [Witter et al., 2007].

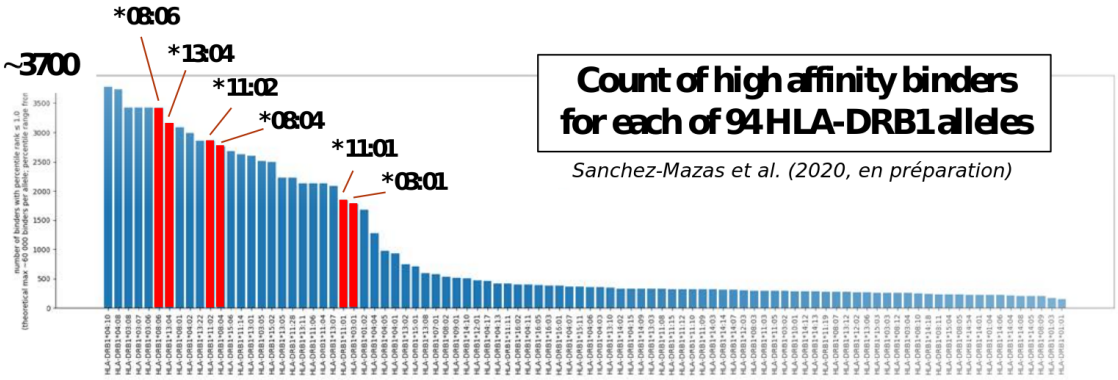


FIGURE 6.2 – Allèles HLA-DRB1 proposés comme meilleurs ligands des peptides (>1% de peptides liés) dérivés du protéome du *Plasmodium falciparum* d'après les prédictions *in silico* de liaison peptidique. La Figure représente la quantité de peptides liés à haute affinité (en ordonnée) par chacun des allèles HLA-DRB1 (en abscisse) selon les prédictions. Les allèles signalés en rouge sont les allèles retrouvés dans la population Mandenka du Chapitre 2. Figure reproduite avec l'autorisation de A. Sanchez-Mazas.

Sur 63'363 peptides de 15 acides-aminés [13], il est apparu que HLA-DRB1*13:04 faisait partie des meilleurs ligands pour ces peptides, c'est-à-dire liant *in-silico* plus de 1% des peptides avec une haute affinité, en l'occurrence plus de 3'000 peptides sur les 63'363. Ainsi il est possible de penser que la conversion allélique ayant abouti à HLA-DRB1*13:04 ait donné un avantage sélectif contre la malaria. Cette maladie étant fortement présente en Afrique de l'ouest (voir carte en Figure 4.30), cet avantage sélectif expliquerait la fréquence élevée de cet allèle. Sous cette seconde hypothèse, la fréquence élevée de l'haplotype formé par HLA-DQA1*05:01∼DQB1*03:19∼DRB1*13:04 s'expliquerait alors par une protection conférée contre la malaria par HLA-DRB1*13:04 et entraînant les allèles des loci HLA-DQA1 et HLA-DQB1, ou une double protection apportée contre l'onchocercose (par HLA-DQ), si celle-ci venait à se confirmer, et la malaria (HLA-DR). Il faut noter que cette même étude a aussi identifié HLA-DRB1*11:02 comme l'un des meilleurs ligands de *P. falciparum*, quoique prédit comme liant légèrement moins de peptides à haute affinité. Il semble donc que la conversion allélique n'aurait alors pas créé la fonction de protection de HLA-DRB1*13:04 vis-à-vis de la malaria, mais aurait amélioré celle de DRB1*11:02.

Sur les 94 allèles prédits comme meilleurs ligands de *P. falciparum* (Figure 6.2), six sont retrouvés dans la population Mandenka étudiée au Chapitre 2, avec une fréquence cumulée de 66% (HLA-DRB1*13:04 : 29%, DRB1*11:01 : 9%, DRB1*11:02 : 9%, DRB1*03:01 : 9%, DRB1*08:06 : 5%, DRB1*08:04 : 5%). L'analyse d'association entre les fréquences alléliques et la prévalence de la malaria réalisée dans le Chapitre 4 a identifié trois séquences DRB1-Exon2 comme ayant un rôle potentiel de défense contre la malaria. Ces trois séquences, DRB1*3144, DRB1*3149 et DRB1*3155 (Figure 4.32) peuvent correspondre à, respectivement, HLA-DRB1*08:06, DRB1*03:02 et DRB1*11:02 et chacun de ces trois allèles fait partie de la liste des meilleurs ligands de *P. falciparum* (Figure 6.2). Il est à noter que HLA-DRB1*13:04 n'est pas retrouvé dans l'analyse du Chapitre 4 alors qu'il apparaît ici comme l'un des meilleurs ligands. Cela peut s'expliquer par le fait que HLA-DRB1*13:04 est un allèle présentant une distribution géographique

13. Obtenus par une fenêtre glissante de 15 acides-aminés sur 56 protéines du *P. falciparum*.

très restreinte (voir page 287 pour la discussion sur la répartition géographique de cet allèle) et n'est donc pas apparu comme associé à la prévalence de *P. falciparum*, l'analyse ayant été menée sur une aire géographique large (la ceinture du Sahel et l'Afrique du nord).
Finalement, il est à noter qu'une association entre HLA-DRB1 et la malaria avait déjà été rapportée dans une étude de 1991 [Hill et al., 1991]. Dans cette étude, Hill *et al.* avaient comparé les profils génétiques des loci HLA de classe II pour 3'500 enfants en Gambie, infectés par la malaria, et avaient mis en évidence une plus forte fréquence de HLA-DRB1*13:02 chez les enfants montrant peu ou pas de symptômes comparé aux enfants montrant une forme sévère de malaria, suggérant alors une protection de cet allèle contre la maladie.
La fréquence des allèles prédits comme meilleurs ligands de *P. falciparum*, leur fréquence dans la population Mandenka et l'identification de trois des séquences DRB1-Exon2 du Chapitre 4 renforcent l'hypothèse d'un rôle de HLA-DRB1 dans la protection contre la malaria.

L'étude menée par Sanchez-Mazas *et al.* en 2017 a étudié les profils de liaison des acides-aminés des allèles HLA-B les plus fréquents dans des populations africaines (fréquence d'au moins 15% dans une des populations) [Sanchez-Mazas et al., 2017]. Il ressort que les allèles HLA-B*53:01 et B*78:01 (identifiés dans l'étude comme allèles potentiellement protecteurs de la malaria), B*35:01 (suggéré comme étant protecteur de la malaria dans une étude menée au Ghana [Yamazaki et al., 2011]) et B*39:10 (possédant une similarité de liaison peptidique avec HLA-B*53:01 [Yagüe et al., 1998]) montrent tous une affinité pour la proline dans la poche B du site de reconnaissance de l'antigène. Cette affinité est retrouvée dans plusieurs autres allèles HLA-B présents dans la population Mandenka, tels que HLA-B*51:01 et B*07:02, mais aussi les deux allèles HLA-B les plus fréquents, HLA-B*35:01 et B*78:01. Au total, la fréquence cumulée de tous ces allèles chez les Mandenkalu est de 36.6% et, parmi les cinq allèles HLA-B les plus fréquents dans cette population, quatre sont proposés comme protecteurs vis-à-vis de la malaria (HLA-B*35:01, *78:01, *15:03 et *53:01). Cette observation, ainsi que l'importante hétérozygotie à ce locus ($H = 0.94$, la plus élevée des huit loci étudiés, voir Table 2.7) et l'excès d'hétérozygotes au test d'Ewen-Watterson-Slatkin (Table 2.10), suggèrent un mécanisme de sélection balancée, exercée par la malaria, au locus HLA-B des Mandenkalu.

La pression de sélection de la malaria n'est toutefois pas exclusive aux seuls loci HLA-B et -DRB1, puisque l'analyse menée dans le Chapitre 4, sur l'association des fréquences alléliques avec la prévalence de cette maladie, a identifié des allèles appartenant à trois autres loci de classe II, à savoir HLA-DQA1, -DQB1 et -DPB1. Pour ces trois loci, les séquences DQA1*5 (correspondant aux allèles HLA-DQA1*04:01 et DQA1*04:02), DQB1*2902 (correspondant aux allèles HLA-DQB1*03 et incluant HLA-DQB1*03:19) et DPB1*66 (correspondant aux allèles HLA-DPB1*01:01:01, DPB1*162:01 ou DPB1*733:01) ont montré une fréquence corrélée avec la prévalence de la malaria liée à *P. falciparum* (même en tenant compte d'un effet géographique dans la distribution des fréquences alléliques, voir Figures 4.31 et 4.32).
Premièrement, bien que DQA1-Exon2 et DQB1-Exon2 soient en déséquilibre de liaison global chez 25 des 27 populations étudiées dans le Chapitre 4 (Figure 4.13 et Table 4.10), l'haplotype formé par les séquences DQA1*5 et DQB1*2902 n'est jamais retrouvé en déséquilibre de liaison dans ces populations, cela indiquant qu'il est impossible que la corrélation identifiée pour un de ces allèles soit due à un déséquilibre de liaison avec le

second allèle.
Ensuite, il est intéressant de noter que DQB1*2902 apparaît comme l'une des séquences associées à la malaria, puisque cette séquence peut correspondre à HLA-DQB1*03:19 (ou HLA-DQB1*03:01, voir matériel supplémentaire S45). Cette association renforcerait l'hypothèse émise page 295 concernant la haute fréquence de l'haplotype étendu de classe II chez les Mandenkalu, puisque HLA-DQB1*03:19 serait soit protecteur à la fois vis-à-vis de la malaria et vis-à-vis de l'onchocercose, soit protecteur vis-à-vis de la malaria plutôt que vis-à-vis de l'onchocercose, la première hypothèse n'étant que peu soutenue du fait de l'absence de mortalité connue liée directement à cette maladie.
Une étude de 2001, menée au Gabon sur une cohorte de 229 enfants, a associé DQB*03:01 à une protection contre la malaria, en observant une plus importante fréquence de HLA-DQB1*03:01 dans le groupe des enfants ayant une réponse immunitaire forte (via les INF-γ), donc cliniquement protégés contre la malaria, que dans les autres groupes [Migot-Nabias et al., 2001]. Une autre étude, menée au Brésil sur 276 individus naturellement exposés à la malaria, a elle aussi identifié une réponse immunitaire plus importante des porteurs de HLA-DQB1*03 lorsque leurs plasmas étaient exposés à une protéine spécifique ($PvMSP-3$) de *Plasmodium vivax* (le principal parasite responsable de la malaria en Amérique du sud) [Lima-Junior et al., 2012]. Ces deux études, ainsi que les résultats de notre travail, soutiennent alors l'hypothèse d'un rôle de DQB1*03:01/03:19 (les deux allèles possédant un même exon 2 et donc un site de reconnaissance de l'antigène identique) dans la défense immunitaire contre la malaria.

Finalement, l'association forte entre la fréquence de la séquence DPB1*66 et la prévalence de la malaria (Figure 4.31) est un résultat intéressant, puisque, à notre connaissance, seules deux études avaient jusque là rapporté un lien entre HLA-DPB1 et la malaria. La première est celle de May *et al.*, discutée dans la section 4.4, qui a rapporté une fréquence plus élevée de HLA-DPB1*01:01 [14] (ainsi que de HLA-DQA1*04:01, correspondant à la séquence DQA1*05 du Chapitre 4) parmi les individus montrant une forme bénigne de la malaria, comparé à ceux présentant une forme sévère [May et al., 1999]. La seconde concerne le développement d'un vaccin contre la malaria et un test chez 71 individus Thai, qui a mis en évidence une réponse en anticorps plus importante chez les porteurs de HLA-DPB1*05:01 [Stephens et al., 1995].
Le peu d'études rapportant une association entre HLA-DPB1 et la malaria est d'autant plus étonnant que DPB1*66 montre l'une des plus fortes associations avec la malaria dans notre étude, mais qu'il semble aussi être sous sélection positive chez les Mossi, Gourmantché et Gourounsi du Burkina Faso. Cette observation est confortée par le déficit en hétérozygotes lié à une très forte fréquence de DPB1*66 (voir les Figures 4.6, 4.11 et 4.14), les plus faibles valeurs du D de Tajima observées pour ces populations, par rapport aux autres populations elles aussi exposées (Table 4.14), et l'association négative entre la prévalence de *P. falciparum* avec les valeurs du D de Tajima et les Θ_π des codons ARS (Équations 4.1 et 4.2).
De précédentes études menées sur HLA-DPB1 avaient alors conclu à la neutralité de ce locus. En 2001, l'étude de Sanchez-Mazas analysant 22 populations réparties mondialement (à l'exception des Amériques) avait conclu à la neutralité du locus HLA-DPB1 au vu de ses distributions de fréquences alléliques (en forme de L) et de l'absence de résultats significatifs au test d'Ewens-Watterson [Sanchez-Mazas, 2001]. L'étude de Solberg *et al.* de 2008, menée sur 497 populations réparties mondialement (incluant les Amériques), était arrivée à la même conclusion sur la base de tests d'homozygotie

14. Pouvant correspondre à la séquence DPB1*66.

[Solberg et al., 2008]. De plus, l'étude de Buhler et Sanchez-Mazas, réalisée sur les séquences nucléotidiques des exons 2 de 4'911 individus provenant de 56 populations (pour HLA-DPB1) de tous les continents, avait mis en évidence une plus forte différenciation des populations au locus HLA-DPB1 et le rejet de la neutralité au test du D de Tajima pour seulement 4 des 56 populations (dont trois populations avec un D supérieur à 0) [Buhler and Sanchez-Mazas, 2011]. Les auteurs ont alors conclu à la neutralité de ce locus tout en interprétant les valeurs du D de Tajima significatives comme la trace d'une ancienne sélection balancée, hypothèse supportée par des preuves de sélection balancée sur certains sites du domaine $\beta 1$ de la molécule.
L'analyse, au Chapitre 5, de la distribution de l'information (entropie) des différentes régions géniques (introns et exons) a mis en évidence une plus faible information des régions de HLA-DPB1 (Figure 5.5), à l'exception de l'analyse de l'information mutuelle (Figure 5.6) montrant une quantité d'information importante à l'exon 2 de HLA-DPB1 et plus faible à l'exon 3 (la plus faible parmi les huit loci étudiés).
Ces résultats peuvent s'interpréter en considérant que HLA-DPB1 évolue de manière neutre dans la plupart des populations (tel que mis en évidence par les trois études à large échelle précédemment citées), tout en présentant des signaux d'une ancienne sélection balancée (l'entropie élevée de l'exon 2) et des signaux de sélection récente dans les populations d'Afrique de l'ouest et centrale, probablement due à la pression particulièrement forte de sélection exercée par la malaria.

En conclusion, en plus de HLA-DRB1, les principaux loci de classe II (HLA-DQA1, -DQB1 et -DPB1) montrent, eux aussi la présence d'allèles ayant un possible rôle protecteur vis-à-vis de la malaria.

Pour conclure cette partie relative à la pression de sélection causée par la malaria, nous venons de voir qu'une sélection naturelle exercée par le pathogène responsable de cette maladie sur le système HLA n'est vraisemblablement pas limitée à un seul locus mais touche plusieurs loci de classe I et II (HLA-B, -DRB1, -DQA1, -DQB1 et -DPB1). De plus, il est important de noter que tous les loci et toutes les populations ne sont pas impactées de la même manière. En effet, si la sélection balancée apparaît comme prédominante pour les loci HLA-B et -DRB1, HLA-DPB1 montre des signaux de sélection positive chez les populations les plus exposées au parasite.

Finalement, pour conclure l'ensemble de cette section dédiée à la sélection naturelle opérant sur les loci HLA, nous soutenons l'idée que cette sélection est plurielle et variée. Nous avons vu que, bien qu'opérant généralement sur les régions en contact avec les peptides antigéniques, la sélection n'agit pas toujours de la même façon entre les différentes régions des différents gènes, tels que les exons 2 et 3 de HLA-A et -B. Nous avons aussi vu que cette sélection prenait des formes différentes selon les loci, certains évoluant sous sélection balancée (HLA-B et -DRB1), tandis que d'autres sous sélection positive (HLA-DPB1 pour certaines des populations d'Afrique de l'ouest), d'autres enfin montrant aussi des changements dans les régimes de sélection (HLA-DQA1). Finalement, nous avons aussi déduit qu'une pression de sélection exercée par un seul pathogène pouvait toucher plusieurs loci et de plusieurs manières, mais aussi que certains allèles (tel que HLA-DQB1*03:19) montraient des traces de sélection imputables à plusieurs pathogènes.

4.2 Effets de la démographie

Une étude, menée par Di et Sanchez-Mazas en 2014, visant à caractériser la diversité génétique des populations d'Asie de l'est avait mis en évidence une plus importante contribution des effets démographiques (par rapport à la sélection naturelle) dans la distribution du polymorphisme HLA de cette région [Di and Sanchez-Mazas, 2014], ces résultats ayant aussi été obtenus par une étude des loci MHC-DQB et -DRB du primate *Microcebus berthae* menée en 2015 [Pechouskova et al., 2015]. La démographie peut agir sur la diversité des populations de plusieurs manières, à savoir des expansions, contractions ou fusions de populations, mais aussi par des effets fondateurs répétés, tels que mis en évidence dans le peuplement mondial après la sortie d'Afrique [Deshpande et al., 2009]. Cette dernière partie de la discussion vise à synthétiser et expliquer les différents résultats de ce travail de doctorat à la lumière des effets démographiques.

Selon la théorie neutraliste de l'évolution, formulée par Kimura en 1968 [Kimura, 1968, Kimura, 1991], la force évolutive majeure agissant sur l'ADN est la dérive génétique. La dérive génétique est le changement de fréquences des allèles d'une génération à l'autre par le simple échantillonnage aléatoire de ces allèles dans une population de taille finie. L'effet de la dérive génétique varie donc en fonction de la taille efficace des populations et, plus cette dernière est réduite, plus l'effet de la dérive est important (on parle alors de dérive génétique rapide).
Plusieurs populations, étudiées dans le Chapitre 2, montrent des signaux qui pourraient être interprétés comme de la dérive génétique rapide. Le locus DRB1-Exon2 est caractérisé par une sélection balancée de type avantage de l'hétérozygote (voir page 292 pour la discussion de la sélection à ce locus). Ainsi, à ce locus, un excès d'homozygotes pourrait être le signal d'une dérive génétique rapide. La Figure 4.14 montre un excès d'homozygotes pour les Arabes Rashaida du Soudan et les Peuls du Mali au test d'Ewens-Watterson-Slatkin. Ces deux populations montrent, de plus, des valeurs du D de Tajima significativement supérieures à zéro pour les codons non-ARS de DRB1-Exon2, signal aussi visible pour les Imazighen de Figuig. Une valeur du D de Tajima supérieure à zéro est soit le résultat d'une sélection balancée, soit le résultat d'une contraction de population (voir la Section 6.5 pour le détail de ce test) et la sélection semblant surtout agir sur les codons ARS de HLA-DRB1, une valeur de D de Tajima supérieure à zéro aux codons non-ARS pourrait donc être interprétée comme un signal démographique.
Une étude de Vangenot *et al.*, comparant la diversité génétique du MHC/HLA de quatre cohortes de chimpanzés avec celle de 89 populations humaines, a mis en évidence que les populations évoluant sous dérive génétique rapide présentaient un déséquilibre de liaison global plus important que celles évoluant sous dérive génétique lente [Vangenot et al., 2020]. Ces résultats sont similaires à ceux présentés dans la revue de Tishkoff *et al.*, montrant un déséquilibre de liaison plus important pour les populations non-africaines par rapport aux populations africaines, ces dernières n'ayant pas subi le goulot d'étranglement génétique consécutif à la sortie d'Afrique et ayant, pour beaucoup, maintenu depuis lors des tailles efficaces de populations importantes [Tishkoff and Kidd, 2004]. Bien que cette analyse ait été réalisée sur des SNP à l'échelle du génome (et non sur une région génomique de taille plus réduite telle que la région des loci HLA de classe II) et pour des événements bien plus anciens (la sortie d'Afrique étant datée de 100'000 à 150'000 ans [Scally and Durbin, 2012, López et al., 2015]), plusieurs signaux de déséquilibre de liaison observées dans les populations du chapitre 4 suggèrent une hypothèse similaire. Les populations Beja Hadendoa et Arabes Rashaida du Soudan, ainsi que les Imazighen de Figuig, les Dangaléat du Tchad, les Peuls et Mandenka du

Sénégal montrent toutes un déséquilibre de liaison global entre DRB1-Exon2∼DQA1-Exon2, DQA1-Exon2∼DQB1-Exon2 et DQB1-Exon2∼DPB1-Exon2 (Figure 4.13). Ce déséquilibre de liaison incluant le locus DPB1-Exon2 est d'autant plus intéressant que ce dernier présente une importante distance génétique avec les trois autres loci, car séparé de ces derniers par une région où la recombinaison est élevée, dû à un ou plusieurs points chauds de recombinaison [Martin et al., 1995]. Ce résultat est toutefois à nuancer, car les Mandenkalu du Sénégal montrent aussi un déséquilibre de liaison, or cette population ne montre pas de signaux de dérive génétique rapide (voir page 54). De plus, le déséquilibre impliquant DPB1-Exon2 pour les Dangaléat du Tchad pourrait être imputé à plusieurs évènements d'absorption de groupes de populations et d'isolement culturel (voir Section 4.3).
La dérive génétique rapide va aussi causer la disparition des allèles les moins fréquents dans une population, visible sur les distributions de fréquences alléliques par quelques allèles à haute fréquence et une absence de variants rares. Les populations de Beja Hadendoa et d'Arabes Rashaida du Soudan montrent plusieurs allèles à fortes fréquences aux loci DRB1-Exon2 (DRB1*3136 [15], fréquences de respectivement 36 et 35%, voir Figure 4.8) ainsi que DQB1-Exon2 (DQB1*2901 [16], 49 et 57%, voir Figure 4.10). Ces fréquences alléliques sont responsables de la position extrême de ces populations sur les MDS de ces mêmes loci (Figures 4.20 et 4.24). Nous avons montré précédemment que ces deux loci évoluaient sous sélection balancée (voir page 4.1). Or, cette dernière est supposée diminuer la différenciation des populations, telle que mesurée par F_{ST} ou Θ_w [Aguilar and Garza, 2006, Biedrzycka and Radwan, 2008, Brandt et al., 2018]. La position de ces populations sur les MDS de DRB1-Exon2 et DQB1-Exon2, et les fortes fréquences observées pour certains allèles de ces loci peut alors s'expliquer par une dérive génétique rapide agissant sur ces populations. L'étude menée en 2017 par Sanchez-Mazas *et al.* a, elle aussi, montré des fréquences importantes de certains allèles aux loci HLA-A et -B de ces mêmes populations (HLA-A*02 à plus de 45%, B*50 et B*51 à 15 et 26% ainsi que B*07 à 17% chez les Arabes Rashaida) [Sanchez-Mazas et al., 2017].
Il semble alors que le mode de vie nomade de ces populations, caractérisé par un isolement et un petit effectif, ait entraîné une dérive génétique rapide, responsable d'un plus fort déséquilibre de liaison global et d'une différenciation plus importante de ces populations.

