Impact of genetic variation on gene expression and cellular phenotypes

Real, Aline

UNIVERSITÉ DE GENÈVE

Département de Médecine

Département de Médecine génétique et Développement

FACULTÉ DE MÉDECINE

Prof. Jörg D. Seebach

Prof. Emmanouil T. Dermitzakis

NEWCASTLE UNIVERSITY

Bioscience Institute

FACULTY OF MEDICAL SCIENCES

Dr. Ana Viñuela

Impact of genetic variation on gene expression and cellular phenotypes

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 Génomique et santé numérique

par

Aline Réal

de

Aoste (Italie)

Thèse N° 182

GENÈVE

2022

LIFE SCIENCES PhD SCHOOL GENEVA

DOCTORAT ÈS SCIENCES EN SCIENCES DE LA VIE DES FACULTÉS DE MÉDECINE ET DES SCIENCES MENTION GÉNOMIQUE ET SANTÉ DIGITALE

Thèse de Mme Aline Real

intitulée :

« Impact of genetic variation on gene expression and cellular phenotypes »

Les Facultés de médecine et des sciences, sur le préavis de Monsieur Jorg Dieter SEEBACH, Professeur ordinaire et directeur de thèse (Département de médecine), Madame Ana VINUELA, Professeure assistante et co-directrice de thèse (Institut de Biosciences, Université de Newcastle, U.K), Monsieur Emmanouil DERMITZAKIS, Professeur ordinaire et co-directeur de thèse (Département de médecine génétique et développement), Monsieur Guillaume ANDREY, Professeur assistant et Président du jury (Département de médecine génétique et développement), Madame Kerrin SMALL, Docteure (King's College, Londres, Angleterre) autorisent l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 26 août 2022

Thèse - 182 -

Le Doyen
Faculté de médecine

Le Décanat
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".

TABLE OF CONTENTS

Acknowledgments

I would like to express my gratitude first to my thesis directors: Prof. Manolis Dermitzakis and Prof. Jörg Seebach. Thank you for providing me with the opportunity to work on such stimulating projects, surrounded by a wonderful team of scientists. I am grateful for your encouragements and feedbacks which helped me accomplished more than I thought was possible and matured professionally and personally. I am deeply grateful to Ana Viñuela who has been closely supervising my work throughout the last year and has been a supportive presence throughout the whole PhD. You worked by my side helping me to gain confidence in myself and you played a substantial role in my growth as a scientist.

A big thank to Gisella Puga Yung, who followed my progress embodying that interest and enthusiasm that make great scientists. With your persistence, you pushed me out of my comfort zone and encouraged me "to never give up" and to always trust in my abilities.

I would like to thank the members of my thesis committee, Prof. Kerrin Small and Prof. Guillaume Andrey, I am grateful for the time they put into revising and evaluating my thesis. Guillaume, I would like to express my gratitude for your help during this last challenging year and for your insightful scientific perspectives.

A special thank goes to Ana, Gisella, Andrew, and Anna who assist me with the writing and the revision of the present thesis and figures. Thank you, Anna, for your friendship and for stepping up when I most needed it. Thanks to your positive vibes, I ended up loving those coding session in which we worked side by side. Thank you, Andrew, for your precious help, questioning and challenging my scientific thinking extensively contribute to my research achievements. If I was able to accomplish this journey and will soon be able to start another, I also owe it to the friendship and the precious guidance of the magic Ana-Andrew team. To my special friend Ambra, a simple thank will never be enough to sum up what we shared during these years of PhD. Tradition like the "Big Mardi" were essential to steam off, discuss science outside the office and realize how much we enjoy what we are doing. *E' stato un viaggio difficile, emozionante ed immensamente gratificante che sono fiera di aver potuto condividere con te*. To Marta, thank you for your friendship, kindness, and team player spirit, you have always been there in the good and the bad moments, listening to me, advising me, and making me feel better. Thanks for including me among your closest friends and never let me feel lonely and for your help with some of the thesis figures. Thank you Rebecca, for your affection and for letting me rely on your unique coffee-trust moments. Ancilla, thank you for all you did for me as

the special funpopgen mum. Merci! From the very beginning, you made me feel welcome. Christelle, thank you for your support in the experimental part of both my projects. Nikos, thank you for your assistance with codes and for always being willing to help. Maris, Halit, Josefina, Diana, Luciana, Pauline, Chloé, Maria, Sara, Thanos, Thao, Viktoriia, Daniela, Théo, David, Dafni, and Sulaiman, working and discussing with you has been a pleasure, thanks for making me feel like I'm a part of a big family. Thank you also to the Andrey's lab members for welcoming me in their lab meeting and their group during this last year.

A very special thanks to all my friends, the ones that have been in my life from the beginning and the new ones who I was lucky to meet in Geneva. *A Karen, Fanny e ai "Fantastici 5" grazie di aver sempre creduto in me dandomi la forza di seguire i miei sogni. Katja, merci pour ton amitié sincère et ta présence précieuse.*

Last but not least, thanks to my family who has always been there for me and has been my rock even from afar. *Maman, grazie per aver sempre trovato il modo di trasmettermi la tua forza e perseveranza, per avermi ricordato che nessun ostacolo è insormontabile e nessun obbiettivo irraggiungibile. Ton amour m'a soutenue et réconfortée tout au long de ces importantes années. Grazie a Massimo per essere diventato parte integrante della famiglia ed essere sempre disponibile e presente nella quotidianità come nei momenti importanti. Merci à ma grande famille de Genève, Eric, Eva, Max, Stella et Grami, vous avez été une source d'inspiration et de confort continue tout au long de mon doctorat, je suis très reconnaissante d'avoir partagé ces importantes années avec vous. Eric, nos 'brainstorming' sur la terrasse resteront toujours des moments précieux à garder dans mon cœur.*

E per concludere grazie al mio big brother Mat. Mi hai instillato la passione per il mio lavoro sin dai primi anni di questo percorso. Sei sempre la persona a cui rivolgermi per ispirazione, consigli, sfoghi e decisioni importanti; grazie di avermi spronato a raggiungere i miei obbiettivi con il tuo esempio, la tua guida e il tuo incondizionato affetto.

Un ultimo ringraziamento lo dedico a chi è al mio fianco ogni giorno, nel mio cuore.

Abbreviations

7-AAD	7-amino-actinomycin D
ASE	allele-specific expression
ASTS	allele-specific transcript structure
AS	alternative splicing
BMI	body mass index
bp	base pairs
BRCA1/2	breast cancer gene 1 and 2
BSA	bovine serum albumin
DNA	deoxyribonucleic acid
cDNA	complementary deoxyribonucleic acid
CEU	Utah residents with Northern and Western European ancestry
CNV	copy-number variation
CTV	cell tracer violet
dRNA-seq	direct RNA-sequencing
EBV	Epstein-Barr Virus
eQTL	expression quantitative trait loci
ESE	exonic splicing enhancers
ESS	exonic splicing silencers
FACS	fluorescence-activated single cell sorting
FasL	Fas ligand
FBS	fetal bovine serum
FDR	false discovery rate
FIN	Finnish in Finland
FSC	forward scatter
GBR	British from England and Scotland
GEUVADIS	Genetic European Variation in Disease
GTEx	Genotype-Tissue Expression
GWAS	Genome-Wide Association Studies

Indels	insertion-deletion mutations
ISE	intronic splicing enhancers
ISS	intronic splicing silencers
Kb	Kilobases
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
LINC00539	Long Intergenic Non-Protein Coding RNA 539
lincRNA	long intergenic non-coding ribonucleic acid
lncRNA	long non-coding ribonucleic acid
MAF	minor allele frequency
MFI	mean fluorescent intensity
mRNA	messenger ribonucleic acid
NGS	next generation sequencing
NHGRI	National Human Genome Research Institute
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PB	pacific blue
POLE4	DNA Polymerase Epsilon 4
Pre-mRNA	precursor-messenger ribonucleic acid
PSI	percentage spliced
PTL	post-transplant lymphoma
PTLD	post transplantation lymphoproliferative disorder
qPCR	quantitative polymerase chain reaction (real-time)
RNA	ribonucleic acid
RNAPolII	RNA polymerase II
RNA-seq	ribonucleic acid sequencing
rRNA	ribosomal ribonucleic acid
RPKM	reads per kilobases per million
SKA3	Spindle And Kinetochore Associated Complex Subunit 3
SMRT	single molecule real-time

SNP	single-nucleotide polymorphism
snRNP	small nuclear ribonucleoprotein
sQTL	splice quantitative traits loci
SRA	short-read archive
SS	splice site
SSC	size scatter
SV	structural variant
T	thymine
TCGA	The Cancer Genome Atlas
TF	transcriptional factor
TNF-R	tumor necrosis factor receptor
TPM	transcripts per million
tRNA	transfer ribonucleic acid
trQTL	transcript quantitative trait loci
TSI	Toscani in Italy
TSS	transcription start site
U	uracil
WES	whole exome sequencing
WGS	whole genome sequencing
YRI	Yoruba in Ibada, Nigeria

Abstract

Complex traits and diseases, such as cancer, are determined by a combination of genetic and environmental factors, which makes them difficult to study. While genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex traits and diseases, the causal genes, and therefore, the underlying diseases-driving processes remain mostly unknown. This thesis aims to contribute to close this gap by unravelling the effect of genetic variation on gene expression, as well as on cellular cancer-like phenotypes.

In the first project, the effects of genetic variation on gene expression were addressed using direct long-read RNA sequencing. RNA alternative splicing (AS) regulates gene expression, and ultimately the proteome, by allowing the production of multiple mRNA molecules from a single gene. AS is a key component of gene regulation because it transforms one RNA molecule into multiple transcripts with different structures causing variation in gene expression levels. It was shown that many genome-wide associations for common diseases are affecting splicing processes by mechanisms that are distinct from the ones affecting gene expression. However, technology limitations made it difficult to determine the genetic effects on AS events on a genome-wide scale. To date, the majority of studies have used short-read sequencing technologies, employing proxy procedures to determine the structure, quantification, and alternative splicing characteristics of transcripts. During the past few years of the post-genomics era, we have been witnessing the emergence of new technologies specifically developed to fill these gaps, such as long-read sequencing. Direct RNA long-read sequencing of 60 Lymphoblastoid cell lines (LCLs) was used to detect annotated and novel transcripts and to identify genetic variants affecting transcript expression and structure. Previously identified expression quantitative trait loci (eQTL) effects, i.e. were characterized more closely by the identification of the transcripts affected by the eQTL among all the ones produced by the same gene. Moreover, we discovered novel transcripts QTLs (trQTL) not included in the eQTLs previously reported using short-read sequencing.

In the second project, the effects of genetic variation on cancer-like phenotypes were studied by designing a population-level study. Historically cancer believed to be linked to somatic driver mutations but, in recent years, it has become increasingly clear from GWAS that non-coding regulatory drivers also play a critical role in cancer development and progression. Specifically, inherited genetic variants (germline variants) might increase the risk of developing cancer. To gain a deeper understanding of the regulatory germline contribution to cancer genetics, in particular to identify which target genes are involved, 87 genetically different LCLs were tested to assess cells' proliferation, apoptosis, and chemotaxis using functional *in-vitro* assays. The aim was to investigate whether specific variants and genes are associated with these three cancer-like phenotypes, i.e. whether they affected the cell line to proliferate, resist apoptosis, and migrate in response to chemical stimuli. Despite the relatively small sample size (n = 87), we were able to identify a significant genetic variant associated with the replication index phenotype and multiple putative genes mediating this genetic effect. This SNP was already reported as an eQTL affecting the long intergenic non-coding RNA (lincRNA) *LINC00539*, which was previously linked with tumor immune response in lung cancer.

The evidence from GWAS that non-coding genetic variants are associated with splicing, and ultimately with cancer, underlines the need to focus on filling the gap still present between variation in the non-coding genome and downstream effects on gene expression. With this thesis, I attempted to obtain a deeper understanding of the genetic mechanisms linking genetic variants, splicing and complex traits such as cancer. Integrating genetic information with gene expression data, generated with the novel long-read sequencing technology, phenotype and functional data, might open new way of study complex diseases, and provide new avenues for diagnostic and therapeutic approaches.

Résumé

Les traits complexes et les maladies, comme par exemple le cancer, sont déterminés par une combinaison de facteurs génétiques et environnementaux qui les rend difficiles à étudier.

Alors que les études GWAS ont identifié des milliers de variantes génétiques associées à des traits et des maladies, les gènes responsables et donc les processus sous-jacents à l'origine de la maladie restent pour la plupart inconnus. Cette thèse vise à fournir une contribution pour combler ces lacunes en dévoilant l'effet des variations génétiques sur l'expression génique ainsi que sur les phénotypes cellulaires de type cancéreux.

Au cours du premier projet, nous étudions les effets des variations génétiques sur l'expression des gènes en utilisant le séquençage direct, à longue lecture (long-reads) de l'ARN. Le mécanisme d'épissage alternatif de l'ARN régule l'expression des gènes et enfin du protéome en permettant la production de plusieurs molécules d'ARN messager à partir d'un seul gène. C'est un élément clé de la régulation génique car il transforme une molécule d'ARN en plusieurs transcrits avec différentes structures causant ainsi une variation des niveaux d'expression des gènes. Il a déjà été démontré que, pour des maladies courantes, de nombreuses associations à l'échelle du génome influencent le processus d'épissage par des mécanismes distincts de ceux qui agissent sur l'expression des gènes. Cependant, les limites de la technologie ont rendu difficile, pour les chercheurs, la détermination des effets génétiques sur les événements d'épissage alternatif à l'échelle du génome. Jusqu'à présent, dans la plupart des études, on a utilisé des technologies de séquençage à courtes séquences (short-reads), en utilisant des procédures de proxy pour déterminer la structure, la quantification et les caractéristiques d'épissage alternatif des transcrits. Dans l'ère post-génomique, au cours des dernières années, nous avons assisté à l'apparition de nouvelles technologies de séquençages spécifiquement développées pour combler ces manques, comme le séquençage à longue lecture. Dans cette étude, nous utilisons le séquençage direct d'ARN à longue lecture de 60 lignées cellulaires lymphoblastoïdes (LCL) pour détecter aussi bien les transcrits déjà annotés que les nouveaux de même que pour

identifier les variantes génétiques affectant l'expression et la structure de ces transcrits. Nous avons été à même de caractériser avec précision les effets de eQTL (locus de caractères quantitatifs) précédemment identifiés, c.-à-d. identifier les transcrits affectés par l'eQTL parmi tous ceux produits par le même gène. De plus, nous avons découvert de nouveaux transcrits QTLs (trQTL) non inclus dans les eQTLs précédemment décrites avec le séquençage à courte lecture (short-reads).

Dans le deuxième projet, nous étudions les effets de la variation génétique sur des phénotypes qui recapitulent des caractéristiques du cancer en concevant une étude au niveau de la population. Le cancer est généralement décrit comme une maladie liée aux mutations somatiques mais, durant ces dernières années, il est devenu de plus en plus clair, à partir de GWAS, que les facteurs régulateurs non-codants jouent également un rôle crucial dans le développement et la progression du cancer. Plus précisément, l'expression des gènes impliqués dans le développement de la tumeur peut être modifiée par des variantes génétiques héréditaires (variantes germinales) qui augmenteront le risque de développer un cancer. Afin de mieux comprendre la contribution des variantes germinales pour la génétique du cancer, et en particulier d'identifier les gènes cibles impliqués, nous avons mis en place trois essais fonctionnels *en vitro* sur 87 LCL génétiquement différentes pour évaluer la prolifération des cellules, l'apoptose et la chimiotaxie. Nous avons investigué pour savoir si des variants génétiques et des gènes spécifiques étaient associés à ces trois phénotypes cancéreux et s'ils affectaient la capacité de prolifération des lignées cellulaire, de résistance à l'apoptose et de migration en réponse à des stimuli chimiques. Malgré la petite taille de la population étudiée (n = 87), nous avons été en mesure d'identifier une variante génétique associée au phénotype de l'indice de réplication cellulaire ainsi que de multiples gènes putatifs comme médiateurs de cet effet génétique. Les évidences produites par les GWAS que les variantes génétiques non codantes sont associées à l'épissage et finalement à des maladies comme le cancer, soulignent la nécessité de se concentrer sur le comblement de l'écart encore présent entre la variation du génome non codant et son effet sur l'expression des gènes.

Avec cette thèse, nous avons tenté de combler cette lacune en intégrant l'information génétique avec les données d'expression génétique générées grâce à la nouvelle technologie de séquençage à longue lecture et les données phénotypes afin d'obtenir une compréhension plus approfondie des mécanismes génétiques reliant les variantes génétiques, l'épissage et les traits complexes comme le cancer.

Introduction

1. The Human Genome

The human genome contains the instructions for all the cells of an organism to build a functional human being. The genetic information is stored in the genome in the form of deoxyribonucleic acid molecules, or DNA, and is inherited from one's parents, ensuring the passing of the information through generations. Humans are diploid organisms, meaning each person inherits two copies of the human genome, one from each parent, and most of the DNA is organized as tightly coiled structure divided into 23 chromosome pairs. The size of the human genome is ~3.2 billion per haploid genome [1] and is composed of DNA sequences that are responsible for initiating and regulating the activity of genes and the production of proteins. Based on the type of information it codes, the human genome can be divided into *coding* and *non-coding* regions. The *coding* DNA is ~1.2% of the total human genome and it is composed of *genes,* i.e. DNA sequences that contain all the information necessary for the production of proteins. The other ~98.8% of the genome is known as the *non-coding* genome, mostly having a regulatory function without coding for any protein. In any individual human, the vast majority of the genetic sequence in all of the cells is identical; which adapts cells to their role and environment is differences in the regulation of *gene expression*, the process of converting DNA instructions into functional proteins.

In this section, I will provide a brief overview of the process of gene expression with a focus on how gene expression is quantified with high throughput techniques and new technologies. Then I will explain the genetic variation and how, together with gene expression, this can be used to understand complex traits and diseases. Furthermore, I will give an overview of how genetic variation is associated with diseases, with special emphasis on cancer, and describe how the regulation of gene expression may be a cause of susceptibility to cancer development. Finally, I will outline the aims of the present thesis.

2. Gene expression

Gene expression is the process in which the information encoded in a gene is *transcribed* from a DNA template to a molecule of RNA. This RNA can be "read" and *translated* into a sequence of amino acids that constitute a functional molecule called protein, or became a functional non-coding RNA [2]. The RNA used as a protein template is called messenger RNA (mRNA), and it is transferred into the cytoplasm before *translation*. Other RNA molecules, such as transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA, and long non-coding RNA (lncRNA), can play a role in gene regulation or have others, not completely understood, functions [3].

2.1 Transcription

Transcription is the first step of gene expression, and it occurs in the nucleus when nascent RNA molecules are produced by copying the gene's DNA (**Figure 1**). It starts by making the DNA more accessible to transcription factors (TFs), activators and co-activators close to the transcriptional start site (TSS) of the gene. This allows the recruitment of RNA polymerase II (RNAPolII), an enzyme which catalyzes the transcription, in the proximity of the TSS. In the process of elongation of the RNA molecule, the RNAPolII moves along the gene producing a single strand precursor-messenger RNA (pre-mRNA) molecule in the direction 5-prime (5'-) to 3-prime (3'-) by copying the DNA molecule [4, 5]. In the meantime, the pre-mRNA undergoes post-transcriptional modifications such as the 5' capping, the 3' cleavage and polyadenylation that protect the molecule from degradation, help transport the molecule to the cytoplasm and translation into a protein [4].

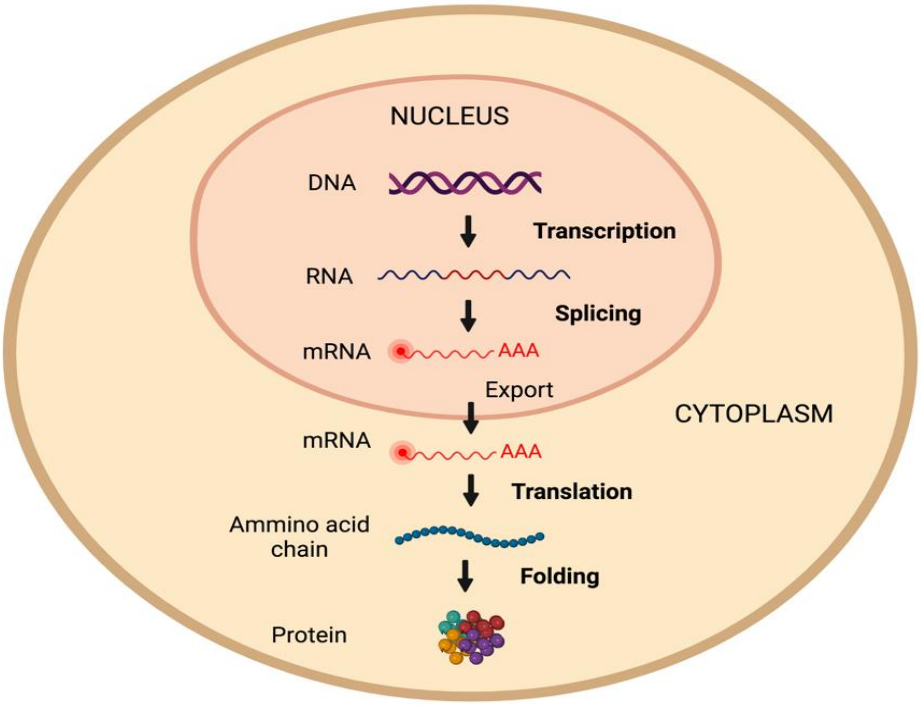


Figure 1. Transmission of information from DNA to proteins. One strand of DNA is transcribed into RNA, which then undergoes a maturation process involving splicing and the addition of poly-A tails. The generated mRNA is exported from the nucleus to the cytoplasm where it is translated into proteins. (Created with BioRender.com).

2.2 Alternative splicing

The pre-mRNA is the first form of RNA created through transcription and it consists of exons, RNA sequences that encode the protein sequence information, and introns, *non-coding* sections of RNA trasncript. During splicing, the pre-mRNA molecules are transformed into mature mRNAs by removing the introns and ligating the exons together into transcripts [6]. Transcription and splicing are closely intertwined, occurring almost simultaneously in the nucleus, but splicing can also take place after transcription is completed and before the pre-mRNA is exported to the cytoplasm [7]. Throughout the mRNA splicing process, a single gene can generate multiple different transcripts, and in turn multiple different proteins, by selecting different combinations of exons. This process is known as *alternative splicing* [8].

To splice pre-mRNA, two major chemical reactions must occur; firstly the 5' exon must be cleaved from the intron and secondly, the intron must be removed, and the 5' and 3' exons joined. This process involves a complex interaction between local acting elements (*cis*-elements) in the pre-mRNA and protein-RNA distal acting factors (*trans*-factor) that bind to the *cis*-elements. The *spliceosome* is the complex of proteins responsible for the splicing of mRNA and catalyzes the removal of the introns from pre-mRNA. The spliceosome contains a large number of different proteins such as the small nuclear ribonucleoproteins (snRNPs) [9]. Among these snRNPs, the U1, U2, U4, U5, and U6 are involved in splicing by binding to specific sequences on the pre-mRNA and by recruiting other co-factors to form the spliceosome (see also **Figure 2**).

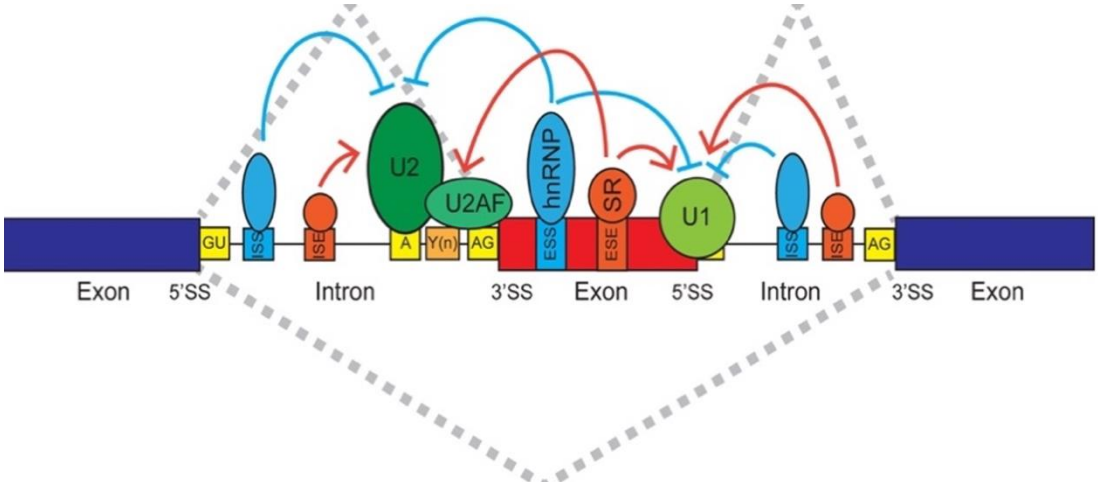


Figure 2. Multiple proteins interact to constitute the spliceosome to regulate the splicing of pre-mRNA molecules. The first and last two nucleotides of an intron are the highly conserved sequences GU and AG in the 5' and 3' splice sites, respectively (in yellow). The splicing process starts with the recognition of the 5' splice site (SS) by the U1 snRNP complex (light green), while the branch site is recognized by the U2 snRNP complex (deep green) and the U2AF proteins are recognized by the 3' splice site and polypyrimidine tract (green). Exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs) are pre-mRNA *cis*-regulatory motifs that recruit various RNA-binding proteins (e.g., SR and hnRNP proteins) to regulate alternative splicing. Blue boxes and red boxes represent the spliced and the alternatively spliced exons, respectively. The dashed grey lines show the three different splicing events that can occur between the three exons represented in the figure. (Taken from [10]).

Basic splicing events that can produce different transcripts include exon skipping, alternative 5’ and 3’ splice sites, intron retention, mutually exclusive exons, and use of alternative first or last exons [10] (**Figure 3**). Combinations of these alternative splicing events result in a variety of mRNA isoforms, RNA molecules that originate from the same locus but have different exon compositions and lengths. These different isoforms may code for different forms of proteins and therefore exhibit distinct regulatory functions within the cell [11]. Thus, alternative splicing offers a powerful tool for eukaryotic organisms to enhance their capability to regulate gene expression. As discussed in Section 4.3, abnormal alternative splicing can play a role in disease development.

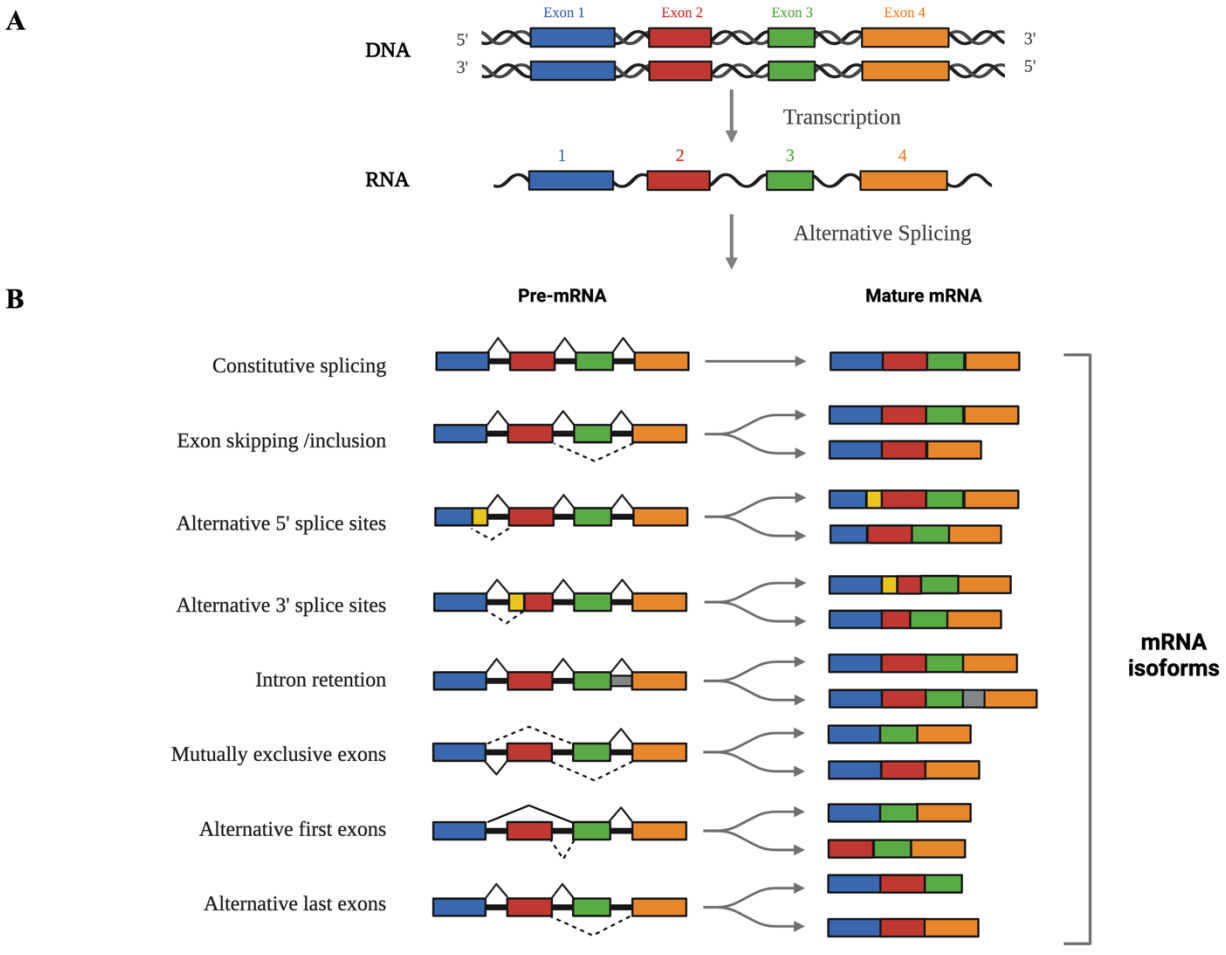


Figure 3. Overview of basic alternative splicing. (A) A pre-mRNA derived from a double-strand DNA shows four different exons. **(B)** Basic splicing events. Blue, red, green, and orange boxes represent different exons undergoing alternative splicing. The two yellow boxes represent part of exons alternative spliced, and the grey boxes are a part of intron retained in the mRNA isoform (Created with BioRender.com).

2.3 Translation

After maturation, the mRNA molecule moves to the cytoplasm where the ribosome initiates the *translation* process. The mRNA sequence is read and translated from a nucleic acid sequence to a polypeptide chain of amino acids to synthesize a protein. The genetic code is read in triplets of nucleotides called *codons*. Each codon specifies one of 20 amino acids, which are the molecular building blocks for proteins. The tRNA are molecules that read the codon sequence and ensure that the correct amino acid is inserted in the polypeptide chain [12]. After the amino acid chain is generated, a vast diversity of post-translational modifications may happen, with the most common being attaching different molecules such as lipids (lipidation), carbohydrates (glycosylation), phosphoryl groups (phosphorylation) to the protein chain [12].

3. Gene expression quantification

DNA sequence is constant across different cells and tissues and are the differences in gene regulation that allows cells to perform their specific role in a given context. Because of this, it is of interest to quantify the process of gene expression, i.e., the quantity of mRNA a particular tissue sample or a single cell produce within a specific context. Indeed, the ability to accurately quantify how many mRNA molecules each gene produces at a whole-genome scale has become an essential instrument in many different biological and clinical applications. Since the initial sequencing of the human genome, different technologies have emerged to do so.

3.1. Short-read RNA sequencing

Short-read RNA sequencing (RNA-seq) is a Next-Generation Sequencing technology (NGS) widely used to profile the transcriptome, identifying and quantifying transcripts [13]. RNA-seq offers a powerful way to analyze global gene expression patterns, capable of identifying novel transcripts and gene fusions [14-17]. In RNA-seq, a library of complementary DNA (cDNA) fragments is created with adapters attached to one or both ends of RNA molecules. Every molecule is then sequenced to

generate short fragments from one end (single-end sequencing) or both ends (paired-end sequencing) [13]. These fragments are called *"reads"* and have a typical length ranging between 30 and 300 base pairs (bp), and the technology usually produces 20 – 100 million reads per sample [18].

To quantify the expression of each gene, the short-read sequences can be aligned to a reference genome or transcriptome (an accepted representation of the human genome sequence) or can be assembled *de novo* without a reference. This first step of alignment to a reference sequence is called *mapping*, and quantification of gene expression is produced by counting the number of reads that map to a particular gene. This technology has shown high specificity, sensitivity, and resolution. Furthermore, RNA-seq has a high degree of agreement with other transcriptomic techniques, such as quantitative real-time PCR (qPCR), both at absolute and relative levels of gene expression measurements [19]. Nowadays, the vast majority of methods for RNA-seq analysis were developed for the established Illumina short-reads sequencing platform [18]. Moreover, short-read sequencing supports a large range of applications such as *(i)* quantification of mRNA, rRNA and tRNA; *(ii)* identification of differential gene expression; *(iii)* detection of splicing events; and *(iv)* capture of post-transcriptional modifications, mainly polyadenylation and 5' capping, among others [20]. Notably, more than 95% of the published RNA-seq data on the Short Read Archive (SRA) was generated with the Illumina short-read sequencing technology [21].

The common practice of sequencing thousands of human genomes has shown that the genome is highly complex and contains a variety of long repetitive elements, copy number alterations, and structural variations that can be relevant to evolution, adaptation, and diseases [22]. However, RNA-seq using short-read technologies suffers from the inability to sequence long stretches of RNA, because it requires fragmentation and amplification of cDNA strands before sequencing. These short sequences are then assembled using computer algorithms that require substantial overlap between cDNA fragments to aid matching to the reference genome to identify their exact location. This process is complicated in the case of complex genomes containing repetitive regions, as the produced short-

read sequences can match multiple regions of the genome, making their exact location uncertain. Even when using sophisticated bioinformatic algorithms, it is often impossible to map or assemble short reads originating from regions with repetitive sequences, gene fusion or sequences consisting of multiple homologous elements within the genome [23, 24]. Thus, sequencing a highly complex and repetitive genome, such the human, can be challenging with these technologies [25]. Additionally, when multiple mRNA isoforms are generated by alternative splicing at a single locus, the process of short-read assembly to distinguish different isoforms can be difficult and error-prone, making it impossible to resolve connectivity between distant exons that are never represented on the same fragment [26, 27].

3.2. Long-read RNA sequencing

One proposed solution to the difficulties of sequencing complex and repetitive genomes and identifying mRNA isoforms is long-read sequencing, also called third-generation sequencing. Using this technology, it is possible to sequence very long DNA/ RNA stretches of the order of 30,000 bp. In general, long-reads provide superior performance for detecting transcriptomic structure and isoforms than short-reads, due to their ability to span the entire length of the transcript. This is important because, as revealed by a recent long-reads profile of the human transcriptome, novel isoforms of spliced genes can contribute more than 10% of the total number of reads sequenced [28]. When quantifying expression, long-reads allow us to determine the exact splice events which connect exons, something that is difficult with short-reads [29].

Currently, two long-read technologies dominate the field: PacBio single-molecule real-time sequencing (PacBio SMRT or PacBio) and Oxford Nanopore Technologies (ONT) sequencing [25], having different types of chemistry for production of long reads (< 30Kb) and ultra-long reads (> 1Mb) [25], (**Figure 4**). Both technologies produce reads that can cover the whole span of repetitive regions of the genome, but present essential differences affecting their read lengths, accuracy, and throughput [30].

3.2.1. Pacific Biosciences (PacBio)

PacBio permits the real-time measurement of fluorescent nucleotide incorporation during the elongation of the replicated DNA strand from a non-amplified single-strand template. Thus, in the process of DNA synthesis, incorporated nucleotides are detected via the accompanying fluorophore that is released and dispelled when the phosphate chain is cleaved. The PacBio technology generates reads between 10 – 25 Kb in length, far exceeding the read lengths generated by Illumina [31].

3.2.2. Oxford Nanopore Technology (ONT)

Alternatively, ONT long-read technology is based on sequencing a single linear DNA or RNA molecule that can be megabases long [32, 33]. In ONT sequencing, double-strand DNA or RNA molecules are attached to a sequencing adapter that is linked to a motor protein. The mixture is then loaded into flow cells, where hundreds or thousands of nanopores are embedded into a synthetic membrane. There, a motor protein separates the double-strand and drives the negative DNA or RNA strand through the pore, creating a disruption of the ionic current inside the pore as the molecule passes through. These disruptions to the current are recorded and '*translated*' into the base sequences in real-time [33]. Unlike other technologies, with ONT is possible to sequence directly native DNA or RNA molecules without PCR amplification and retro-transcription commonly required for other NGS technologies. In addition, read lengths of ONT can outstrip those of PacBio by at least an order of magnitude, with a typical length of 10 – 100 Kb and the base calls are 87 – 98% on average, depending on the algorithms used [34, 35].

Overall, and at the time of writing this thesis, ONT was more cost-effective for generating extremely long reads than PacBio, and also provided the ability to directly sequence DNA and RNA molecules, enabling the detection of base modifications and novel transcripts.

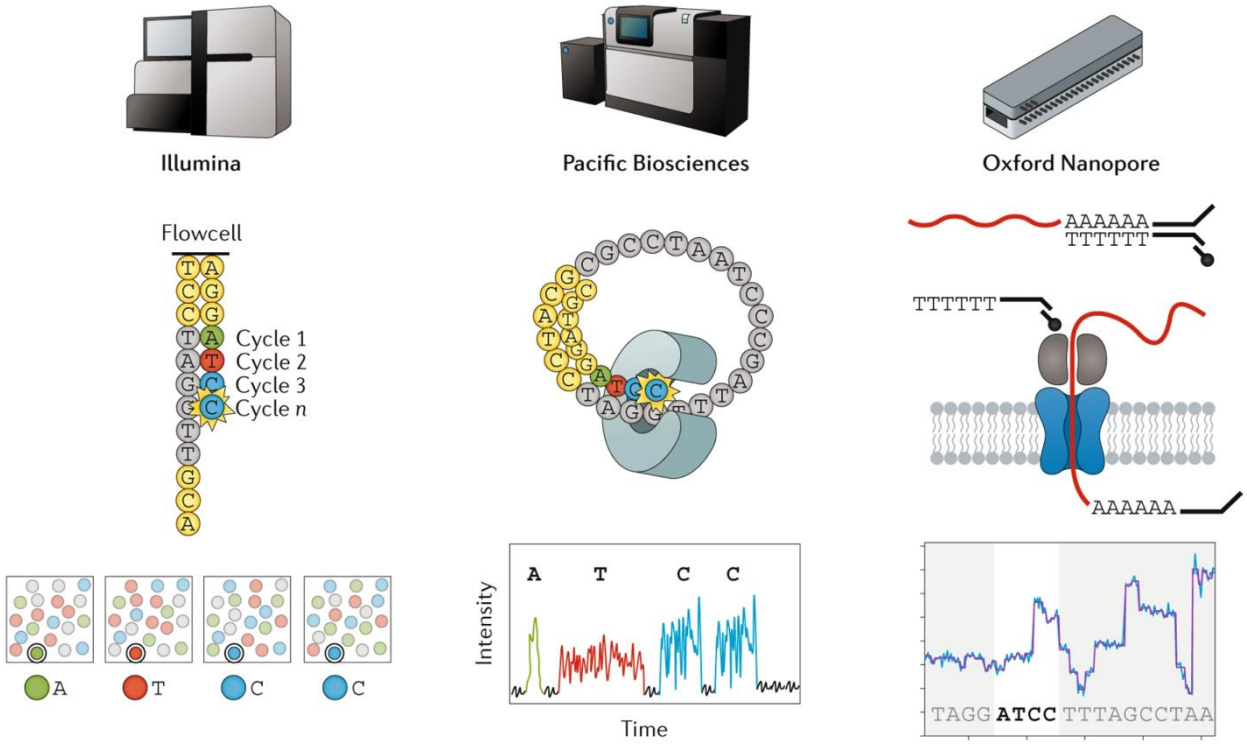


Figure 4. Workflows for Illumina, PacBio and ONT technologies. The Illumina workflow (left), following library preparation, individual cDNA molecules are clustered on a flowcell for sequencing by synthesis with fluorescently labelled nucleotides. In each sequencing round the DNA strand is elongated and fluorophores can be detected by imaging, where 50 – 500 bp reads can be obtained. The Pacific Biosciences workflow (middle panel), molecules are loaded into nano wells on a sequencing chip, where they are bound to immobilized polymerase. Fluorescently labeled nucleotides are incorporated into the growing strand fluoresce and detected, resulting in reads of up to 25 kilobases (Kb). The ONT workflow (right panel), the individual nucleic acid molecules are loaded into a flow cell, where motor proteins are docked to nanopores. RNA strands move through the nanopore due to the motor protein, causing a change in current that is processed into sequencing reads of 1 – 10 Kb (Taken from [18]).

3.2.3. Limitations of long-read sequencing

Long-read sequencing still presents some limitations compared to short-read sequencing, including lower accuracy per read [36]. Nanopore technology has a high error rate due to the inability to control the speed at which DNA/ RNA molecules pass through the pore, introducing systematic errors. In contrast, sequencing using PacBio produces random errors that can be partially corrected using a

circular consensus sequencing that forces DNA to pass through the waveguide chip several times. This technology can generate highly accurate reads of at least 99.8%, similar to short reads platforms [37, 38]. For short-reads, the error rate can be as low as 0.1%, but platform, chemistry, context, stringency of filtering, and other factors all influence read accuracy [39].

3.3. Direct RNA sequencing (dRNA-seq)

ONT can directly sequence native nucleic acid molecules for the detection of natural modifications in both DNA and RNA [40-43]. For direct-RNA sequencing (dRNA-seq) library preparation, a short cDNA strand is synthesized and an RNA – cDNA hybrid duplex serves to stabilize the RNA molecule. However, only the RNA strand passes through the pore during the sequencing process. The average accuracy for dRNA-seq is normally around 85 – 87% [44], but new bioinformatics tools can improve accuracy up to 98%, as reported by Oxford Nanopore [45]. As already mentioned, dRNA-seq avoids steps of PCR amplification and cDNA conversion, which means 6mA (RNA N^6-Methyladenine) modifications can be directly detected [41]. Moreover, dRNA-seq can measure full-length gene isoforms [46-48]. **Table 1** briefly summarizes the main characteristics for short-read Illumina technologies and dRNA long-read ONT technologies, highlighting the advantages, disadvantages and main applications of both technologies.

Table 1. Main characteristics of the long-read ONT dRNA-seq technologies compared to Illumina short-reads

Sequencing technology	Platform	Advantages	Disadvantages	Key applications
Short-read cDNA	Illumina	• Very high throughput: 100 - 1,000 times more reads per run then long-read platform • Biases and errors well understood • Huge catalogue of methods and analysis workflow	• Sample preparation include PCR, reverse transcription and size selection add biases • Limited isoforms detection and quantification • Limited transcripts discovery	Nearly all RNA-seq methods have been developed for short-red cDNA sequencing
Long-read Direct RNA	ONT	• Long-reads up to 20kb, full length transcript captured • *De novo* transcriptome analysis • Sample preparation without PCR and reverse transcriptions, reduce biases • RNA base modification detected	• Low to medium throughput: 50,000 to 1 x 10^6 reads per run • Sample preparation and sequencing bias not well understood for now • Need of high-quality RNA	• Well suited for isoforms and fusion transcript discovery, *de novo* transcriptome, and complex transcripts analysis • RNA modification detection (methylation)

4. Human genetic variation

Genetic variation is the difference in the DNA sequence between individuals of the same species. The human genomes of any two individuals are approximately 99.9% identical; and 90% of the DNA variation at the population level can be attributed to single-nucleotide polymorphisms (SNPs) that are the variations most commonly observed in populations [49]. Variants that are inherited are known as *germline variants* (e.g. present in the sperm or egg), while non-inherited variants that are generated during the lifespan of an individual are known as *somatic variants* or *somatic mutations* [50]. Somatic mutations that occur in germ cells and germline variants that are passed from parents to their offspring both affecting population dynamics. In genomics, the presence of two or more variant forms of a specific DNA sequence among individuals is referred to as a *polymorphism*. The different versions of a polymorphism located in the same coordinates of two DNA molecules are called alleles.

One of the main questions arising after the first draft of the human genome was "how does the human genome vary across individuals?" One early attempt to answer this was the HapMap consortium, which genotyped individuals from four different human populations (Japanese, Chinese, Yoruba and individuals with European ancestry) at a set of common SNPs known at the time [49, 51, 52]. With the development of NGS, it became possible to study not only known SNPs, but the entire genome. These advances led in 2008 to the 1000 Genomes Project, with the aim to sequence 1000 individuals from 14 populations from all over the world. The gathered data generated a catalog of human genetic variation and captured 98% of genetic variants which are common in human populations [53, 54]. The final phase of the 1000 Genomes project characterized 2,504 individuals from 26 populations, identifying 84.7 million SNPs, 3.6 million short indels (insertion-deletion mutations less than 1Kb in length), and 60,000 structural variants (larger than indels) [53]. The project concluded that genetic variants with low frequencies in particular populations were geographically distinct, and different populations have different profiles of common and rare genetic variation [53].

These studies provided accessible genotype information, which has been used to study population genetics and to better understand and characterize the genetic basis of disease. Much of this work was facilitated by the fact that the samples were collected completely anonymously, meaning that data and samples could be shared freely, with minimal privacy concerns. Similarly, the GEUVADIS consortium produced gene expression data which could also be accessed with no restrictions. GEUVADIS undertook pilot mRNA and small RNA sequencing projects on the data from lymphoblastoid cell lines (LCLs) produced from 462 samples collected by the 1000 Genomes project, from five different populations (CEU – Utah residents (CEPH) with Northern and Western European ancestry, FIN – Finnish in Finland, GBR – British from England and Scotland, TSI – Toscani in Italy and YRI – Yoruba in Ibadan, Nigeria) [55].

A detailed overview on the relevance of LCLs for population-based genetic studies will be given in Section 6, with a specific focus on its use as *in vitro* model to mimic cancer features. In particular, LCLs were also chosen as the experimental model in the two studies presented in the Result section of this thesis.

4.1.Complex traits and diseases

Complex genetic diseases are those caused by multiple genetic and environmental factors (e.g., diet, climate, lifestyle), with patterns of inheritance more complex than the single-gene model of Mendelian diseases [56-58]. To explain how genetic variation contributes to complex traits and diseases, the “Common Disease/ Common Variants” hypothesis proposed that, at the population level, the largest contribution to the genetic risk of developing a disease is caused by the effects of common variants [59]. This model was supported by results from HapMap and 1000 Genomes finding that most of the genetic differences within a population are driven by common genetic variants (Minor-allele frequency (MAF) > 5%) [49, 54].

Genome-Wide Association Studies (GWAS) are used to identify genomic variants that are statistically associated with a risk of disease or a particular trait [60]. More specifically, GWAS rely on the survey of the genomes of many people to determine whether there are genetic variants that occur more frequently in the people with a certain disease or trait (cases in blue in **Figure 5**) than in those without the disease or trait (controls in red in **Figure 5**). Typically, around 10 million SNPs from across the genome are tested for these differences in frequencies [61]. Manhattan plots, where each dot represents a different SNP, are commonly used to present GWAS results, with the *x*-axis being a chromosomal position and the *y*-axis the level of significance of the association (–log10 of the *p*-value) (**Figure 5B**).

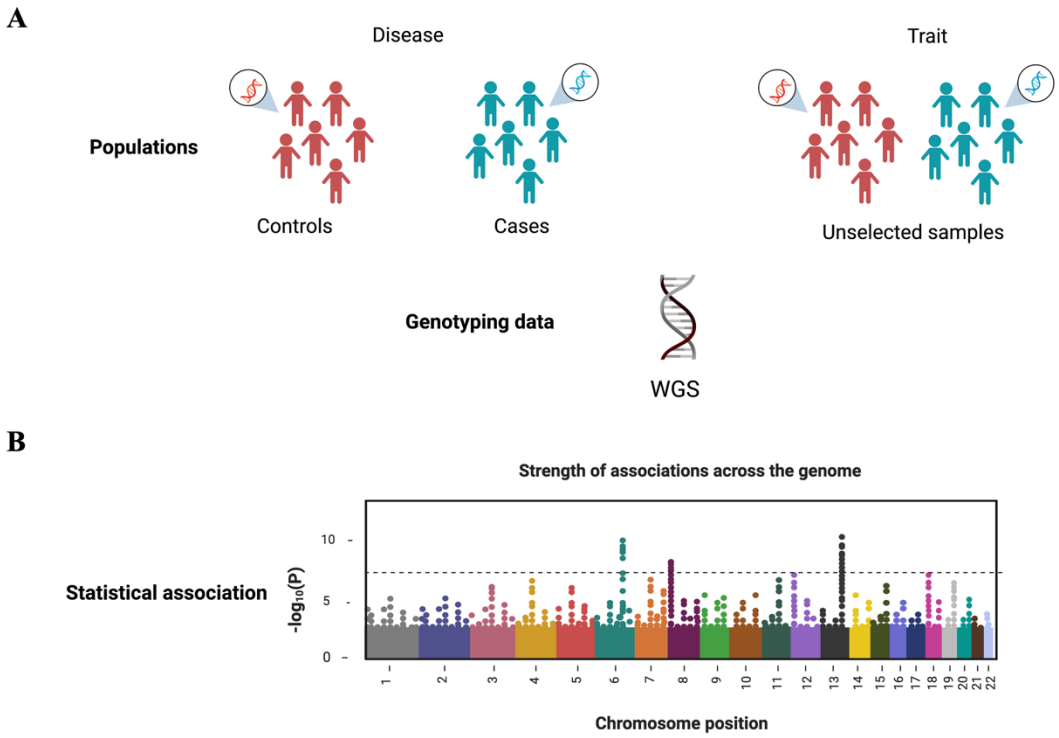


Figure 5. Study design for Genome-Wide Association Studies (GWAS). (A) The study populations where the genotyping data is extracted involves controls and cases for disease or unselected samples in the case of traits. **(B)** GWAS use spaced proxies for association tests across the genome and represented by Manhattan plots where each dot identifies a different SNP in its genomic position and shows the –log10 (*p*-value) for the association between the SNP and the trait. The threshold for significance is indicated as a dotted line (usually, *p*-value threshold of 5×10^{-8}) (Created with BioRender.com).

4.2.Expression Quantitative Trait Loci

While GWAS identify genetic variants associated with complex traits/ diseases, when these variants fall in the non-coding region it can be difficult to identify the biological mechanisms by which they affect disease risk. This is known to occur for 88% of the SNPs associated with whole-organism phenotype, as they are located in non-coding regions of the genome [62]. A useful approach to connect GWAS variants with diseases is to study how genetic variants affect molecular cellular phenotypes, such as gene expression, which in turn could have downstream effects on disease risk. Gene expression can be viewed as a quantitative trait that is highly heritable, representing a cell's state or a cellular phenotype [63]. A better understanding of how genetic variants affect cellular phenotypes may provide insights into the development of intermediate organ-tissue phenotypes, and ultimately whole organism phenotypes (**Figure 6**) [63]. Moreover, it is difficult to dissect the impact of genetic variants of whole-organism phenotypes when moving from cellular phenotypes because genetics have a more direct effect on gene expression than on tissue/organ or whole-organism phenotypes (**Figure 6**) [63].