Le Chapitre 2 montre aussi deux exemples de signaux démographiques, visibles sur l'exon 2 de HLA-A. Comme discuté dans la Section 4.1, l'exon 2 de HLA semble être moins soumis à la pression de sélection liée aux pathogènes et montrerait alors davantage d'effets liés à la démographie.
Ainsi, pour les Mandenkalu, la position des codons ARS de l'exon 2 de HLA-A sur l'ACP (en bas, Figure 2.8), associée aux faibles valeurs du D de Tajima ($D^{ARS}_{exon2} = 0.57$ et $D^{non-ARS}_{exon2} = -0.89$, Figure 2.7) traduirait non pas un effet de sélection (balayage sélectif, voir Section 6.5) mais un effet d'expansion démographique, cohérent avec les observations d'une importante diversité génétique pour d'autres marqueurs [Poloni et al., 1995, Graven et al., 1995, Martinson et al., 1995, Dard et al., 1996, Currat et al., 2002, Sabbagh et al., 2008].
De la même manière, les valeurs plus élevées de D de Tajima observées pour HLA-A exon 2 des Cham comparé aux Mandenkalu ($D^{ARS}_{exon2} = 2.35$ et $D^{non-ARS}_{exon2} = 0.99$, Figure 2.7) serait le signal d'une contraction démographique plutôt que d'une sélection balancée. Une étude menée sur des populations vietnamiennes, incluant les Cham, a observé une chute

15. Correspondant à HLA-DRB1*07, voir annexe S-44.
16. Correspondant à HLA-DQB1*02, voir annexe S-44.

de la taille efficace de la population Cham (estimée à l'aide de *Extended Bayesian Skyline Plot* [Heled and Drummond, 2008] sur les génomes mitochondriaux) il y a 1'000 ans. Cette chute démographique coïncide avec l'événement appelé Nam tiến, l'expansion du territoire du Vietnam vers le sud entre le 10ème et 18ème siècle, caractérisée par de nombreuses guerres et une diminution de la population du Vietnam [Pischedda et al., 2017].

Un autre effet démographique visible pour la population Cham serait lié à l'origine des Cham. L'hypothèse actuelle la plus probable pour expliquer l'origine de cette population est l'hypothèse démique (voir page 1) issue de la linguistique, les Cham étant l'une des deux seules populations d'Asie continentale du sud-est (avec les Moken) locutrices de langues de la famille austronésienne. Cette hypothèse avance que les Cham seraient issus d'une fusion de populations entre une population austronésienne migrante et une population locale d'Asie continentale du sud-est [Thurgood, 1999, Higham, 2002, Southworth et al., 2004, Bellwood, 2007, Peng et al., 2010].
La Figure 2.14, représentant le déséquilibre haplotypique pour les huit loci étudiés chez les Cham, montre deux grands réseaux d'allèles en déséquilibre (signifiés en bleu et rouge sur la figure). Pour comparaison, la Figure 2.15 montre le réseau d'allèles en déséquilibre de liaison haplotypique pour la population Mandenka. Cette dernière figure montre un seul réseau d'haplotypes, principalement formé par HLA-DRB1*13:04∼DQA1*05:01 ∼DQB1*03:19, dont l'origine est discutée dans la Section 4.1. La comparaison de ces deux figures illustre les différences fondamentales entre ces deux populations. Les Mandenka étant une grande population africaine n'ayant pas subi (à notre connaissance) de fusion démographique telle que supposée chez les Cham, leur déséquilibre de liaison est principalement attribuable à l'effet de la sélection naturelle. Les Cham, quant à eux, sont une population non-africaine, ayant subi plusieurs événements démographiques majeurs et il semble, qu'à l'instar des autres populations d'Asie de l'est, ce soient les effets démographiques qui prédominent dans leur profil génétique HLA.

Pour conclure sur cette partie liée aux effets de la démographie, nous venons de voir que la démographie apparaît elle aussi comme une force majeure dans la distribution du polymorphisme HLA. Toutefois, les effets confondants de la sélection naturelle et de la démographie sont difficiles à démêler, particulièrement lorsque l'on étudie des régions des gènes HLA fortement sous sélection. Les effets démographiques s'observent alors par une analyse plus fine de ces régions, telle que l'analyse séparée des exons 2 et 3 de certains gènes de classe I qui permet de mieux séparer les effets de la sélection et de la démographie, voire un découpage encore plus fin des exons en codons ARS et non-ARS qui, dans le cas de DRB1-Exon2, permet d'identifier des effets démographiques sur un locus sous forte sélection balancée.

Chapitre 7

Conclusion

En conclusion de ces trois études présentées, nous allons faire un état des avancées et des questions encore en suspens au terme de ce travail de doctorat. Dans l'introduction générale, nous avons vu la grande diversité du système HLA et présenté les différentes théories et hypothèses sur les mécanismes évolutifs qui sont à l'origine de cette diversité. Nous avons alors émis plusieurs questions sur cette diversité, questions qui se regroupent en trois grands axes.

Le premier axe de ces recherches s'intéresse la génération de ce polymorphisme : quels sont, d'une part, les mécanismes qui ont mené aux gènes HLA que l'on connaît actuellement et, d'autre part, les mécanismes qui ont généré les variants au sein de ces gènes.
Les analyses du Chapitre 5 ont montré que les gènes de classe I semblaient tous avoir une origine commune, qui remonterait à un gène MHC ancestral de classe I, cohérent avec le modèle de « naissance et mort » de la génération des gènes HLA. Ces mêmes analyses ont aussi montré que les gènes de classe II, au contraire, se regroupaient en deux groupes selon s'ils codent pour la chaîne α ou β des molécules HLA de classe II.
Ensuite, les études des Chapitres 2 et 4 ont permis d'étudier l'étendue de la mutation ponctuelle et de la recombinaison comme facteur de génération du polymorphisme HLA. Nous avons vu, au travers de plusieurs exemples, comment les nouveaux sites variables étaient générés par des mutations ponctuelles de l'ADN et comment cette variabilité était échangée entre les allèles, grâce à la recombinaison intragénique. Nous avons alors eu un aperçu de l'importance de la recombinaison avec l'allèle HLA-DRB1*13:04, qui serait apparu à la suite d'une recombinaison intragénique, probablement en Afrique de l'ouest. Cet exemple a illustré, d'une part, l'importance quantitative de la recombinaison en montrant que le fragment échangé n'est pas rare parmi les séquences de HLA-DRB1 et, d'autre part, l'importance qualitative de la recombinaison, puisque ce phénomène aurait permis d'améliorer la fonction d'un allèle pré-existant vis-à-vis de la présentation de certains peptides antigéniques.

Le second axe de recherche concerne la répartition du polymorphisme HLA entre les différents groupes humains. Nous avons vu, au travers de plusieurs exemples, que les flux géniques entre les populations constituaient une part notable de la répartition des allèles HLA dans les populations. Ces flux géniques peuvent alors être de plusieurs magnitudes, allant d'un simple partage de quelques allèles entre des populations coexistant sur une même aire géographique, à une homogénéisation de populations qui ne sont alors plus différenciées. Nous avons aussi étudié l'origine des Cham, pour lesquels les analyses de

déséquilibre de liaison ont révélé la présence de deux grands réseaux d'allèles, qui sont de probables signaux d'une fusion entre deux groupes de populations d'origines différentes.
De plus, l'étude des flux de gènes entre les populations de la bande du Sahel a révélé une différence est-ouest des populations, qui apparaît liée à l'histoire de ces populations. Les populations de l'Afrique de l'ouest sont différenciées des populations de l'est du Sahel et montrent deux groupes distincts, caractérisés soit par des allèles qui leur sont propres (HLA-DRB1*13:04 pour les populations du Sénégal), soit par des fréquences de certains allèles beaucoup plus élevées (telle que la séquence DPB1*66 [1] au Burkina Faso). Au contraire, les populations de l'est du Sahel (régions d'Afrique de l'est, centrale et du Soudan), forment un groupe à part, au sein duquel les populations montrent davantage de diversité et beaucoup moins de différences. Cette structure serait ici due à des liens plus importants entre ces populations, mais aussi des mouvements plus complexes entre les populations, puisque cette région voit évoluer plusieurs populations pastoralistes nomades, dont certaines sont arrivées il y a moins de 200 ans depuis la péninsule arabique.

Le troisième et dernier grand axe de recherche porte sur les facteurs qui font varier ces polymorphismes entre les populations. Ces facteurs sont les deux grandes forces majeures de l'évolution, à savoir la sélection naturelle, au travers des pressions exercées par l'environnement sur les individus, et les forces stochastiques liées aux histoires démographiques des populations, telles que la dérive génétique. Les études présentées dans les différents chapitres ont illustré les différents effets et modalités de la sélection.
L'étude des Mandenkalu du Sénégal a montré que le locus HLA-B semblait évoluer sous un régime de sélection balancée, puisque plusieurs des allèles les plus fréquents auraient un rôle protecteur vis-à-vis de la malaria. Au contraire, pour cette même population, les allèles de classe II présentent un haplotype étendu, de HLA-DRB1 à HLA-DQB1, qui aurait subi un balayage sélectif récent. Cet haplotype aurait été sélectionné car apportant un avantage sélectif face à au moins une, et peut-être deux maladies fortement présentes dans cette région, la malaria et l'onchocercose. Cette étude a aussi apporté des preuves supplémentaires de la cible de la sélection, en montrant que ce sont particulièrement les codons qui codent pour le site de reconnaissance de l'antigène (ARS) qui abritent la plus importante part de la diversité des gènes HLA.
L'étude des populations du Sahel a mis en évidence des grandes tendances évolutives pour les loci de classe II. On peut citer le cas de HLA-DRB1, dont l'extrême diversité montre nettement une sélection balancée de type avantage de l'hétérozygote. Au contraire de HLA-DRB1, HLA-DPB1 révèle des traces d'une ancienne sélection balancée, tout en évoluant actuellement de manière neutre, à l'exception de quelques populations où la malaria semble causer une sélection directionnelle sur un des allèles de ce locus. Finalement, il nous faut aussi citer le cas de HLA-DQA1 qui montre des signaux complexes, voire contradictoires, qui s'expliqueraient, là aussi, par un changement de régime de sélection, passant d'une ancienne sélection balancée à une sélection directionnelle récente.

Les effets démographiques, et par eux la dérive génétique, apparaissent aussi, dans ces études, comme une composante majeure de la variation de la diversité du HLA. Cependant, eu égard à la fonction première des gènes HLA, ces effets sont plus difficiles à caractériser. En effet, les fortes pressions de sélection qui agissent sur les loci HLA, notamment sur les exons 2 étudiés au Chapitre 4 vont avoir des effets confondants avec ceux de la démographie. Il est alors nécessaire de réaliser une analyse plus fine de ces régions, notamment en séparant les codons codant pour les ARS de ceux ne codant

1. Pouvant correspondre aux allèles HLA-DPB1*01:01:01, -DPB1*162:01 ou -DPB1*733:01.

pas pour les ARS. Ces derniers, moins soumis à la sélection, sont alors plus à même de révéler des signaux démographiques. C'est ainsi qu'ont pu être observées des traces de contraction démographique pour plusieurs populations (semi-)nomades d'Afrique, visibles sur les codons non-ARS de l'exon 2 de HLA-DRB1, ce locus évoluant pourtant sous une forte sélection balancée.
Nous avons aussi vu que, dans d'autres circonstances, c'est la comparaison des différents exons qui permet d'obtenir des indices sur les évènements démographiques des populations. Par exemple, pour les Cham du Vietnam, une contraction démographique est observée lorsque l'on compare la diversité des exons 2 et 3 de HLA-A. En effet, l'exon 2 de HLA-A est moins décisif dans la reconnaissance des peptides antigéniques (au contraire de l'exon 3) et, à l'instar des codons non-ARS, il va être plus enclin à révéler des signaux démographiques. Au contraire, ce même exon chez les Mandenkalu porte la marque d'une expansion démographique.

Il nous faut finalement parler de l'avenir de ces recherches. Les deux études présentées aux Chapitres 2 et 4 visaient, aussi, à démontrer la faisabilité et l'intérêt, d'une part, des études génomiques à large échelle, plus particulièrement en Afrique et, d'autre part, des séquençages de gènes complets pour l'étude des populations. Les résultats obtenus ici se sont montrés prometteurs et encouragent à poursuivre dans cette voie. C'est le sujet du nouveau programme de recherche porté par Alicia Sanchez-Mazas et son équipe, qui va, à l'aide de collaborations internationales, réaliser un séquençage complet de 12 gènes HLA pour 50 populations africaines, qui incluent certaines de celles analysées au Chapitre 4. Ce projet inclut aussi les séquençages de SNP à l'échelle du génome afin de contrôler les effets dus à la démographie de ceux dus à la sélection. Ce projet vise à explorer la diversité HLA de populations jusque-là peu ou pas étudiées pour les polymorphismes HLA, mais aussi de gènes sur lesquels l'attention s'est encore peu portée, comme le gène HLA-DRA.

Ce doctorat est un travail qui se situe à la frontière entre l'anthropologie, l'immunogénétique et la bio-informatique. Une grande partie des résultats de ce travail n'ont été rendus possibles que par l'adoption, par la communauté HLA, des nouvelles méthodes de séquençage. L'intérêt de ces méthodes réside dans le volume des séquençages qu'elles produisent. Cela permet d'intégrer, au sein d'une seule et même étude de génétique des populations, l'information moléculaire de plusieurs milliers d'individus. De plus, les innovations dans le domaine du séquençage ADN, qui émergent sur une base presque annuelle, permettent de séquencer des régions encore plus grandes des gènes pour ces milliers d'individus.
Il est évident que dans les prochaines années, l'étude du HLA, déjà captivante de par la complexité et la diversité de cette région, va devenir encore plus passionnante grâce à ces nouvelles technologies. Ces avancées technologiques vont permettre d'affiner les modèles et théories évolutives de cette région, tout en caractérisant plus précisément la diversité de ce polymorphisme, en explorant notamment les régions introniques et régulatrices de ces gènes. Cette accélération de la recherche sur le HLA va s'accompagner aussi de nouveaux défis, techniques et scientifiques à résoudre. Cela va concerner la mise au point d'outils adaptés pour étudier de tels volumes de données, mais aussi probablement d'une mise à jour de la nomenclature des allèles HLA, afin de pouvoir gérer le nombre croissant (exponentiellement) de nouveaux allèles.

Finalement, ce n'est qu'au travers de la mise en commun des données, à l'aide de

bases de données ouvertes telles que IPD-IMGT/HLA, mais aussi, de la publication des résultats dans des journaux en libre accès, que les recherches sur le HLA et leurs découvertes sauront trouver leur plein potentiel.

Chapitre 8

Bibliographie

[Adamek et al., 2015] Adamek, M., Klages, C., Bauer, M., Kudlek, E., Drechsler, A., Leuser, B., Scherer, S., Opelz, G., and Tran, T. H. (2015). Seven novel HLA alleles reflect different mechanisms involved in the evolution of HLA diversity : Description of the new alleles and review of the literature. *Human Immunology*, 76(1) : 30–35.

[Aguilar and Garza, 2006] Aguilar, A. and Garza, J. C. (2006). A comparison of variability and population structure for major histocompatibility complex and microsatellite loci in California coastal steelhead (Oncorhynchus mykiss Walbaum) : steelhead MHC. *Molecular Ecology*, 15(4) : 923–937.

[Aidoo et al., 2002] Aidoo, M., Terlouw, D. J., Kolczak, M. S., McElroy, P. D., ter Kuile, F. O., Kariuki, S., Nahlen, B. L., Lal, A. A., and Udhayakumar, V. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet*, 359(9314) : 1311–1312.

[Alexander et al., 2009] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9) : 1655–1664.

[Allen et al., 2018] Allen, E. S., Yang, B., Garrett, J., Ball, E. D., Maiers, M., and Morris, G. P. (2018). Improved accuracy of clinical HLA genotyping by next-generation DNA sequencing affects unrelated donor search results for hematopoietic stem cell transplantation. *Human Immunology*, 79(12) : 848–854.

[Allison, 1954] Allison, A. C. (1954). Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *BMJ*, 1(4857) : 290–294.

[Allsopp et al., 1991] Allsopp, C., Hill, A., Kwiatkowski, D., Hughes, A., Bunce, M., Taylor, C., Pazmany, L., Brewster, D., McMichael, A., and Greenwood, B. (1991). Sequence analysis of HLA-Bw53, a common West African allele, suggests an origin by gene conversion of HLA-B35. *Human Immunology*, 30(2) : 105–109.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) : 403–410.

[Alvarez-Busto et al., 2007] Alvarez-Busto, J., García-Etxebarria, K., Herrero, J., Garin, I., and Jugo, B. M. (2007). Diversity and evolution of the MHC-DRB1 gene in the two endemic Iberian subspecies of Pyrenean chamois, Rupicapra pyrenaica. *Heredity*, 99(4) : 406–413.

[Amadou, 1999] Amadou, C. (1999). Evolution of the MHC class I region : the framework hypothesis. *Immunogenetics*, 49(4) : 362–367.

[Amos et al., 1965] Amos, D. B., Russell, P. S., National Research Council (U.S.). Division of Medical Sciences, and National Research Council (U.S.) Committee on Tissue Transplantation (1965). Histocompatibility Testing : Report of a Conference and Workshop. In *Washington DC : National Academy of Sciences – National Research Council*, Durham (USA).

[Andersson et al., 1987] Andersson, G., Larhammar, D., Widmark, E., Servenius, B., Peterson, P. A., and Rask, L. (1987). Class II genes of the human major histocompatibility complex. Organization and evolutionary relationship of the DR beta genes. *Journal of Biological Chemistry*, 262(18) : 8748–8758.

[Andersson and Mikko, 1995] Andersson, L. and Mikko, S. (1995). Generation of MHC Class II Diversity by Intra- and Intergenic Recombination. *Immunological Reviews*, 143(1) : 5–12.

[Andrien et al., 1993] Andrien, M., Tiercy, J.-M., Defleur, V., Bouillenne, C., Toungouz, M., Jeannet, M., and Dupont, E. (1993). HLA-B locus DNA typing : Detection of B*7801 and seven additional alleles by BW6-specific exon 2 amplification. *Tissue Antigens*, 42(5) : 480–487.