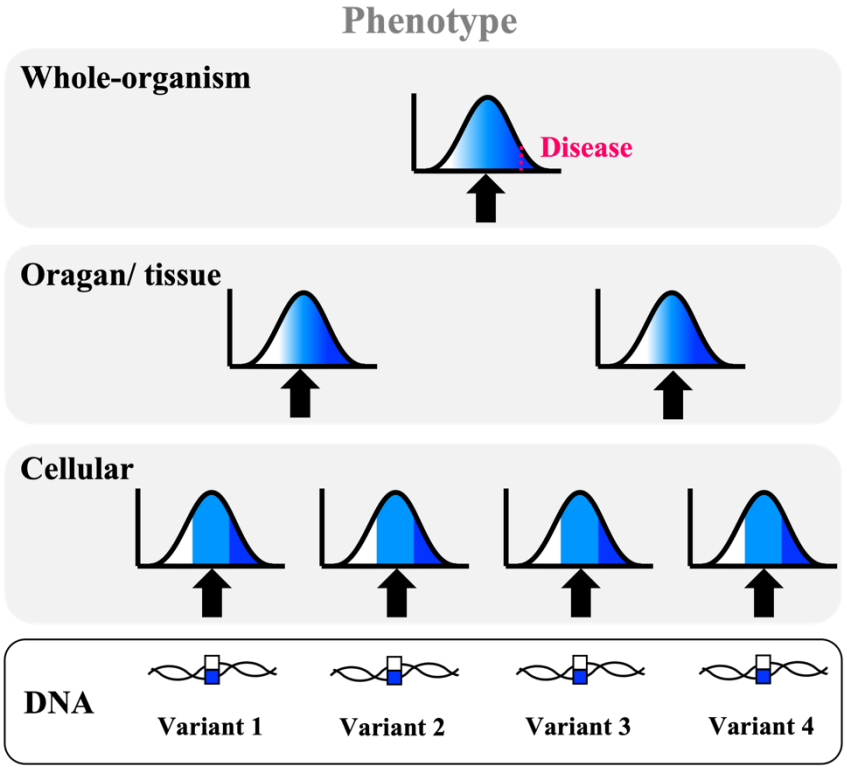


Figure 6. Whole-organism phenotype is a sum of molecular phenotypes affected by all DNA variants. Four common not linked genetic variants (numbered from 1 to 4) affect cellular, organ-tissue and whole-organism resulting in seven different normally distributed phenotypes. The white-blue gradient in the normal distribution histograms reflects the effects of each of the four DNA variants (white and blue). However, the ability to dissect genetic variants associated with phenotypes decreases as moving from cellular to whole-organism traits making the distinction blurrier. In the case of a giving disease (dashed pink line), it can be interpreted as the end of the continuous phenotypic spectrum (Adapted from [63]).

Variants that are associated with gene expression are known as expression quantitative trait loci (eQTLs) [64]. eQTL studies typically involves measuring gene expression in tissue samples taken from hundreds or thousands of individuals and testing these expression levels for association with an individual's genotype (**Figure 7**) [64].

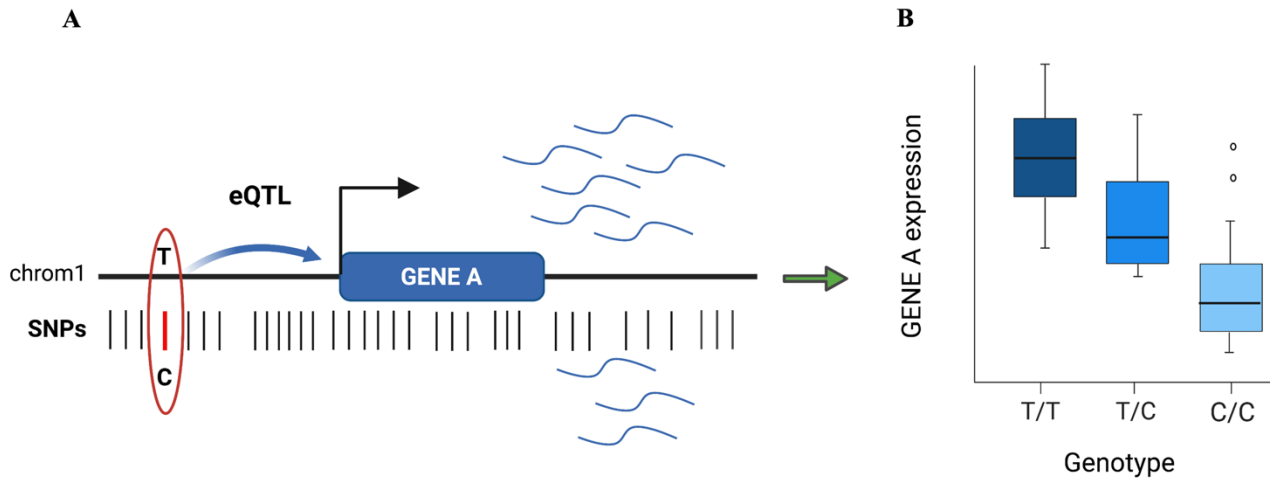


Figure 7. Schematic representation of expression quantitative trait loci (eQTLs). **(A)** In this illustration a given SNP of interest (red oval) is in the proximity of *gene A* (blue square) that codes for the corresponding mRNA molecules. **(B)** Box-plot showing the expression of *gene A* by the genotypes of the individuals and representing the effect of the eQTL on the expression levels. In this example, the homozygous T allele gives higher gene expression as the C allele (Created with BioRender.com).

The first eQTL study in humans was carried out in 2004 in LCLs by Morley *et al.* who showed that almost one-third of the genes tested had an eQTL [65]. Over the years, a large number of eQTL studies were performed to uncover the genetic basis of gene expression. The eQTLs were often

located in regulatory elements of the genome and in promoters near the TSS of the associated gene [55, 66-75]. Further studies investigated the degree of sharing *cis*-eQTLs across different tissues, have found that between 69% – 80% eQTLs were tissue-specific, meaning they show a unique activity in a particular tissue [67]. Tissue-specific eQTLs were located further away from the TSS of the gene they were associated as opposed to those found to be shared across tissues. This suggests that genetic variants in promoters are more likely to exert a shared effect across tissues than genetic variants in enhancers [67, 76]. To further study how genetic variants affect gene expression and regulation across human tissues, the Genotype-Tissue Expression (GTEx) Consortia was established. They described genetic effects on gene expression in 49 human tissues, finding that almost every gene has at least one eQTL [77].

eQTL mapping studies became an important tool for the functional interpretation of GWAS results, providing a way to identify candidate causal genes mediating a GWAS association, and to determine the type of cell or tissue most likely to be involved in a disease [72, 78]. However, despite these studies it remains a challenge to identify which genes causally drive disease. An overlap between eQTL and GWAS signals could be explained by three scenarios: *(i)* two independent causal SNPs are in LD with each other (linkage), *(ii)* one SNP affects the trait through affecting the expression of the gene (causality), and *(iii)* a single-causal SNP has an independent effect on gene expression and the trait (pleiotropy) [79]. It is only in the second case where the causal gene is implicated. Co-localization tests are designed to distinguish between these three cases [80]. In general, eQTL studies led to a better understanding of transcriptional regulation of gene expression, and this is beginning to translate into a deeper comprehension of complex traits and diseases.

4.3. Quantitative trait loci affecting splicing and transcripts

Over the last decade, as a result of transcriptome profiling of large cohorts of genotyped individuals, genetic variants affecting alternative splicing (AS) have been identified (known as splicing quantitative trait loci or sQTLs) [55, 77, 81-83]. Researchers have used sQTL analyses in a variety

of experimental settings to gain insight into the role of splicing plays in mediating GWAS associations. Traits studied include adipose-related traits, Alzheimer's disease, schizophrenia, and breast cancer [84]. In particular, it was shown that sQTLs might contribute to complex traits and diseases and play a similarly important role in disease development as variants from eQTLs [85]. Several studies have examined the effects of genetics on transcriptomic variation in recent years [55, 71, 73, 85-88], but the complexities of reconstructing transcripts from short reads made it difficult to fully understand the genetic regulation of splicing. In contrast to gene expression levels, which are most often represented by a single value per gene, alternatively spliced genes can be represented in a variety of different ways, each of them suited to identify different types of events. The main phenotypes generally used in sQTL analysis include: *(i)* exon expression level [71]; *(ii)* transcripts ratio, the ratio for each transcript over all possible transcripts [55, 86]; and *(iii)* percent exon inclusion (PSI), or how often is a given exon included *versus* excluded from a gene [73, 87, 88].

However, current methods to detect splicing based on short-read data can only identify AS using sequencing reads that span from one exon to another (e.g., splice junctions) or using computational reconstructions of possible transcripts. With the advantages of long-read sequencing technologies, spanning the entire length of a transcript with one single read, a more accurate measurement of transcript expression can be achieved. From there, single transcript abundance can be used as a quantitative phenotype to determine the association with SNPs. These transcripts eQTLs (trQTLs) will add a new piece of information to the complex landscape of gene expression regulation. In particular, sQTLs are genetic variants which affect a particular splicing event that has an effect on the exon composition of the generated transcripts. In contrast, trQTLs specifically affect the expression of a particular transcripts among those produced by AS events. Before the advent of long-read sequencing technologies, it was not possible to detect trQTLs; if the SNP had an effect on the most abundant transcripts, then this genetic effect would have been captured by an eQTL analysis but if a less abundant transcript was affected by the SNP then the effect may be missed using eQTL

analysis, an expression phenotype that summarizes all transcripts expression. **Figure 8** illustrates two examples of sQTL and trQTL for the same gene.

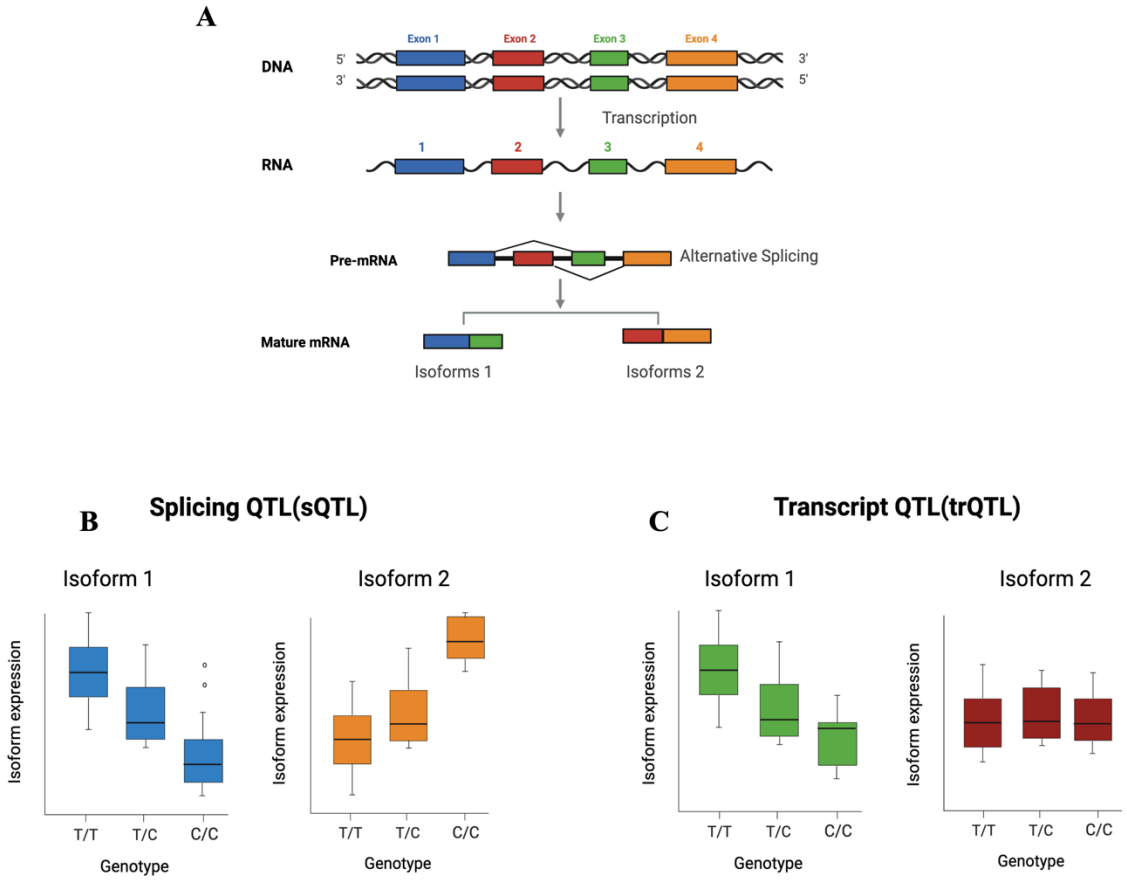


Figure 8. Representation of an eQTL effects on splicing and transcripts. (A) Two alternative splicing events produce two different mRNA isoforms for the same gene. Different boxes represent different exons. **(B)** Box-plots representing the effect of the sQTL on the isoform expression. The SNP is influencing the splicing events that change the production of the isoforms 1 and 2. In the first example (blue plot) T/T genotype gives the higher isoforms expression, while C/C the lower; in the second example (orange plot) C/C genotype gives the higher isoforms expression while T/T the lower. **(C)** Box-plots representing the effect of the trQTL only on the expression on the isoforms 1 without any effects on the expression of the isoform 2. In the first example (green plot) T/T genotype gives the higher isoforms expression, while C/C the lower; in the second example (red plot) the genotype did not have any effect on the expression of the isoform (Created with BioRender.com).

5. The genetics of cancer

Cancer is a complex chronic genetic disease in which cells proliferate abnormally and spread in the body, escaping the immune surveillance [89]. The importance of genetics in cancer development was suggested by Theodor Boveri in the 19th century after observing chromosomal aberrations in cancer-dividing cells under the microscope [90]. However, the word cancer refers to hundreds of different disease types which share similar primary properties, namely abnormal cell proliferation [89]. In addition, cancer is characterized by a high degree of cellular plasticity and heterogeneity, evolving at the genetic, phenotypic, and pathological level [91].

5.1. Mutations in cancer

Historically, cancer was seen as a disease caused by the acquisition of somatic mutations in individual cells. Mutations with detrimental effects are generally eliminated by *purifying* natural selection, while those that are beneficial because they increase cell fitness – i.e., the ability to survive and reproduce – are selected and favored by evolutionary processes (positive selection). The effects of the vast majority of these beneficial mutations are generally mild and do not have a discernible consequence on the organism phenotypes affecting cell division, cells death or other cellular phenotypes [90]. However, other mutations may enhance the proliferation or invasive potential of the cells, giving them the ability to rapidly grow, colonize other tissues, and spread to other sites in the body by metastasis [90]. These are considered *"driver"* mutations and affect genes known as *"cancer genes"*, resulting in the promotion of tumorigenesis. Other mutations called *"passengers"* do not act directly on the tumorigenic process, but instead hitch-hike the driver mutations that were earlier acquired by the cells [90]. Passenger mutations can be present in cancer genes even if they do not directly promote tumorigenesis. Genes affected by passengers' mutations can be of two different types: *oncogenes* and *tumor-suppressor genes*. Oncogenes are genes with the potential to cause cancer. They are highly activated by mutations (gain of function mutations), giving selective advantages to the cell. Tumor suppressor genes, on the other hand, are inactivated by mutations (loss of function mutations), giving

proliferative advantages to mutated cancer cells [92]. In general, the mutational signature between oncogenes and tumor suppressor genes is different, with recurrent mutations at specific amino acids for oncogenes, while tumor suppressors show a higher fraction of truncating mutations, that lead to the production of a shorter version of the protein [93].

For centuries, genetic predisposition has been recognized as an important factor in cancer development and progression, but this has only recently been the subject of extensive investigation [94, 95]. The Cancer Genome Atlas (TCGA, http://cancer.sanger.ac.uk/census) is a large study that has produced a molecular characterization of over 20,000 primary cancer and matched normal samples, spanning 33 cancer types. The TCGA shows that nearly 1% of human genes harbor mutations that recur in cancer, most of them are acquired somatically and approximately 20% of these variants are inherited as germline variation. Germline genetic variants known to drive cancer are sporadically distributed over the genome and restricted to a small group of genes [96]. Therefore, the investigation of germline genetic variants has mainly focused on known cancer genes, including tumor suppressors, DNA repair, oncogenic signaling pathways and cell cycle genes [97]. For example, individuals with *Lynch syndrome* who are first-degree relatives of patients diagnosed with colorectal cancer carry defects that prevent DNA repair which can leads to the accumulation of somatic mutations resulting in colon cancer [98]. In addition, cancer types such as breast and ovarian cancers have been directly associated with germline mutations in the breast cancer susceptibility gene type 1 and type 2 (*BRCA1 and BRCA2*) [99, 100]. *BRCA1/2* are tumor suppressor genes that repair DNA damage, and when mutated they no longer can prevent malignancies occurring. Women who carry mutations in these genes have an increased lifetime risk of 84% and 39% of developing breast and ovarian cancers [99]. *BRCA2* mutations are also associated with predisposition to pancreas, prostate cancer, and melanoma [101-103]. The germline mutations in *BRCA1/2* seem to promote cancer development at younger age, with the cancer being more aggressive and the patient having a poorer prognosis compared to the sporadic breast cancer due to somatic mutations [99, 104].

Tumorigenesis is a complex process and a deeper investigation of the interplay between the germline genomic contribution and somatic mutations may uncover genetic mechanisms responsible for cancer other than those that are linked to already known genes.

GWAS in cancer have shown that most cancer susceptibility loci, like those for other diseases, are common variants which individually only moderately affect disease risk and that these variants are mostly found in non-coding genomic regions (**Figure 9**) [105, 106]. As a result, genetic susceptibility to cancer is characterized by unequal levels of risk and prevalence of predisposition alleles, and the loci identified so far explain only a small portion of the familiar risk of many types of cancer [107]. Germline variants affecting cancer risk can be characterized as those with high penetrance, often found in rare familial cancer cases, which seem to be directly involved in tumorigenesis [108] and variants with low penetrance which may modulate the outcomes of other somatic variants with smaller effects [109].

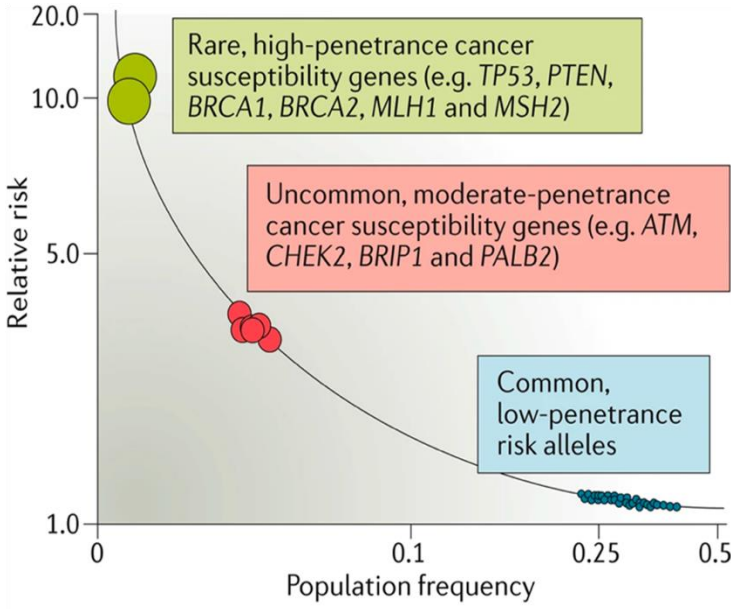


Figure 9. Genetic architecture of cancer risk. An analysis of the findings of genome-wide association studies (GWAS) depicts the low relative risks (RRs) associated with common, low-penetrance genetic variants. The RR is moderate for rare, moderate-penetrance genetic variants (such as *ATM*) and checkpoint kinase 2 (*CHEK2*); and higher for rare, high-penetrance genetic variants (such as pathogenic mutations in *BRCA1* and *BRCA2* in hereditary breast and ovarian cancer) (Taken from [107]).

6. Lymphoblastoid cell lines as a model for cancer

LCLs are actively proliferating immortalized B-cells, transformed with Epstein-Barr Virus (EBV) and derived from human resting B-lymphocytes. LCLs provide a nearly unlimited source of individual's genetic material as well as other biomolecules; they are easy to keep in culture, manipulate, and have a stable karyotype, as mentioned before. Since LCL samples collected by the 1000 Genomes Project were derived from anonymous donors, the genotype information could become publicly available. Over the last decade, LCLs have been extensively used for functional and molecular studies. In particular, a large number of population genetic studies used LCLs to show how genetic variation affects basic cellular phenotypes, i.e., to better understand the genetic basis of disease [54, 63, 67, 110]. In population and family-based studies, LCLs played a key role in providing the basis for research into rare genetic diseases. Furthermore, as previously mentioned, important collections of genetically different LCLs generated over the years have proved to be a suitable model to identify associations between genetic and transcriptomic signatures, as well as phenotypic changes [111].

Since LCLs are generated through EBV-transformation of primary B-lymphocytes, LCLs are also a suitable *in vitro* model to study EBV-associated B-cell lymphomas [112]. EBV immortalizes human B-cells by pushing the cells to proliferate indefinitely and control their ability to resist apoptosis. Post-transplantation lymphoproliferative disorder (PTLD) is a group of diseases characterized by polyclonal lymphoid proliferations or aggressive monoclonal lymphomas [113]. For the majority of PTLD cases, 60 – 80%, are associated to EBV infection and represent a rare, but well-described, complication associated to solid organ transplantation. Due to pharmacologic immune suppression in transplanted patients, EBV-transformed B-lymphocytes can grow in patients with PTLD. The result is an abnormal growth and spreading of white blood cells; in particular, of B-lymphocytes infected by EBV, causing complications that can range from benign noncancerous tissue overgrowth due to the overproduction of the B-cells (hyperplasia) to malignant lymphoma or post-transplant lymphomas

(PTLs). PTLD can affect transplant recipients either because of reactivation of dormant EBV or because the patient becomes infected with the virus for the first-time following transplantation. PTLDs can form at all stages of the B-lymphocyte differentiation pathway, including some that are pathogenetically related to other EBV-associated B-cell lymphomas. In addition, 40% of cases of classic Hodgkin's lymphoma, one of the two lymphoma categories, can also be associated with EBV. Classic Hodgkin's lymphoma is an uncommon cancer marked by the presence of Reed-Sternberg cells. Reed-Sternberg cells are mature big B-lymphocytes that start overgrowing, becoming larger with more than one nucleus. These malignant B-cells flow in the lymphatic system, a network of blood vessels and glands located throughout the body; migrate and accumulate in the lymph nodes causing the swelling and abnormal growth throughout the body. Finally, the 'endemic' Burkitt's lymphoma is also caused by EBV infection and is most common in African children. Burkitt's lymphoma is a non-Hodgkin's B-cells lymphoma, characterized by a mass of small malignant B-cells fused with non-neoplastic macrophages. Burkitt's lymphoma is fast-growing and highly aggressive, containing one of the three chromosomal translocations (i.e. t(8:14), t(2:8), t(8:22)), causing the dysregulation of the c-Myc oncogene responsible for changes in proliferation, differentiation, metabolism, and apoptosis of cancer cells [114].

7. Cancer-like phenotypes

In 2000, Hanahan and Weinberg suggested that six crucial alterations in the physiology of cancer cells collectively develop into malignant growth: *(i)* limitless replication potential, *(ii)* self-sufficiency in growth signals, *(iii)* non-sensitivity to growth-inhibitory (anti-growth) signals, *(iv)* evasion of programmed cell death (apoptosis), *(v)* tissue invasion and metastasis *(vi)* sustained angiogenesis [89] (**Figure 10**). These phenotypic abnormalities were called the *six hallmarks of cancer*, and in each of these physiologic changes occurring during tumor development, a new capability is acquired to breach the anticancer defenses hardwired into cells and tissues. These six hallmarks are shared by most, perhaps all, types of human tumors. Unfortunately, some of these

cancer-like phenotypes are not easy to recapitulate *in vitro*. However, functional assays for cell proliferation, cell apoptosis and cell chemotaxis allow the detection and measurement of these cancer features in the lab. In particular, an *in vitro* proliferation assay can assess the self-sufficiency in growth signals, the insensitivity to inhibitory-growth signals and the limitless replication potential. *In vitro*, apoptosis assays measure the evasion or resistance to apoptosis. Finally, *in vitro* chemotaxis assays measure the migration of cells in response to stimuli, which could partially recapitulate tissue invasion and metastasis (**Figure 10**).

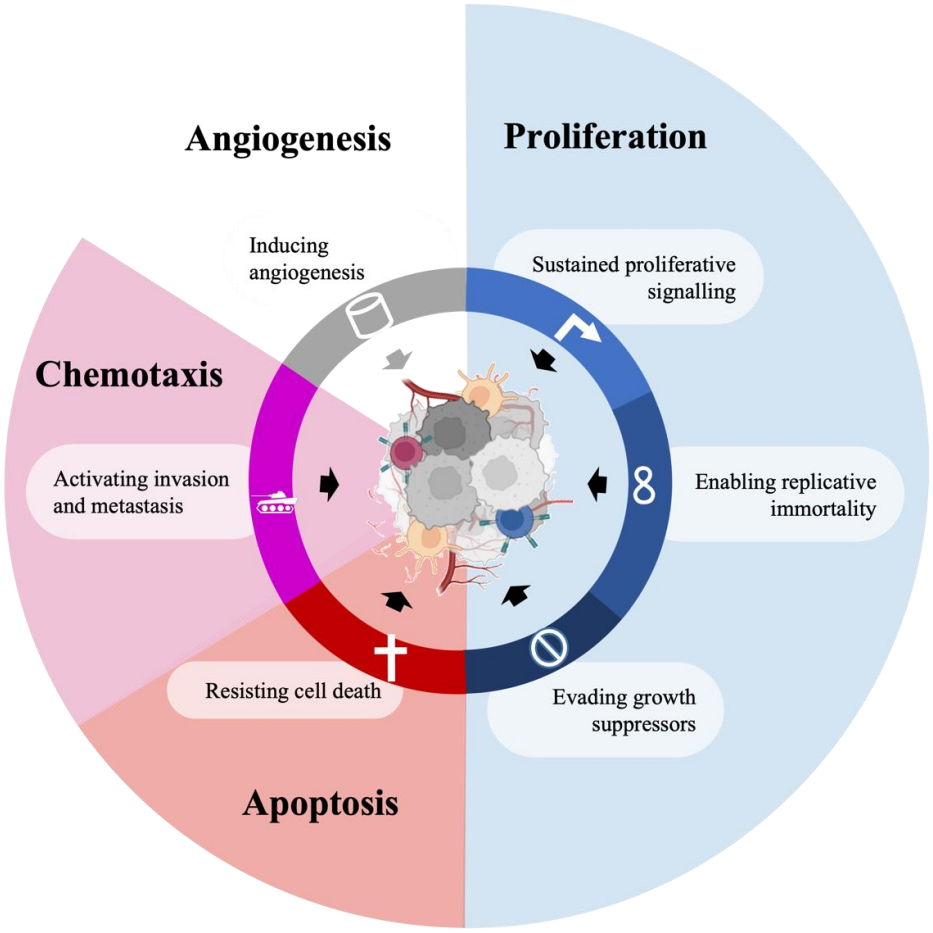


Figure 10. Hallmarks of cancer. This illustration summarizes the six hallmark capabilities of cancer proposed by Hanahan, organizing principles for understanding neoplastic disease complexity. The cancer capabilities include *(i)* sustained proliferation signals, *(ii)* ability of replicative immortality and *(iii)* evasion of growth suppressors leading to abnormal *Proliferation* (blue), *(iv)* resistance to cell death interfering with *Apoptosis* (red); *(v)* active invasion and metastasis defined as *Chemotaxis* (magenta); finally, *(vi)* promotion of the vascularization of the solid tumors, *Angiogenesis* (grey) (Adapted from [89]).