[Apps et al., 2013] Apps, R., Qi, Y., Carlson, J. M., Chen, H., Gao, X., Thomas, R., Yuki, Y., Del Prete, G. Q., Goulder, P., Brumme, Z. L., Brumme, C. J., John, M., Mallal, S., Nelson, G., Bosch, R., Heckerman, D., Stein, J. L., Soderberg, K. A., Moody, M. A., Denny, T. N., Zeng, X., Fang, J., Moffett, A., Lifson, J. D., Goedert, J. J., Buchbinder, S., Kirk, G. D., Fellay, J., McLaren, P., Deeks, S. G., Pereyra, F., Walker, B., Michael, N. L., Weintrob, A., Wolinsky, S., Liao, W., and Carrington, M. (2013). Influence of HLA-C Expression Level on HIV Control. *Science*, 340(6128) : 87–91.

[Aris-Brosou and Excoffier, 1996] Aris-Brosou, S. and Excoffier, L. (1996). The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution*, 13(3) : 494–504.

[B. Patillon et al., 2014] B. Patillon, P. Luisi, E. S. Poloni, S. Boukouvala, P. Darlu, E. Genin, and A. Sabbagh (2014). A Homogenizing Process of Selection Has Maintained an“Ultra-Slow” Acetylation NAT2 Variant in Humans. *Human Biology*, 86(3) : 185.

[Ba et al., 2015] Ba, A., Beley, S., Chiaroni, J., Bailly, P., and Silvy, M. (2015). RH diversity in Mali : characterization of a new haplotype RHD*DIVa/RHCE*ceTI(D2) : RH diversity in mali. *Transfusion*, 55(6pt2) : 1423–1431.

[Bai et al., 2014] Bai, Y., Ni, M., Cooper, B., Wei, Y., and Fury, W. (2014). Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics*, 15(1) : 325.

[Bamshad and Wooding, 2003] Bamshad, M. and Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics*, 4(2) : 99–110.

[Band et al., 1998] Band, M., Larson, J., Womack, J., and Lewin, H. (1998). A Radiation Hybrid Map of BTA23 : Identification of a Chromosomal Rearrangement Leading to Separation of the Cattle MHC Class II Subregions. *Genomics*, 53(3) : 269–275.

[Barone et al., 2015] Barone, J. C., Saito, K., Beutner, K., Campo, M., Dong, W., Goswami, C. P., Johnson, E. S., Wang, Z.-X., and Hsu, S. (2015). HLA-genotyping of clinical specimens using Ion Torrent-based NGS. *Human Immunology*, 76(12) : 903–909.

[Barquera et al., 2020] Barquera, R., Collen, E., Di, D., Buhler, S., Teixeira, J., Llamas, B., Nunes, J. M., and Sanchez-Mazas, A. (2020). Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. *HLA*, page tan.13956.

[Belich et al., 1992] Belich, M. P., Madrigal, J. A., Hildebrand, W. H., Zemmour, J., Williams, R. C., Luz, R., Petzl-Erler, M. L., and Parham, P. (1992). Unusual HLA-B alleles in two tribes of Brazilian Indians. *Nature*, 357(6376) : 326–329.

[Bellwood, 2007] Bellwood, P. (2007). *Prehistory of the Indo-Malaysian Archipelago*. (ACT)ANU E Press, Canberra.

[Bellwood et al., 2006] Bellwood, P., Fox, J. J., and Tryon, D. (2006). The Austronesians in history : common origins and diverse transformations. In *The Austronesians : historical and comparative perspectives*, pages 1–14. Canberra : ANU E Press.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1) : 289–300.

[Bentley et al., 2009] Bentley, G., Higuchi, R., Hoglund, B., Goodridge, D., Sayer, D., Trachtenberg, E. A., and Erlich, H. A. (2009). High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens*, 74(5) : 393–403.

[Benzécri, 1973] Benzécri, J.-P., editor (1973). *L'analyse des correspondances*. Number 2 in L'analyse des données. Dunod, Paris.

[Bergström et al., 1998] Bergström, T. F., Josefsson, A., Erlich, H. A., and Gyllensten, U. (1998). Recent origin of HLA-DRB1 alleles and implications for human evolution. *Nature Genetics*, 18(3) : 237–242.

[Bernatchez and Landry, 2003] Bernatchez, L. and Landry, C. (2003). MHC studies in nonmodel vertebrates : what have we learned about natural selection in 15 years ? : MHC studies in nonmodel vertebrates. *Journal of Evolutionary Biology*, 16(3) : 363–377.

[Bernstein et al., 1996] Bernstein, R. M., Schluter, S. F., Bernstein, H., and Marchalonis, J. J. (1996). Primordial emergence of the recombination activating gene 1 (RAG1) : sequence of the complete shark gene indicates homology to microbial integrases. *Proceedings of the National Academy of Sciences*, 93(18) : 9454–9459.

[Bernus, 1993] Bernus, E. (1993). *Touaregs nigériens : unité culturelle et diversité régionale d'un peuple pasteur*. Mémoires ORSTOM. Éditions l'Harmattan, Paris.

[Betti et al., 2009] Betti, L., Balloux, F., Amos, W., Hanihara, T., and Manica, A. (2009). Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proceedings of the Royal Society B : Biological Sciences*, 276(1658) : 809–814.

[Bhatt et al., 2015] Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., and Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526(7572) : 207–211.

[Bidwell et al., 1988] Bidwell, J. L., Bidwell, E. A., Savage, D. A., Middleton, D., Klouda, P. T., and Bradley, B. A. (1988). A DNA-RFLP typing system that positively identifies serologically well-defined and ill-defined HLA-DR and DQ alleles, including DRw10 :. *Transplantation*, 45(3) : 640–646.

[Biedrzycka and Radwan, 2008] Biedrzycka, A. and Radwan, J. (2008). Population fragmentation and major histocompatibility complex variation in the spotted suslik, *Spermophilus suslicus*. *Molecular Ecology*, 17(22) : 4801–4811.

[Bingulac-Popovic et al., 1997] Bingulac-Popovic, J., Figueroa, F., Sato, A., Talbot, W. S., Johnson, S. L., Gates, M., Postlethwait, J. H., and Klein, J. (1997). Mapping of Mhc class I and class II regions to different linkage groups in the zebrafish, Danio rerio. *Immunogenetics*, 46(2) : 129–134.

[Bitarello et al., 2016] Bitarello, B. D., Francisco, R. d. S., and Meyer, D. (2016). Heterogeneity of dN/dS Ratios at the Classical HLA Class I Genes over Divergence Time and Across the Allelic Phylogeny. *Journal of Molecular Evolution*, 82(1) : 38–50.

[Bjorkman et al., 1987a] Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. (1987a). The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature*, 329(6139) : 512–518.

[Bjorkman et al., 1987b] Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. (1987b). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*, 329(6139) : 506–512.

[Blanc et al., 1990] Blanc, M., Sanchez-Mazas, A., Van Blyenburgh, N. H., Sevin, A., Pinson, G., and Langaney, A. (1990). Interethnic Genetic Differentiation : GM Polymorphism in Eastern Senegal. *Am J Hum Genet.b*, 46 : 383–392.

[Blasczyk et al., 2004] Blasczyk, R., Kotsch, K., and Wehling, J. (2004). The Nature of Polymorphism of the HLA Class I Non-Coding Regions and Their Contribution to the Diversification of HLA. *Hereditas*, 127(1-2) : 7–9.

[Boström et al., 2012] Boström, S., Giusti, P., Arama, C., Persson, J.-O., Dara, V., Traore, B., Dolo, A., Doumbo, O., and Troye-Blomberg, M. (2012). Changes in the levels of cytokines, chemokines and malaria-specific antibodies in response to Plasmodium falciparum infection in children living in sympatry in Mali. *Malaria Journal*, 11(1) : 109.

[Bougeard and Dray, 2018] Bougeard, S. and Dray, S. (2018). Supervised Multiblock Analysis in *R* with the **ade4** Package. *Journal of Statistical Software*, 86(1).

[Bourgeon et al., 2017] Bourgeon, L., Burke, A., and Higham, T. (2017). Earliest Human Presence in North America Dated to the Last Glacial Maximum : New Radiocarbon Dates from Bluefish Caves, Canada. *PLOS ONE*, 12(1) : e0169486.

[Bouvier, 2003] Bouvier, M. (2003). Accessory proteins and the assembly of human class I MHC molecules : a molecular and structural perspective. *Molecular Immunology*, 39(12) : 697–706.

[Boyer et al., 2015] Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., and Coissac, E. (2015). OBITOOLS : a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, pages n/a–n/a.

[Brandt et al., 2018] Brandt, D. Y. C., César, J., Goudet, J., and Meyer, D. (2018). The Effect of Balancing Selection on Population Differentiation : A Study with HLA Genes. *G3 ; Genes|Genomes|Genetics*, 8(8) : 2805–2815.

[Brooks et al., 2005] Brooks, N., Chiapello, I., Lernia, S. D., Drake, N., Legrand, M., Moulin, C., and Prospero, J. (2005). The climate-environment-society nexus in the Sahara from prehistoric times to the present day. *The Journal of North African Studies*, 10(3-4) : 253–292.

[Brown et al., 1993] Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., and Wiley, D. C. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*, 364(6432) : 33–39.

[Bruijn, de and Erdös, 1948] Bruijn, de, N. and Erdös, P. (1948). On a combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 51(10) :1277–1279.

[Bugawan et al., 1999] Bugawan, T., Mack, S., Stoneking, M., Saha, M., Beck, H., and Erlich, H. (1999). HLA class I allele distributions in six Pacific/Asian populations : evidence of selection at the HLA-A locus. *Tissue Antigens*, 53(4) : 311–319.

[Buhler, 2007] Buhler, S. (2007). *Étude du polymorphisme moléculaire des gènes HLA de classes I et II à l'échelle mondiale : analyse de la diversité nucléotidique dans les populations*. PhD thesis, Université de Genève.

[Buhler et al., 2016] Buhler, S., Nunes, J. M., and Sanchez-Mazas, A. (2016). HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*, 68(6-7) : 401–416.

[Buhler and Sanchez-Mazas, 2011] Buhler, S. and Sanchez-Mazas, A. (2011). HLA DNA Sequence Variation among Human Populations : Molecular Signatures of Demographic and Selective Events. *PLoS ONE*, 6(2) : e14643.

[Bunce et al., 1997] Bunce, M., Young, N. T., and Welsh, K. I. (1997). Molecular HLA typing-the brave new world. *Transplantation*, 64(11) : 1505–1513.

[Butte and Kohane, 1999] Butte, A. J. and Kohane, I. S. (1999). Mutual information relevance networks : functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429, Honolulu, Hawaii, USA. World Scientific.

[Campbell and Tishkoff, 2008] Campbell, M. C. and Tishkoff, S. A. (2008). African Genetic Diversity : Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual Review of Genomics and Human Genetics*, 9(1) : 403–433.

[Cann et al., 1987] Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099) : 31–36.

[Cao et al., 2004] Cao, K., Moormann, A., Lyke, K., Masaberg, C., Sumba, O., Doumbo, O., Koech, D., Lancaster, A., Nelson, M., Meyer, D., Single, R., Hartzman, R., Plowe, C., Kazura, J., Mann, D., Sztein, M., Thomson, G., and Fernandez-Vina, M. (2004). Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens*, 63(4) : 293–325.

[Ceman et al., 1995] Ceman, S., Rudersdorf, R. A., Petersen, J. M., and DeMars, R. (1995). DMA and DMB are the only genes in the class II region of the human MHC needed for class II-associated antigen processing. *The Journal of Immunology*, 154(6) : 2545.

[Ceppellini et al., 1955] Ceppellini, B. R., Siniscalco, M., and Smith, C. A. B. (1955). The estimation of gene frequencies in a random-mating population. *Annals of Human Genetics*, 20(2) : 97–115.

[Cereb et al., 2015] Cereb, N., Kim, H. R., Ryu, J., and Yang, S. Y. (2015). Advances in DNA sequencing technologies for high resolution HLA typing. *Human Immunology*, 76(12) : 923–927.

[Chaves et al., 2009] Chaves, L. D., Krueth, S. B., and Reed, K. M. (2009). Defining the Turkey MHC : Sequence and Genes of the B Locus. *The Journal of Immunology*, 183(10) : 6530–6537.

[Chen et al., 1992] Chen, Z. W., McAdam, S. N., Hughes, A. L., Dogon, A. L., Letvin, N. L., and Watkins, D. I. (1992). Molecular cloning of orangutan and gibbon MHC class I cDNA. The HLA-A and -B loci diverged over 30 million years ago. *The Journal of Immunology*, 148(8) : 2547.

[Chessel et al., 2004] Chessel, D., Dufour, A.-B., and Thioulouse, J. (2004). The ade4 package – I : One-table methods. *R News*, 4(1) : 5–10.

[Chevallier, 2015] Chevallier, E. (2015). Diversité génétique des loci HLA-A et HLA-B le long de la frange sahélienne : relation avec l'histoire du peuplement et la prévalence de la malaria en Afrique. Master's thesis, Université de Genève, Genève, Suisse.

[Choudhury et al., 2018] Choudhury, A., Aron, S., Sengupta, D., Hazelhurst, S., and Ramsay, M. (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. *Human Molecular Genetics*, 27(R2) : 209–218.

[Chu et al., 2001] Chu, C.-C., Lin, M., Nakajima, F., Lee, H.-L., Chang, S.-L., Juji, T., and Tokunaga, K. (2001). Diversity of HLA among Taiwan's indigenous tribes and the Ivatans in the Philippines. *Tissue Antigens*, 58(1) : 9–18.

[Clark et al., 2001] Clark, M. S., Snell, P., Elgar, G., Kelly, A., and Shaw, L. (2001). Characterization of the MHC class I region of the Japanese pufferfish (Fugu rubripes). *Immunogenetics*, 52(3-4) : 174–185.

[Cockerham, 1969] Cockerham, C. C. (1969). Variance of Gene Frequencies. *Evolution*, 23(1) : 72–84. Publisher : [Society for the Study of Evolution, Wiley].

[Cockerham, 1973] Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics*, 74(4) : 679.

[Coelho et al., 2005] Coelho, M., Luiselli, D., Bertorelle, G., Lopes, A. I., Seixas, S., Destro-Bisol, G., and Rocha, J. (2005). Microsatellite variation and evolution of human lactase persistence. *Human Genetics*, 117(4) : 329–339.

[Cornish-Bowden, 1985] Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences : rcommendations 1984. *Nucleic Acids Research*, 13(9) : 3021–3030.

[Courtet et al., 2001] Courtet, M., Flajnik, M., and Du Pasquier, L. (2001). Major histocompatibility complex and immunoglobulin loci visualized by in situ hybridization on Xenopus chromosomes. *Developmental & Comparative Immunology*, 25(2) : 149–157.

[Cullen et al., 1997] Cullen, M., Noble, J., Erlich, H., Thorpe, K., Beck, S., Klitz, W., Trowsdale, J., and Carrington, M. (1997). Characterization of recombination in the HLA class II region. *American journal of human genetics*, 60(2) : 397–407.

[Currat et al., 2010] Currat, M., Poloni, E. S., and Sanchez-Mazas, A. (2010). Human genetic differentiation across the Strait of Gibraltar. *BMC Evolutionary Biology*, 10(1) : 237.

[Currat et al., 2002] Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R. M., Clegg, J. B., Langaney, A., and Excoffier, L. (2002). Molecular Analysis of the beta-Globin Gene Cluster in the Niokholo Mandenka Population Reveals a Recent Origin of the beta S Senegal Mutation. *The American Journal of Human Genetics*, 70(1) : 207–223.

[Curtoni et al., 1967] Curtoni, E., Mattiuz, P., and Tossi, R. (1967). Histocompatibility Testing 1967. In -, Torino (IT). Copenhagen : Munksgaard.

[Cáceres and Cáceres-Saez, 2011] Cáceres, M. O. and Cáceres-Saez, I. (2011). Random Leslie matrices in population dynamics. *Journal of Mathematical Biology*, 63(3) : 519–556.

[Dard et al., 1997] Dard, P., Huck, S., Frippiat, J. P., Lefranc, G., Langaney, A., Lefranc, M. P., and Sanchez-Mazas, A. (1997). The IGHG3 gene shows a structural polymorphism characterized by different hinge lengths : sequence of a new 2-exon hinge gene. *Human Genetics*, 99(1) : 138–141.

[Dard et al., 1996] Dard, P., Sanchez-Mazas, A., Dugoujon, J. M., De Lange, G., Langaney, A., Lefranc, M. P., and Lefranc, G. (1996). DNA analysis of the immunoglobulin IGHG loci in a Mandenka population from eastern Senegal : correlation with Gm haplotypes and hypotheses for the evolution of the Ig CH region. *Human Genetics*, 98(1) : 36–47.

[Dard et al., 1992] Dard, P., Schreiber, Y., Excoffier, L., Sanchez-Mazas, A., Shi-Isaac, X., Epelbouin, A., Langaney, A., and Jeannet, M. (1992). [Polymorphism of HLA class I loci HLA-A, -B, -C, in the Mandenka population from eastern Senegal]. *Comptes rendus de l'Academie des sciences. Serie III, Sciences de la vie*, 314(13) : 573–578.

[Dausset, 1958] Dausset, J. (1958). Iso-leuco-anticorps. *Acta Haematologica*, 20(1-4) : 156–166.

[Dausset, 1984] Dausset, J. (1984). The Birth of MAC. *Vox Sanguinis*, 46(4) : 235–237.

[de Groot et al., 2000] de Groot, N. G., Otting, N., Argüello, R., Watkins, D. I., Doxiadis, G. G. M., Madrigal, J. A., and Bontrop, R. E. (2000). Major histocompatibility complex class I diversity in a West African chimpanzee population : implications for HIV research. *Immunogenetics*, 51(6) : 398–409.

[de Groot et al., 2018] de Groot, N. G., Stevens, J. M., and Bontrop, R. E. (2018). Does the MHC Confer Protection against Malaria in Bonobos ? *Trends in Immunology*, 39(10) : 768–771.

[De Santis et al., 2013] De Santis, D., Dinauer, D., Duke, J., Erlich, H. A., Holcomb, C. L., Lind, C., Mackiewicz, K., Monos, D., Moudgil, A., Norman, P., Parham, P., Sasson, A., and Allcock, R. J. N. (2013). 16th IHIW : Review of HLA typing by NGS. *International Journal of Immunogenetics*, 40(1) : 72–76.

[Delfino et al., 2003] Delfino, L., Morabito, A., and Ferrara, G. (2003). HLA-C sequence based typing : nucleotide analysis from exon 1 through exon 8. Identification of a new allele : Cw*0718. *Tissue Antigens*, 62(5) : 418–425.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) : 1–22.

[Deshpande et al., 2009] Deshpande, O., Batzoglou, S., Feldman, M. W., and Luca Cavalli-Sforza, L. (2009). A serial founder effect model for human settlement out of Africa. *Proceedings of the Royal Society B : Biological Sciences*, 276(1655) : 291–300.

[Dey et al., 1992] Dey, A., Thornton, A. M., Lonergan, M., Weissman, S. M., Chamberlain, J. W., and Ozato, K. (1992). Occupancy of upstream regulatory sites in vivo coincides with major histocompatibility complex class I gene expression in mouse tissues. *Molecular and Cellular Biology*, 12(8) : 3590–3599.

[Di and Sanchez-Mazas, 2014] Di, D. and Sanchez-Mazas, A. (2014). HLA variation reveals genetic continuity rather than population group structure in East Asia. *Immunogenetics*, 66(3) : 153–160.

[Di et al., 2015] Di, D., Sanchez-Mazas, A., and Currat, M. (2015). Computer simulation of human leukocyte antigen genes supports two main routes of colonization by human populations in East Asia. *BMC Evolutionary Biology*, 15(1) : 240.

[Dilthey et al., 2013] Dilthey, A., Leslie, S., Moutsianas, L., Shen, J., Cox, C., Nelson, M. R., and McVean, G. (2013). Multi-Population Classical HLA Type Imputation. *PLoS Computational Biology*, 9(2) : e1002877.

[Do and Choi, 2006] Do, J. H. and Choi, D. (2006). Computational Approaches to Gene Prediction. *Journal of Microbiology*, 44(2) : 137–144.

[Doherty and Zinkernagel, 1975] Doherty, P. C. and Zinkernagel, R. M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, 256(5512) : 50–52.

[dos Santos Francisco et al., 2015] dos Santos Francisco, R., Buhler, S., Nunes, J. M., Bitarello, B. D., França, G. S., Meyer, D., and Sanchez-Mazas, A. (2015). HLA supertype variation across populations : new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. *Immunogenetics*, 67(11-12) : 651–663.

[Doxiadis et al., 2008a] Doxiadis, G. G., de Groot, N., de Groot, N. G., Doxiadis, I. I., and Bontrop, R. E. (2008a). Reshuffling of ancient peptide binding motifs between HLA-DRB multigene family members : Old wine served in new skins. *Molecular Immunology*, 45(10) : 2743–2751.