7.1.Proliferation

Proliferation is the ability of a cell to sustain growth and survival, and it represents one of the most fundamental characteristics of cancer cells [89]. Healthy cells carefully regulate the production and release of growth factors and signals to ensure proper cellular homeostasis and to preserve the normal tissue function and structure. In contrast, cancer cells dysregulate those signals and maintain and promote proliferative signals through a wide range of mechanisms that are not yet completely understood. These mechanisms include *(i)* the self-production of ligands for growth factors and the expression of surface/ intracellular/ nuclear receptors to respond to the proliferation factors from stromal cells outside the tumor [115, 116]. *(ii)* The constitutive activation of different gene pathways that promote cellular survival. These two processes are particularly well-regulated by cancer cells to ensure a high proliferation and survival rate, while at the same time avoiding cell senescence and apoptosis.

7.2.Apoptosis

Apoptosis is programmed cell death occurring during embryogenesis, development and ageing to maintain cellular homeostasis in tissues. Apoptosis also plays a role in eliminating damaged cells and can be induced by a wide range of stimuli that can be used as cancer prevention mechanisms [117]. Besides the increased proliferative abilities of cancer cells and the sustaining cell growth, malignant cells often evade cellular programs that induce death or negatively regulate cell proliferation [89]. There are two major circuits of apoptosis regulation: *(i)* the extrinsic apoptosis program that involves for example the Fas ligand/Fas (also called CD95L/CD95), a member of the tumor necrosis factor death receptor subfamily, and *(ii)* the intrinsic apoptotic program of intracellular origin [118]. Both circuits result in the activation of the caspase 8 and 9 pathways, which initiates a cascade of proteolytic activity and results in apoptosis where the cells are dissembled, consumed, and phagocyted. Apoptosis is currently considered a process involving multiple signaling pathways that affect different cellular phenotypes [118] For example, the intrinsic apoptosis is a mitochondria-

dependent pathway that can be initiated by signals such as growth factor withdrawal, DNA damage, endoplasmic reticulum stress, reactive oxygen species overload and replication stress [119]. *In vitro*, the intrinsic apoptosis pathway leads to a complete breakdown of the plasma membrane and the gain of a necrotic morphology of the apoptotic cells. Alternately, the extrinsic apoptosis is a programmed cell death pathway activated by external or internal events of cells. Extrinsic apoptosis is triggered by the perturbations of the extracellular microenvironment [120] and is mostly driven by cellular receptors such as *(i)* death receptors, activated by the binding of the cognate ligand(s), and *(ii)* dependence receptors, activated when the level of their ligand drops below a certain threshold [121].

7.3. Chemotaxis

Cellular migration is the cellular ability to move from one place to another, usually triggered in response to an attractant such as chemokines and growth factors among others [122]. Chemotaxis is the process in which invasion, migration and dissemination of cells is directed by a gradient of chemokines or other stimuli, and is an important property of living cells [123]. In conditions such as cancer, the migration of malignant cells far from the primary lesion site could result in their metastatic dissemination [124]. The initiation of the metastatic process involves chemotaxis, where a complex network of chemokines signals stimulates the cancer cells to move [125]. On the other hand, chemokines also regulate other cancer-related processes, such as tumor cell proliferation and survival, angiogenesis, immune evasion, senescence and metastatic progression [126-128]. More than 20 chemokines and chemokines receptors are linked to chemotaxis in cancer [126]. Because of this, and the complexity of the tumor microenvironment, it is still difficult to measure precisely chemotaxis *in vivo*. However, *in vitro* systems are a useful tool to evaluate the ability of different cell lines to migrate towards chemoattractant/ chemokines.

The study of these cancer-like phenotypes in the context of cancer genetics could provide the suitable framework to dissect the impact of genetic variants on complex traits and diseases, i.e., cancer. Long-read dRNA-seq provides further insights of these regulatory mechanisms.

Scope of the thesis

During the last 20 years, the use of next-generation sequencing technologies provided a great amount of information to analyze the role of human genetic variants in gene regulation and complex diseases, such as cancer. Yet, a variety of fundamental questions remain unanswered including the precise mechanism by which non-coding genetic variants affect gene expression in complex phenotypes in humans and what are the characteristics of these effects. This thesis aims to shed light on these questions by studying how gene expression plays a role in mediating the effect of genetic variants on phenotypes.

For such, the following specific aims (SA) were defined:

SA1: To produce population-based direct long-read RNA sequencing data using 60 genetically different LCLs to evaluate the contribution of genetic variants in altering gene expression, transcript expression, and splicing events. Particularly, to increase the ability to detect and quantify complex transcripts and alternative splicing events across the genome and that the previous short-read technology missed.

SA2: To generate population-based cancer-like *in vitro* phenotypes using 87 genetically different LCLs to understand the link between genetic variants, changes in gene expression and alteration in cancer-like phenotypes. The ultimate goal is to identify putative genes that could predict cancer-like phenotypes.

Results

Article 1.

Gene expression profiling with direct long-read RNA sequencing uncovers functional variation affecting transcripts production.

Aline Réal, Christelle Borel, Nikolaos M. R. Lykoskoufis, G. Puga Yung, Jörg D. Seebach, Andrew Brown, Emmanouil T. Dermitzakis, Ana Viñuela and Anna Ramisch.

This chapter is a preprint of the *in preparation* for submission manuscript

Goal: This study aims is to evaluate the impact of genetic variants on gene expression, transcript expression, and splicing events using population-based long-read RNA sequencing data from 60 genetically different LCLs.

Personal contribution: For this manuscript in preparation, I performed all the experiments from LCLs culture to the direct RNA library preparation and sequencing using ONT. I also performed all the long-read data analysis of the study. For the manuscript, I prepared the figures and wrote the manuscript with comments from the thesis supervisory team and co-authors.

Gene expression profiling with direct long-read RNA sequencing uncovers functional variation affecting transcripts production.

Aline Réal[1,2*], Christelle Borel[1], Nikolaos M. R. Lykoskoufis[1], G. Puga Yung[2], Jörg D. Seebach[2], Andrew Brown[3], Emmanouil T. Dermitzakis[1‡], Ana Viñuela[4‡*] and Anna Ramisch[1,5‡*].

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

[2]Division of Immunology and Allergology, University Hospitals and Medical Faculty, Geneva, Switzerland

[3]Population Health and Genomics, University of Dundee, Dundee, United Kingdom

[4]Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne, UK

[5]Department of Basic Neuroscience, University of Geneva Medical School, Geneva, Switzerland

[‡]These authors contributed equally

* Corresponding authors

Competing Interest Statement

Emmanouil T. Dermitzakis is currently an employee of GSK. The work presented in this manuscript was performed before he joined GSK. All other authors declare no competing interests.

Funding

This work was supported by the Swiss National Science Foundation (FNS ME10662 and ME11559) to ETD.

Keywords

ONT, long-read RNA, LCL, trQTL, eQTL.

Abstract

Expression quantitative trait loci (eQTLs) studies have uncovered thousands of genetic effects acting on gene expression. However, our knowledge of the impact of genetic variants on gene expression is limited by the short-read RNA-seq technologies currently used, as these do not characterize transcripts in their full-length form and do not allow us to explore genetic effects on splicing and isoforms production. To directly measure the impact of genetic variation on transcript abundance, we produced long-read native poly(A) RNA-seq data using the Oxford Nanopore Technologies (ONT) platform for 60 genetically different lymphoblastoid cell lines (LCLs) from the 1000 Genomes/GEUVADIS project. We identified 11,154 protein-coding genes and lncRNAs expressed in at least 50% of the samples, in close agreement with published gene expression quantifications of the same samples based on Illumina short-read sequencing (pairwise sample correlation 0.61 - 0.79). We identified 48,539 transcripts, of which 39% were already annotated based on GENCODE version 19. While 35% of these annotated transcripts were expressed in all of the 60 LCL samples, this was only true for 7% of unannotated transcripts. A genome-wide QTL analysis on the 14,447 annotated transcripts expressed in at least 50% of samples identified 72 transcript QTLs (trQTLs; FDR 5%) of which 71 were not identified as eQTLs from the larger published larger Illumina dataset (317 samples). Using our own gene-level quantifications from long-read data, we detected 45 eQTLs of which 12 were also trQTLs. We observed that genes with eQTLs had a significantly lower number of annotated transcripts than genes with trQTLs (Wilcoxon test *p*-value = 5.23e-05), suggesting that genetic effects on genes with higher transcript diversity were missed using gene-level quantifications. Overall, we were able to identify new trQTLs based on a small number of samples, whose effect on expression was missed when using short-read technology. Hence, we could show that by using long-read data to analyze transcriptomes, we are one step closer to understand the key role of genetic variants on transcription expression and splicing, which many also improve the characterization of disease-associated variants.

Abbreviations

eQTLs, expression quantitative trait loci; ONT, Oxford Nanopore Technologies; LCLs, lymphoblastoid cell lines; trQTL, transcript expression quantitative trait loci; sQTL, splicing quantitative trait loci; GWAS, Genome wide association studies; lncRNAs, long non-coding RNAs; QC, quality control; RPKM, Reads Per Kilobase of transcript per Million; TPM, transcripts per million; FDR, False Discovery Rate.

Introduction

Splicing, a key feature of gene regulation is the molecular mechanism that produces a diversity of mRNA molecules from a single gene [1]. The mechanisms that regulate splicing events play a central role in gene expression regulation; by selecting different combinations of exons, an alternative splicing process can produce multiple transcripts, and consequently multiple proteins [1-3]. However, our knowledge about how an individual's genetic background may affect the splicing process is limited. A main reason is that most studies use short-read sequencing technologies to identify genetic effects acting on gene-level expression phenotype (eQTLs) that summarize all the molecules produced by a gene, independently of their structure. To study transcript structure, transcript expression and alternative splicing, the use of short-read RNA sequencing requires the development of incomplete *in silico* proxy phenotypes [4-7]. Moreover, short-read sequencing approaches still need to rely on cDNA conversion, PCR amplification and post-sequencing transcripts reconstruction, all steps which potentially introduce biases affecting transcript quantifications and the prediction of their structures. Therefore, these gene-level phenotypes are not suitable to identify genetic effects on specific transcript abundances or structure. Using long-read sequencing, it is now possible to analyze transcripts in their native form, reconstruct their precise structures [8, 9], identify novel transcripts [10, 11] and study the allele-specific effects on transcript abundance and structure [12, 13]. Moreover, it allows us to perform quantitative trait loci (QTL) analysis for transcripts (trQTL) and splicing (sQTL) and can be used to identify genetic variants affecting the expression of specific transcripts and the splicing processes that define their structure [12, 14]. In particular, trQTLs identify the effect of genetic variants on transcript abundance while sQTLs detect the effect of genetic variants on splicing events, two highly interconnected phenotypes.

Many genome-wide association signals for common diseases are enriched in genetic variants that drive changes in gene expression and splicing [6, 8]. Analyses in multiple human tissues have shown that sQTLs can have a similar, or possibly even bigger, impact on GWAS traits than eQTLs [15-17].

However, before the advent of long-read sequencing technologies, it was difficult to perform high-throughput trQTLs analyses and explore the role of specific transcripts in disease risk.

Here, we explored the advantages of direct long-read RNA sequencing of transcripts in their native form in a population of 60 lymphoblastoid cell lines (LCLs), using a genetically diverse population of cells. We investigated *(i)* the distribution of annotated and novel isoforms across the population; *(ii)* the effect of genetic variants on directly measured transcript abundance, while *(iii)* characterizing some of the mechanisms by which eQTLs discovered using short-read technology affect gene and transcript expression.

Material and Methods

LCL samples

The 60 LCLs were from the 1000 Genomes Project cohort with European ancestry and from unrelated individuals [18]. All the samples are part of the NHGRI sample repository for human genetic research. All LCLs came from the Coriell Institute for medical research (Camden, New Jersey, USA). LCLs were grown under identical conditions in RPMI 1640 media supplemented with 1% penicillin-streptomycin, 1% L-glutamine and 10% fetal bovine serum (FBS). LCLs were cultured for at least 3 to 4 weeks until their exponential growth phase and had a total concentration of at least 4×10^7 LCLs and a constant high viability (~98%). All cells were tested for mycoplasma contamination (Lonza, MycoAlert mycoplasma detection kit) before being used for the following steps.

RNA extraction, library preparation, and sequencing

From 4×10^7 mycoplasma-free LCLs we obtained total RNA using Trizol Reagent (Invitrogen). Cells were washed twice with 1 × phosphate buffered saline (Invitrogen) to remove all the media and 1ml of Trizol was added per 5-10 $\times 10^6$ cells in each sample, incubated for 5min at room temperature (RT) and transferred to Eppendorf tubes. The rest of the protocol followed the manufacturer's guidelines with the addition of 200µl of chloroform for every ml of Trizol for the phase separation followed by mixing and centrifuging for 15min at 2000 × *g* at 4°C. The phase containing RNA was recovered and transferred in new Eppendorf tubes for RNA precipitation with 500µl of isopropanol every 1ml of Trizol and incubation at RT for 15min followed by centrifugation 20min at 2000 × *g* at 4°C. After RNA precipitation, 70% ethanol was used to wash the pellet and centrifuged 5min at 2000 × *g* at 4°C, supernatant was discarded, and the RNA pellet was air dried for 5 -10min. The pellet was solubilized in 20µl of 0.5% SDS in RNase-free water. No DNAse treatment was applied.

The total RNA was quantified using Qubit Fluorometer 2.0 with the Qubit RNA Broad Range (BR) Assay kit according to the manufacturer's instructions (Thermo Fisher) and Nanodrop to exclude the presence of alcohol and protein contaminants that could interfere with the sequencing, keeping RNA

samples with a ratio OD $_{260/280}$ of at least 1.9 and ratio OD $_{260/230}$ > 1.5. Agilent Bioanalyzer RNA 6000 Nano Kit (Agilent) was used to assess the quality and integrity of RNA. Only samples with RNA Integrity Number (RIN) >8.9 were used for the following steps. The total RNA was then poly-A^+ tailed before the library preparation using the Dynabeads™ mRNA Purification Kit (Thermo Fisher). The poly-A^+ tailed capture step was repeated re-using the same Dynabeads, which increased the enrichment of poly-A^+ tailed RNA and improved the consequent elimination for ribosomal RNA. The final quantification of poly-A^+ tailed RNA was performed by the TapeStation with High Sensitivity RNA ScreenTape (Agilent) to ensure the quasi-total elimination of the 28S and 16S ribosomal peaks.

For the library preparation we used 500ng poly-A^+ tailed RNA in a total volume of 9µl and followed the all the steps of the ONT protocol for the Direct RNA Sequencing Kit (updated version 27/12/2019 nanoporetech.com, cat# SQK-RNA002). The quantification of the library RNA was performed using the Qubit fluorometer DNA HS assay (Thermo Fisher) - recovery aim of ~200ng. Then, before loading the RNA into the FLOW-MIN106D flow cell, the numbers of pores and properly primed pores were checked according to the manufacturer's instructions. Finally, the sequencing was carried on for 72h on the GridION Mk1 sequencing device (ONT) that allows sequencing of a maximum of five samples in each run, one per flowcell.

Pre-processing of RNA sequencing data

Base-calling was performed using the Guppy software (from ONT, version 3.2.10) in the *high accuracy mode*. Guppy used the *fast5* files generated by the ONT Device Control software (MinKNOW), embedded in the GridION sequencing device, as input to *(i)* generate a *fastq* file for each *fast5* file containing the base-called sequences; *(ii)* create base-called *fast5* files. *(iii)* classify *fastq* and *fast5* files into pass / fail folders according to the average quality score of each read (above 7.0); and *(iv)* make summary files for every flow cell sequenced. We applied the specific options suggested in the Direct RNA sequencing protocol taking into consideration the reversed direction of

the sequencing (3′→5′), the presence of uracil instead of thymine, and an optimized strategy for trimming the adapter's raw signal. We used only the passing reads for the following analysis.

Mapping sequences. We concatenated the *fastq* files obtained from base-calling into a single *fastq* file. These *fastq* files were then mapped to the reference human genome GRCh37 (hg19_chr_only_and_herpes.fa) using minimap 2 v2.12 [19]. The calling was performed in a splicing-aware manner with the following options: *minimap2 -a -x splice -k14 -uf* (-a -x splice: splice alignment mode; -uf: force minimap2 to consider the forward transcript strand only; -k14: small k-mer to increase sensitivity to the first or the last exons). Alignment files from minimap2 were converted to *bam* format, sorted and indexed using samtools v1.6 [20].

Data quality control (QC). We applied Nanoplot (version 1.33.0) [21] to produce QC graphs displaying multiple aspects of sequencing raw data; while NanoStat (version1.4.0) was used to obtain a statistical data summary [21]. The pycoQC tool [22] (version2.5.2) served to generate an interactive QC report from base caller's datasets. Specifically, pycoQC uses the sequencing summary file generated by Guppy and the *bam / sam* file to generate a pre / post-alignment QC report.

Gene quantification

We used *featureCounts* to provide the gene-level counts (from Subread v1.6.0) to the genome alignments using exons as the feature type [23]. We used the *-L* argument of *featureCounts* to enable the long-read mode with a minimum overlap of 10 bases *(--minOverlap 10*) and we used GENCODE v19 as the reference annotation [24]. We converted counts to RPKM (Reads Per Kilobase of transcript, per Million mapped reads) using the *rpkm* function of the edgeR package [25].

RNA-seq data and genotype data from external datasets

Curated short-read Illumina RNA-seq data and genotype data of 60 LCLs were used as described in Delaneau et al. [26]. Briefly, gene expression was quantified using QTLtools [27] with GENCODE v19 [24] as the reference gene annotation. Genes were filtered to retain only protein-coding genes

and long non-coding RNAs (lncRNAs) expressed in more than 90% of the samples. The gene expression was quantified using RPKM units for the gene expression quantification in the ONT dataset. The genotype data for these samples, available from either the 1000 Genomes project or the Illumina Human OMNI 2.5M SNP array, were filtered using standard procedures to remove low-quality SNPs. Moreover, the resulting genotype matrix of 317 individuals and 9,255,024 variants was imputed from the 1000 Genomes phase 3 reference panel [18], and poorly imputed variants were removed [26].

Gene expression correlation between Illumina and ONT

Pair-wise gene expression Spearman correlations were computed between Illumina short-reads and direct long-read RNA sequencing ONT from the same 60 LCLs samples using the *rcorr* function in the corrplot R package [28]. Both gene expression datasets included protein-coding genes and lncRNAs expressed in at least 50% of their samples, which were found in both sets (n = 11,542).

Transcripts detection and characterization

To identify transcripts from the native RNA sequences we used FLAIR v1.5 (https://github.com/BrooksLabUCSC/flair). For the analysis, *bam* files obtained using the minimap2 aligner were converted to *bed* format using the bam2bed12.py script provided with FLAIR. FLAIR-*correct* was used to correct the splice-site boundaries of reads. It corrected misaligned splice sites using genome annotations from GENCODE v19 and GRCh37 as the reference genome. Next, the FLAIR-*collapse* command processed the corrected reads, generating a first-pass transcripts set. To do this, FLAIR-*collapse* grouped reads on their splice junction chains and only kept transcripts supported by at least 10 reads and mapping quality >10. At this step, the first-round alignments were split by chromosome due to computational limitations. FLAIR-*quantify* was used to determine transcript levels in all samples where reads aligned to annotated transcripts (GENCODE v19). Transcripts with intron chains not matching any transcripts in the reference annotation (GENCODE v19) were defined as *'novel isoforms'*. The 36,782 transcripts not aligning with any gene in the

GENCODE v19 annotation were excluded from the following analyses. Reads were normalized using a transcripts per million (TPM) normalization. Moreover, mitochondrial transcripts as well as transcripts expressed with less than five TPMs in at least one sample were excluded.

Molecular quantitative trait loci

For each molecular phenotype, gene abundances and transcripts abundances, we identify QTLs using the QTLtools software package (version 1.3.1) [27]. Shortly, all genetic variants within +/– 1Mb of the transcription start site were associated with the phenotypes, and the best-associated SNP (i.e., with the smallest nominal *p*-value) was retained. After that, the nominal *p*-values were adjusted for the number of variants being tested using 1,000 permutations. This is implemented in the *cis* mode of the QTLtools software package [27]. Multiple testing correction across phenotypes was done using the qvalue package in R (version 2.18.0) [29] to identify all significant phenotype-variant pairs at 5% False Discovery Rate (FDR). For gene-eQTL analysis, we tested genes expressed in at least 50% of the samples (n = 13,997). For the transcripts-eQTL analysis, we tested annotated transcripts expressed in at least 50% of the samples (n = 14,447) which corresponded to 9,364 unique genes. For transcript-eQTL analysis, we used the option *-grp* to correct and account for multiple phenotypes (transcripts) per gene. This option performs a permutation pass at the gene group level across all phenotype-SNP pairs per gene to discover gene-level trQTLs.

All QTL analyses included the following covariates: sex, the first three principal components (PCs) from genotypes, and three PCs from expression (genes or transcripts).

Illumina eQTL recapitulation in ONT dataset

The eQTLs already identified using 317 samples from the Illumina RNA-seq data described in Delaneau et al. [26] were used. From these 7,658 significant eQTLs in the Illumina dataset, only 4,169 involved genes and transcripts expressed in at least 50% of the samples that were kept as part of the ONT dataset. To detect how many of the significant short-reads eQTL were also detected in long-read native RNA-seq, we extracted the *p*-values from the same phenotype-SNP pair associations

and calculated $\pi 1$ using the *q*-value package in R [29]. The *q*-value estimation for false discovery rate control R package (version 2.18.0) used a $\lambda = 0.05$ and FDR = 0.05.

Results

General RNA data characteristics

Total RNA was isolated from LCLs of 60 unrelated individuals with European ancestry from the 1000 Genome Project cohort [18]. Samples were sequenced on the GridION ONT platform using the direct RNA sequencing protocol following the experimental workflow represented in **Figure 1A**. After mapping our data to the Reference Genome GRCh37, we measured both, gene abundance and transcript abundance according to the workflow illustrated in **Figures 1B**. The sequence average was 1.87×10^9 nucleotide bases, generating a yield from 1 to 2.5 million raw reads per sample (median = 1.7×10^6) (**Figure 2A**). The median read length distributions were similar among the samples, ranging from 700bp to 900bp (median = 822.5bp) (**Figure 2B)** with a median read quality between 9.5 and 10.6 PHRED-like score (median = 10.1) (**Figure 2C)**; corresponding to a minimum and a maximum read accuracy of 88% and 91%, respectively (median = 90%) (**Figure 2C)**. The percentage of high-quality raw reads aligning to Reference Genome GRCh37 had a median of 96.3% and a spearman correlation of 0.98 (**Figure 2D**). Thus, we detected 11,929 protein-coding genes or lncRNAs expressed in at least 50% of the samples.

Gene expression: a comparison with short-read RNA-seq

For all the 60 LCL samples of the current work, we downloaded gene quantifications based on Illumina short-read RNA-seq data from previous studies [18, 29]. We compared gene expression quantifications from short and long-read on 11,542 protein-coding genes and lncRNAs expressed in 50% of the samples that were also detected in with short-read RNA-seq technology [26] (**Figure 3A**). A pair-wise correlation between samples had a median Spearman correlation between 0.61 and 0.79 (p-value < 2×10^{-286}) (**Figure 3B**, light blue). In addition, 3,170 protein-coding and lncRNA genes were detected only using short-read data. These genes undetected by ONT were low expressed with a median expression of 0.426 RPKM (**Figure 3C**), being 11 times less expressed than the median RPKM expression of the total number of protein-coding and lncRNA genes detected in Illumina (Wilcoxon test p-value < 2.2e-16). We also found that highly expressed genes quantified with long-

read ONT data correlated better with short-read RNA-seq data. These differences suggest a reduced sensitivity of dRNA-seq ONT for capturing low expressed genes compared to short-read Illumina technology, suggesting that direct ONT RNA sequencing requires a larger yield.