[Doxiadis et al., 2012] Doxiadis, G. G., Hoof, I., de Groot, N., and Bontrop, R. E. (2012). Evolution of HLA-DRB Genes. *Molecular Biology and Evolution*, 29(12) : 3843–3853.

[Doxiadis et al., 2008b] Doxiadis, G. G. M., de Groot, N., and Bontrop, R. E. (2008b). Impact of Endogenous Intronic Retroviruses on Major Histocompatibility Complex Class II Diversity and Stability. *Journal of Virology*, 82(13) : 6667–6677.

[Dray and Dufour, 2007] Dray, S. and Dufour, A.-B. (2007). The **ade4** Package : Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22(4).

[Dray et al., 2007] Dray, S., Dufour, A.-B., and Chessel, D. (2007). The ade4 package – II : Two-table and K-table methods. *R News*, 7(2) : 47–52.

[Durbin, 1998] Durbin, R., editor (1998). *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK : New York.

[D'Atanasio et al., 2018] D'Atanasio, E., Trombetta, B., Bonito, M., Finocchio, A., Di Vito, G., Seghizzi, M., Romano, R., Russo, G., Paganotti, G. M., Watson, E., Coppa, A., Anagnostou, P., Dugoujon, J.-M., Moral, P., Sellitto, D., Novelletto, A., and Cruciani, F. (2018). The peopling of the last Green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages. *Genome Biology*, 19(1) : 20.

[Eberhard et al., 2019] Eberhard, D. M., Gary, F. S., and Charles, D. F. (2019). *Ethnologue : Languages of the World*. SIL International, Dallas, Texas, 22 edition.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5 : 113.

[Ejsmond et al., 2010] Ejsmond, M., Babik, W., and Radwan, J. (2010). MHC allele frequency distributions under parasite-driven selection : A simulation model. *BMC Evolutionary Biology*, 10(1) : 332.

[El Mousadik and Petit, 1996] El Mousadik, A. and Petit, R. J. (1996). High level of genetic differentiation for allelic richness among populations of the argan tree [Argania spinosa (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics*, 92(7) : 832–839.

[Eliaou et al., 1989] Eliaou, J.-F., Humbert, M., Balaguer, P., Gebuhrer, L., Amsellem, S., Bétuel, H., Nicolas, J.-C., and Clot, J. (1989). A method of HLA class II typing using non-radioactive labelled oligonucleotides. *Tissue Antigens*, 33(4) : 475–485.

[Enattah et al., 2008] Enattah, N. S., Jensen, T. G., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J. K., Alifrangis, M., Khalil, I. F., Natah, A., Ali, A., Natah, S., Comas, D., Mehdi, S. Q., Groop, L., Vestergaard, E. M., Imtiaz,

F., Rashed, M. S., Meyer, B., Troelsen, J., and Peltonen, L. (2008). Independent Introduction of Two Lactase-Persistence Alleles into Human Populations Reflects Different History of Adaptation to Milk Culture. *The American Journal of Human Genetics*, 82(1) : 57–72.

[Erlich et al., 2011] Erlich, R. L., Jia, X., Anderson, S., Banks, E., Gao, X., Carrington, M., Gupta, N., DePristo, M. A., Henn, M. R., Lennon, N. J., and de Bakker, P. I. (2011). Next-generation sequencing for HLA typing of class I loci. *BMC Genomics*, 12(1) : 42.

[Erlich et al., 2018] Erlich, Y., Shor, T., Pe'er, I., and Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. *Science*, 362(6415) : 690–694.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, pages 226–231.

[Evans and Neff, 2009] Evans, M. L. and Neff, B. D. (2009). Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology*, 18(22) : 4716–4729.

[Excoffier and Lischer, 2010] Excoffier, L. and Lischer, H. E. L. (2010). Arlequin suite ver 3.5 : a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3) : 564–567.

[Excoffier and Schneider, 1999] Excoffier, L. and Schneider, S. (1999). Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proceedings of the National Academy of Sciences*, 96(19) : 10597–10602.

[Excoffier and Slatkin, 1995] Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*.

[Excoffier et al., 1992] Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes : Application to Human Mitochondrial DNA Restriction Data. *Genetics*, 131 : 479–491.

[Fernandez-Viña et al., 1992] Fernandez-Viña, M. A., Falco, M., Sun, Y., and Stastny, P. (1992). DNA typing for HLA class I alleles : I. Subsets of HLA-A2 and of -A28. *Human Immunology*, 33(3) : 163–173.

[Fernando et al., 2008] Fernando, M. M. A., Stevens, C. R., Walsh, E. C., De Jager, P. L., Goyette, P., Plenge, R. M., Vyse, T. J., and Rioux, J. D. (2008). Defining the Role of the MHC in Autoimmunity : A Review and Pooled Analysis. *PLoS Genetics*, 4(4) : e1000024.

[Field et al., 2016] Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., and Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science*, 354(6313) : 760–764.

[Flajnik et al., 1993] Flajnik, M., Kasahara, M., Shum, B., Salter-Cid, L., Taylor, E., and Du Pasquier, L. (1993). A novel type of class I gene organization in vertebrates : a large family of non-MHC-linked class I genes is expressed at the RNA level in the amphibian Xenopus. *The EMBO Journal*, 12(11) : 4385–4396.

[Flajnik et al., 1999] Flajnik, M., Ohta, Y., Namikawa-Yomada, C., and Nonaka, M. (1999). Insight into the primordial MHC from studies in ectothermic vertebrates. *Immunological Reviews*, 167(1) : 59–67.

[Flajnik, 2018] Flajnik, M. F. (2018). A cold-blooded view of adaptive immunity. *Nature Reviews Immunology*, 18(7) : 438–453.

[Flajnik et al., 1991] Flajnik, M. F., Canel, C., Kramer, J., and Kasahara, M. (1991). Which came first, MHC class I or class II ? *Immunogenetics*, 33(5-6) : 295–300.

[Flajnik et al., 1987] Flajnik, M. F., Hsu, E., Kaufman, J. F., and Pasquier, L. D. (1987). Changes in the immune system during metamorphosis of Xenopus. *Immunology Today*, 8(2) : 58–64.

[Flajnik and Kasahara, 2001] Flajnik, M. F. and Kasahara, M. (2001). Comparative Genomics of the MHC. *Immunity*, 15(3) : 351–362.

[Fogel, 2003] Fogel, F. (2003). Figures nubiennes de l'ethnie, de la minorité, de la nation (Égypte, Soudan)1. *Ateliers d'anthropologie*, 26.

[Fort et al., 1998] Fort, M., Stefano, G.-F., Cambon-Thomsen, A., Giraldo-Alvarez, P., Dugoujon, J.-M., Ohayon, E., Scano, G., and Abbal, M. (1998). HLA class II allele and haplotype frequencies in Ethiopian Amhara and Oromo populations. *Tissue Antigens*, 51(4) : 327–336.

[Foundation, 2020] Foundation, F. S. (2020). *GNU Grep 3.4*.

[Francke and Pellegrino, 1977] Francke, U. and Pellegrino, M. A. (1977). Assignment of the major histocompatibility complex to a region of the short arm of human chromosome 6. *Proceedings of the National Academy of Sciences*, 74(3) : 1147–1151.

[Gabriel et al., 2009] Gabriel, C., Danzer, M., Hackl, C., Kopal, G., Hufnagl, P., Hofer, K., Polin, H., Stabentheiner, S., and Pröll, J. (2009). Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. *Human Immunology*, 70(11) : 960–964.

[Galan et al., 2010] Galan, M., Guivier, E., Caraux, G., Charbonnel, N., and Cosson, J.-F. (2010). A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, 11(1) : 296.

[Garrigan and Hedrick, 2001] Garrigan, D. and Hedrick, P. (2001). Class I MHC polymorphism and evolution in endangered California Chinook and other Pacific salmon. *Immunogenetics*, 53(6) : 483–489.

[Garrigan et al., 2007] Garrigan, D., Kingan, S. B., Pilkington, M. M., Wilder, J. A., Cox, M. P., Soodyall, H., Strassmann, B., Destro-Bisol, G., de Knijff, P., Novelletto, A., Friedlaender, J., and Hammer, M. F. (2007). Inferring Human Population Sizes, Divergence Times and Rates of Gene Flow From Mitochondrial, X and Y Chromosome Resequencing Data. *Genetics*, 177(4) : 2195–2207.

[Gerbault et al., 2009] Gerbault, P., Moret, C., Currat, M., and Sanchez-Mazas, A. (2009). Impact of Selection and Demography on the Diffusion of Lactase Persistence. *PLoS ONE*, 4(7) : e6369.

[Glazko, 2003] Glazko, G. V. (2003). Estimation of Divergence Times for Major Lineages of Primate Species. *Molecular Biology and Evolution*, 20(3) : 424–434.

[Goebel et al., 2008] Goebel, T., Waters, M. R., and O'Rourke, D. H. (2008). The Late Pleistocene Dispersal of Modern Humans in the Americas. *Science*, 319(5869) : 1497–1502.

[Goecks et al., 2010] Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy Team, T. (2010). Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8) : R86.

[Goeury et al., 2018a] Goeury, T., Creary, L. E., Brunet, L., Galan, M., Pasquier, M., Kervaire, B., Langaney, A., Tiercy, J.-M., Fernández-Viña, M. A., Nunes, J. M., and Sanchez-Mazas, A. (2018a). Deciphering the fine nucleotide diversity of full HLA class

I and class II genes in a well-documented population from sub-Saharan Africa. *HLA*, 91(1) : 36–51.

[Goeury et al., 2018b] Goeury, T., Creary, L. E., Fernandez-Vina, M. A., Tiercy, J.-M., Nunes, J. M., and Sanchez-Mazas, A. (2018b). Mandenka from Senegal : Next Generation Sequencing typings reveal very high frequencies of particular HLA class II alleles and haplotypes. *HLA*, 91(2) : 148–150.

[Goldman and Yang, 1994] Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5) : 725–736.

[Goodfellow et al., 1975] Goodfellow, P. N., Jones, E. A., Van Heyningen, V., Solomon, E., Bobrow, M., Miggiano, V., and Bodmer, W. F. (1975). The b2-microglobulin gene is on chromosome 15 and not in the HL-A region. *Nature*, 254(5497) : 267–269.

[Gorer, 1937] Gorer, P. (1937). The genetic and antigenic basis of tumour transplantation. *Journal of Pathology and Bacteriology*, 44 : 691–697.

[Grahovac et al., 1993] Grahovac, B., Schönbach, C., Brändle, U., Mayer, W. E., Golubic, M., Figueroa, F., Trowsdale, J., and Klein, J. (1993). Conservative evolution of the Mbc-DP region in anthropoid primates. *Human Immunology*, 37(2) : 75–84.

[Graven et al., 1995] Graven, L., Passarino, G., Semino, O., Boursot, P., Santachiara-Benerecetti, S., Langaney, A., and Excoffier, L. (1995). Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Molecular Biology and Evolution*, 12(2) : 334–345.

[Greenacre, 1984] Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press, London ; Orlando, Fla.

[Grogan et al., 2016] Grogan, K. E., McGinnis, G. J., Sauther, M. L., Cuozzo, F. P., and Drea, C. M. (2016). Next-generation genotyping of hypervariable loci in many individuals of a non-model species : technical and theoretical implications. *BMC Genomics*, 17(1).

[Gruen and Weissman, 1997] Gruen, J. R. and Weissman, S. M. (1997). Evolving view of the major histocompatibility complex. *Blood*, 90 : 4252–4265.

[Grundschober et al., 1997] Grundschober, C., Rufer, N., Sanchez-Mazas, A., Madrigal, A., Jeannet, M., Roosnek, E., and Tiercy, J.-M. (1997). Molecular characterization of HLA-C incompatibilities in HLA-ABDR-matched unrelated bone marrow donor-recipient pairs. *Tissue Antigens*, 49(6) : 612–623.

[Gunther et al., 2010] Gunther, S., Schlundt, A., Sticht, J., Roske, Y., Heinemann, U., Wiesmuller, K.-H., Jung, G., Falk, K., Rotzschke, O., and Freund, C. (2010). Bidirectional binding of invariant chain peptides to an MHC class II molecule. *Proceedings of the National Academy of Sciences*, 107(51) : 22219–22224.

[Gurdasani et al., 2015] Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., Doumatey, A. P., Asiki, G., Seeley, J., Sisay-Joof, F., Jallow, M., Tollman, S., Mekonnen, E., Ekong, R., Oljira, T., Bradman, N., Bojang, K., Ramsay, M., Adeyemo, A., Bekele, E., Motala, A., Norris, S. A., Pirie, F., Kaleebu, P., Kwiatkowski, D., Tyler-Smith, C., Rotimi, C., Zeggini, E., and Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534) : 327–332.

[Gyllensten et al., 1991] Gyllensten, U. B., Sundvall, M., and Erlich, H. A. (1991). Allelic diversity is generated by intraexon sequence exchange at the DRB1 locus of primates. *Proceedings of the National Academy of Sciences*, 88(9) : 3686–3690.

[Güldemann and Stoneking, 2008] Güldemann, T. and Stoneking, M. (2008). A Historical Appraisal of Clicks : A Linguistic and Genetic Population Perspective. *Annual Review of Anthropology*, 37(1) : 93–109.

[Günther and Walter, 2001] Günther, E. and Walter, L. (2001). The major histocompatibility complex of the rat (Rattus norvegicus). *Immunogenetics*, 53(7) : 520–542.

[Hajjej et al., 2015] Hajjej, A., Almawi, W. Y., Hattab, L., El-Gaaied, A., and Hmida, S. (2015). HLA Class I and Class II Alleles and Haplotypes Confirm the Berber Origin of the Present Day Tunisian Population. *PLOS ONE*, 10(8) : e0136909.

[Hansen, 2006] Hansen, J. A. (2006). *Immunobiology of the human MHC : proceedings of the 13th International Histocompatibility Workshop and Conference and the XIII International Congress of Histocompatibility and Immunogenetics Vol. 1, Vol. 1,*. P.J. Schmidt A/S ; IHWG Press, Vojens ; [s.l. OCLC : 794547846.

[Harich et al., 2010] Harich, N., Costa, M. D., Fernandes, V., Kandil, M., Pereira, J. B., Silva, N. M., and Pereira, L. (2010). The trans-Saharan slave trade - clues from interpolation analyses and high-resolution characterization of mitochondrial DNA lineages. *BMC Evolutionary Biology*, 10(1) : 138.

[Hashimoto et al., 1990] Hashimoto, K., Nakanishi, T., and Kurosawa, Y. (1990). Isolation of carp genes encoding major histocompatibility complex antigens. *Proceedings of the National Academy of Sciences*, 87(17) : 6863–6867.

[Hausser and Strimmer, 2009] Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, 10 : 1469–1484.

[Hawley and Kidd, 1995] Hawley, M. E. and Kidd, K. K. (1995). HAPLO : A Program Using the EM Algorithm to Estimate the Frequencies of Multi-site Haplotypes. *Journal of Heredity*, 86(5) : 409–411.

[Hayashi, 1992] Hayashi, K. (1992). PCR-SSCP : A method for detection of mutations. *Genetic Analysis : Biomolecular Engineering*, 9(3) : 73–79.

[He et al., 2012] He, J.-D., Peng, M.-S., Quang, H. H., Dang, K. P., Trieu, A. V., Wu, S.-F., Jin, J.-Q., Murphy, R. W., Yao, Y.-G., and Zhang, Y.-P. (2012). Patrilineal Perspective on the Austronesian Diffusion in Mainland Southeast Asia. *PLoS ONE*, 7(5) : e36437.

[Hedrick, 1998] Hedrick, P. W. (1998). Balancing selection and MHC. *Genetica*, 104(3) : 207–214.

[Hedrick, 2002] Hedrick, P. W. (2002). Pathogen resistance and genetic variation at MHC loci. *Evolution*, 56(10) : 1902–1908.

[Hedrick et al., 2000] Hedrick, P. W., Parker, K. M., Gutiérrez-Espeleta, G. A., Rattink, A., and Lievers, K. (2000). Major histocompatibility complex variation in the arabian oryx. *Evolution*, 54(6) : 2145–2151.

[Heled and Drummond, 2008] Heled, J. and Drummond, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8(1) : 289.

[Higham, 2002] Higham, C. (2002). *Early cultures of Mainland Southeast Asia.* River Books, Bangkok (Thailand).

[Hijmans, 2019a] Hijmans, R. J. (2019a). *geosphere : Spherical Trigonometry*. R package version 1.5-10.

[Hijmans, 2019b] Hijmans, R. J. (2019b). *raster : Geographic Data Analysis and Modeling*. R package version 2.9-5.

[Hill et al., 1992a] Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Taylor, T. E., Yates, S. N., Anstey, N. M., Wirima, J. J., Brewster, D. R., McMichael, A. J., and Molyneux, M. E. (1992a). Extensive genetic diversity in the HLA class II region of Africans, with a focally predominant allele, DRB1*1304. *Proceedings of the National Academy of Sciences*, 89(6) : 2277–2281.

[Hill et al., 1991] Hill, A. V. S., Allsopp, C. E. M., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., and Greenwood, B. M. (1991). Common West African HLA antigens are associated with protection from severe malaria. *Nature*, 352(6336) : 595–600.

[Hill et al., 1992b] Hill, A. V. S., Elvin, J., Willis, A. C., Aidoo, M., Allsopp, C. E. M., Gotch, F. M., Ming Gao, X., Takiguchis, M., Greenwood, B. M., Townsend, A. R. M., McMichael, A. J., and Whittle, H. C. (1992b). Molecular analysis of the association of HLA-B53 and resistance to severe malaria. *Nature*, 360(6403) : 434–439.

[Hilton et al., 2015] Hilton, H. G., Guethlein, L. A., Goyos, A., Nemat-Gorgani, N., Bushnell, D. A., Norman, P. J., and Parham, P. (2015). Polymorphic HLA-C Receptors Balance the Functional Characteristics of *KIR* Haplotypes. *The Journal of Immunology*, 195(7) : 3160–3170.

[Hinton and Roweis, 2002] Hinton, G. and Roweis, S. (2002). Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 857–864, Cambridge, MA, USA. MIT Press.

[Hoelzel et al., 1999] Hoelzel, A. R., Stephens, J. C., and O'Brien, S. J. (1999). Molecular genetic diversity and evolution at the MHC DQB locus in four species of pinnipeds. *Molecular Biology and Evolution*, 16(5) : 611–618.

[Hohjoh et al., 2001] Hohjoh, H., Tokunaga, K., Chida, S., and Hirai, M. (2001). Haplotype-specific sequence encoding the protein kinase, interferon-inducible double-stranded RNA-dependent activator in the human leukocyte antigen classII region. *Immunogenetics*, 52(3-4) : 186–194.

[Holcomb et al., 2011] Holcomb, C. L., Höglund, B., Anderson, M. W., Blake, L. A., Böhme, I., Egholm, M., Ferriola, D., Gabriel, C., Gelber, S. E., Goodridge, D., Hawbecker, S., Klein, R., Ladner, M., Lind, C., Monos, D., Pando, M. J., Pröll, J., Sayer, D. C., Schmitz-Agheguian, G., Simen, B. B., Thiele, B., Trachtenberg, E. A., Tyan, D. B., Wassmuth, R., White, S., and Erlich, H. A. (2011). A multi-site study using high-resolution HLA genotyping by next generation sequencing. *Tissue Antigens*, 77(3) : 206–217.

[Holm, 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2) : 65–70.

[Holmes et al., 2013] Holmes, J. C., Holmer, S. G., Ross, P., Buntzman, A. S., Frelinger, J. A., and Hess, P. R. (2013). Polymorphisms and tissue expression of the feline leukocyte antigen class I loci FLAI-E, FLAI-H, and FLAI-K. *Immunogenetics*, 65(9) : 675–689.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6) : 417–441.

[Hublin et al., 2017] Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., and Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, 546(7657) : 289–292.

[Huchard et al., 2010] Huchard, E., Knapp, L. A., Wang, J., Raymond, M., and Cowlishaw, G. (2010). MHC, mate choice and heterozygote advantage in a wild social primate : MHC and reproductive strategies in baboons. *Molecular Ecology*, pages no–no.

[Hughes and Nei, 1988] Hughes, A. L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186) : 167–170.

[Hughes and Nei, 1989a] Hughes, A. L. and Nei, M. (1989a). Ancient interlocus exon exchange in the history of the HLA-A locus. *Genetics*, 122(3) : 681.

[Hughes and Nei, 1989b] Hughes, A. L. and Nei, M. (1989b). Evolution of the major histocompatibility complex : independent origin of nonclassical class I genes in different groups of mammals. *Molecular Biology and Evolution*.

[Hughes and Nei, 1989c] Hughes, A. L. and Nei, M. (1989c). Nucleotide substitution at major histocompatibility complex class II loci : evidence for overdominant selection. *Proceedings of the National Academy of Sciences*, 86(3) : 958–962.