Identification of annotated transcripts and discovery of novel transcripts

Using FLAIR [30], we detected a total of 48,539 transcripts across 14,962 genes supported by at least 10 reads and with a mapping quality >10 and five TPM in at least one sample. Among these, 61.9% were novel transcripts (n = 30,041) while 38.1% were already annotated (n = 18,498), underlying the ability of long-read direct RNA sequencing to identify a larger proportion of novel transcripts. We observed that 35% of annotated isoforms, but only 7% of novel ones, were expressed in all the 60 LCLs (**Figure 4 A-B**). The evaluated genes had on average 1.6 isoforms. However, 2% of expressed genes had more than 10 novel transcripts and more than 5 annotated transcripts (**Figure 4 C-D**).

eQTLs and trQTLs discovery

We performed a genome-wide eQTL analysis of long-read native RNA-seq data in *cis* for the 13,996 protein-coding and lncRNA genes expressed in at least 50% of the samples (see **Material and Methods** for details). We discovered 45 significant eQTLs (FDR 5%), a considerably lower number compared to the 536 eQTLs that were discovered using 60 short-read Illumina RNA-seq LCL samples [26]. This result suggests that the low gene coverage of our long-read data (1.7 million reads per sample) is a limitation for eQTL discovery. However, long-read native RNA-seq data also allows us to identify genetic variants affecting transcript abundance. We performed a genome-wide QTL analysis on 14,447 annotated transcripts from 9,364 unique genes that were expressed in at least 50% of the samples, detecting 72 trQTLs (FDR 5%). The vast majority of the trQTLs were from protein-coding genes (75%), and only 47% of trQTLs affected the most expressed transcripts per gene. Transcripts with significant trQTLs had between 3 and 144 annotated exons, with a mean of 29 exons. Among the genes with a trQTL, we observed two examples with no eQTLs identified with the ONT dataset, but both cases having a significant eQTL in the Illumina short-read dataset different to the

trQTLs in ONT meaning that a different SNP is mediating the genetic effect on gene expression compared to transcript expression. These genes were *BCL2A1* and *EIF5A*, being in the Illumina dataset under the genetic effect of the SNP rs2163005 (*p*-value = $3.37891e^{-07}$) and rs28636077 (*p*-value = $7.999e^{-58}$), respectively. The *BCL2A1* gene encodes a member of the BCL-2 protein family. This family of proteins acts as an anti- and pro-apoptotic regulator involved in a wide range of cellular functions, including embryonic development, homeostasis, and tumorigenesis. *BCL2A1* gene has three exons and generates two known transcripts, both detected in ONT dataset (**Figure 5A**). The SNP rs8025803 was also identified as a trQTL (*p*-value = $1.08601e^{-09}$) in the ONT dataset affecting the expression of the less expressed transcript (ENST00000335661.6, green in **Figure 5B**). While the association between the SNP rs8025803 and the other transcript (ENST00000267953.3) was not significant (*p*-value = 0.98, blue in **Figure 5B**). Our second example is the *EIF5A* gene, which codes for a protein involved in the translation elongation process and has an important function at the level of mRNA turnover. The gene has eight exons, and six annotated transcripts in our dataset (**Figure 5C**). The SNP rs4796398 is a trQTL affecting the expression of the most abundant transcript of the gene (ENST00000336458.8, represented in green in **Figure 5D)** and had a *p*-value = $1.36902e^{-09}$. The other expressed transcripts of the gene were not affected by the trQTL (**Figure 5D**).

Overall, we observed that using the same number of samples in Illumina and ONT dataset (n = 60), we could find more QTLs using transcripts (n = 72) than gene-level quantifications (n = 45), and only 12 SNP-gene associations were both eQTLs and trQTLs (**Figure 6**). Among the 72 trQTLs, 11 trQTLs affected the most expressed transcript per gene; only one trQTL, for *NDUFS5*, involved the less expressed transcript. The *NDUFS5* gene is an NADH Dehydrogenase [Ubiquinone] Iron-Sulfur Protein 5 that due to alternative splicing expresses two different annotated transcripts, both of them detected in our samples (**Figure 7A**). The lead SNP rs12043492 was detected as a significant trQTL for one the transcripts of the *NDUFS5* gene (ENST00000372969.3, *p*-value = $8.64797e^{-08}$, **Figure 7B**) while the other transcript of the same gene (ENST00000372967.3) was not associated (*p*-value

= 0.24, **Figure 7C**). However, the same SNP rs12043492 was also an eQTL for *NDUFS5* gene (*p*-value = 5.26765e^{-08}, **Figure 7D**).

On the other hand, of the 45 genes with signification eQTLs, 19 (42.4%) also has a significant trQTL. However, for 7 of these 19 genes the SNP affecting the gene expression (eQTL) and the SNP affecting the transcript expression (trQTL) is not the same; only 12 genes, as previously mentioned, shared the same SNP for eQTL and trQTL (**Figure 6**). The 19 genes affected by both an eQTL and a trQTL did not show a significant difference in terms of the number of exons, the number of annotated transcripts or the number of novel transcripts compared to the 45 genes with eQTLs in ONT dataset (Wilcoxon test *p*-value = 0.92, 0.34 and 0.51 respectively). Even when we compared these 19 genes with the genes not associated with any trQTL we would not detect any significant difference in terms of the number of exons, the number of annotated and novel transcripts detected (Wilcoxon test *p*-value = 0.84, 0.13, 0.3, respectively). Additionally, we observed that the genes affected by eQTL (n = 45) had a lower number of annotated transcripts than those genes affected by trQTLs (n = 72, Wilcoxon test *p*-value = 5.23e-05), suggesting that genetic effects on genes with higher transcripts diversity are missed using gene-level summary quantifications and eQTL analyses.

Long-read sequencing informs of the role of splicing in eQTLs

We then assessed the level of replication of our long-read sequencing QTL results with short-read eQTLs previously published using a cohort of 317 LCL samples. We found that five of the 45 long-read eQTLs overlapped with Illumina eQTLs, but only one of the 72 trQTL was also detected as an eQTL in both, the ONT and Illumina dataset **(Figure 8B-C**). This trQTL, rs12366 (*p*-value = 2.27959e^{-08}), involved the expression of the gene *POLE4* which is a DNA Polymerase Epsilon Subunit 4 with seven already annotated transcripts. The affected transcript ENST00000483063.1 was also the most expressed transcript of the two that we detected **(Figure 8A, D-E)**.

Gene-level eQTLs detect genetic effects on gene expression abundances without identifying possible effects on splicing or specific transcript abundance. However, the identification of eQTLs or trQTLs

is strongly dependent on statistical power, i.e., sample size. To bypass the limitations of our study, we evaluated the characteristic of the 7,658 eQTLs discovered using 317 Illumina short-read LCL samples [26] in our ONT dataset. Here we focused on 4,157 Illumina eQTLs that are associated to gene expression and expressed in at least 50% of the 60 long-read samples. After extracting the *p*-values for all SNP-gene and SNP-transcripts pairs and performing multiple testing corrections, we found that 7.7% of Illumina eQTLs were significant ONT eQTLs (*p*-value < 0.05, $\pi_1 = 0.22$), and 5% of Illumina eQTLs were significant ONT trQTLs (*p*-value < 0.05, $\pi_1=0.24$). Among the 507 genes associated with either an eQTL or a trQTL (eGenes), only 13 genes – SNPs pairs were both eQTLs and trQTLs (**Figure 9**). For the remaining genes, 192 had only significant trQTLs (37.9%) while 302 had only eQTLs (59.6%). Our results highlight the ability of long-read sequencing to capture genetic effect on gene expression, while identifying specific effects on transcript expression.

Discussion

In this study, we generated a direct-RNA long-read sequencing dataset from 60 genetically different LCLs of European ancestry. This unique dataset allowed us to investigate the genetic regulation of transcript abundance, offering a chance to understand the role of splicing as a regulatory mechanism of gene expression [15, 16]. Because the reconstruction of the structure of the transcript from short-read sequencing is challenging, it is difficult to capture and comprehend how events such as gene splicing may play a role in disease [4, 6, 7, 31]. We overcame those limitations using long-read technology while identifying which of the known genetic effects on gene expression in LCLs reflected the regulation in the abundance of specific transcripts.

We first observed that the median read length distributions and the median read quality were similar to other publications using long-reads and different cell lines [32]. Compared to short-read datasets, we reported comparable gene-level expression quantifications in the same LCL samples [26]. However, we were unable to detect low expressed genes with long-read direct RNA-seq, highlighting the need for technological improvements to produce a higher sequencing yield. As a consequence, we were limited in the identification of eQTLs with direct RNA long-read sequencing compared to Illumina short-read RNA-seq, identifying fewer eQTLs (8.4% of the total Illumina eQTLs) in comparison using the same 60 samples for both technologies. We attributed this limitation to the reduced coverage obtained with ONT direct RNA-seq, which agrees with previous studies [32, 33]. In particular, previous work from Soneson et al. [32], already suggested that the low sequencing depth of ONT long-read sequencing could result in a reduced number of detected genes and a decreased quantification accuracy. Together with our observations, this shows that addressing these limitations is important when designing population studies aiming to dissect the interplay between a genetic variant and gene expression variation using long-read sequencing technologies.

Nevertheless, we identified a large proportion of novel transcripts (61.9% of the transcripts detected). Overall, we identified one transcript per gene for most genes, often already annotated, whereas

additional novel transcripts were discovered in complex genes with a large number of already annotated transcripts. These results are in line with previous findings by Glinos et al. [12] who sequenced whole-genome transcriptome using ONT across multiple tissues. The novel transcripts were the results of multiple alternative splicing patterns which were difficult to characterize with previous short-read technologies in the attempt to understand the functional consequences of different splicing changes using single exon junctions [34]. In particular, based on short-reads, it is difficult to determine the connectivity of exons because the data is highly fragmented [35]. By sequencing full-length transcript molecules, we are now able to provide more information on exon connectivity and on the effects of alternative splicing events that result in a wide range of different transcripts [36-38]. However, work is further needed to precisely characterize the large number of novel transcripts. For example, some of these novel transcripts may be the result of mRNA molecules being fragmented during the sample processing and sequencing, while others may be the result of biological errors. Investigating the biological relevance of the "novel" transcripts, if any, would be essential to provide a better annotation of gene isoforms.

We decided to focus only on annotated transcripts when addressing the question of the impact of genetic variants on transcript abundance, as these annotated transcripts allowed a better comparison with published short-read datasets. Significantly, we could discover genetic effects on gene expression that short-read technology missed, even considering the small sample size of our study and the larger multiple testing burden of testing multiple transcripts per gene. Specifically, we discovered 72 trQTLs that significantly affected the expression of the transcripts. We observed that when trQTL's effects involved the most expressed transcript of a gene, their effect was also captured by the eQTL analysis, but if the trQTL involved one of the less expressed transcripts, the effect would not be re-capitulated using gene-level eQTL analysis. Moreover, among the 72 trQTLs discovered, 60 did not overlap with any eQTLs found in our study or previously discovered with Illumina short-read sequencing [26]. From this, we conclude that long-read sequencing identifies the specific

transcript affected by the genetic variant, providing more accurate information about the underlying processes involved in expression regulation. Ultimately, even if our study did not discover novel eQTLs, we were able to report that 7.4% of significant eQTLs identified with short-reads were the results of changes in the expression of specific transcripts (trQTLs), which helps us to untangle the nature of the genetic regulation.

Our study reinforces the importance of analyzing the transcriptome not only at the gene level but also with a detailed understanding of specific transcript expression [39]. In light of the relevant role that genetic variants play in transcript abundance and structure, we expect that a high-resolution transcriptome characterization based on long-read data will be one of the main strategies for identifying the underlying mechanisms of disease-associated variants [15, 16, 40, 41]. From our study, we conclude that direct RNA-seq using long-reads will not increase dramatically the ability to discover genetic effects on gene expression, as shown by the low number of eQTLs we found, but it will greatly contribute to our understanding and characterization of the mechanism of eQTL effects.

References

1. Park, E., et al., *The Expanding Landscape of Alternative Splicing Variation in Human Populations.* Am J Hum Genet, 2018. **102**(1): p. 11-26.
2. Kelemen, O., et al., *Function of alternative splicing.* Gene, 2013. **514**(1): p. 1-30.
3. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
4. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
5. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.
6. Teng, M., et al., *A benchmark for RNA-seq quantification pipelines.* Genome Biol, 2016. **17**, 74. doi: 10.1186/s13059-016-0940-1.
7. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.
8. Sedlazeck, F.J., et al., *Piercing the dark matter: bioinformatics of long-range sequencing and mapping.* Nat Rev Genet, 2018. **19**(6): p. 329-346.
9. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome Biol, 2020. **21**(1): 30. doi: 10.1186/s13059-020-1935-5.
10. Weirather, J.L., et al., *Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis.* F1000Res, 2017. **6**:100. doi: 10.12688/f1000research.10571.2. eCollection 2017.
11. Anvar, S.Y., et al., *Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing.* Genome Biol, 2018. **19**(1): 46. doi: 10.1186/s13059-018-1418-0.
12. Glinos, D.A., Garborcauskas, G., Hoffman, P. et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature (2022). doi: 10.1038/s41586-022-05035-y.
13. Tilgner, H., et al., *Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events.* Nat Biotechnol, 2015. **33**(7): p. 736-42.
14. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population.* Nature, 2010. **464**(7289): p. 773-7.
15. Li, Y.I., et al., *RNA splicing is a primary link between genetic variation and disease.* Science, 2016. **352**(6285): p. 600-4.
16. Consortium, G., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.
17. Garrido-Martín, D., et al., *Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome.* Nat Commun, 2021. **12**(1): 727. doi: 10.1038/s41467-020-20578-2.
18. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
19. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
20. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2):giab008. doi: 10.1093/gigascience/giab008.
21. De Coster, W., et al., *NanoPack: visualizing and processing long-read sequencing data.* Bioinformatics, 2018. **34**(15): p. 2666-2669.
22. Leger, *pycoQC, interactive quality control for Oxford Nanopore Sequencing.* Journal of Open Source Software, 2019. 4(34), 1236, https://doi.org/10.21105/joss.01236.

23. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.
24. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
25. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
26. Delaneau, O., et al., *Chromatin three-dimensional interactions mediate genetic effects on gene expression.* Science, 2019. **364**(6439). doi: 10.1126/science.aat8266.
27. Delaneau, O., et al., *A complete tool set for molecular QTL discovery and analysis.* Nat Commun, 2017. **8**: 15452. doi: 10.1038/ncomms15452.
28. Wei, T., & Simko, V., *R package "corrplot": Visualization of a Correlation Matrix.* (Version 0.92), *https://github.com/taiyun/corrplot*. 2021.
29. Storey, J.D., et al., *qvalue: Q-value estimation for false discovery rate control.* R package version, 2015. **2**(0): p. 10.18129.
30. Tang, A.D., et al., *Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.* Nat Commun, 2020. **11**(1): 1438. doi: 10.1038/s41467-020-15171-6.
31. Arnold, M., et al., Global burden of cutaneous melanoma *attributable to ultraviolet radiation in 2012.* Int J Cancer, 2018. **143**(6): p. 1305-1314.
32. Soneson, C., et al., *A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes.* Nat Commun, 2019. **10**(1): 3359. doi: 10.1038/s41467-019-11272-z.
33. Workman, R.E., et al., *Nanopore native RNA sequencing of a human poly(A) transcriptome.* Nature Methods, 2019. **16**(12): p. 1297-1305.
34. Li, Y.I., et al., *Annotation-free quantification of RNA splicing using LeafCutter.* Nat Genet, 2018. **50**(1): p. 151-158.
35. Steijger, T., et al., *Assessment of transcript reconstruction methods for RNA-seq.* Nature Methods, 2013. **10**(12): p. 1177-1184.
36. Bolisetty, M.T., G. Rajadinakaran, and B.R. Graveley, *Determining exon connectivity in complex mRNAs by nanopore sequencing.* Genome Biol, 2015. **16**: 204. doi: 10.1186/s13059-015-0777-z.
37. Sharon, D., et al., *A single-molecule long-read survey of the human transcriptome.* Nat Biotechnol, 2013. **31**(11): p. 1009-14.
38. Bueno, R., et al., *Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations.* Nat Genet, 2016. **48**(4): p. 407-16.
39. Alasoo, K., et al., *Genetic effects on promoter usage are highly context-specific and contribute to complex traits.* Elife, 2019. **8**:e41673. doi: 10.7554/eLife.41673.
40. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
41. Gandal, M.J., et al., *Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder.* Science, 2018. **362**(6420):eaat8127. doi: 10.1126/science.aat8127.

Figure legends

Figure 1. Scheme depicting the different workflows used in the study.

(A) Overview of the experimental workflow and analysis pipelines used for direct RNA-seq with ONT. The three major steps are indicated by the blue titles. LCLs cultures were in the exponential growth phase and free of mycoplasma contamination. The preparation of RNA for the library consisted: of total RNA extraction with Trizol, enrichment of polyA$^+$ RNA, generation of the library with SQK-RNA002 and loading of the libraries in the FLOW-MINI106D flow cell. Samples were run for 72h and analysis was done afterwards as described. **(B)** Analysis workflow from raw reads alignment, mapping, quality control, gene quantification (left) and transcript quantification (right). Reads per kilobases per million (RPKM) are used for gene quantification (left). The transcript identification and quantification was performed using the FLAIR pipeline (right). Transcript per million (TPM) was used for transcripts quantification. The Reference Genome used was the GRCh 37 with the gene annotation GENCODE v19 for both gene and transcript quantification.

Figure 2. Quality control and summary statistics of the ONT sequencing data.

Histograms representing the **(A)** median number of reads, **(B)** median reads length, and **(C)** median read quality across the LCL population of the ONT sequencing data. **(D)** Scatter-plot comparing the number of raw reads and number of aligned reads in the 60 LCL samples. A median of 96.3% of raw reads was mapped correctly to the Reference Genome (GRCh 37) with a spearman correlation of $\rho = 0.98$.

Figure 3. Comparison of protein-coding and lncRNA gene expression between ONT and Illumina short-reads.

(A) Venn diagram showing the protein-coding genes and lncRNA expressed in 50% of the sample by ONT and Illumina short-read technologies. In the samples, the expression of 11,154 genes and lncRNA were detected by both technologies. In ONT, 3,170 genes were not detected and 387 were missed by Illumina. **(B)** Pair-wise Spearman correlation between the overall gene expression in the

same 60 LCLs samples among and between technologies. The correlation observed inside the same technology, median ρ = 0.906 and 0.95 in ONT and Illumina, respectively. Whereas the correlation in gene expression between the two technologies was median ρ = 0.695. **(C)** Histogram of the total number of protein-coding and lncRNA genes detect using Illumina technology (blue) and the distribution of the 3,170 protein-coding and lncRNA genes detected in Illumina but not with ONT (yellow). In red, are the two medians of the two distributions, 5.23 for the blue histogram and 0.42 for the yellow one. *p*-value < 2.2e-16 (Wilcoxon test).

Figure 4. Distribution of annotated and novel isoforms in the LCLs population.

Histograms showing the distribution of **(A)** annotated and **(B)** novel isoforms expressing at least 5 transcripts per million (TPM) in at least one sample in the LCLs population. Whereas the other histograms represent the distribution for **(C)** annotated and **(D)** novel isoforms per expressed genes.

Figure 5. Significant trQTLs for *BCL2A* and *EIF5A* transcripts.

Schemes representing the gene model for splicing and most abundant transcripts structure of **(A)** *BCL2A* and **(C)** *EIF5A* genes. Box-plot representing the genotype (0, 1 and 2 for homozygous for the reference allele, heterozygous and homozygous for the alternate allele, respectively) *versus* transcript expression expressed transcripts per million (TPM) and where outliers are shown as black dots. In **(B)**, the effect of the SNP rs8025803 on the expression of the BCL2A transcripts. rs8025803 is a significant trQTL only for the ENST00000335661.6 transcript expression (*p*-value = 1.08e-09, in green). In **(D)** the effect of the SNP rs4796398 on the expression of the *EIF5A* transcripts. rs4796398 is a significant trQTL only for the transcript ENST00000336458.8 (*p*-value = 1.36e-09, green) but not for the other 4 transcripts.

Figure 6. The SNP – gene overlap between datasets obtained with different technologies.

(A) Venn diagram representing the eQTL – gene overlap between eQTLs affecting gene expression in ONT, Illumina, and QTL affecting transcripts expression (trQTLs) determined by ONT. Only one

eQTL-gene association overlapped across the three analysis [27]. Red highlights the 60 trQTL - ONT that did not overlap with any eQTLs - ONT and eQTL – Illumina.

Figure 7. Significant eQTLs for the gene *NDUFS5* and its transcript.

(A) Scheme representing the gene and most abundant transcripts structures of *NDUFS5*. Box-plots representing the genotype (0, 1 and 2 for homozygous for the reference allele, heterozygous and homozygous for the alternate allele, respectively) *versus* gene expression or transcript expression are shown as Reads per kilobases per million (RPKM) and transcripts per million (TPM), respectively; and where outliers are shown as black dots. The SNP rs12043492 is an eQTL for both, gene expression and transcript expression. rs12043492 is **(B)** a significant trQTL for the ENST00000372967.3 (green) (*p*-value = 8.64e-08), but **(C)** no effect of the second transcript ENST00000372969.3 (blue). In **(D)** showing the effect of the SNP rs12043492 on the *NDUFS5* gene expression (eQTL) (*p*-value = 5.26e-08).

Figure 8. Significant eQTLs for the gene *POLE4* and derived transcripts.

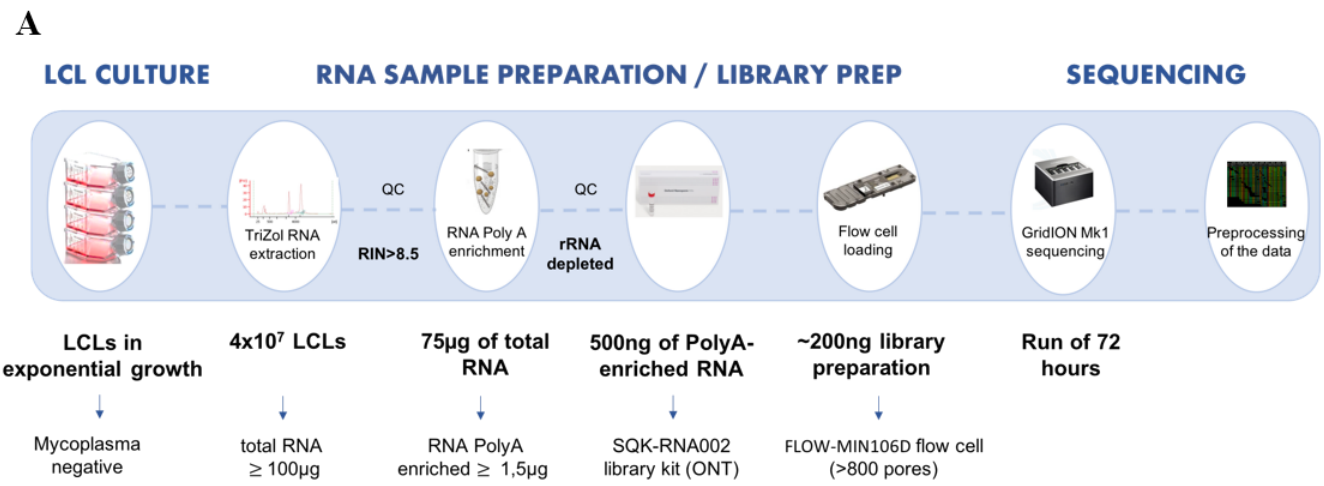
(A) Scheme representing the gene and most abundant transcripts structures for the gene *POLE4*. Box-plots representing the genotype (0, 1 and 2 for homozygous for the reference allele, heterozygous and homozygous for the alternate allele, respectively) *versus* gene expression or transcript expression are shown as Reads per kilobases per million (RPKM) transcripts per million (TPM), respectively; and where outliers are shown as black dots. The effect of the SNP rs12366 on *POLE4* gene expression measured by **(B)** Illumina (*p*-value = 1.04e-38) and **(C)** ONT dataset (*p*-value = 2.98e-08). The SNP rs12366 is also a significant trQTL for one transcript of the gene *POLE4*, **(D)** ENST00000483063.1 (*p*-value= 2.27e-08), but **(E)** not for the alternative ENST00000465242.1, (*p*-value = 0.17).

Figure 9. The SNP – gene overlap between Illumina eQTL and eQTLs, and trQTLs in the ONT dataset.

(A) Venn diagram representing the eQTL – gene overlap between eQTLs affecting gene expression in ONT data, eQTLs affecting gene expression in Illumina and eQTLs affecting transcripts expression (trQTLs).

Figures

Figure 1



B

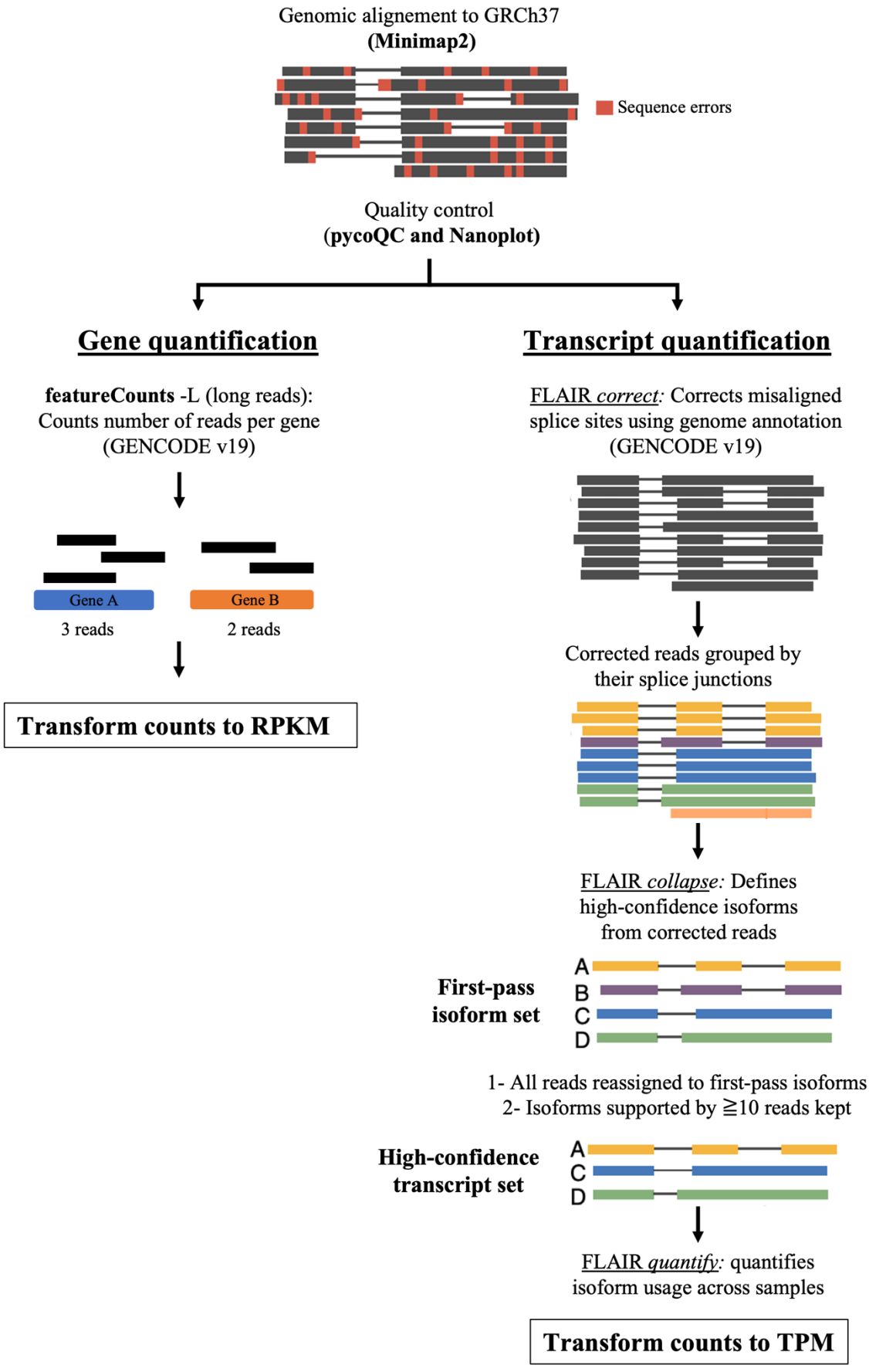

Genomic alignement to GRCh37
(Minimap2)
Sequence errors
Quality control
(pycoQC and Nanoplot)
Gene quantification
Transcript quantification
featureCounts -L (long reads):
Counts number of reads per gene
(GENCODE v19)
Gene A
Gene B
3 reads
2 reads
Transform counts to RPKM
FLAIR correct: Corrects misaligned
splice sites using genome annotation
(GENCODE v19)
Corrected reads grouped by
their splice junctions
FLAIR collapse: Defines
high-confidence isoforms
from corrected reads
First-pass
isoform set
A
B
C
D
1- All reads reassigned to first-pass isoforms
2- Isoforms supported by ≧10 reads kept
High-confidence
transcript set
A
C
D
FLAIR quantify: quantifies
isoform usage across samples
Transform counts to TPM

Figure 2

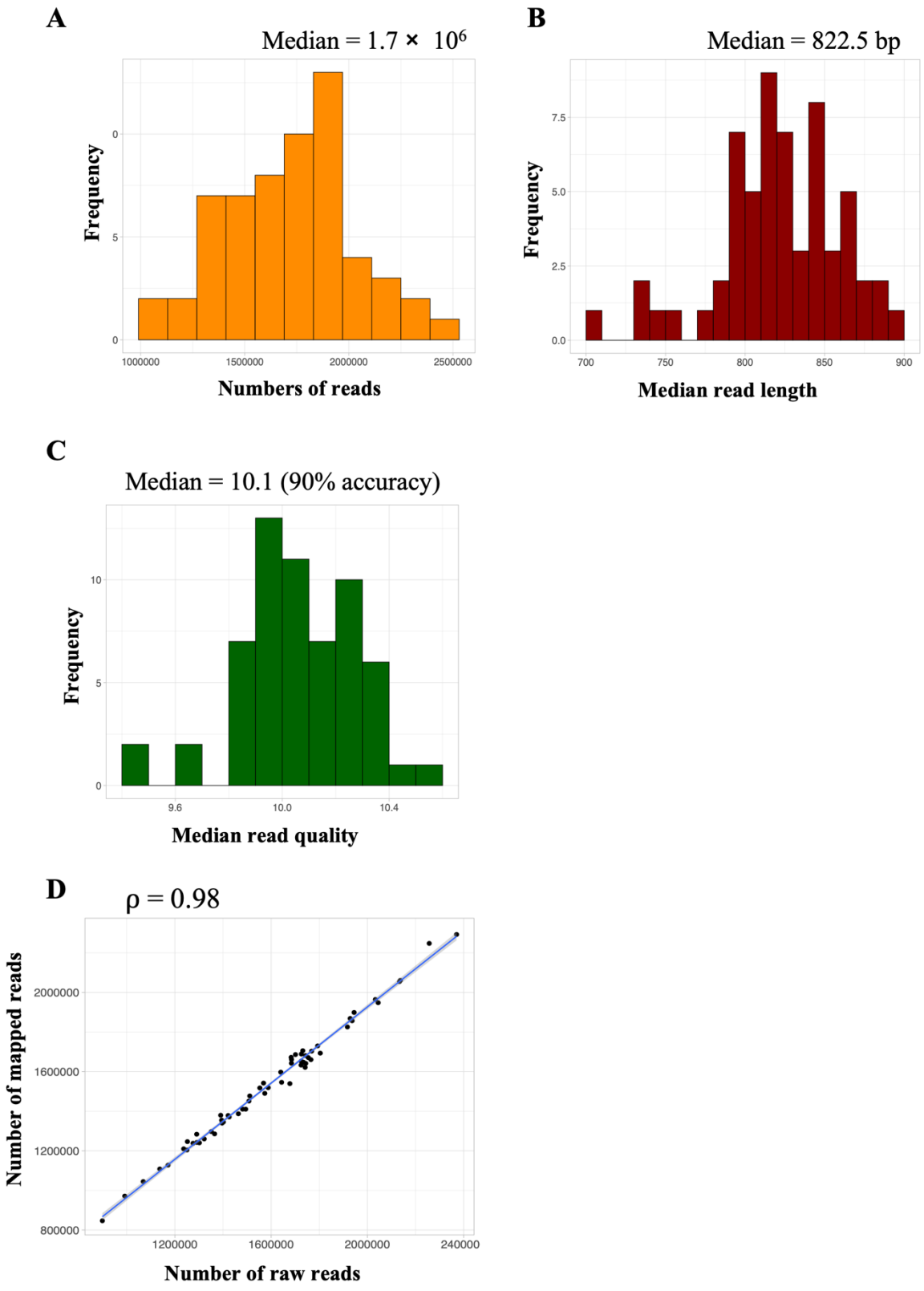
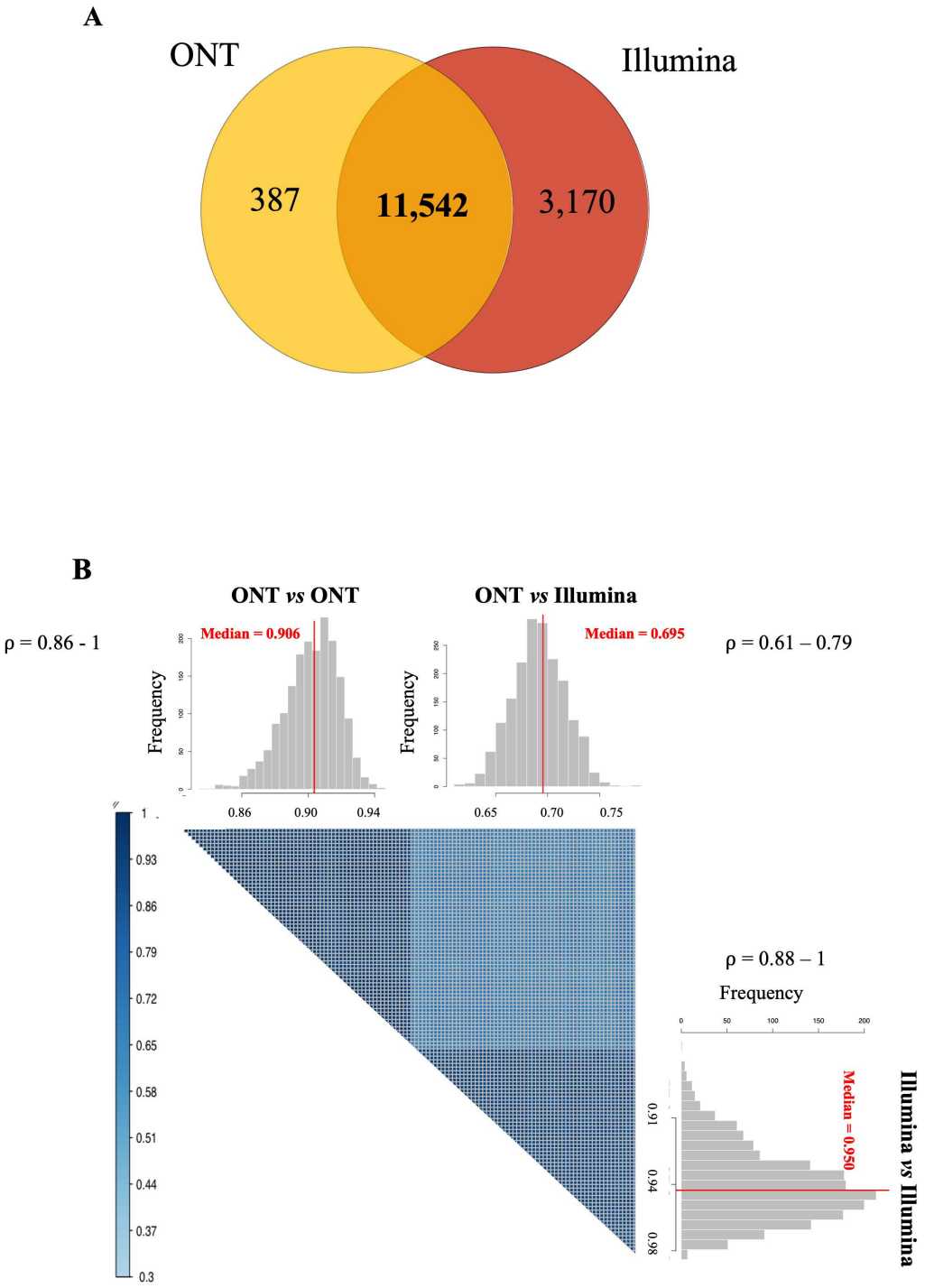

A
Median = 1.7 × 10⁶
Frequency
Numbers of reads
B
Median = 822.5 bp
Frequency
Median read length
C
Median = 10.1 (90% accuracy)
Frequency
Median read quality
D
ρ = 0.98
Number of mapped reads
Number of raw reads

Figure 3


A
ONT
Illumina
387
11,542
3,170
B
ONT vs ONT
ONT vs Illumina
ρ = 0.86 - 1
Median = 0.906
Median = 0.695
ρ = 0.61 – 0.79
Frequency
Frequency
0.86
0.90
0.94
0.65
0.70
0.75
1
0.93
0.86
0.79
0.72
0.65
0.58
0.51
0.44
0.37
0.3
ρ = 0.88 – 1
Frequency
Median = 0.950
Illumina vs Illumina

C

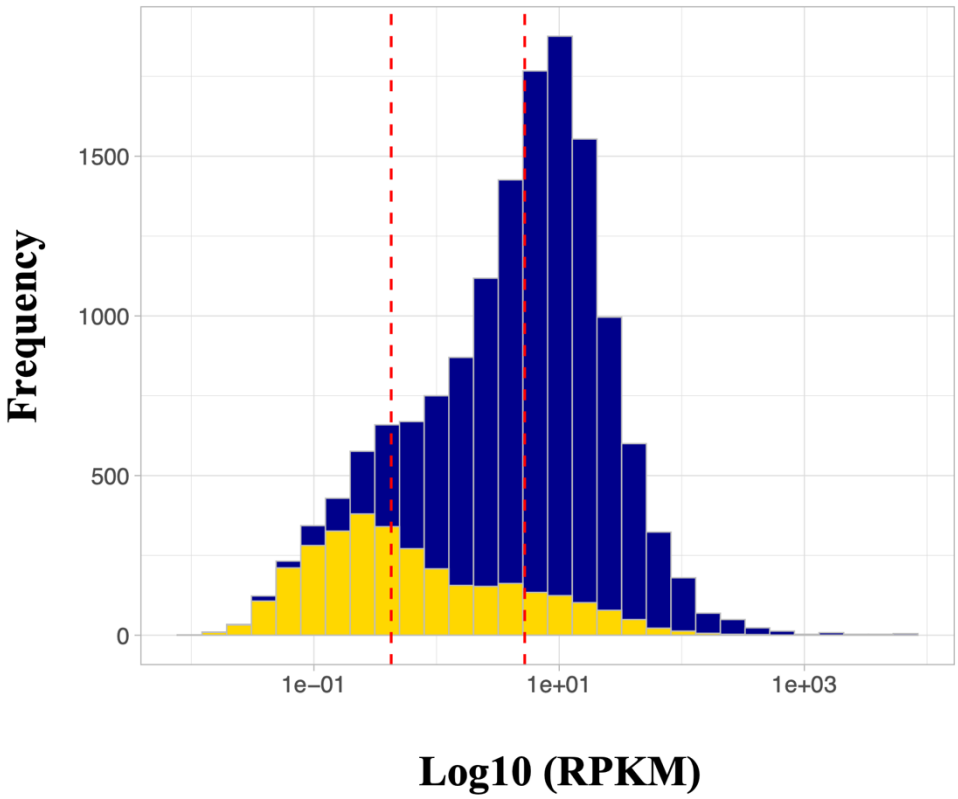


Figure 4

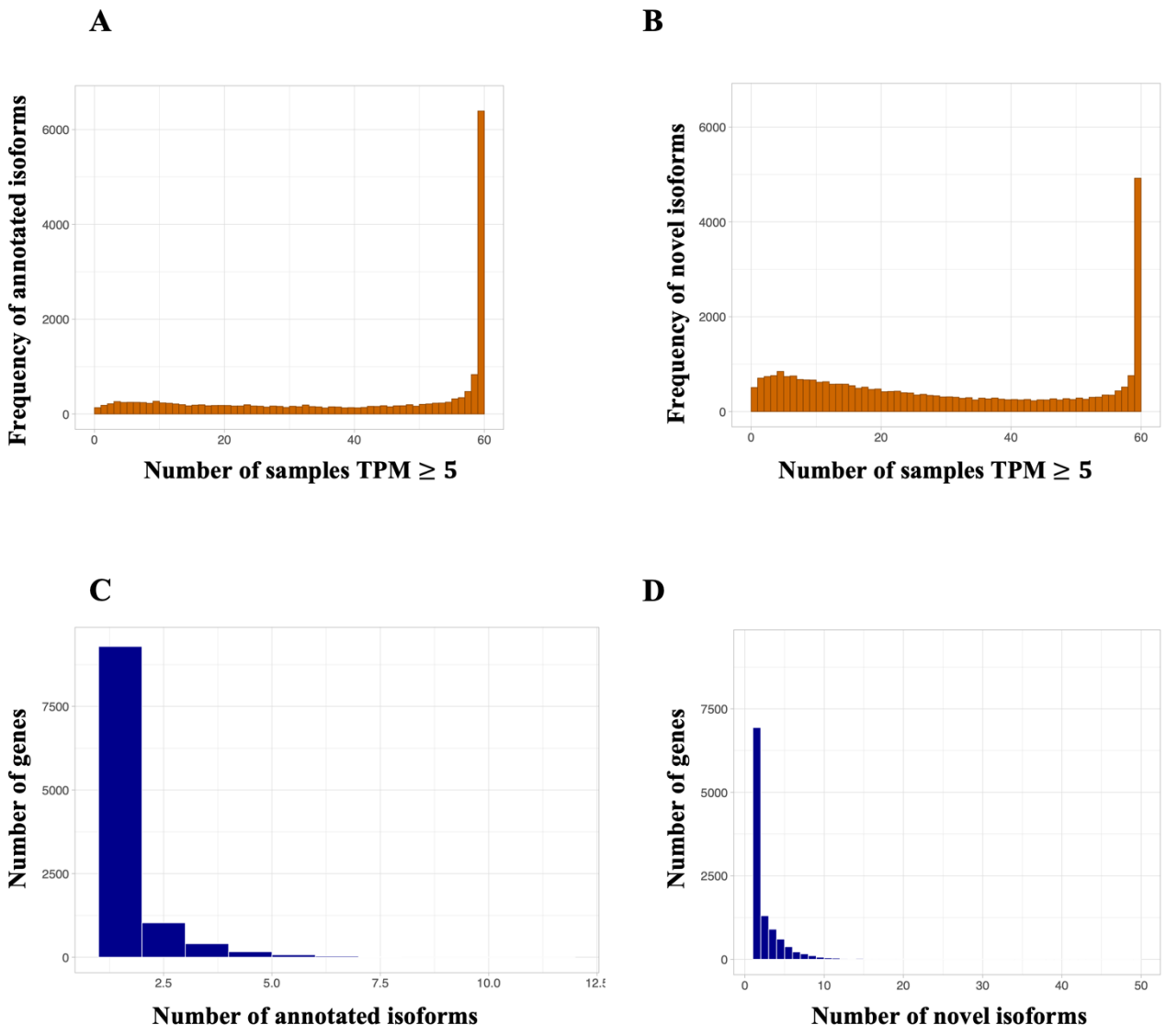
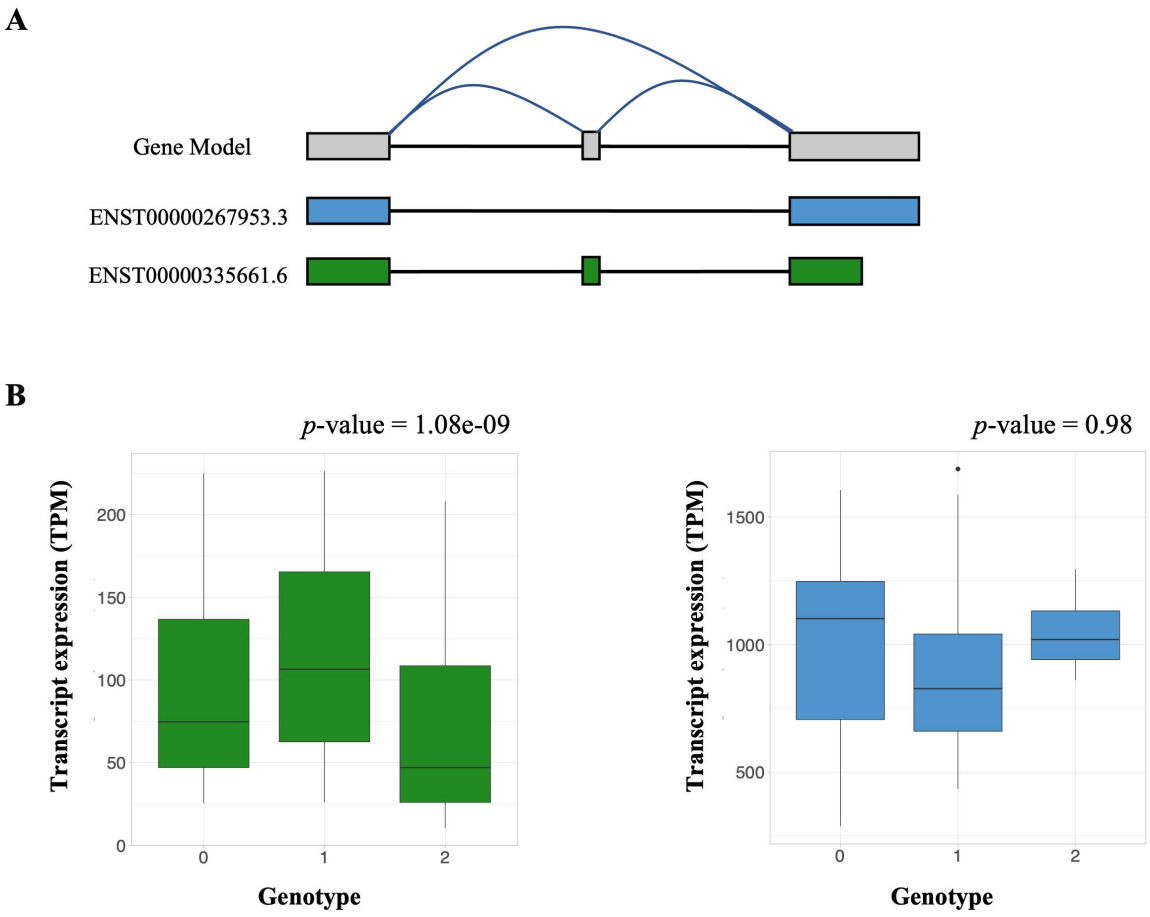
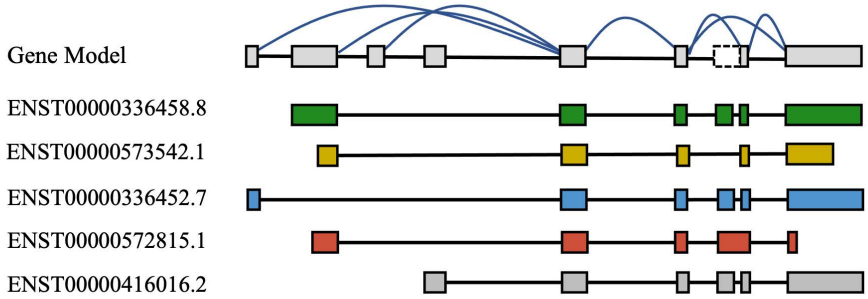

A
Frequency of annotated isoforms
6000
4000
2000
0
0
20
40
60
Number of samples TPM ≥ 5
B
Frequency of novel isoforms
6000
4000
2000
0
0
20
40
60
Number of samples TPM ≥ 5
C
Number of genes
7500
5000
2500
0
2.5
5.0
7.5
10.0
Number of annotated isoforms
D
Number of genes
7500
5000
2500
0
0
10
20
30
40
50
Number of novel isoforms

Figure 5


A
Gene Model
ENST00000267953.3
ENST00000335661.6
B
p-value = 1.08e-09
Transcript expression (TPM)
200
150
100
50
0
0
1
2
Genotype
p-value = 0.98
Transcript expression (TPM)
1500
1000
500
0
1
2
Genotype

C


Gene Model
ENST00000336458.8
ENST00000573542.1
ENST00000336452.7
ENST00000572815.1
ENST00000416016.2

D

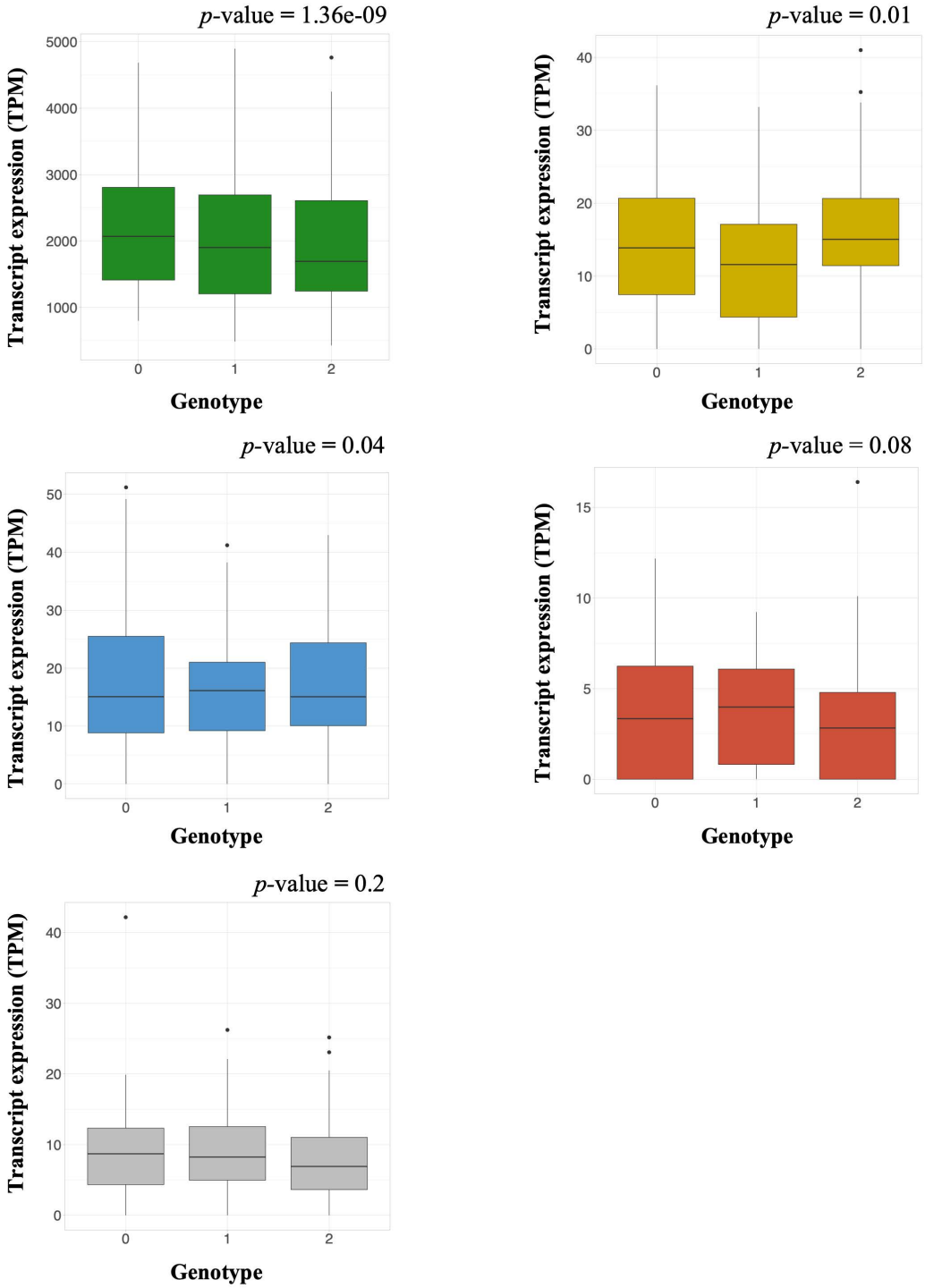

p-value = 1.36e-09
p-value = 0.01
p-value = 0.04
p-value = 0.08
p-value = 0.2
Transcript expression (TPM)
Genotype