[Hughes and Nei, 1990] Hughes, A. L. and Nei, M. (1990). Evolutionary relationships of class II major-histocompatibility-complex genes in mammals. *Molecular Biology and Evolution*.

[Hughes and Nei, 1992] Hughes, A. L. and Nei, M. (1992). Maintenance of MHC polymorphism. *Nature*, 355(6359) : 402–403.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib : A 2d graphics environment. *Computing in Science & Engineering*, 9(3) :90–95.

[Hurlbert, 1971] Hurlbert, S. H. (1971). The Nonconcept of Species Diversity : A Critique and Alternative Parameters. *Ecology*, 52(4) : 577–586.

[IHWG, 1965] IHWG (1965). Histocompatibility Testing 1965. In -, Leiden (NL). Copenhagen : Munksgaard.

[Inotai et al., 2015] Inotai, D., Szilvasi, A., Benko, S., Boros-Major, A., Illes, Z., Bors, A., Kiss, K. P., Rajczy, K., Gelle-Hossó, A., Buhler, S., Nunes, J. M., Sanchez-Mazas, A., and Tordai, A. (2015). HLA genetic diversity in Hungarians and Hungarian Gypsies : complementary differentiation patterns and demographic signals revealed by HLA-A, -B and -DRB1 in Central Europe : HLA genetic diversity in Hungary. *Tissue Antigens*, 86(2) : 115–121.

[Ioannidu et al., 2001] Ioannidu, S., Walter, L., Dressel, R., and Günther, E. (2001). Physical Map and Expression Profile of Genes of the Telomeric Class I Gene Region of the Rat MHC. *The Journal of Immunology*, 166(6) : 3957–3965.

[IUPAC-IUB, 1970] IUPAC-IUB (1970). Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents. *Biochemistry*, 9(20) : 4022–4027.

[Jaratlerdsiri et al., 2014] Jaratlerdsiri, W., Deakin, J., Godinez, R. M., Shan, X., Peterson, D. G., Marthey, S., Lyons, E., McCarthy, F. M., Isberg, S. R., Higgins, D. P., Chong, A. Y., John, J. S., Glenn, T. C., Ray, D. A., and Gongora, J. (2014). Comparative Genome Analyses Reveal Distinct Structure in the Saltwater Crocodile MHC. *PLoS ONE*, 9(12) : e114631.

[Jeffreys et al., 2001] Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2) : 217–222.

[Joshi et al., 2006] Joshi, P. K., Lele, N., and Agarwal, S. P. (2006). Entropy as an indicator of fragmented landscape. *Current Science*, 91(3) : 276–278. Publisher : Temporary Publisher.

[Juang and Rabiner, 1991] Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov Models for Speech Recognition. *Technometrics*, 33(3) : 251–272.

[Kasahara et al., 1995] Kasahara, M., Flajnik, M. F., Ishibashi, T., and Natori, T. (1995). Evolution of the major histocompatibility complex : a current overview. *Transplant Immunology*, 3(1) : 1–20.

[Katoh, 2002] Katoh, K. (2002). MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14) : 3059–3066.

[Kaufman, 1988] Kaufman, J. (1988). Vertebrates and the evolution of the major histocompatibility complex (MHC) class I and class II molecules. In *Verh. Dtsch. Zool. Ges*, volume 81 of *Verhandlungen der Deutschen Zoologischen Gesellschaft*, pages 131–144. Fischer, New York, NY. Journal Abbreviation : Verh. Dtsch. Zool. Ges.

[Kaufman et al., 1999] Kaufman, J., Milne, S., Göbel, T. W. F., Walker, B. A., Jacob, J. P., Auffray, C., Zoorob, R., and Beck, S. (1999). The chicken B locus is a minimal essential major histocompatibility complex. *Nature*, 401(6756) : 923–925.

[Kaur et al., 2017] Kaur, G., Gras, S., Mobbs, J. I., Vivian, J. P., Cortes, A., Barber, T., Kuttikkatte, S. B., Jensen, L. T., Attfield, K. E., Dendrou, C. A., Carrington, M., McVean, G., Purcell, A. W., Rossjohn, J., and Fugger, L. (2017). Structural and regulatory diversity shape HLA-C protein expression levels. *Nature Communications*, 8(1) : 15924.

[Kelley et al., 2005] Kelley, J., Walter, L., and Trowsdale, J. (2005). Comparative genomics of major histocompatibility complexes. *Immunogenetics*, 56(10) : 683–695.

[Khakoo et al., 2000] Khakoo, S. I., Rajalingam, R., Shum, B. P., Weidenbach, K., Flodin, L., Muir, D. G., Canavez, F., Cooper, S. L., Valiante, N. M., Lanier, L. L., and Parham, P. (2000). Rapid Evolution of NK Cell Receptor Systems Demonstrated by Comparison of Chimpanzees and Humans. *Immunity*, 12(6) : 687–698.

[Kimura, 1968] Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, 217(5129) : 624–626.

[Kimura, 1977] Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608) : 275–276.

[Kimura, 1991] Kimura, M. (1991). The neutral theory of molecular evolution : A review of recent evidence. *The Japanese Journal of Genetics*, 66(4) : 367–386.

[Klein and Sato, 1998] Klein and Sato (1998). Birth of the Major Histocompatibility Complex. *Scandinavian Journal of Immunology*, 47(3) : 199–209.

[Klein and Figueroa, 1986] Klein, J. and Figueroa, F. (1986). Evolution of the major histocompatibility complex. Critical Reviews in Immunology. *Critical Reviews in Immunology*, 6(4) : 295–386.

[Klein et al., 1993a] Klein, J., O'hUigin, C., Figueroa, F., Mayer, W. E., and Klein, D. (1993a). Different modes of Mhc evolution in primates. *Molecular Biology and Evolution*, 10(1) : 48–59.

[Klein et al., 1993b] Klein, J., Ono, H., Klein, D., and O'hUigin, C. (1993b). The Accordion Model of Mhc Evolution. In *Progress in Immunology Vol. VIII*, pages 137–143. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Klein et al., 1993c] Klein, J., Satta, Y., O'hUigin, C., and Takahata, N. (1993c). The Molecular Descent of the Major Histocompatibility Complex. *Annual Review of Immunology*, 11(1) : 269–295.

[Klein et al., 1993d] Klein, J., Satta, Y., Takahata, N., and O'hUigin, C. (1993d). Trans-specific Mhc polymorphism and the origin of species in primates. *Journal of medical primatology*, 22(1) : 57–64. Place : Denmark.

[Komatsu-Wakui et al., 1999] Komatsu-Wakui, M., Tokunaga, K., Ishikawa, Y., Kashiwase, K., Moriyama, S., Tsuchiya, N., Ando, H., Shiina, T., Geraghty, D. E., Inoko, H., and Juji, T. (1999). MIC-A polymorphism in Japanese and a MIC-A-MIC-B null haplotype. *Immunogenetics*, 49(7-8) : 620–628.

[Kourilsky, 1983] Kourilsky, P. (1983). Genetic exchanges between partially homologous nucleotide sequences : possible implications for multigene families. *Biochimie*, 65(2) : 85–93.

[Krijthe, 2015] Krijthe, J. H. (2015). *Rtsne : T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.15.

[Kruskal, 1964a] Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1) : 1–27.

[Kruskal, 1964b] Kruskal, J. B. (1964b). Nonmetric multidimensional scaling : A numerical method. *Psychometrika*, 29(2) : 115–129.

[Kryazhimskiy and Plotkin, 2008] Kryazhimskiy, S. and Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS genetics*, 4(12) : e1000304–e1000304.

[Kuba et al., 2003] Kuba, R., Lentz, C., Somda, C. N., am Main, U. F., and de Ouagadougou, U., editors (2003). *Histoire du peuplement et relations interethniques au Burkina Faso*. Hommes et sociétés. Karthala, Paris.

[Kulichová et al., 2017] Kulichová, I., Fernandes, V., Deme, A., Nováčková, J., Stenzl, V., Novelletto, A., Pereira, L., and Černý, V. (2017). Internal diversification of non-Sub-Saharan haplogroups in Sahelian populations and the spread of pastoralism beyond the Sahara. *American Journal of Physical Anthropology*, 164(2) : 424–434.

[Kulski et al., 2000] Kulski, J. K., Gaudieri, S., and Dawkins, R. L. (2000). Transposable elements and the metamerismatic evolution of the HLA class I region. In Kasahara, M., editor, *Major Histocompatibility Complex*, pages 158–177. Springer Japan, Tokyo.

[Kulski et al., 1999] Kulski, J. K., Gaudieri, S., Martin, A., and Dawkins, R. L. (1999). Coevolution of PERB11 (MIC) and HLA Class I Genes with HERV-16 and Retroelements by Extended Genomic Duplication. *Journal of Molecular Evolution*, 49(1) : 84–97.

[Kulski et al., 2002] Kulski, J. K., Shiina, T., Anzai, T., Kohara, S., and Inoko, H. (2002). Comparative genomic analysis of the MHC : the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunological Reviews*, 190(1) : 95–122.

[Kumar et al., 2016] Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7 : Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33(7) : 1870–1874.

[Kumánovics et al., 2002] Kumánovics, A., Madan, A., Qin, S., Rowen, L., Hood, L., and Fischer Lindahl, K. (2002). Quod erat faciendum : sequence analysis of the H2-D and H2-Q regions of 129/SvJ mice. *Immunogenetics*, 54(7) : 479–489.

[Kurtz et al., 2004] Kurtz, J., Kalbe, M., Aeschlimann, P. B., Häberli, M. A., Wegner, K. M., Reusch, T. B. H., and Milinski, M. (2004). Major histocompatibility complex diversity influences parasite resistance and innate immunity in sticklebacks. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 271(1535) : 197–204.

[Laird et al., 2000] Laird, D. J., De Tomaso, A. W., Cooper, M. D., and Weissman, I. L. (2000). 50 million years of chordate evolution : Seeking the origins of adaptive immunity. *Proceedings of the National Academy of Sciences*, 97(13) : 6924–6926.

[Lalueza-Fox et al., 2005] Lalueza-Fox, C., Castresana, J., Sampietro, L., Marquès-Bonet, T., Alcover, J. A., and Bertranpetit, J. (2005). Molecular dating of caprines using ancient DNA sequences of Myotragus balearicus, an extinct endemic Balearic mammal. *BMC Evolutionary Biology*, 5(1) : 70.

[Le Than, 2007] Le Than, L. (2007). Exclusion d'apparentés par ré-échantillonnage d'un échantillon de population : application à l'estimation de la diversité génétique chez les Mandenka du Sénégal oriental. Master's thesis, Université de Genève, Genève, Suisse.

[Lee et al., 1990] Lee, K. W., Hurley, C. K., Hartzman, R., and Johnson, A. H. (1990). The Complexity of DRw6 and DR5 Haplotypes in American Blacks Demonstrated by Serology, Cellular Typing, and Restriction Fragment Length Polymorphism Analysis. *Human Immunology*, 29 : 202–219.

[Leffler et al., 2017] Leffler, E. M., Band, G., Busby, G. B. J., Kivinen, K., Le, Q. S., Clarke, G. M., Bojang, K. A., Conway, D. J., Jallow, M., Sisay-Joof, F., Bougouma, E. C., Mangano, V. D., Modiano, D., Sirima, S. B., Achidi, E., Apinjoh, T. O., Marsh, K., Ndila, C. M., Peshu, N., Williams, T. N., Drakeley, C., Manjurano, A., Reyburn, H., Riley, E., Kachala, D., Molyneux, M., Nyirongo, V., Taylor, T., Thornton, N., Tilley, L., Grimsley, S., Drury, E., Stalker, J., Cornelius, V., Hubbart, C., Jeffreys, A. E., Rowlands, K., Rockett, K. A., Spencer, C. C. A., Kwiatkowski, D. P., and Malaria Genomic Epidemiology Network (2017). Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, 356(6343) : eaam6393.

[Lenormand et al., 2012] Lenormand, C., Bausinger, H., Gross, F., Signorino-Gelo, F., Koch, S., Peressin, M., Fricker, D., Cazenave, J.-P., Bieber, T., Hanau, D., de la Salle, H., and Tourne, S. (2012). *HLA-DQA2* and *HLA-DQB2* Genes Are Specifically Expressed in Human Langerhans Cells and Encode a New HLA Class II Molecule. *The Journal of Immunology*, 188(8) : 3903–3911.

[Leslie, 1945] Leslie, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika*, 33(3) : 183–212.

[Levine and Yang, 1994] Levine, J. E. and Yang, S. Y. (1994). SSOP typing of the Tenth International Histocompatibility Workshop reference cell lines for HLA-C alleles. *Tissue Antigens*, 44(3) : 174–183.

[Li et al., 2014] Li, S., Schlebusch, C., and Jakobsson, M. (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proceedings of the Royal Society B : Biological Sciences*, 281(1793) : 20141448.

[Li et al., 2011] Li, S. S., Wang, H., Smith, A., Zhang, B., Zhang, X. C., Schoch, G., Geraghty, D., Hansen, J. A., and Zhao, L. P. (2011). Predicting multiallelic genes using unphased and flanking single nucleotide polymorphisms. *Genetic Epidemiology*, 35(2) : 85–92.

[Li et al., 1985] Li, W., Wu, C., and Luo, C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*.

[Li, 1993] Li, W.-H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution*, 36(1) : 96–99.

[Lighten et al., 2014a] Lighten, J., van Oosterhout, C., and Bentzen, P. (2014a). Critical review of NGS analyses for de novo genotyping multigene families. *Molecular Ecology*, 23(16) : 3957–3972.

[Lighten et al., 2014b] Lighten, J., van Oosterhout, C., Paterson, I. G., McMullan, M., and Bentzen, P. (2014b). Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources*, 14(4) : 753–767.

[Lima-Junior et al., 2012] Lima-Junior, J. C., Rodrigues-da Silva, R. N., Banic, D. M., Jiang, J., Singh, B., Fabrício-Silva, G. M., Porto, L. C. S., Meyer, E. V. S., Moreno, A., Rodrigues, M. M., Barnwell, J. W., Galinski, M. R., and de Oliveira-Ferreira, J. (2012). Influence of HLA-DRB1 and HLA-DQB1 Alleles on IgG Antibody Response to the P. vivax MSP-1, MSP-3a and MSP-9 in Individuals from Brazilian Endemic Area. *PLoS ONE*, 7(5) : e36419.

[Lind et al., 2010] Lind, C., Ferriola, D., Mackiewicz, K., Heron, S., Rogers, M., Slavich, L., Walker, R., Hsiao, T., McLaughlin, L., D'Arcy, M., Gai, X., Goodridge, D., Sayer, D., and Monos, D. (2010). Next-generation sequencing : the solution for high-resolution, unambiguous human leukocyte antigen typing. *Human Immunology*, 71(10) : 1033–1042.

[Lindo et al., 2016] Lindo, J., Huerta-Sánchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., Cybulski, J. S., Willerslev, E., DeGiorgio, M., and Malhi, R. S. (2016). A time transect of exomes from a Native American population before and after European contact. *Nature Communications*, 7(1) : 13175.

[Liu et al., 2006] Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *The American Journal of Human Genetics*, 79(2) : 230–237.

[Lokki et al., 2011] Lokki, A. I., Järvelä, I., Israelsson, E., Maiga, B., Troye-Blomberg, M., Dolo, A., Doumbo, O. K., Meri, S., and Holmberg, V. (2011). Lactase persistence genotypes and malaria susceptibility in Fulani of Mali. *Malaria Journal*, 10(1) : 9.

[Lombard et al., 2006] Lombard, Z., Brune, A. E., Hoal, E. G., Babb, C., Van Helden, P. D., Epplen, J. T., and Bornman, L. (2006). HLA class II disease associations in southern Africa. *Tissue Antigens*, 67(2) : 97–110.

[Long et al., 1995] Long, J. C., Williams, R. C., and Urbanek, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *American journal of human genetics*, 56(3) : 799–810.

[Lulli et al., 2009] Lulli, P., Mangano, V. D., Onori, A., Batini, C., Luoni, G., Sirima, B. S., Nebie, I., Chessa, L., Petrarca, V., and Modiano, D. (2009). HLA-DRB1 and -DQB1 loci in three west African ethnic groups : Genetic relationship with sub-Saharan African and European populations. *Human Immunology*, 70(11) : 903–909.

[López et al., 2015] López, S., Van Dorp, L., and Hellenthal, G. (2015). Human Dispersal Out of Africa : A Lasting Debate. *Evolutionary Bioinformatics*, 11s2.

[Mack et al., 2000] Mack, S., Bugawan, T., Moonsamy, P., Erlich, J., Trachtenberg, E., Paik, Y., Begovich, A., Saha, N., Beck, H., Stoneking, M., and Erlich, H. (2000). Evolution of Pacific/Asian populations inferred from HLA class II allele frequency distributions. *Tissue Antigens*, 55(5) : 383–400.

[Mack, 2015] Mack, S. J. (2015). A gene feature enumeration approach for describing HLA allele polymorphism. *Human Immunology*, 76(12) : 975–981.

[Madden, 1995] Madden, D. R. (1995). The Three-Dimensional Structure of Peptide-MHC Complexes. *Annual Review of Immunology*, 13(1) : 587–622.

[Maiers et al., 2019] Maiers, M., Halagan, M., Gragert, L., Bashyal, P., Brelsford, J., Schneider, J., Lutsker, P., and Louzoun, Y. (2019). GRIMM : GRaph IMputation and matching for HLA genotypes. *Bioinformatics*, 35(18) : 3520–3523.

[Major et al., 2013] Major, E., Rigó, K., Hague, T., Bérces, A., and Juhos, S. (2013). HLA Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data. *PLoS ONE*, 8(11) : e78410.

[Malmstrøm et al., 2016] Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., Baalsrud, H. T., Nederbragt, A. J., Hanel, R., Salzburger, W., Stenseth, N. C., Jakobsen, K. S., and Jentoft, S. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, 48(10) : 1204–1210.

[Manica et al., 2007] Manica, A., Amos, W., Balloux, F., and Hanihara, T. (2007). The effect of ancient population bottlenecks on human phenotypic variation. *Nature*, 448(7151) : 346–348.

[Mantel, 1967] Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27 :209–220.

[Marsh et al., 2010] Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Fernández-Viña, M., Geraghty, D. E., Holdsworth, R., Hurley, C. K., Lau, M., Lee, K. W., Mach, B., Maiers, M., Mayr, W. R., Müller, C. R., Parham, P., Petersdorf, E. W., Sasazuki, T., Strominger, J. L., Svejgaard, A., Terasaki, P. I., Tiercy, J. M., and Trowsdale, J. (2010). Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 75(4) : 291–455.

[Martin et al., 2018] Martin, A. R., Teferra, S., Möller, M., Hoal, E. G., and Daly, M. J. (2018). The critical needs and challenges for genetic architecture studies in Africa. *Current Opinion in Genetics & Development*, 53 : 113–120.

[Martin et al., 1995] Martin, M., Mann, D., and Carrington, M. (1995). Recombination rates across the HLA complex : use of microsatellites as a rapid screen for recombinant chromosomes. *Human Molecular Genetics*, 4(3) : 423–428.

[Martinson et al., 1995] Martinson, J. J., Excoffier, L., Swinburn, C., Boyce, A. J., Harding, R. M., Langaney, A., and Clegg, J. B. (1995). High diversity of alpha-globin haplotypes in a Senegalese population, including many previously unreported variants. *American Journal of Human Genetics*, 57(5) : 1186–1198.

[May et al., 1999] May, J., Meyer, C., Kun, J., Lell, B., Luckner, D., Dippmann, A., Bienzle, U., and Kremsner, P. (1999). HLA Class II Factors Associated with *Plasmodium falciparum* Merozoite Surface Antigen Allele Families. *The Journal of Infectious Diseases*, 179(4) : 1042–1045.

[Mayor, 2011] Mayor, A. (2011). *Traditions céramiques dans la boucle du Niger : ethnoarchéologie du peuplement au temps des empires précoloniaux*. Africa Magna Verlag, Frankfurt am Main. OCLC : 1101343597.

[Mayor et al., 2014] Mayor, N. P., Robinson, J., Ranade, S., Eng, K., Wallis-Jones, S., McWhinnie, A. J., Bultitude, W. P., Midwinter, W., Bowman, B., Hepler, L., Braund, H., Madrigal, J. A., Latham, K., and Marsh, S. G. (2014). OR57 : generation of 252 HLA class I genomic sequences in a single sequencing reaction using DNA barcodes and single molecule real-time (SMRT) DNA sequencing technology. *Human Immunology*, 75 : 49.

[McAdam et al., 1994] McAdam, S. N., Boyson, J. E., Liu, X., Garber, T. L., Hughes, A. L., Bontrop, R. E., and Watkins, D. I. (1994). A uniquely high level of recombination at the HLA-B locus. *Proceedings of the National Academy of Sciences*, 91(13) : 5893–5897.