Figure 6

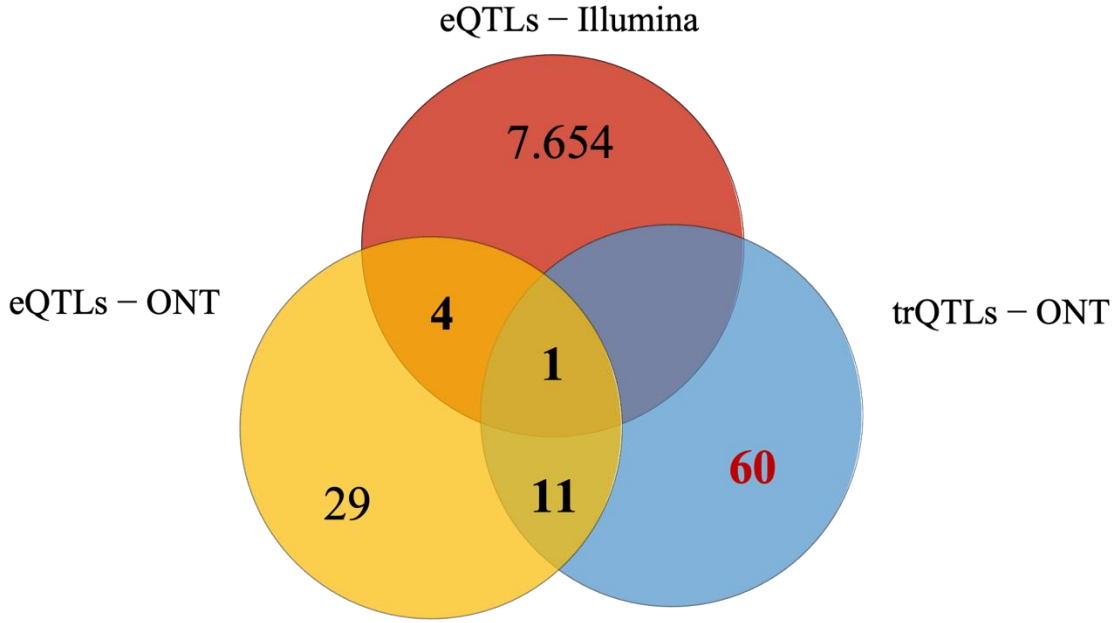


Figure 7

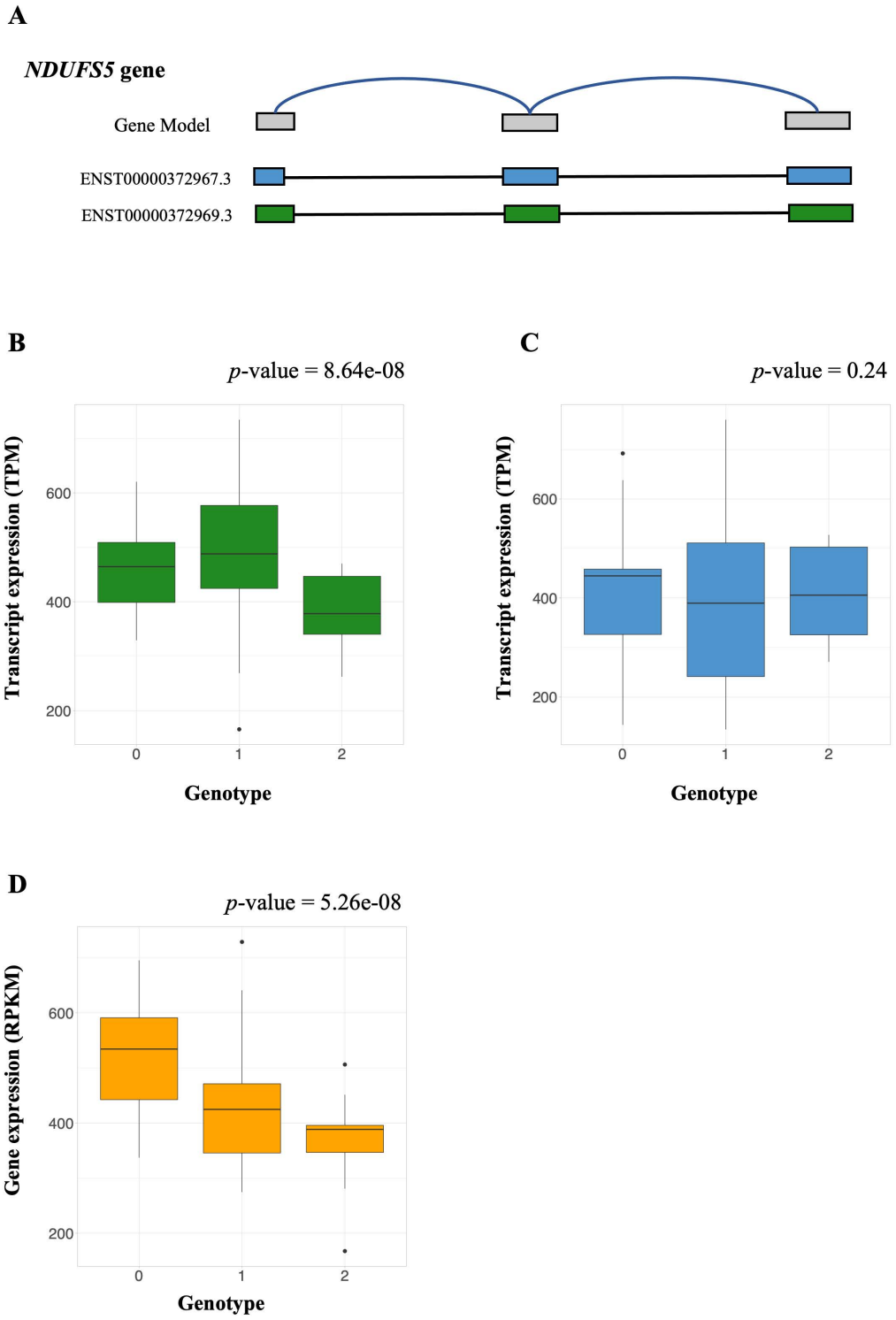

A
NDUFS5 gene
Gene Model
ENST00000372967.3
ENST00000372969.3
B
p-value = 8.64e-08
Transcript expression (TPM)
600
400
200
0
1
2
Genotype
C
p-value = 0.24
Transcript expression (TPM)
600
400
200
0
1
2
Genotype
D
p-value = 5.26e-08
Gene expression (RPKM)
600
400
200
0
1
2
Genotype

Figure 8

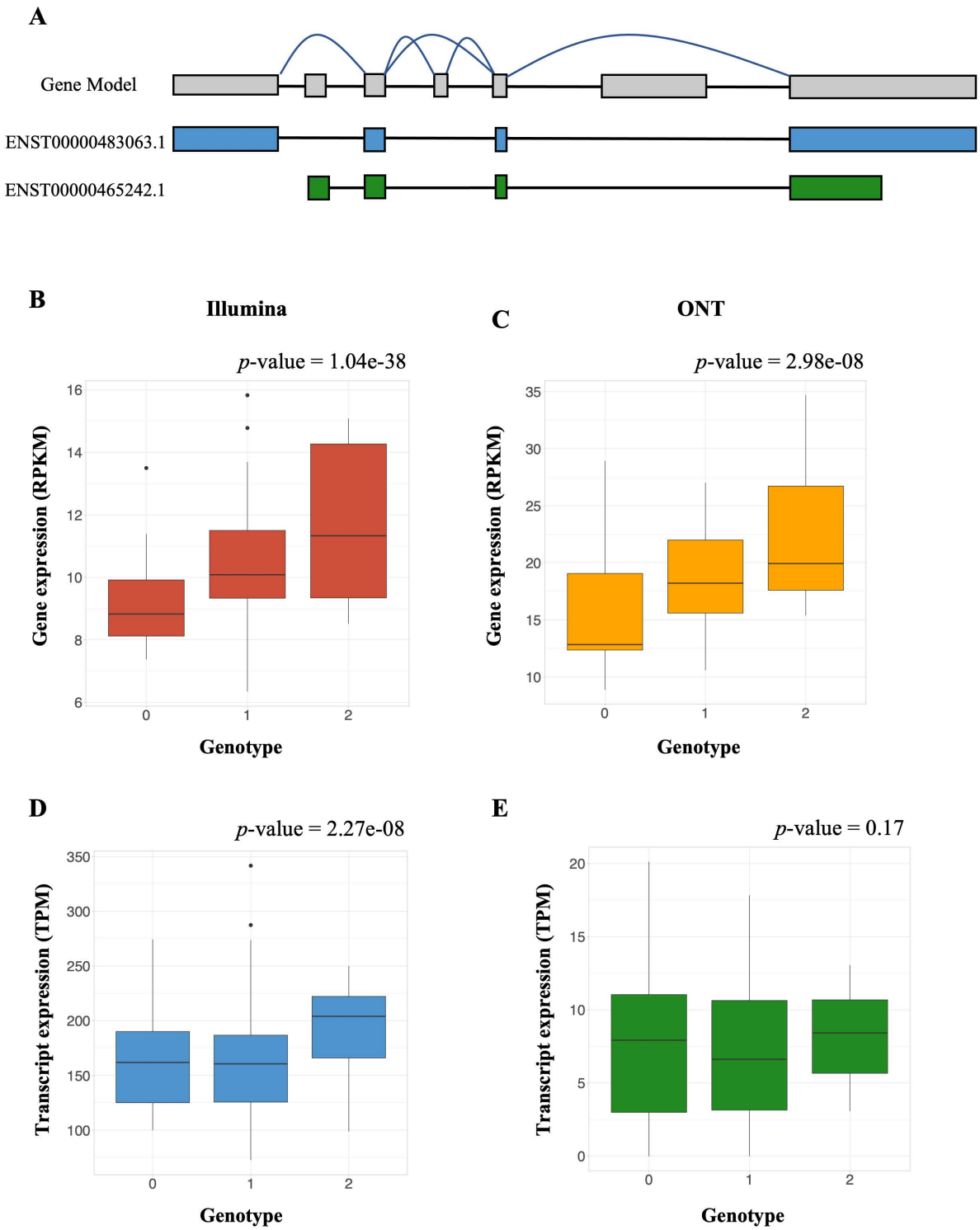
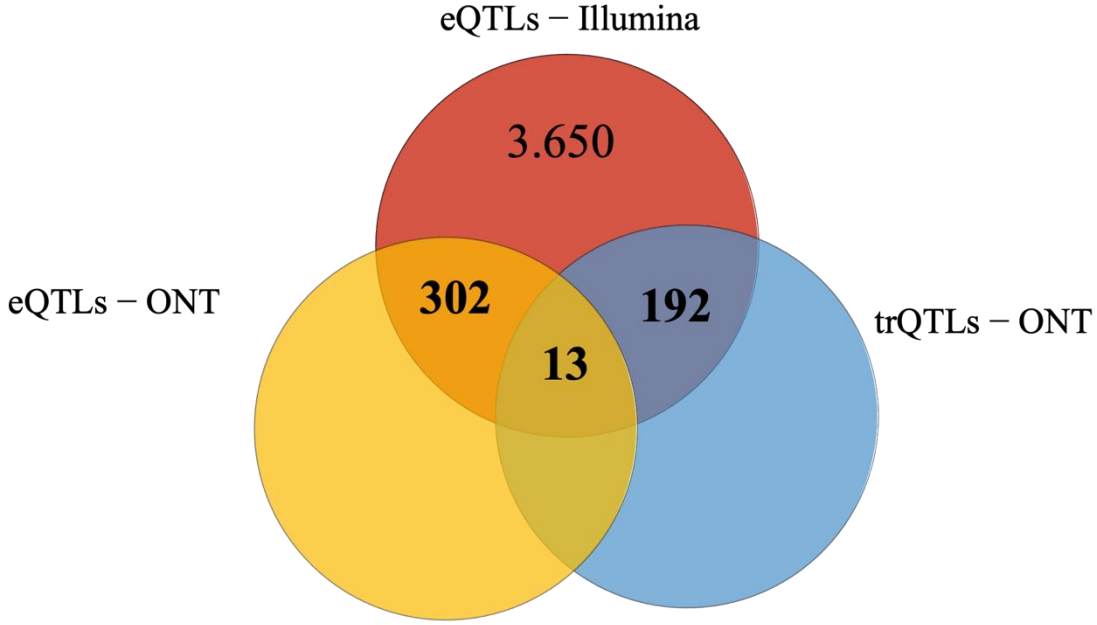

A
Gene Model
ENST00000483063.1
ENST00000465242.1
B
Illumina
p-value = 1.04e-38
Gene expression (RPKM)
Genotype
C
ONT
p-value = 2.98e-08
Gene expression (RPKM)
Genotype
D
p-value = 2.27e-08
Transcript expression (TPM)
Genotype
E
p-value = 0.17
Transcript expression (TPM)
Genotype

Figure 9



Article 2.

Dissecting the genetic contribution in cancer-related phenotypes using EBV-transformed lymphoblastoid cell lines.

Aline Réal, Gisella Puga Yung, Christelle Borel, Anna Ramisch, Andrew Brown, Emmanouil T. Dermitzakis, Jörg D. Seebach and Ana Viñuela.

This chapter is a preprint of the *in preparation* for submission manuscript

Goal: The aim of this study is to generate population-based cancer-like *in vitro* phenotypes using 87 genetically different LCLs to understand the link between germline variation, changes in gene expression and alteration in cancer-like phenotypes. The ultimate goal is to identify putative genes that could predict cancer-like phenotypes.

Personal contribution: For this manuscript in preparation, I performed all the experimental part of the study including the culture of LCLs, the design of the experimental assays in collaboration with the co-authors of the manuscript and the execution of the assays *in vitro*. I also performed the bioinformatic part and the data analysis under the supervision of the co-directors and director of my PhD. In terms of the manuscript itself, I prepared the figures and wrote the manuscript and fruitful discussions were worked out with co-authors of the study during the analysis and drafting, writing, and revisions of the manuscript.

Dissecting the genetic contribution in cancer-related phenotypes using EBV-transformed lymphoblastoid cell lines.

Aline Réal,[1,2*] Gisella Puga Yung,[1] Christelle Borel,[2] Anna Ramisch,[3] Andrew Brown,[4] Emmanouil T. Dermitzakis,[2] Jörg D. Seebach[1‡] and Ana Viñuela[5‡*].

[1] Division of Immunology and Allergology, University Hospitals and Medical Faculty, Geneva, Switzerland

[2b] Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

[3] Department of Basic Neuroscience, University of Geneva Medical School, Geneva, Switzerland

[4] Population Health and Genomics, University of Dundee, Dundee, United Kingdom

[5] Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne, UK

‡ These authors contributed equally

* Corresponding authors

Competing Interest Statement

Emmanouil T. Dermitzakis is currently an employee of GSK. The work presented in this manuscript was performed before he joined GSK. All other authors declare no competing interests.

Funding

This work was supported by the FNS ME10662 and ME11559S to ETD and the Swiss National Science Foundation (SNSF, # CRSII5_198577) to JS.

Keywords

Cancer-like phenotypes, germline variations, LCL, GWAS.

Abstract

The conventional wisdom that cancer is a disease of somatic coding mutations was disproved by GWAS studies, which have revealed that non-coding regulatory factors also play a significant role in cancer development and progression. While coding mutations have been thoroughly studied, the contribution of non-coding variants in cancer development remains poorly understood and has only recently been the subject of investigation. In this study, we aimed to gain a deeper understanding of the non-coding genetic contribution to the development of cancer, and in particular identifying which genes are involved in this process. We developed an *in vitro* model in which we measured six cancer phenotypes related to proliferation, apoptosis and chemotaxis in 87 genetically different lymphoblastoid cell lines (LCLs). We performed Genome Wide Association Studies (GWAS) and found one significant association (pvalue <5e-08) with a phenotype which measured the cells' replication index. This variant rs12865307 is an eQTL for the long intergenic non-coding RNA (lincRNA) *LINC00539*, which was previously linked with tumor immune response in lung cancer. In addition, the variant is also associated with the expression of one of four annotated transcripts of the gene *SKA3*, which has been implicated in the promotion of proliferation and cell migration in different cancer types and thus, is another putative cancer driver gene. Finally, at a lower significance threshold (p-value $<10^{-4}$) we discovered 3,974 variants associated with cancer phenotypes that also had effects on gene expression, potentially implicating 175 genes. In summary, this study demonstrates the potential of using the genetic variations of LCLs as an experimental system to explore and dissect the contribution of the non-coding germline variants to cancer risk, particularly for EBV-related cancer types.

Abbreviations

LCLs, lymphoblastoid cell lines; GWAS, Genome wide association studies; lncRNAs, long non-coding RNAs; WES, whole exome sequencing; EBV, Epstein-Barr virus; FACS, fluorescence-activated cell sorter; CTV, CellTrace Violet dye; hFasL, human Fas ligand; FDR, False Discovery Rate; eQTLs, expression quantitative trait loci.

Introduction

Cancer has been typically categorized as a disease of somatic coding mutations, with these mutations often caused by environmental factors such as sun exposure for many skin cancers and smoking for lung cancers [1-5]. In the late fifties, researchers provided evidence that UV light induces damage to cell DNA, causing the initial type of somatic mutations which leads to UV-induced skin cancer [6]. This provided the first evidence of a mutational signature left behind in cancer cells as a result of being exposed to carcinogens. More recently, next generation sequencing has allowed to look systematically, in an unbiased way, at the somatic mutations that are present in different cancers. Early studies applied whole exome sequencing (WES) to tumor samples and healthy tissue to detect the mutations only present in the tumor and identify the genes in which these mutations fell [7]. This identified *TP53* as a gene which was frequently found to harbor mutations across a wide range of different cancer types [7, 8] including skin cancer, hepatocellular carcinomas and squamous-cell carcinomas. *TP53* was also found to contain mutations in lung cancer samples, possible due to smoke exposure [9]. Independent studies of different cancer types have found a variety of genes to frequently display coding mutations, with tumor samples showing more mutations than corresponding healthy tissue, demonstrating the importance of somatic mutations to the development of cancer and providing a tool by which genes involved in particular cancers can be identified.

An orthogonal approach to studying cancer can be found in GWAS. Instead of looking for mutations in the DNA of tumor cells, these studies investigate the relationship between cancer development and the inherited germline genome, comparing the frequencies of genetic variants between individuals who developed cancer to controls who did not. In this way researchers can identify genetic variants that contribute to cancer predisposition [10, 11]. Unlike the previous study design, GWAS are typically not restricted to studying only the genome that codes for proteins, indeed, many of the cancer risk variants that have been discovered lie outside the coding region of the genome and are thus assumed to be involved in the development of cancer by affecting transcription [12, 13]. Combining

this information with the previous evidence from exome studies has led to the proposal of the two-hit hypothesis: rare variants or somatic mutations with high penetrance may directly cause tumorigenesis (for example, as observed in familial cancer or discovered in exome studies [13]) whereas non-coding germline variants with low penetrance may modulate the effects of these somatic/rare variants [10]. Germline variants could increase the risk of developing cancer by modulating the expression of genes involved in tumor development, while the necessary changes at the protein level in genes responsible for cancer development are either rare (likely due to selection) or due to somatic mutations.

To study the impact of germline variants on cancer related phenotypes and investigate the genes mediating their effects we can use experimental models. As models for cancer research, cell lines derived from cancer patients are commonly used, since they are easy to growth and maintain *in vitro* and they are presumed to carry the same genomic and epigenomic changes as the tumors of origin [14]. The use of cell lines as a cancer model is cost-effective, convenient, and allows high-throughput screening [14-16]. In addition, it is possible to recapitulate and distinguish between some of the main cancer features by studying these cell lines *in vitro* [17]. *In vitro* functional assays can look at different aspects of cancer, such as the cell's ability to replicate without limit, self-sufficiency from growth signals, its resistance to inhibiting signals, as well as the ability of the cells to evade apoptosis and to migrate to distant tissues in response to stimuli. Together, these assays represent a powerful tool to generate measurable cancer-like phenotypes in a controlled environment. Once the phenotypes are generated and quantified, they can be associated with genetic variants to discover mechanisms of action for germline variants associated with cancer and can also be used to highlight some of the genes mediating their effects.

For this purpose, LCLs are one model system that has already been extensively used in functional and molecular studies, meaning that large amounts of genetic and molecular data have already been generated with them, providing insights into a number of complex diseases including cancer [18-20]. LCLs were produced from a large number of individuals, providing genetic variation across the

system that can be used to test for association with derived cancer like phenotypes, and the transcriptome has also been well characterized to aid in identifying mediating genes [21]. While cancer cell lines could be viewed as a more suitable model given, they are expected to contain the same epigenetic and genetic signature of cancer, these are usually established from tumors collected from at most a few individuals, meaning that the genetic diversity is not present to perform population-based genetic studies. In addition, as LCLs are EBV-transformed B cells, they provide an *in vitro* model to study a number of different cancers, including EBV-associated B-cell lymphomas [22] such as PTLD [23], Hodgkin's lymphoma [24] as well as B-cell related cancers such as chronic lymphocytic leukemia (CLL).

In this work, we aim to gain a deeper understanding of the regulatory germline contribution to cancer, in particular, to identify which target genes are involved in mediating the variants activity. We performed a population-based functional *in-vitro* screening on 87 genetically different LCLs to produce phenotypic data assessing the: *(i)* proliferation phenotypes that measure the self-sufficiency in growth signals, sensitivity to growth-inhibitory (anti-growth) signals and the unlimited replication potential; *(ii)* apoptotic phenotypes that quantify the evasion of programmed cell death, and *(iii)* chemotaxis phenotypes that identify the tissue invasion typical of some cancers. We combine this *in vitro* generated phenotypic data with available RNA-seq and genotype data to look for genetic variants associated with these phenotypes and the genes which mediate these associations.

Material and Methods

Reagents, buffers, and culture media

To culture of the LCL cell lines we used RPMI 1640 media that consisted of RPMI 1640 media (Gibco) supplemented with 100IU/ml / 100µg/ml of penicillin/ streptomycin (P/ S) (Gibco), 200mM L-glutamine (BioSwisstec AG) and 10% fetal bovine serum (FBS, Life Technologies, lot number 42G9295K). For the chemotaxis, we used a specific chemotaxis medium consisting of RPMI-1640 media supplemented with 100IU/ml / 100µg P/S, 200mM L-glutamine and 1% and bovine serum albumin (BSA, Sigma- Aldrich). This medium was used to resuspend the cells before the chemotaxis trans-well assay.

Dulbeccos's phosphate-buffered saline 1× (PBS) (Sigma- Aldrich) was used to wash the cell pellet before FACS analysis. The cell pellets before flow cytometry data acquisition were then resuspended in a FACS buffer composed of 1× PBS, 0.1% BSA and 0.05% sodium azide (Sigma- Aldrich). For the apoptosis assay, we used a specific buffer to resuspend the cell pellet before flow cytometry analysis composed of Annexin V binding buffer 1× prepared according to the manufacturer directions from the commercially available product (Thermo Fisher). This buffer facilitates the binding of annexin V to phosphatidylserine in the apoptosis assays.

Lymphoblastoid cell lines (LCLs)

All LCLs are part of the NHGRI (National Human Genome Research Institute) sample repository for human genetic research. The LCL samples came from five different sources. *(i)* 14 LCLs derived from umbilical cord from newborns of Western European origin born at the maternity ward of the University of Geneva Hospital, for which pregnancies were full term or near full term (38-41 weeks). For each sample informed consent was obtained after an interview of the mother with a trained nurse. The project was approved by the Geneva University Hospital Ethics Committee [20]. *(ii)* 22 LCLs that were part of a previous study and originated from B cells of Northern and Western European ancestry [25]. *(iii)* 33 LCLs from the 1000 Genomes Project [26]. *(iv)* 13 LCLs from GEUVADIS cohort sample [27]. *(v)* 5 LCLs from the HapMap project [28]. Altogether, this gives 87 LCLs, all

from individuals of European ancestry. LCLs are from unrelated individuals and grown under identical conditions in a complete RPMI medium. Two randomly chosen LCLs were used to establish the *in vitro* LCLs culture growth conditions and ideal day of splitting. To assess the concentration for seeding cells were plated at different concentrations (from 0.0375×10^6 to 1.2×10^6 cells/ml) and counted after 3 and 4 days with the best concentration found to be 0.2×10^6 cells/ml. To assess the ideal day for splitting cells were plated at 0.2×10^6 cells/ml and followed in culture for 7 days, with the most suitable day found to be day 3. LCLs were kept for at least 3 weeks in culture until they all reach a similar exponential growth phase and constant high viability (~98%). From the 87 LCLs, the RNA-seq data (HiSeq2000 sequencing system, Illumina) and genotype data were already available from a previous study [29].

Cancer-like phenotype assays

Proliferation was measured at 24h, 48h, and 72h by fluorescence-activated cell sorter (FACS) without previous cell cycle synchronization. Cell lines were split the day before the assay. Next, LCLs were suspended at 1×10^6 cells/ml in 1× PBS and labelled with 0.625µM CellTrace™ Violet dye (CTV, ThermoFisher Scientific). After 20min of labelling in the dark at room temperature (RT), LCLs were washed with 5ml of complete RPMI medium and left alone for 5min to remove any excess dye. The cells were then pelleted by centrifugation at 300*g* for 7min and resuspended in a fresh pre-warmed complete RPMI medium. Immediately, 0.2×10^6 labelled LCLs were plated in 96-well flat-bottom tissue culture plates (Corning-Costar) in a total volume of 200µL per well.

At each time-point and before flow cytometry using an Attune NxT Cytometer coupled with autosampler (Thermo Fisher) measurement, 100µl of FACS buffer and 10µl of 7-AAD viability dye were added to the samples. 7-AAD allows the distinction between live cells and dead cells. Finally, to calculate the absolute counts of cells, a pre-defined number of beads CountBright™ (ThermoFisher Scientific) were added directly to the cells for a total initial volume of 310µL into a 96-well flat-bottom plate without cell washing steps. The FACS data were analyzed using the FlowJo software

(version 10.5.3, LLC). The gating strategy consisted of the identification of LCLs on forward scatter-area (FSC-A) *versus* side scatter-area (SSC-A) dot-plots, followed by the selection of single cells in both, high *versus* width FCS and SSC plots. At last, dot-plots for 7-AAD *versus* CTV allowed us to identify live cells (7-AAD^{-}) and violet cell tracer positive cell (CTV^{+}) (**Figure 1A-B**).

The proliferation parameters assessed were: *(i) Replication index,* defined as the average of absolute cell number at each time point giving an estimate of the overall proliferation ability. *(ii) Proliferation index*, is the ratio of final absolute cell count at 72h to the starting absolute cell count at 24h, representing the fold of expansion during culture, which gives the intensity of proliferation potential. *(iii) Fold-dye dilution* using the average of the mean fluorescent intensity (MFI) between each time point and day 1. Day 1 was used to set the zero generation since it is well known that during the first 24 hours the fluorescence from the CTV decreased rapidly due to the efflux of the dye not bound to the cell membrane. Additionally, it was necessary to normalize CTV dye intensity and the starting number of cells at each time point by dividing the obtained values for the respective values at 0h to set the baseline for each LCL.