[McClelland et al., 2003] McClelland, E. E., Penn, D. J., and Potts, W. K. (2003). Major Histocompatibility Complex Heterozygote Superiority during Coinfection. *Infection and Immunity*, 71(4) : 2079–2086.

[McConnell et al., 1998] McConnell, T. J., Godwin, U. B., Norton, S. F., Nairn, R. S., Kazianis, S., and Morizot, D. C. (1998). Identification and Mapping of Two Divergent, Unlinked Major Histocompatibility Complex Class II B Genes in Xiphophorus Fishes. *Genetics*, 149(4) : 1921.

[McDougall et al., 2005] McDougall, I., Brown, F. H., and Fleagle, J. G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433(7027) : 733–736.

[Meglécz et al., 2011] Meglécz, E., Piry, S., Desmarais, E., Galan, M., Gilles, A., Guivier, E., Pech, N., and Martin, J.-F. (2011). SESAME (SEquence Sorter & AMplicon Explorer) : genotyping based on high-throughput multiplex amplicon sequencing. *Bioinformatics*, 27(2) : 277–278.

[Messaoudi, 2002] Messaoudi, I. (2002). Direct Link Between mhc Polymorphism, T Cell Avidity, and Diversity in Immune Defense. *Science*, 298(5599) : 1797–1800.

[Messoussi et al., 2019] Messoussi, M., Hajjej, A., Ammar Elgaaied, A. B., Almawi, W. Y., Arnaiz-Villena, A., Hmida, S., and Fadhlaoui-Zid, K. (2019). HLA Class II Allele and Haplotype Diversity in Libyans and Their Genetic Relationships with Other Populations. *Immunological Investigations*, 48(8) : 875–892.

[Meyer et al., 1994] Meyer, C. G., Gallin, M., Erttmann, K. D., Brattig, N., Schnittger, L., Gelhaus, A., Tannich, E., Begovich, A. B., Erlich, H. A., and Horstmann, R. D. (1994). HLA-D alleles associated with generalized disease, localized disease, and putative immunity in Onchocerca volvulus infection. *Proceedings of the National Academy of Sciences*, 91(16) : 7515–7519.

[Meyer et al., 2018] Meyer, D., C. Aguiar, V. R., Bitarello, B. D., C. Brandt, D. Y., and Nunes, K. (2018). A genomic perspective on HLA evolution. *Immunogenetics*, 70(1) : 5–27.

[Meyer et al., 2006] Meyer, D., Single, R. M., Mack, S. J., Erlich, H. A., and Thomson, G. (2006). Signatures of Demographic History and Natural Selection in the Human Major Histocompatibility Complex Loci. *Genetics*, 173(4) : 2121–2142.

[Meyer and Thomson, 2001] Meyer, D. and Thomson, G. (2001). How selection shapes variation of the human major histocompatibility complex : a review. *Annals of Human Genetics*, 65(1) : 1–26.

[Migot-Nabias et al., 2001] Migot-Nabias, F., Luty, A., Minh, T., Fajardy, I., Tamouza, R., Marzais, F., Charron, D., Danzé, P.-M., Renaut, A., and Deloron, P. (2001). HLA alleles in relation to specific immunity to Liver Stage Antigen-1 from Plasmodium falciparum in Gabon. *Genes & Immunity*, 2(1) : 4–10.

[Mikko and Andersson, 1995] Mikko, S. and Andersson, L. (1995). Low major histocompatibility complex class II diversity in European and North American moose. *Proceedings of the National Academy of Sciences*, 92(10) : 4259–4263.

[Miller and Withler, 2004] Miller, K. M. and Withler, R. E. (2004). Mhc Diversity in Pacific Salmon : Population Structure and Trans-Species Allelism. *Hereditas*, 127(1-2) : 83–95.

[Miller and Love, 1989] Miller, M. J. and Love, E. J., editors (1989). *Parasitic diseases : treatment and control*. CRC Press, Boca Raton, Fla. Meeting Name : International Congress on Tropical Medicine and Malaria.

[Miller et al., 1996] Miller, M. M., Goto, R. M., Taylor, R. L., Zoorob, R., Auffray, C., Briles, R. W., Briles, W. E., and Bloom, S. E. (1996). Assignment of Rfp-Y to the chicken major histocompatibility complex/NOR microchromosome and evidence for high-frequency recombination associated with the nucleolar organizer region. *Proceedings of the National Academy of Sciences*, 93(9) : 3958–3962.

[Milner, 2001] Milner, M., C. (2001). Genetic organization of the human MHC class III region. *Frontiers in Bioscience*, 6(1) : d914.

[Modiano et al., 1996] Modiano, D., Petrarca, V., Sirima, B. S., Nebie, I., Diallo, D., Esposito, F., and Coluzzi, M. (1996). Different response to Plasmodium falciparum malaria in West African sympatric ethnic groups. *Proceedings of the National Academy of Sciences*, 93(23) : 13206–13211.

[Mona et al., 2008] Mona, S., Crestanello, B., Bankhead-Dronnet, S., Pecchioli, E., Ingrosso, S., D'Amelio, S., Rossi, L., Meneguz, P. G., and Bertorelle, G. (2008). Disentangling the effects of recombination, selection, and demography on the genetic variation at a major histocompatibility complex class II gene in the alpine chamois. *Molecular Ecology*, 17(18) : 4053–4067.

[Monaco et al., 2019] Monaco, A., Amoroso, N., Bellantuono, L., Lella, E., Lombardi, A., Monda, A., Tateo, A., Bellotti, R., and Tangaro, S. (2019). Shannon entropy approach reveals relevant genes in Alzheimer's disease. *PLOS ONE*, 14(12) : e0226190.

[Moonsamy et al., 2013] Moonsamy, P. V., Williams, T., Bonella, P., Holcomb, C. L., Höglund, B. N., Hillman, G., Goodridge, D., Turenchalk, G. S., Blake, L. A., Daigle, D. A., Simen, B. B., Hamilton, A., May, A. P., and Erlich, H. A. (2013). High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array™ system for simplified amplicon library preparation : High throughput HLA 454 sequencing using the Fluidigm Access Array™ system. *Tissue Antigens*, 81(3) : 141–149.

[Moorjani et al., 2016] Moorjani, P., Amorim, C. E. G., Arndt, P. F., and Przeworski, M. (2016). Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences*, 113(38) : 10607–10612.

[Morel et al., 1990] Morel, C., Zwahlen, F., Jeannet, M., Mach, B., and Tiercy, J. M. (1990). Complete analysis of HLA-DQB1 polymorphism and DR-DQ linkage disequilibrium by oligonucleotide typing. *Human Immunology*, 29(1) : 64–77.

[Muirhead, 2001] Muirhead, C. A. (2001). Consequences of population structure on genes under balancing selection. *Evolution*, 55(8) : 1532–1541.

[Mungall et al., 2003] Mungall, A. J., Palmer, S. A., Sims, S. K., Edwards, C. A., Ashurst, J. L., Wilming, L., Jones, M. C., Horton, R., Hunt, S. E., Scott, C. E., Gilbert, J. G. R., Clamp, M. E., Bethel, G., Milne, S., Ainscough, R., Almeida, J. P., Ambrose, K. D., Andrews, T. D., Ashwell, R. I. S., Babbage, A. K., Bagguley, C. L., Bailey, J., Banerjee, R., Barker, D. J., Barlow, K. F., Bates, K., Beare, D. M., Beasley, H., Beasley, O., Bird, C. P., Blakey, S., Bray-Allen, S., Brook, J., Brown, A. J., Brown, J. Y., Burford, D. C., Burrill, W., Burton, J., Carder, C., Carter, N. P., Chapman, J. C., Clark, S. Y., Clark, G., Clee, C. M., Clegg, S., Cobley, V., Collier, R. E., Collins, J. E., Colman, L. K., Corby,

N. R., Coville, G. J., Culley, K. M., Dhami, P., Davies, J., Dunn, M., Earthrowl, M. E., Ellington, A. E., Evans, K. A., Faulkner, L., Francis, M. D., Frankish, A., Frankland, J., French, L., Garner, P., Garnett, J., Ghori, M. J. R., Gilby, L. M., Gillson, C. J., Glithero, R. J., Grafham, D. V., Grant, M., Gribble, S., Griffiths, C., Griffiths, M., Hall, R., Halls, K. S., Hammond, S., Harley, J. L., Hart, E. A., Heath, P. D., Heathcott, R., Holmes, S. J., Howden, P. J., Howe, K. L., Howell, G. R., Huckle, E., Humphray, S. J., Humphries, M. D., Hunt, A. R., Johnson, C. M., Joy, A. A., Kay, M., Keenan, S. J., Kimberley, A. M., King, A., Laird, G. K., Langford, C., Lawlor, S., Leongamornlert, D. A., Leversha, M., Lloyd, C. R., Lloyd, D. M., Loveland, J. E., Lovell, J., Martin, S., Mashreghi-Mohammadi, M., Maslen, G. L., Matthews, L., McCann, O. T., McLaren, S. J., McLay, K., McMurray, A., Moore, M. J. F., Mullikin, J. C., Niblett, D., Nickerson, T., Novik, K. L., Oliver, K., Overton-Larty, E. K., Parker, A., Patel, R., Pearce, A. V., Peck, A. I., Phillimore, B., Phillips, S., Plumb, R. W., Porter, K. M., Ramsey, Y., Ranby, S. A., Rice, C. M., Ross, M. T., Searle, S. M., Sehra, H. K., Sheridan, E., Skuce, C. D., Smith, S., Smith, M., Spraggon, L., Squares, S. L., Steward, C. A., Sycamore, N., Tamlyn-Hall, G., Tester, J., Theaker, A. J., Thomas, D. W., Thorpe, A., Tracey, A., Tromans, A., Tubby, B., Wall, M., Wallis, J. M., West, A. P., White, S. S., Whitehead, S. L., Whittaker, H., Wild, A., Willey, D. J., Wilmer, T. E., Wood, J. M., Wray, P. W., Wyatt, J. C., Young, L., Younger, R. M., Bentley, D. R., Coulson, A., Durbin, R., Hubbard, T., Sulston, J. E., Dunham, I., Rogers, J., and Beck, S. (2003). The DNA sequence and analysis of human chromosome 6. *Nature*, 425(6960) : 805–811.

[Murdock, 1959] Murdock, G. P. (1959). *Africa, Its Peoples and Their Culture History.* McGraw-Hill Book Company.

[Muse and Gaut, 1994] Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5) : 715–724.

[Musolf et al., 2004] Musolf, K., Meyer-Lucht, Y., and Sommer, S. (2004). Evolution of MHC-DRB class II polymorphism in the genus Apodemus and a comparison of DRB sequences within the family Muridae (Mammalia : Rodentia). *Immunogenetics*, 56(6).

[Mytilineos et al., 1997] Mytilineos, J., Christ, U., Lempert, M., and Opelz, G. (1997). Comparison of typing results by serology and polymerase chain reaction with sequence-specific primers for HLA-Cw in 650 individuals. *Tissue Antigens*, 50(4) : 395–400.

[Nachman and Crowell, 2000] Nachman, M. W. and Crowell, S. L. (2000). Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics*, 156(1) : 297.

[Nei, 1987] Nei, M. (1987). *Molecular evolutionary genetics.* Columbia University Press, New York.

[Nei and Gojobori, 1986] Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution.*

[Nei et al., 1997] Nei, M., Gu, X., and Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, 94(15) : 7799–7806.

[Nei and Hughes, 1992] Nei, M. and Hughes, A. L. (1992). Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In *In : Tsuji K, Aizawa M, Sasazuki T, editors.* Oxford University Press, Oxford UK.

[Nei and Kumar, 2000] Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics.* Oxford University Press, Oxford ; New York.

[Nevo and Beiles, 1992] Nevo, E. and Beiles, A. (1992). Selection for class IIMhc heterozygosity by parasites in subterranean mole rats. *Experientia*, 48(5) : 512–515.

[Nielsen and Yang, 2003] Nielsen, R. and Yang, Z. (2003). Estimating the Distribution of Selection Coefficients from Phylogenetic Data with Applications to Mitochondrial and Viral DNA. *Molecular Biology and Evolution*, 20(8) : 1231–1239.

[Nowak et al., 1992] Nowak, M. A., Tarczy-Hornoch, K., and Austyn, J. M. (1992). The optimal number of major histocompatibility complex molecules in an individual. *Proceedings of the National Academy of Sciences*, 89(22) : 10896–10899.

[Nunes, 2005] Nunes, J. M. (2005). *Counting genes*. Doctorate Thesis, University of Porto, Portugal.

[Nunes, 2007] Nunes, J. M. (2007). Tools for analysing ambiguous HLA data. *Tissue Antigens*, 69 : 203–205.

[Nunes, 2016] Nunes, J. M. (2016). Using UNIFORMAT and GENE[RATE] to Analyze Data with Ambiguities in Population Genetics. *Evolutionary Bioinformatics*, page 19.

[Nunes et al., 2014] Nunes, J. M., Buhler, S., Roessli, D., Sanchez-Mazas, A., and the HLA-net 2013 collaboration (2014). The *HLA-net GENE[RATE]* pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. *Tissue Antigens*, 83(5) : 307–323.

[Oh et al., 1993] Oh, S.-H., Fleischhauer, K., and Yang, S. Y. (1993). Isoelectric focusing subtypes of HLA-A can be defined by oligonucleotide typing. *Tissue Antigens*, 41(3) : 135–142.

[Ohno, 1999] Ohno, S. (1999). Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. *Seminars in Cell & Developmental Biology*, 10(5) : 517–522.

[Ohta and Kimura, 1970] Ohta, T. and Kimura, M. (1970). Development of associative overdominance through linkage disequilibrium in finite populations. *Genetical Research*, 16(2) : 165–177.

[Ohta et al., 2000] Ohta, Y., Okamura, K., McKinney, E. C., Bartl, S., Hashimoto, K., and Flajnik, M. F. (2000). Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. *Proceedings of the National Academy of Sciences*, 97(9) : 4712–4717.

[Oksanen et al., 2019] Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2019). *vegan : Community Ecology Package*. R package version 2.5-6.

[Oliphant, 2006] Oliphant, T. (2006). NumPy : A guide to NumPy. USA : Trelgol Publishing.

[Oliver et al., 2008] Oliver, M. K., Lambin, X., Cornulier, T., and Piertney, S. B. (2008). Spatio-temporal variation in the strength and mode of selection acting on major histocompatibility complex diversity in water vole (*Arvicola terrestris*) metapopulations. *Molecular Ecology*.

[Oomen et al., 2013] Oomen, R. A., Gillett, R. M., and Kyle, C. J. (2013). Comparison of 454 pyrosequencing methods for characterizing the major histocompatibility complex of nonmodel species and the advantages of ultra deep coverage. *Molecular Ecology Resources*, 13(1) : 103–116.

[Otsu, 1979] Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1) : 62–66.

[Otting et al., 1998] Otting, N., Doxiadis, G., Versluis, L., de Groot, N., Anholts, J., Verduin, W., Rozemuller, E., Claas, F., Tilanus, M., and Bontrop, R. (1998). Characterization and distribution of Mhc-DPB1 alleles in chimpanzee and rhesus macaque populations. *Human Immunology*, 59(10) : 656–664.

[O'Hanlon et al., 2016] O'Hanlon, S. J., Slater, H. C., Cheke, R. A., Boatin, B. A., Coffeng, L. E., Pion, S. D. S., Boussinesq, M., Zouré, H. G. M., Stolk, W. A., and Basáñez, M.-G. (2016). Model-Based Geostatistical Mapping of the Prevalence of Onchocerca volvulus in West Africa. *PLOS Neglected Tropical Diseases*, 10(1) : e0004328.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking : Bringing order to the web. In *WWW 1999*.

[Pamilo and Bianchi, 1993] Pamilo, P. and Bianchi, N. (1993). Evolution of the Zfx and Zfy genes : rates and interdependence between the genes. *Molecular Biology and Evolution*.

[Parham, 1988] Parham, P. (1988). Function and polymorphism of human leukocyte antigen-A,B,C molecules. *The American Journal of Medicine*, 85(6) : 2–5.

[Parham et al., 1995] Parham, P., Adams, E. J., and Arnett, K. L. (1995). The Origins of HLA-A,B,C Polymorphism. *Immunological Reviews*, 143(1) : 141–180.

[Parham et al., 1997] Parham, P., Arnett, K., Adams, E., Little, A.-M., Tees, K., Barber, L. D., Marsh, S., Ohta, T., Markow, T., and Petzl-Erler, M. L. (1997). Episodic evolution and turnover of HLA-B in the indigenous human populations of the Americas. *Tissue Antigens*, 50(3) : 219–232.

[Pasquier, 2016] Pasquier, M. (2016). Étude des polymorphismes HLA-DQA1 et HLA-DQB1 dans des populations africaines du Sahel. Master's thesis, Université de Genève, Genève, Suisse.

[Pasvol et al., 1978] Pasvol, G., Weatherall, D. J., and Wilson, R. J. M. (1978). Cellular mechanism for the protective effect of haemoglobin S against P. falciparum malaria. *Nature*, 274(5672) : 701–703.

[Paulsson et al., 2002] Paulsson, K. M., Kleijmeer, M. J., Griffith, J., Jevon, M., Chen, S., Anderson, P. O., Sjögren, H.-O., Li, S., and Wang, P. (2002). Association of Tapasin and COPI Provides a Mechanism for the Retrograde Transport of Major Histocompatibility Complex (MHC) Class I Molecules from the Golgi Complex to the Endoplasmic Reticulum. *Journal of Biological Chemistry*, 277(21) : 18266–18271.

[Payne and Rolfs, 1958] Payne, R. and Rolfs, M. R. (1958). Fetomaternal Leukocyte Incompatibility. *Journal of Clinical Investigation*, 37(12) : 1756–1763.

[Payne et al., 1964] Payne, R., Trip, M., Weigle, J., Bodmer, W. F., and Bodmer, J. G. (1964). A new leukocyte iso-antigen system in man. *Cold Spring Harbor Symposia on Quantitative Biology*, 29 : 285–295.

[Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) : 559–572.

[Pechouskova et al., 2015] Pechouskova, E., Dammhahn, M., Brameier, M., Fichtel, C., Kappeler, P. M., and Huchard, E. (2015). MHC class II variation in a rare and ecological specialist mouse lemur reveals lower allelic richness and contrasting selection patterns compared to a generalist and widespread sympatric congener. *Immunogenetics*, 67(4) : 229–245.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

[Peng et al., 2010] Peng, M.-S., Quang, H. H., Dang, K. P., Trieu, A. V., Wang, H.-W., Yao, Y.-G., Kong, Q.-P., and Zhang, Y.-P. (2010). Tracing the Austronesian Footprint in Mainland Southeast Asia : A Perspective from Mitochondrial DNA. *Molecular Biology and Evolution*, 27(10) : 2417–2430.

[Penn, 2002] Penn, D. J. (2002). The Scent of Genetic Compatibility : Sexual Selection and the Major Histocompatibility Complex. *Ethology*, 108(1) : 1–21.

[Penn et al., 2002] Penn, D. J., Damjanovich, K., and Potts, W. K. (2002). MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences*, 99(17) : 11260–11264.

[Pereira et al., 2010] Pereira, L., Černý, V., Cerezo, M., Silva, N. M., Hájek, M., Vašíková, A., Kujanová, M., Brdička, R., and Salas, A. (2010). Linking the sub-Saharan and West Eurasian gene pools : maternal and paternal heritage of the Tuareg nomads from the African Sahel. *European Journal of Human Genetics*, 18(8) : 915–923.

[Perelman et al., 2011] Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P. C., Silva, A., O'Brien, S. J., and Pecon-Slattery, J. (2011). A Molecular Phylogeny of Living Primates. *PLoS Genetics*, 7(3) : e1001342.

[Piertney and Oliver, 2006] Piertney, S. B. and Oliver, M. K. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity*, 96(1) : 7–21.

[Pimtanothai et al., 2001] Pimtanothai, N., Hurley, C., Leke, R., Klitz, W., and Johnson, A. (2001). HLA-DR and -DQ polymorphism in Cameroon. *Tissue Antigens*, 58(1) : 1–8.

[Piontkivska, 2003] Piontkivska, H. (2003). Birth-and-Death Evolution in Primate MHC Class I Genes : Divergence Time Estimates. *Molecular Biology and Evolution*, 20(4) : 601–609.

[Pischedda et al., 2017] Pischedda, S., Barral-Arca, R., Gómez-Carballa, A., Pardo-Seco, J., Catelli, M. L., Álvarez Iglesias, V., Cárdenas, J. M., Nguyen, N. D., Ha, H. H., Le, A. T., Martinón-Torres, F., Vullo, C., and Salas, A. (2017). Phylogeographic and genome-wide investigations of Vietnam ethnic groups reveal signatures of complex historical demographic movements. *Scientific Reports*, 7(1) : 12630.

[Ploegh and Watts, 1998] Ploegh, H. and Watts, C. (1998). Antigen overview. *Current Opinion in Immunology*, 10(1) : 57–58.

[Podgorná et al., 2015] Podgorná, E., Diallo, I., Vangenot, C., Sanchez-Mazas, A., Sabbagh, A., Černý, V., and Poloni, E. S. (2015). Variation in NAT2 acetylation phenotypes is associated with differences in food-producing subsistence modes and ecoregions in Africa. *BMC Evolutionary Biology*, 15(1) : 263.

[Poloni et al., 1995] Poloni, E. S., Excoffier, L., Mountain, J. L., Langaney, A., and Cavalli-Sforza, L. L. (1995). Nuclear DNA polymorphism in a Mandenka population from Senegal : Comparison with eight other human populations. *Annals of Human Genetics*, 59(1) : 43–61.

[Poloni et al., 2009] Poloni, E. S., Naciri, Y., Bucho, R., Niba, R., Kervaire, B., Excoffier, L., Langaney, A., and Sanchez-Mazas, A. (2009). Genetic Evidence for Complexity

in Ethnic Differentiation and History in East Africa. *Annals of Human Genetics*, 73(6) : 582–600.

[Priehodová et al., 2014] Priehodová, E., Abdelsawy, A., Heyer, E., and Černý, V. (2014). Lactase Persistence Variants in Arabia and in the African Arabs. *Human Biology*, 86(1) : 7–18.

[Prugnolle et al., 2005a] Prugnolle, F., Manica, A., and Balloux, F. (2005a). Geography predicts neutral genetic diversity of human populations. *Current Biology*, 15(5) : R159–R160.

[Prugnolle et al., 2005b] Prugnolle, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V., and Balloux, F. (2005b). Pathogen-Driven Selection and Worldwide HLA Class I Diversity. *Current Biology*, 15(11) : 1022–1027.

[Qutob et al., 2012] Qutob, N., Balloux, F., Raj, T., Liu, H., Marion de Procé, S., Trowsdale, J., and Manica, A. (2012). Signatures of historical demography and pathogen richness on MHC class I genes. *Immunogenetics*, 64(3) : 165–175.

[R Core Team, 2020] R Core Team (2020). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Radtkey et al., 1996] Radtkey, R. R., Becker, B., Miller, R. D., Riblet, R., and Case, T. J. (1996). Variation and evolution of Class I Mhc in sexual and parthenogenetic geckos. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 263(1373) : 1023–1032.

[Radwan et al., 2020] Radwan, J., Babik, W., Kaufman, J., Lenz, T. L., and Winternitz, J. (2020). Advances in the Evolutionary Understanding of MHC Polymorphism. *Trends in Genetics*, 36(4) : 298–311.

[Ranciaro et al., 2014] Ranciaro, A., Campbell, M., Hirbo, J., Ko, W.-Y., Froment, A., Anagnostou, P., Kotze, M., Ibrahim, M., Nyambo, T., Omar, S., and Tishkoff, S. (2014). Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *The American Journal of Human Genetics*, 94(4) : 496–510.

[Reche and Reinherz, 2003] Reche, P. A. and Reinherz, E. L. (2003). Sequence variability analysis of human class I and class II MHC molecules : functional and structural correlates of amino acid polymorphisms. *Journal of Molecular Biology*, 331(3) : 623–641.

[Reece et al., 2007] Reece, J. B., Lachaîne, R., and Bosset, M. (2007). *Biologie*. Pearson Education France, Paris. OCLC : 853141326.

[Reed and Tishkoff, 2006] Reed, F. A. and Tishkoff, S. A. (2006). African human diversity, origins and migrations. *Current Opinion in Genetics & Development*, 16(6) : 597–605.

[Reich et al., 2001] Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834) : 199–204.

[Renquin et al., 2001] Renquin, J., Sanchez-Mazas, A., Halle, L., Rivalland, S., Jaeger, G., Mbayo, K., Bianchi, F., and Kaplan, C. (2001). HLA class II polymorphism in Aka Pygmies and Bantu Congolese and a reassessment of HLA-DRB1 African diversity. *Tissue Antigens*, 58(4) : 211–222.

[Reynolds et al., 1983] Reynolds, J., Weir, B. S., and Cockerham, C. C. (1983). Estimation of the coancestry coefficient : basis for a short-term genetic distance. *Genetics*, 105(3) : 767–779.

[Richardson and Westerdahl, 2003] Richardson, D. S. and Westerdahl, H. (2003). MHC diversity in two Acrocephalus species : the outbred Great reed warbler and the inbred Seychelles warbler. *Molecular Ecology*, 12(12) : 3523–3529.

[Richman et al., 2003a] Richman, A. D., Herrera, L. G., and Nash, D. (2003a). Evolution of MHC Class II Eβ Diversity Within the Genus Peromyscus. *Genetics*, 164(1) : 289.

[Richman et al., 2003b] Richman, A. D., Herrera, L. G., Nash, D., and Schierup, M. H. (2003b). Relative roles of mutation and recombination in generating allelic polymorphism at an MHC class II locus in *Peromyscus maniculatus*. *Genetical Research*, 82(2) : 89–99.

[Rico et al., 2015] Rico, Y., Morris-Pocock, J., Zigouris, J., Nocera, J. J., and Kyle, C. J. (2015). Lack of Spatial Immunogenetic Structure among Wolverine (Gulo gulo) Populations Suggestive of Broad Scale Balancing Selection. *PLOS ONE*, 10(10) : e0140170.

[Robinson et al., 2015] Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. (2015). The IPD and IMGT/HLA database : allele variant databases. *Nucleic Acids Research*, 43(D1) : D423–D431.

[Rogers, 1985] Rogers, J. (1985). Mouse histocompatibility-related genes are not conserved in other mammals. *The EMBO Journal*, 4(3) : 749–753.

[Roosnek et al., 2015] Roosnek, E., Guyot, N., Vu, N. V., Walker, P., Tiercy, J.-M., Chapuis, B., Chalandon, Y., Peyrard, T., Zubler, R., Villard, J., Ferrari, S., Buhler, L., Scherly, D., and Baumgart, C. (2015). Hla typing. `https://medweb4.unige.ch/immunologie/home/HSC/donor/HLA_typing/`. Consulté : 16/03/2020.

[Rousset, 2008] Rousset, F. (2008). Inferences from Spatial Population Genetics. In Balding, D. J., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*, pages 945–979. John Wiley & Sons, Ltd, Chichester, UK.

[Sabbagh et al., 2011] Sabbagh, A., Darlu, P., Crouau-Roy, B., and Poloni, E. S. (2011). Arylamine N-Acetyltransferase 2 (NAT2) Genetic Diversity and Traditional Subsistence : A Worldwide Population Survey. *PLoS ONE*, 6(4) : e18507.

[Sabbagh et al., 2008] Sabbagh, A., Langaney, A., Darlu, P., Gérard, N., Krishnamoorthy, R., and Poloni, E. S. (2008). Worldwide distribution of NAT2 diversity : Implications for NAT2 evolutionary history. *BMC Genetics*, 9(1) : 21.

[Sabeti, 2006] Sabeti, P. C. (2006). Positive Natural Selection in the Human Lineage. *Science*, 312(5780) : 1614–1620.

[Saiki et al., 1986] Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1986). Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. *Nature*, 324(6093) : 163–166.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4) : 406–425. Place : United States.

[Salamon et al., 1999] Salamon, H., Klitz, W., Easteal, S., Gao, X., Erlich, H. A., Fernandez-Viña, M., Trachtenberg, E. A., McWeeney, S. K., Nelson, M. P., and Thomson, G. (1999). Evolution of HLA Class II Molecules : Allelic and Amino Acid Site Variability Across Populations. *Genetics*, 152(1) : 393.

[Salter-Cid et al., 1998] Salter-Cid, L., Nonaka, M., and Flajnik, M. F. (1998). Expression of MHC Class Ia and Class Ib During Ontogeny : High Expression in Epithelia and Coregulation of Class Ia and lmp7 Genes. *The Journal of Immunology*, 160(6) : 2853.

[Sanchez-Mazas, 2001] Sanchez-Mazas, A. (2001). African diversity from the HLA point of view : influence of genetic drift, geography, linguistics, and natural selection. *Human Immunology*, 62(9) : 937–948.

[Sanchez-Mazas, 2007] Sanchez-Mazas, A. (2007). An apportionment of human HLA diversity. *Tissue Antigens*, 69 : 198–202.

[Sanchez-Mazas et al., 2005] Sanchez-Mazas, A., Jacques, G., and Sagart, L. (2005). Comparing linguistic and genetic relationships among East Asian populations : a study of the Rh and GM polymorphisms. In *The Peopling of East Asia : Putting Together Archaeology, Linguistics and Genetics*, pages 252–272. Routledge Curzon, London and New York.

[Sanchez-Mazas et al., 2012] Sanchez-Mazas, A., Lemaitre, J.-F., and Currat, M. (2012). Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 367(1590) : 830–839.

[Sanchez-Mazas and Meyer, 2014] Sanchez-Mazas, A. and Meyer, D. (2014). The Relevance of HLA Sequencing in Population Genetics Studies. *Journal of Immunology Research*, 2014 : 1–12.

[Sanchez-Mazas et al., 2017] Sanchez-Mazas, A., Černý, V., Di, D., Buhler, S., Podgorná, E., Chevallier, E., Brunet, L., Weber, S., Kervaire, B., Testi, M., Andreani, M., Tiercy, J.-M., Villard, J., and Nunes, J. M. (2017). The HLA-B landscape of Africa : Signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. *Molecular Ecology*, 26(22) : 6238–6252.

[Santamaria et al., 1993] Santamaria, P., Lindstrom, A. L., Boyce-Jacino, M. T., Myster, S. H., Barbosa, J. J., Faras, A. J., and Rich, S. S. (1993). HLA class I sequence-based typing. *Human Immunology*, 37(1) : 39–50.

[Saper et al., 1991] Saper, M., Bjorkman, P., and Wiley, D. (1991). Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. *Journal of Molecular Biology*, 219(2) : 277–319.

[Sato et al., 2000] Sato, A., Figueroa, F., Murray, B. W., Málaga-Trillo, E., Zaleska-Rutczynska, Z., Sultmann, H., Toyosawa, S., Wedekind, C., Steck, N., and Klein, J. (2000). Nonlinkage of major histocompatibility complex class I and class II loci in bony fishes. *Immunogenetics*, 51(2) : 108–116.

[Satta et al., 1996a] Satta, Y., Mayer, W. E., and Klein, J. (1996a). Evolutionary relationship of HLA-DRB genes inferred from intron sequences. *Journal of Molecular Evolution*, 42(6) : 648–657.

[Satta et al., 1996b] Satta, Y., Mayer, W. E., and Klein, J. (1996b). HLA-DRB intron 1 sequences : Implications for the evolution of HLA-DRB genes and haplotypes. *Human Immunology*, 51(1) : 1–12.

[Satta et al., 1993] Satta, Y., O'hUigin, C., Takahata, N., and Klein, J. (1993). The synonymous substitution rate of the major histocompatibility complex loci in primates. *Proceedings of the National Academy of Sciences*, 90(16) : 7480–7484.

[Scally and Durbin, 2012] Scally, A. and Durbin, R. (2012). Revising the human mutation rate : implications for understanding human evolution. *Nature Reviews Genetics*, 13(10) : 745–753.

[Schaschl et al., 2004] Schaschl, H., Goodman, S. J., and Suchentrunk, F. (2004). Sequence analysis of the MHC class II DRB alleles in Alpine chamois (Rupicapra r. rupicapra). *Developmental & Comparative Immunology*, 28(3) : 265–277.

[Schaschl et al., 2005] Schaschl, H., Suchentrunk, F., Hammer, S., and Goodman, S. J. (2005). Recombination and the origin of sequence diversity in the DRB MHC class II locus in chamois (Rupicapra spp.). *Immunogenetics*, 57(1-2) : 108–115.

[Schaschl et al., 2012] Schaschl, H., Suchentrunk, F., Morris, D. L., Slimen, H., Smith, S., and Arnold, W. (2012). Sex-specific selection for MHC variability in Alpine chamois. *BMC Evolutionary Biology*, 12(1) : 20.

[Schaschl et al., 2006] Schaschl, H., Wandeler, P., Suchentrunk, F., Obexer-Ruff, G., and Goodman, S. J. (2006). Selection and recombination drive the evolution of MHC class II DRB diversity in ungulates. *Heredity*, 97(6) : 427–437.

[Schlebusch et al., 2017] Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., and Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, 358(6363) : 652–655.

[Schloss et al., 2009] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur : Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23) : 7537–7541.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(4) :623–656.

[Shiina et al., 2007] Shiina, T., Briles, W. E., Goto, R. M., Hosomichi, K., Yanagiya, K., Shimizu, S., Inoko, H., and Miller, M. M. (2007). Extended Gene Map Reveals Tripartite Motif, C-Type Lectin, and Ig Superfamily Type Genes within a Subregion of the Chicken MHC-B Affecting Infectious Disease. *The Journal of Immunology*, 178(11) : 7162–7172.

[Shiina et al., 2009] Shiina, T., Hosomichi, K., Inoko, H., and Kulski, J. K. (2009). The HLA genomic loci map : expression, interaction, diversity and disease. *Journal of Human Genetics*, 54(1) : 15–39.

[Shiina et al., 2004] Shiina, T., Shimizu, S., Hosomichi, K., Kohara, S., Watanabe, S., Hanzawa, K., Beck, S., Kulski, J. K., and Inoko, H. (2004). Comparative Genomic Analysis of Two Avian (Quail and Chicken) MHC Regions. *The Journal of Immunology*, 172(11) : 6751–6763.

[Shiina et al., 2012] Shiina, T., Suzuki, S., Ozaki, Y., Taira, H., Kikkawa, E., Shigenari, A., Oka, A., Umemura, T., Joshita, S., Takahashi, O., Hayashi, Y., Paumen, M., Katsuyama, Y., Mitsunaga, S., Ota, M., Kulski, J. K., and Inoko, H. (2012). Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers : Super high-resolution DNA typing of HLA loci. *Tissue Antigens*, 80(4) : 305–316.

[Shriner and Rotimi, 2018] Shriner, D. and Rotimi, C. N. (2018). Genetic history of Chad. *American Journal of Physical Anthropology*, 167(4) : 804–812.

[Shriner et al., 2015] Shriner, D., Tekola-Ayele, F., Adeyemo, A., and Rotimi, C. N. (2015). Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. *Scientific Reports*, 4(1) : 6055.

[Sidney et al., 1996] Sidney, J., Grey, H. M., Kubo, R. T., and Sette, A. (1996). Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. *Immunology Today*, 17(6) : 261–266.

[Sidney et al., 2008] Sidney, J., Peters, B., Frahm, N., Brander, C., and Sette, A. (2008). HLA class I supertypes : a revised and updated classification. *BMC Immunology*, 9(1) : 1.

[Single et al., 2007] Single, R. M., Martin, M. P., Gao, X., Meyer, D., Yeager, M., Kidd, J. R., Kidd, K. K., and Carrington, M. (2007). Global diversity and evidence for coevolution of KIR and HLA. *Nature Genetics*, 39(9) : 1114–1119.

[Skoglund and Reich, 2016] Skoglund, P. and Reich, D. (2016). A genomic view of the peopling of the Americas. *Current Opinion in Genetics & Development*, 41 : 27–35.

[Slade and McCallum, 1992] Slade, R. W. and McCallum, H. I. (1992). Overdominant vs. frequency-dependent selection at MHC loci. *Genetics*, 132(3) : 861.

[Slatkin, 1991] Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, 58(2) : 167–175.

[Slatkin, 1993] Slatkin, M. (1993). Isolation by Distance in Equilibrium and Non-Equilibrium Populations. *Evolution*, 47(1) : 264–279. Publisher : [Society for the Study of Evolution, Wiley].

[Slatkin, 1994] Slatkin, M. (1994). An exact test for neutrality based on the Ewens sampling distribution. *Genetical Research*, 64(1) : 71–74.

[Slatkin, 1995] Slatkin, M. (1995). Hitchhiking and associative overdominance at a microsatellite locus. *Molecular Biology and Evolution*, 12(3) : 473–480.

[Slatkin, 1996] Slatkin, M. (1996). A correction to the exact test based on the Ewens sampling distribution. *Genetical Research*, 68(3) : 259–260.

[Smulders et al., 2003] Smulders, M., Snoek, L., Booy, G., and Vosman, B. (2003). Complete loss of MHC genetic diversity in the Common Hamster (Cricetus cricetus) population in The Netherlands. Consequences for conservation strategies. *Conservation Genetics*, 4(4) : 441–451.

[Solberg et al., 2008] Solberg, O. D., Mack, S. J., Lancaster, A. K., Single, R. M., Tsai, Y., Sanchez-Mazas, A., and Thomson, G. (2008). Balancing selection and heterogeneity across the classical human leukocyte antigen loci : A meta-analytic review of 497 population studies. *Human Immunology*, 69(7) : 443–464.

[Solheim et al., 2007] Solheim, W., Bulbeck, D., and Flavel, A. (2007). *Archaeology and culture in Southeast Asia : unraveling the Nusantao*. University of the Philippines Press, Quezon City (Phillippines).

[Sommer, 2005] Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in Zoology*, 2(1) : 16.

[Sommer et al., 2013] Sommer, S., Courtiol, A., and Mazzoni, C. J. (2013). MHC genotyping of non-model organisms using next-generation sequencing : a new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, 14(1) : 542.

[Southworth et al., 2004] Southworth, W., Glover, I., and Bellwood, P. (2004). *The coastal states of Champa, Southeast Asia : from prehistory to history*. Routledge Curzon, London.

[Souza et al., 2020] Souza, A. S., Sonon, P., Paz, M. A., Tokplonou, L., Lima, T. H. A., Porto, I. O. P., Andrade, H. S., Silva, N. d. S. B., Veiga-Castelli, L. C., Oliveira, M. L. G., Sadissou, I. A., Massaro, J. D., Moutairou, K. A., Donadi, E. A., Massougbodji, A., Garcia, A., Ibikounlé, M., Meyer, D., Sabbagh, A., Mendes-Junior, C. T., Courtin, D., and Castelli, E. C. (2020). HLA-C genetic diversity and evolutionary insights in two samples from Brazil and Benin. *HLA*, page tan.13996.

[Spurgin and Richardson, 2010] Spurgin, L. G. and Richardson, D. S. (2010). How pathogens drive genetic diversity : MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B : Biological Sciences*, 277(1684) : 979–988.

[St John et al., 2012] St John, J. A., Braun, E. L., Isberg, S. R., Miles, L. G., Chong, A. Y., Gongora, J., Dalzell, P., Moran, C., Bed'Hom, B., Abzhanov, A., Burgess, S. C., Cooksey, A. M., Castoe, T. A., Crawford, N. G., Densmore, L. D., Drew, J. C., Edwards, S. V., Faircloth, B. C., Fujita, M. K., Greenwold, M. J., Hoffmann, F. G., Howard, J. M., Iguchi, T., Janes, D. E., Khan, S. Y., Kohno, S., de Koning, A. J., Lance, S. L., McCarthy, F. M., McCormack, J. E., Merchant, M. E., Peterson, D. G., Pollock, D. D., Pourmand, N., Raney, B. J., Roessler, K. A., Sanford, J. R., Sawyer, R. H., Schmidt, C. J., Triplett, E. W., Tuberville, T. D., Venegas-Anaya, M., Howard, J. T., Jarvis, E. D., Guillette, L. J., Glenn, T. C., Green, R. E., and Ray, D. A. (2012). Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology*, 13(1) : 415.

[Star and Jentoft, 2012] Star, B. and Jentoft, S. (2012). Why does the immune system of Atlantic cod lack MHC II ? *BioEssays*, 34(8) : 648–651.

[Star et al., 2011] Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lanzén, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, O., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R. B., Tina, K. G., Espelund, M., Nepal, C., Previti, C., Karlsen, B. O., Moum, T., Skage, M., Berg, P. R., Gjøen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S. D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S. W., Stenseth, N. C., and Jakobsen, K. S. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, 477(7363) : 207–210.

[Stephens et al., 1995] Stephens, H. A. F., Brown, A. E., Chandanayingyong, D., Webster, H. K., Sirikong, M., Longta, P., Vangseratthana, R., Gordon, D. M., Lekmak, S., and Rungruang, E. (1995). The presence of the HLA class. II allele DPB1*0501 in ethnic Thais correlates with an enhanced vaccine-induced antibody response to a malaria sporozoite antigen. *European Journal of Immunology*, 25(11) : 3142–3147.

[Stern and Wiley, 1994] Stern, L. J. and Wiley, D. C. (1994). Antigenic peptide binding by class I and class II histocompatibility proteins. *Structure*, 2(4) : 245–251.

[Steven et al., 2000] Steven, G. M., Peter, P., and Linda, D. B. (2000). *The HLA FactsBook*. Elsevier.

[Stringer and Andrews, 1988] Stringer, C. and Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science*, 239(4845) : 1263–1268.

[Stutz and Bolnick, 2014] Stutz, W. E. and Bolnick, D. I. (2014). Stepwise Threshold Clustering : A New Method for Genotyping MHC Loci Using Next-Generation Sequencing Technology. *PLoS ONE*, 9(7) : e100587.

[Suarez et al., 2003] Suarez, C., Cardenas, P., Llanos-Ballestas, E., Martinez, P., Obregon, M., Patarroyo, M., and Patarroyo, M. (2003). alpha1 and alpha2 domains of Aotus MHC Class I and Catarrhini MHC Class Ia share similar characteristics. *Tissue Antigens*, 61(5) : 362–373.

[Sun et al., 2015] Sun, H., Yang, Z., Lin, K., Liu, S., Huang, K., Wang, X., Chu, J., and Huang, X. (2015). The Adaptive Change of HLA-DRB1 Allele Frequencies Caused by Natural Selection in a Mongolian Population That Migrated to the South of China. *PLOS ONE*, 10(7) : e0134334.

[Tajima, 1989a] Tajima, F. (1989a). The effect of change in population size on DNA polymorphism. *Genetics*, 123(3) : 597–601.

[Tajima, 1989b] Tajima, F. (1989b). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3) : 585.

[Tajima, 1993] Tajima, F. (1993). Measurement of DNA polymorphism. In *Mechanisms of Molecular Evolution. Introduction to Molecular Paleopopulation Biology*, pages 37–59. Takahata, N. and Clark, A.G., Tokyo, Sunderland, MA :Japan Scientific Societies Press, Sinauer Associates, Inc.

[Takahashi, 2000] Takahashi, K. (2000). Origins and divergence times of mammalian class II MHC gene clusters. *Journal of Heredity*, 91(3) : 198–204.

[Takahata, 1990] Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proceedings of the National Academy of Sciences*, 87(7) : 2419–2423.

[Takahata and Nei, 1990] Takahata, N. and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, 124(4) : 967.

[Takahata and Satta, 1998] Takahata, N. and Satta, Y. (1998). Selection, convergence, and intragenic recombination in HLA diversity. *Genetica*, 102/103 : 157–169.

[Takahata et al., 1992] Takahata, N., Satta, Y., and Klein, J. (1992). Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics*, 130(4) : 925.

[Terado et al., 2003] Terado, T., Okamura, K., Ohta, Y., Shin, D.-H., Smith, S. L., Hashimoto, K., Takemoto, T., Nonaka, M. I., Kimura, H., Flajnik, M. F., and Nonaka, M. (2003). Molecular Cloning of C4 Gene and Identification of the Class III Complement Region in the Shark MHC. *The Journal of Immunology*, 171(5) : 2461–2466.

[The 1000 Genomes Project Consortium, 2015] The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571) : 68–74.

[The International HapMap Consortium, 2003] The International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968) : 789–796.

[The MHC sequencing consortium, 1999] The MHC sequencing consortium (1999). Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401(6756) : 921–923.

[The UK10K Consortium, 2015] The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571) : 82–90.

[Thorstenson et al., 2018] Thorstenson, Y. R., Creary, L. E., Huang, H., Rozot, V., Nguyen, T. T., Babrzadeh, F., Kancharla, S., Fukushima, M., Kuehn, R., Wang, C., Li, M., Krishnakumar, S., Mindrinos, M., Fernandez Viña, M. A., Scriba, T. J., and Davis, M. M. (2018). Allelic resolution NGS HLA typing of Class I and Class II loci and haplotypes in Cape Town, South Africa. *Human Immunology*, 79(12) : 839–847.

[Thurgood, 1999] Thurgood, G. (1999). *From ancient Cham to modern dialects : two thousand years of language contact and change*. University of Hawaii Press, Honolulu (HI).

[Thursz et al., 1997] Thursz, M. R., Thomas, H. C., Greenwood, B. M., and Hill, A. V. (1997). Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nature Genetics*, 17(1) : 11–12.

[Tiercy et al., 1994] Tiercy, J. M., Djavad, N., Rufer, N., Speiser, D. E., Jeannet, M., and Roosnek, E. (1994). Oligotyping of HLA-A2, -A3, and -B44 subtypes. Detection of subtype incompatibilities between patients and their serologically matched unrelated bone marrow donors. *Human Immunology*, 41(3) : 207–215.

[Tiercy et al., 1989] Tiercy, J. M., Gorski, J., Bétuel, H., Freidel, A. C., Gebuhrer, L., Jeannet, M., and Mach, B. (1989). DNA typing of DRw6 subtypes : correlation with

DRB1 and DRB3 allelic sequences by hybridization with oligonucleotide probes. *Human Immunology*, 24(1) : 1–14.

[Tiercy et al., 1988] Tiercy, J. M., Gorski, J., Jeannet, M., and Mach, B. (1988). Identification and distribution of three serologically undetected alleles of HLA-DR by oligonucleotide.DNA typing analysis. *Proceedings of the National Academy of Sciences*, 85(1) : 198–202.

[Tiercy et al., 1992] Tiercy, J. M., Sanchez-Mazas, A., Excoffier, L., Shi-Isaac, X., Jeannet, M., Mach, B., and Langaney, A. (1992). HLA-DR Polymorphism in a Senegalese Mandenka Population : DNA Oligotyping and Population Genetics of DRB I Specificities. *American Journal of Human Genetics*, 51 : 592–608.

[Tishkoff and Kidd, 2004] Tishkoff, S. A. and Kidd, K. K. (2004). Implications of biogeography of human populations for 'race' and medicine. *Nature Genetics*, 36(S11) : S21–S27.

[Tishkoff et al., 2009] Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., and Williams, S. M. (2009). The Genetic Structure and History of Africans and African Americans. *Science*, 324(5930) : 1035–1044.

[Torcia et al., 2008] Torcia, M. G., Santarlasci, V., Cosmi, L., Clemente, A., Maggi, L., Mangano, V. D., Verra, F., Bancone, G., Nebie, I., Sirima, B. S., Liotta, F., Frosali, F., Angeli, R., Severini, C., Sannella, A. R., Bonini, P., Lucibello, M., Maggi, E., Garaci, E., Coluzzi, M., Cozzolino, F., Annunziato, F., Romagnani, S., and Modiano, D. (2008). Functional deficit of T regulatory cells in Fulani, an ethnic group with low susceptibility to Plasmodium falciparum malaria. *Proceedings of the National Academy of Sciences*, 105(2) : 646–651.

[Triska et al., 2015] Triska, P., Soares, P., Patin, E., Fernandes, V., Cerny, V., and Pereira, L. (2015). Extensive Admixture and Selective Pressure Across the Sahel Belt. *Genome Biology and Evolution*, 7(12) : 3484–3495.

[Trowsdale, 1995] Trowsdale, J. (1995). Both man & bird & beast : comparative organization of MHC genes. *Immunogenetics*, 41(1) : 1–17.

[Trowsdale, 2002] Trowsdale, J. (2002). The gentle art of gene arrangement : the meaning of gene clusters. *Genome Biology*, 3(3) : comment2002.1.

[Turner et al., 1998] Turner, S., Ellexson, M. E., Hickman, H. D., Sidebottom, D. A., Fernández-Viña, M., Confer, D. L., and Hildebrand, W. H. (1998). Sequence-Based Typing Provides a New Look at HLA-C Diversity. *The Journal of Immunology*, 161(3) : 1406.

[Valdes et al., 1999] Valdes, A. M., McWeeney, S. K., Meyer, D., Nelson, M. P., and Thomson, G. (1999). Locus and population specific evolution in HLA class II genes. *Annals of Human Genetics*, 63(1) : 27–43.

[Van Den Bussche et al., 2002] Van Den Bussche, R. A., Ross, T. G., and Hoofer, S. R. (2002). Genetic Variation at a Major Histocompatibility Locus within and Among Populations of White-Tailed Deer (Odocoileus Virginianus). *Journal of Mammalogy*, 83(1) : 31–39.

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 : 2579–2605.

[Van Rood et al., 1958] Van Rood, J. J., Eernisse, J. G., and Van Leeuwen, A. (1958). Leucocyte Antibodies in Sera from Pregnant Women. *Nature*, 181(4625) : 1735–1736.

[van Rood and Van Leeuwen, 1963] van Rood, J. J. and Van Leeuwen, A. (1963). Leukocyte grouping : a method and its applications. *Journal of Clinical Investigation*, 42 : 1382–1390.

[Vangenot et al., 2020] Vangenot, C., Nunes, J. M., Doxiadis, G. M., Poloni, E. S., Bontrop, R. E., de Groot, N. G., and Sanchez-Mazas, A. (2020). Similar patterns of genetic diversity and linkage disequilibrium in Western chimpanzees (Pan troglodytes verus) and humans indicate highly conserved mechanisms of MHC molecular evolution. *BMC Evolutionary Biology*, 20(1) : 119.

[Velten et al., 2008] Velten, F., Rogel-Gaillard, C., Renard, C., Chardon, P., Pontarotti, P., Tazi-Ahnini, R., and Vaiman, M. (2008). A first map of the porcine major histocompatibility complex class I region. *Tissue Antigens*, 51(2) : 183–194.

[Vilches and Parham, 2002] Vilches, C. and Parham, P. (2002). Kir : Diverse, Rapidly Evolving Receptors of Innate and Adaptive Immunity. *Annual Review of Immunology*, 20(1) : 217–251.

[Vitti et al., 2013] Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, 47(1) : 97–120.

[von Salomé et al., 2007] von Salomé, J., Gyllensten, U., and Bergström, T. F. (2007). Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. *Immunogenetics*, 59(4) : 261–271.

[Wade et al., 2007] Wade, J. A., Katovich Hurley, C., Takemoto, S. K., Thompson, J., Davies, S. M., Fuller, T. C., Rodey, G., Confer, D. L., Noreen, H., Haagenson, M., Kan, F., Klein, J., Eapen, M., Spellman, S., and Kollman, C. (2007). HLA mismatching within or outside of cross-reactive groups (CREGs) is associated with similar outcomes after unrelated hematopoietic stem cell transplantation. *Blood*, 109(9) : 4064–4070.

[Wakeland et al., 1990] Wakeland, E. K., Boehme, S., She, J. X., Lu, C.-C., McIndoe, R. A., Cheng, I., Ye, Y., and Potts, W. K. (1990). Ancestral polymorphisms of MHC class II genes : Divergent allele advantage. *Immunologic Research*, 9(2) : 115–122.

[Wang et al., 2012] Wang, B., Ekblom, R., Strand, T. M., Portela-Bens, S., and Höglund, J. (2012). Sequencing of the core MHC region of black grouse (Tetrao tetrix) and comparative genomics of the galliform MHC. *BMC Genomics*, 13(1) : 553.

[Watkins et al., 1992] Watkins, D. I., McAdam, S. N., Liu, X., Strang, C. R., Milford, E. L., Levine, C. G., Garber, T. L., Dogon, A. L., Lord, C. I., Ghim, S. H., Troup, G. M., Hughes, A. L., and Letvin, N. L. (1992). New recombinant HLA-B alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. *Nature*, 357(6376) : 329–333.

[Watterson, 1996] Watterson, G. (1996). Motoo Kimura's Use of Diffusion Theory in Population Genetics. *Theoretical Population Biology*, 49(2) : 154–188.

[Wegner et al., 2004] Wegner, K. M., Kalbe, M., Schaschl, H., and Reusch, T. B. (2004). Parasites and individual major histocompatibility complex diversity—an optimal choice ? *Microbes and Infection*, 6(12) : 1110–1116.

[Wegner et al., 2003] Wegner, K. M., Reusch, T. B. H., and Kalbe, M. (2003). Multiple parasites are driving major histocompatibility complex polymorphism in the wild : Multiple parasites drive MHC polymorphism. *Journal of Evolutionary Biology*, 16(2) : 224–232.

[Weir and Cockerham, 1984] Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6) : 1358–1370.

[White et al., 2009] White, T. D., Asfaw, B., Beyene, Y., Haile-Selassie, Y., Lovejoy, C. O., Suwa, G., and WoldeGabriel, G. (2009). Ardipithecus ramidus and the Paleobiology of Early Hominids. *Science*, 326(5949) : 64–64, 75–86.

[Wilkinson et al., 2011] Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R. D., Yang, Z., and Tavaré, S. (2011). Dating Primate Divergences through an Integrated Analysis of Palaeontological and Molecular Data. *Systematic Biology*, 60(1) : 16–31.

[Winter and Long, 1997] Winter, C. C. and Long, E. O. (1997). A single amino acid in the p58 killer cell inhibitory receptor controls the ability of natural killer cells to discriminate between the two groups of HLA-C allotypes. *Journal of Immunology (Baltimore, Md. : 1950)*, 158(9) : 4026–4028.

[Witter et al., 2007] Witter, K., Mautner, J., Albert, T., Zahn, R., and Kauke, T. (2007). HLA-DQB1*0319, a novel HLA-DQB1 allele, shows strong haplotype association to HLA-DRB1*1102. *Tissue Antigens*, 70(1) : 73–75.

[Wolfe, 2001] Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, 2(5) : 333–341.

[Worley et al., 2010] Worley, K., Collet, J., Spurgin, L. G., Cornwallis, C., Pizzari, T., and Richardson, D. S. (2010). MHC heterozygosity and survival in red junglefowl : MHC AND SURVIVAL IN JUNGLEFOWL. *Molecular Ecology*, 19(15) : 3064–3075.

[Wright, 1949] Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, 15(1) : 323–354.

[Wright, 1965] Wright, S. (1965). The interpretation of population structure by f-statistics with special regard to systems of mating. *Evolution*, 19(3) : 395–420.

[Yagüe et al., 1998] Yagüe, J., Vázquez, J., and Castro, J. L. (1998). A single amino acid change makes the peptide specificity of B*3910 unrelated to B*3901 and closer to a group of HLA-B proteins including the malaria-protecting allotype HLA-B53. *Tissue Antigens*, 52(5) : 416–421.

[Yamazaki et al., 2011] Yamazaki, A., Yasunami, M., Ofori, M., Horie, H., Kikuchi, M., Helegbe, G., Takaki, A., Ishii, K., Omar, A. H., Akanmori, B. D., and Hirayama, K. (2011). Human leukocyte antigen class I polymorphisms influence the mild clinical manifestation of Plasmodium falciparum infection in Ghanaian children. *Human Immunology*, 72(10) : 881–888.

[Ye et al., 2012] Ye, Q., He, K., Wu, S.-Y., and Wan, Q.-H. (2012). Isolation of a 97-kb Minimal Essential MHC B Locus from a New Reverse-4D BAC Library of the Golden Pheasant. *PLoS ONE*, 7(3) : e32154.

[Yoshida et al., 1992] Yoshida, M., Kimura, A., Numano, F., and Sasazuki, T. (1992). Polymerase-chain-reaction-based analysis of polymorphism in the HLA-B gene. *Human Immunology*, 34(4) : 257–266.

[Yuhki, 2003] Yuhki, N. (2003). Comparative Genome Organization of Human, Murine, and Feline MHC Class II Region. *Genome Research*, 13(6) : 1169–1179.

[Yuhki et al., 2007] Yuhki, N., Beck, T., Stephens, R., Neelam, B., and O'Brien, S. J. (2007). Comparative Genomic Structure of Human, Dog, and Cat MHC : HLA, DLA, and FLA. *Journal of Heredity*, 98(5) : 390–399.

[Yuhki et al., 2008] Yuhki, N., Mullikin, J. C., Beck, T., Stephens, R., and O'Brien, S. J. (2008). Sequences, Annotation and Single Nucleotide Polymorphism of the Major Histocompatibility Complex in the Domestic Cat. *PLoS ONE*, 3(7) : e2674.

[Zagalska-Neubauer et al., 2010] Zagalska-Neubauer, M., Babik, W., Stuglik, M., Gustafsson, L., Cichoń, M., and Radwan, J. (2010). 454 sequencing reveals extreme complexity of the class II Major Histocompatibility Complex in the collared flycatcher. *BMC Evolutionary Biology*, 10(1) : 395.

[Zangenberg et al., 1995] Zangenberg, G., Huang, M.-M., Arnheim, N., and Erlich, H. (1995). New HLA–DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. *Nature Genetics*, 10(4) : 407–414.

[Zemmour and Parham, 1992] Zemmour, J. and Parham, P. (1992). Distinctive polymorphism at the HLA-C locus : implications for the expression of HLA-C. *The Journal of Experimental Medicine*, 176(4) : 937–950.

[Zheng et al., 2014] Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., and Weir, B. S. (2014). HIBAG—HLA genotype imputation with attribute bagging. *The Pharmacogenomics Journal*, 14(2) : 192–200.

[Zhu et al., 1996] Zhu, X., Zhao, X., Burkholder, W. F., Gragerov, A., Ogata, C. M., Gottesman, M. E., and Hendrickson, W. A. (1996). Structural Analysis of Substrate Binding by the Molecular Chaperone DnaK. *Science*, 272(5268) : 1606–1614.

[Černý et al., 2018] Černý, V., Kulichová, I., Poloni, E. S., Nunes, J. M., Pereira, L., Mayor, A., and Sanchez-Mazas, A. (2018). Genetic history of the African Sahelian populations. *HLA*, 91(3) : 153–166.

[Černý et al., 2011] Černý, V., Pereira, L., Musilová, E., Kujanová, M., Vašíková, A., Blasi, P., Garofalo, L., Soares, P., Diallo, I., Brdička, R., and Novelletto, A. (2011). Genetic Structure of Pastoral and Farmer Populations in the African Sahel. *Molecular Biology and Evolution*, 28(9) : 2491–2500.

Liste des Figures

Page

Liste des Tables

Liste des Équations

Page

Liste des communications

— **2016 :** Congrès annuel de l'EFI [1] (Kos, Grèce) : "From SSO to NGS in HLA population studies : the case of Mandenka (Senegal)" [poster et présentation orale] ;
— **2017 :** Congrès annuel de l'EFI (Mannheim, Allemagne) : "The use of entropy to describe HLA molecular information" [présentation orale]
— **2017 :** Conférence annuelle "Biology 17" (Bern, Suisse) : "From SSO to HTS in HLA population studies : Diving into the fine nucleotide diversity of HLA genes", [poster]
— **2018 :** Conférence annuelle "Biology 18" (Neuchâtel, Suisse) : "Deciphering the fine nucleotide diversity of full HLA genes in a sub-Saharan African population", [poster]
— **2018 :** Congrès annuel de l'EFI (Venise, Italie) : "Markov chain modelling brings new clues on HLA evolutionary history", [poster]
— **2019 :** Geneva Science Seminar (Genève, Suisse) "In-Silico analysis of HLA genes in the context of Population Genetics" [présentation orale]
— **2019 :** Congrès annuel de l'EFI (Lisbonne, Portugal) : "Fine-scale DNA sequence and Linkage Disequilibrium analysis of 11 HLA genes suggests a dual origin of the Vietnamese Cham : an example of population study based on NGS" [poster et présentation orale]

1. *European Federation for Immunogenetics*

Liste des matériels supplémentaires

Chapitre 2 :
— S-21 : Diversité en acides aminés
— S-22 : Statistiques de diversité et tests de neutralité sélective
— S-23 : Haplotypes en déséquilibre de liaison
— S-24 : Donneurs potentiels pour la conversion allélique
— S-25 : Analyses en Composantes Principales (axes 1 et 3)
— S-26 : Fréquences alléliques
— S-27 : Comparaison des typages
— S-28 : Génotypes

Chapitre 3 :
— S-31 : Séquences de référence
— S-32 : Fréquence des dinucléotides pour HLA-DRB1 exon 2
— S-33 : Numéros d'accession GENBANK
— S-34 : Fichier de résultat de MADaM
— S-35 : Rapport des réplicats
— S-36 : Séquences des variants identifiés

Chapitre 4 :
— S-41 : Sequences nucléotidiques
— S-42 : Corrélations des variables avec la pfpr2000
— S-43 : Fréquences alléliques
— S-44 : Correspondances exons2/alleles
— S-45 : Déséquilibres de liaison
— S-46 : Tests d'équilibre de Hardy-Weinberg
— S-47 : Tests de Tajima
— S-48 : Analyses Factorielles des Correspondances

Chapitre 5 :
— S-51 : Numéro d'accession GENBANK des séquençes de chimpanzés
— S-52 : Entropies estimée à chaque locus
— S-53 : Modèles lineaires de la distribution de l'entropie
— S-54 : Noms des allèles de chacun des groupes d'exon 2 de DPB1

Le matériel supplémentaire est disponible à l'adresse suivante : `https://gitlab.unige.ch/Thomas.Goeury/phd-thesis-2020`.