To get the *proliferation index* and *replication index* phenotypes we used the absolute cell count. The absolute number for CTV^{+} and 7-AAD^{-} cells (live proliferating cells) were calculated by the following equation:

$$\boldsymbol{Absolute\ cell\ count}\left(\frac{cells}{\mu l}\right) = \frac{(Cell\ count \times Counting\ beads\ volume)}{(Counting\ beads\ count \times Cell\ volume)} \times Counting\ beads\ concentration$$

Finally, for the *Fold dye dilution* proliferation phenotype, the MFI was calculated at each time point without the need for absolute cell counting.

Apoptosis was measured *in vitro* using different media conditions and 96-well flat-bottom tissue culture plates (Corning-Costar) where cells were plated at a concentration of 0.2×10^6 cells/well in a complete RPMI medium. The apoptotic phenotypes observed after culture for 18h at 37°C in a 5%

CO_2 incubator were: i) *Apoptosis index no hFas-induced* (the intrinsic apoptosis) that was measured without any stimuli. ii) *Apoptosis index hFas-induced* (the extrinsic apoptosis) that was measured after incubation with 10µg/µl of the anti CD95 (APO-1/Fas) monoclonal antibody (clone EOS9.1, eBioscience). This CD95 monoclonal antibody is an agonist of the human FAS death cell receptor, when bound to the receptor it activates the signaling cascade transduction leading to cell death signals. At the endpoint, LCLs were washed with 1× PBS and pelleted by centrifugation at 300*g* for 7min and RT, before FACS analysis. After spinning, cells were then resuspended in 200µl of 1× Annexin V binding buffer and stained with 1:2 dilution of Annexin V conjugated to Pacific Blue (-PB, ThermoFisher Scientific) and incubated for 15min at RT according to the manufacturer instructions. Ten µl of 7-AAD viability dye was used to distinguish between early/late apoptotic cells and dead cells. Finally, samples were acquired by flow cytometry as described before with the last modification in the gating strategy, after single-cell inclusion. Then, the dot-plot of 7-AAD *versus* Annexin V-PB distinguishes the following conditions: *(i)* live cells (defined as 7-AAD$^-$Annexin V-PB$^-$ population); *(ii)* early apoptotic cells (7-AAD$^-$Annexin V-PB$^+$ cells); *(iii)* late apoptotic cells (7-AAD$^+$Annexin V-PB$^+$), and *(iv)* dead cells (7AAD$^+$Annexin V PB$^-$) (**Figure 1A, 1C**). Early and late apoptotic LCLs were pooled together to define the percentage of total apoptotic cells used in the two phenotypes analyzed. The *apoptosis index* was calculated as a ratio of the absolute cell number undergoing apoptosis over the starting absolute number of cells.

The chemotaxis assay was performed in a 96-well trans-well plate with a 5.0µm pore size polycarbonate membrane (Corning-Costar). The LCLs cells were split 24h before the assay and kept in an RPMI complete medium. For the assay, the cells were washed, counted to reach a working concentration of 1.5×10^6 cells/ml and suspended in an RPMI chemotaxis medium. For each of the three experimental conditions, in the lower part of each of the trans-well chambers, we placed 240µl of *(i)* RPMI chemotaxis medium (negative control); *(ii)* RPMI chemotaxis medium + 10% FBS (chemoattractant solution); and *(iii)* 1.1×10^5 cells in RPMI chemotaxis medium + 10% FBS (positive

control). Then, on the top of the chamber (trans-well), 1.1×10^5 LCLs were added, with the sole exception of the condition positive control (condition *(iii)*) where the cells are placed in the lower compartment. The plate was then incubated for 2h at 37°C in 5% CO_2 [30]. Migrated cells were stained with 10µl of 7-AAD and incubated for 10min to quantify by flow cytometry as described above. After single-cell gating, the living LCLs were counted by excluding those dead $7AAD^+$ cells (**Figure 1A**). The absolute cell count was performed in the same way as the proliferation assays. The percentage of migrating cells was calculated as the percentage of the absolute number of the cells able to migrate in the lower compartment of the trans-well divided by the total number of cells plated (condition *(iii)*).

Flow cytometry-based functional assays

Proliferation, apoptosis and chemotaxis phenotypes were analyzed by flow cytometry using the same first three gating strategies (**Figure 1A)** followed by specific ones according to each phenotype as shown by representative experiments in **Figure 1**. The gating strategy of LCLs consisted first in discarding cellular debris by using forward scatter-area (FSC-A) *versus* size scatter-area (SSC-A) dot-plots and selecting by size and granularity both populations of interest, beads (proliferation and migration only) and LCLs (indicated within the plot, far left **Figure 1A**). Single cells were selected by consequently plotting SSC-width (SSC-W) *versus* SSC-high (SSC-H) and forward side scatter-height (FSC-H) *versus* FCS-W in (**Figure 1A**). The final population of single-cell LCLs was next evaluated separately for the three phenotypes. Proliferation and chemotaxis measurement required the definition of an additional gate in the Yellow laser-area (YL1-A) *versus* Red laser-area (RL1-A) dot-plot corresponding to the counting beads (**Figure 1A left**) that are represented in red in the FSC-A *versus* SSC-A plot. Proliferation was obtained from gating on live cells ($7AAD^-$) plotting 7-AAD *versus* forward FSC-H, and live cells were used to quantify the number of cells at each time point to derive proliferation and replication index (**Figure 1B right**). Then, on live cells, CTV intensity dilution was assessed over the three-time points and used to define the fold dye dilution phenotype (**Figure 1B**). Apoptosis required to set quadrants in Annexin V-PB *versus* 7AAD dot-plots. Here, the

dead (upper-left), late apoptotic (upper-right), early apoptotic (lower-right) and living LCLs (lower-left) are indicated, including the percentage of cells belonging to each quadrant (**Figure 1C**). For chemotaxis, we gate on live cells (as in the proliferation) (**Figure 1B right**) which were used for quantification of migrated cells with the help of cells counting beads.

Statistical Analysis of phenotype analyses

Flow cytometry data analysis was done using FlowJo (version 10.5.3, LLC). The absolute cell numbers from all assays were divided by the absolute cell number at time zero (t0) as an internal data normalization for each cell line. Technical covariates considered were: batch effect, reflecting the date of each experiment, cell origin (the source of the population of the LCLs samples) (**S1 and S2** for the effect of technical covariates used) [20, 25, 26]; and sex. These were identified using a linear regression model associating each phenotype with each of the covariates available. Spearman correlations between phenotypes were performed using the *rcorr* function from the *corrplot* R package (version 0.92) [31].

RNA-seq data and genotype data from external datasets

We used short-read Illumina RNA-seq data and genotype data from 317 LCLs (including the 87 LCLs in this study) created as described in Delaneau et al. [29]. Briefly, available RNA-seq gene expression data were quantified using QTLtools [32] with GENCODE v19 [33] as reference annotation and genes were filtered to only retain protein-coding genes and long intergenic non-coding RNA (lincRNAs) expressed in more than 90% of the samples. The genotype data for these samples, available from either the 1000 Genomes project or the Illumina Human OMNI 2.5M SNP array [29], were filtered using standard procedures to remove low-quality SNPs. The resulting genotype matrix of 317 individuals and 6,785,28 variants was imputed from the 1000 Genomes phase 3 reference panel [26] and poorly imputed variants were removed [29].

GWAS

We performed a genome-wide association analysis to associate the 6,785,282 SNPs to the six *in vitro* phenotypes. We used the *gwas* function implemented in QTLtool [32] and covariates to correct the

phenotype data, the GWAS threshold considered for significant associations was 5e-8. This analysis included the following covariates: sex, the first three principal components (PCs) from genotypes, batch for the *in vitro* phenotypic assays and origin of the cells.

Genetic associations

For gene expression phenotype, we mapped eQTLs using the *QTLtools* package, nominal pass in *cis* [34]. Briefly, all genetic variants within +/– 1 Mb of the transcription start site of the gene were identified, and all the associations were retained. The *QTLtools* package implements this in its *cis* nominal mode [34]. The *R/qvalue* package was used to correct for multiple tests across phenotypes using a false discovery rate (FDR) of 5% [35] (http://github.com/jdstorey/qvalue).

All eQTLs analyses included the following covariates: sex, the first three principal components (PCs) from genotypes, and 50 PCs from expression (genes or transcripts).

Gene Ontology enrichment analysis and pathway mapping

For the pathway analysis conducted in this study, the pathway information from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) were tested using the online KEGG pathway database [36-38]. Basically, we tested our candidate genes associated with different cancer-like phenotypes using the online KEGG Mapper for pathway identification [36-38]. The Gene Ontology enrichment analysis was performed on the same gene sets directly from the GOC website (https://optical.org/)[39, 40] that used the analysis tool from the PANTHER Classification System [41].

Gene expression association with phenotypes

We used a linear regression model written with *R* (version 3.6.2) to associate each protein-coding and lncRNA genes (n = 14,246) to each cancer-like phenotype independently (6 linear regressions per gene). The analyses included the following covariates: sex, cell origin, the origin of sequencing data, and batch for the phenotypic assays. Multiple testing correction was performed across phenotypes was performed using the qvalue package in R (version 2.18.0) [35]

Results

Differences in cellular phenotypes across cell lines

We assessed the proliferation ability of the 87 genetically different LCLs using *(i)* the cells' replication index; *(ii)* the cells' proliferation index and *(iii)* fold-dye dilution of the violet cell tracer dye across time (see **Material and Methods**). Overall, the LCL populations showed small differences in terms of cell proliferation ability (replication index, range 0.156 – 2.265) and cell expansion (proliferation index, range 0.560 – 3.927) (**Table 1**). We detected a positive correlation between all the three proliferation phenotypes, with a Spearman correlation ρ of 0.6; 0.51; and 0.65 for proliferation *versus* replication index; replication index *versus* fold dye dilution; and proliferation index *versus* fold dye dilution respectively (*p*-values from 0.0001 to $-8.78e^{-10}$) (**Figure 2A, 2B**), showing that the three proliferation phenotypes are all highly correlated.

The ability of LCLs to resist both intrinsic (non-induced) and extrinsic (ligand-induced) apoptosis was also investigated. The engagement of the hFas ligand (hFasL) to the hFas death receptor (hFasR) expressed on the cell's surface is known to induce cell death [42]. Among the extrinsic apoptosis induction receptors, hFasL was used as the expression of this gene was homogeneous across all the 87 LCLs, based on previously published RNA-seq [27, 29] (see **Material and Methods**). To capture both intrinsic and extrinsic apoptosis we therefore assayed both *(i)* stimulation with an agonist of the human Fas ligand (hFasL, CD95L) and *(ii)* without stimulation. After the 18h in culture, there was an increase in apoptosis when LCLs were stimulated with the agonist of hFasL, with a median apoptotic index of 2.5, compared to the 1.9 without stimulation. Of note, for 6.8% of cells the induced apoptosis was lower than the spontaneous, underlying an ability to reduce their spontaneous apoptosis but also resist the induced one (**Table 1**). Furthermore, we observed a strong positive Spearman correlation ($\rho = 0.82$, *p*-value = $2.2e^{-16}$) between the spontaneous and the Fas-induced apoptosis (**Figure 2A, 2C**).

Chemotaxis in the 87 LCLs was assessed using a trans-well migration assay with 10% FBS as a chemoattractant. We used FBS in order to ensure that our experiment was robust to the choice of chemokine used [43, 44]. Nearly 50% of the study population showed a small percentage of migratory capacity (< 1.5% migrating cells among the total number of cells), and 10% of the total population showed a more pronounced migration toward FBS attraction (< 5%) (**Table1)**. Nonetheless, we did not observe any association between chemotaxis and proliferation and apoptosis phenotypes, as shown in **Figure 2A**. An example of correlation between chemotaxis and replication index in shown in **Figure 2D**.

Finally, we detected a negative Spearman correlation ($\rho = -0.13$ to -0.03, *p*-value = 0.23 to 0.70) between all the three proliferation phenotypes measured and the spontaneous and induced apoptosis index (**Figure 2A, 2E**).

Sex association with phenotypes

Complex human phenotypes often manifest in a sex-specific manner. Cancer types, in particular, are known to exhibit gender differences not only in incidence and prevalence (https://seer.cancer.gov) but also in tumor growth rate and treatment responses [45]. To evaluate the influence of sex on LCLs proliferation, apoptosis and chemotaxis we performed a linear regression model between the six cancer-like phenotypes and sex inferred from genotypes in LCLs. Of the proliferation phenotypes, both proliferation index and fold dye dilution measurements were correlated with sex ($\rho = 0.25$ and 0.28, *p*-value = 0.017 and 0.018, respectively), suggesting sex differences in the ability of LCLs to proliferate (**Figure 3A, 3B)**. On the other hand, no significant correlation was present between sex and chemotaxis (*p*-value > 0.3) and neither between sex and apoptosis phenotypes *(p*-value > 0.4) (**Figure 2A and S3**).

Genetic variants associated with cancer phenotypes

Next, we evaluated the association between genetic variants across the genome and the phenotypes by performing six genome-wide association studies (GWAS) using the *gwas* option of QTLtool [32].

We detected only one significant association below the GWAS threshold, defined as a *p*-value <5e-8 (**Figure 4**, red line). The SNP rs12865307 was associated with replication index (*p*-value = $1.6510^{-8,}$ **Figure 4A)**. A number of nearby SNPs in moderate LD with rs12865307 (R2>0.4) also showed evidence of association with replication index, though not to the genome-wide threshold, reducing the risk that this association is a false positive due to genotyping errors (**Figure 5A and S4**). Genes in a 1MB window around this SNP, potential candidates as mediating the effect in cis, are shown in **Figure 5B**.

To identify putative candidate genes mediating the significant SNP rs1286530 effect on replication index, we looked at publicly available genetic associations between the SNP and the gene expression of nearby genes using the same LCLs [32, 34]. We found 12 genes significantly associated after controlling for multiple testing (**Table 2,** FDR<5%), the most significant association was with the lincRNA *LINC00539* (*p*-value = $3.84e^{-14}$). This SNP-gene association was already described as an eQTL in the GTEx eQTL catalog (https://gtexportal.org). *LINC00539* was found in a recent study to be associated with tumor immune response in lung adenocarcinoma and lung squamous cell carcinoma [46], making it a potential gene candidate for further investigation.

Splicing effect on rs12865307 target genes

Using the long-read direct RNA sequencing data we produced in another study (see **Article 1**), we investigated the role of splicing on the gene expression of the 12 candidate genes associated with rs12865307. We detected at least one annotated transcript for 8 of the 12 target genes listed in **Table 2**. In particular, the *SKA3* gene, coding for the spindle and kinetochore associated complex subunit 3 which regulates microtubule attachment to the kinetochores during mitosis, had four different annotated transcripts. The protein encoded by this gene localizes to the outer kinetochore and may be crucial for proper cell division and chromosome segregation, both traits related to cell replication. We tested for association between the *SKA3* gene and its four transcripts and rs12865307. While the SNP was not a significant eQTL for the *SKA3* gene (*p*-value =0.597), we still detected an association

with the expression of three of the transcripts represented in **Figure 6** (blue, yellow and red). *SKA3* is known to promote cell proliferation and migration in cervical cancer by activating the PI3K/Akt signaling pathway [47]. Another multi-omics analysis identified *SKA3* as a candidate oncogene associated with poor prognosis in lung adenocarcinomas [48] and it was also proposed as a prognostic biomarker associated with immune infiltration in bladder cancer [49].

GWAS – eQTLs overlap

Since the small sample size (n = 87) limited our power to detect significant GWAS hits for all the six phenotypes under investigation, we decided to lower the threshold to a *p*-value < 10^{-4} and investigate the variants which were significant at this threshold (from 357 to 1,003 variants depending on phenotype; **Table 3**). We performed an overlap between these GWAS-hits (*p*-value < 10^{-4}) and previously published SNP-gene expression associations [32]. We defined significant eQTLs amongst these variants as those whose association with expression passed multiple testing correction (FDR<5%). **Table 3** summarizes these results. We observed that the proliferation index phenotype showed the highest number of significant eQTLs genes (eGenes, n = 64) while chemotaxis the lowest (n = 14). Among the 176 significant unique eGenes detected in the analysis described above, 84.8% were not shared by more than one phenotype. However, one gene was associated with a variant linked to both the apoptosis phenotypes and the fold dye dilution proliferation phenotype, and 27% of the genes linked to an apoptosis-hFAS induced variant were also linked to a not induced phenotype. Among these 20 genes linked to the two apoptosis phenotypes (**S5** and **Figure 7**), we saw an enrichment in the lysosome pathways (FDR = 3.0e-02, fold enrichment = 29.7; KEGG database[36]); a function which plays an important role in the progression of apoptosis in cancer conditions [50, 51]. A Gene Ontology enrichment analysis identified other apoptosis-related pathways that were enriched in our set of apoptotic genes, as shown in **Table 3**. Moreover, the two genes linked to the proliferation index and apoptosis index (no hFAS induced) (**Figure 7**) included *IGF2BP3*, which was already associated with pancreatic cancer (The Human Protein Atlas). *IGF2BP3* is described as one of the KH-domain-containing RNA-binding protein (RBP) and has been reported to promote invasiveness

and metastatic properties of pancreatic ductal adenocarcinoma cells if overexpressed [52]. The other gene is the *MALSU1*, which codes for the Mitochondrial Assembly Of Ribosomal Large Subunit 1. *MALSU1* is essential for mitochondrial ribosome function and mitochondrial translation, and could prevent premature association of the 28S and 39S ribosomal subunits during ribosome biogenesis. *MALSU1* may also be involved in the assembly and/or regulation of the mitochondrial ribosome large subunit [53]. Finally, the only gene shared between proliferation index, fold dye dilution and apoptosis index (no hFAS induced) phenotypes is the *GAA* (alpha-glucosidase), a gene previously reported as an unfavorable prognostic marker for colorectal cancer (*p*-value < 0.001 Human Protein Atlas).

Gene expression associated with cancer phenotypes

To investigate the direct association between gene expression and cancer-like phenotypes, we performed a linear regression between protein-coding genes and lncRNAs (n = 14,246) expressed in LCLs samples and the six phenotypes derived from the functional assays. We detected target genes with a *p*-value < 0.05 (**Figure 8**), but none of them survived multiple testing corrections (FDR 5%). Therefore, due to the small sample size and the complexity of the phenotypes investigated we could not perform a direct association between gene expression and cancer-like phenotypes.

Discussion

In this study, we produced *in vitro* functional data for three cancer phenotypes: proliferation, apoptosis and chemotaxis. Combining our functional data with genotype and expression data from previous studies [20, 25, 29] we investigated the genetic contribution to cancer phenotypes and differences in gene expression associated to them. In particular, we aimed to identify putative candidate genes that could mediate the effect of genetic variants on cancer-related phenotypes. Cancer is caused by a combination of several types of mutations which leads to the emergence of different abnormalities, such as excessive proliferation, migration and the avoidance of apoptosis [17, 54]. Even if cancer has been considered for a long time a disease caused by somatic mutations, GWAS studies have highlighted the important contribution of germline non-coding variants in cancer development [55, 56]. The six phenotypes we investigated summarize *in vitro* three of the main characteristics of cancer cells and provide measurements of cancer properties.

We chose LCLs as an *in vitro* model for multiple reasons: *(i)* the different cell lines have been derived from a well-studied population with multiple published genetic and gene expression studies [18-20, 25, 29] ; *(ii)* they are easy to grow in culture and therefore suitable for large population-based studies; and *(iii)* since LCLs are immortalized cell lines, they are an almost unlimited source of genetic and gene expression material. These factors mean that it is possible to study these cellular phenotypes from a population perspective, using the available data. All in all, this particular population of LCLs is a suitable model for identifying the molecular relationships between genetic and transcriptomic signatures, and measurable phenotypes [21].

Our first analysis identified a strong correlation between the proliferation and the apoptosis phenotypes, while it showed a negative correlation between all the proliferation and the apoptosis phenotypes. In addition, we identified significant associations between two of the proliferation phenotypes and sex, while other phenotypes did not show significant associations in our study. It has already been observed that complex human phenotypes, including diseases, present sex differences

[57]. Different theories attributed sex differences to different hormones, the presence of sex chromosomes, genotype-by-sex effects, differences in behavior and environmental exposures between men and women, but their mechanisms and underlying biology remain poorly understood [57]. In cancer, sex differences were reported for the incidence in specific cancer types as well as different prognosis and responses to treatments [58, 59]. Our findings show that sex may influence the proliferation capability of cells, which has been proposed as an important factor in cancer. Moreover, given the strong correlation we observed between proliferation phenotypes and apoptosis phenotypes, we cannot discard the possibility that larger sample size studies may also find differences with other cancer-like phenotypes. Therefore, our work points to potential cellular mechanisms underlying the cancer risk and the progression differences observed between men and women.

Given the prominent role of genetics in the risk of cancer, we wanted to investigate the effect of genetic variants on cancer-like phenotypes. Moreover, using available gene expression we aimed to identify genes mediating their effect. However, we observed that our study was underpowered to find genome-wide associations with many of the cancer phenotypes. This is not surprising, given that complex phenotypes are often defined by the effect of thousands of variants with small effect sizes [60]. Nevertheless, we were able to identify a significant GWAS association, rs12865307, for the replication index, one of the proliferation phenotypes. Furthermore, using publicly available eQTL results from a much larger study (https://gtexportal.org) we reported a genome-wide significant eQTL for the rs12865307 SNP and the lincRNA *LINC00539* (Long Intergenic Non-Protein Coding RNA 539). This gene is known to correlate with tumor immune responses in lung adenocarcinomas and lung squamous cell carcinomas [46]. Altogether, our results indicate that cancer-like cellular phenotypes show a genetic complexity comparable to whole-organism traits often studied using GWAS, which likely requires studies with thousands of samples to uncover their genetics. However, we were able to identify one genome-wide significant association for a proliferation phenotype and provide a putative candidate gene that mediates its activity.

Recent eQTL studies showed that many SNPs often influence the expression of multiple genes [61, 62]. Therefore, we decided to explore in more detail the associations between the significant rs12865307 SNP and the expression of genes nearby. We identified 12 genes in the vicinity of the SNP. Of those genes only the lincRNA *LINC00539* previously mentioned had a significant association with the SNP after multiple testing corrections. However, among these genes only we highlighted the the *SKA3* gene since it is the only one producing more than two isoforms. Three among the four transcripts, even if not affected by a transcript-QTL, shown a genetic association with the GWAS SNP that could be further explored in a larger dataset with increased statistical power. *SKA3* is known to promote cell proliferation and migration in cervical cancer [47] and has also been proposed as a prognostic biomarker associated with immune infiltration in bladder cancer [49] and lung adenocarcinomas [48]. In conclusion, by integrating different data types and technologies we could investigate the specific mechanism of genetic variants affecting slicing in cancer-related phenotypes.

Our study aimed to better understand the underlying molecular mechanisms of cancer by studying cancer-like phenotypes. We integrated genetic variation, gene expression and *in vitro* cancer-like phenotypes and found that the complexities of these phenotypes are comparable to whole-organism phenotypes. Therefore, future studies should further employ methods commonly used in GWAS, to build a direct link between functional features and cancer genetic variants.

.

Tables

Table 1. Summary statistics of the six analyzed phenotypes.

Phenotype	Range	Mean	Median	Standard deviation	Variance
Replication index	0.156 – 2.265	1.207	1.304	0.481	0.231
Proliferation index	0.560 – 3.927	1.998	1.93	0.746	0.557
Fold tracer dye dilution (CTV)	1.228 – 6.966	3.733	3.778	1.06	1.124
Apoptosis index	0.811 – 10.935	2.493	1.945	1.718	2.95
Apoptosis index, anti-hFAS-induced	0.137 – 14.399	2.946	2.504	1.993	3.974
Chemotaxis (% of the total)	0.015 – 19.120	2.147	1.209	3.01	9.062

Abbreviations. CTV, CellTrace violet dye; hFAS, human FAS.

Table 2. Gene candidates to mediate the effect of the GWAS hit.

List of the 12 gene candidates for the mediation of the effect of the GWAS for the replication index. The SNP rs12865307 was associated with replication index with a *p*-value of 1.66e-08. The *p*-value association is the *p*-value of the SNP-gene association in the eQTL analysis.

Gene	***p*-value association**	**Multiple transcripts (# of transcripts)**
LINC00539	3.84E-14	YES
ZDHHC20	0.0111112	NO
EEF1AKMT1	0.134562	YES
XPO4	0.154231	NO
LATS2	0.181377	YES
SKA3	0.597652	YES, (4)
MRPL57	0.684558	NO
SAP18	0.863979	YES
CRYL1	0.874091	YES
IFT88	0.888233	YES
MICU2	0.90972	YES
IL17D	0.992938	NO

Abbreviations. LINC00539, Long Intergenic Non-Protein Coding RNA 539; *ZDHHC20*, Zinc Finger DHHC-Type Palmitoyltransferase 20; *EEF1AKMT1*, EEF1A Lysine Methyltransferase 1; *XPO4*, Exportin 4; *LATS2*, Large Tumor Suppressor Kinase 2; *SKA3*, Spindle And Kinetochore Associated Complex, Subunit 3; *MRPL57*, Mitochondrial Ribosomal Protein L57; *SAP18*, Sin3A Associated Protein 18; *CRYL1*, Crystallin Lambda 1; *IFT88*, Intraflagellar Transport 88; *MICU2*, Mitochondrial Calcium Uptake 2; *IL17D*, Interleukin 17D.

Table 3. Summary of the number of GWAS – hits overlapping eQTLs

Phenotype	GWAS *p*-value < 10e-4	GWAS – eQTLs-gene	Significant eQTLs (5%FDR)	Unique genes
Proliferation index	1,003	14,180	887	64
Replication index	453	5,044	102	23
Dye dilution	641	9,258	111	23
Apoptosis	762	10,982	186	36
Apoptosis hFAS	758	9,405	185	42
Chemotaxis	357	4,349	36	14

Abbreviations. eQTL, expression quantitative trait loci; FDR, False discovery rate; GWAS, Genome-wide association studies; hFAS, human FAS.

References

1. Arnold, M., et al., *Global burden of cutaneous melanoma attributable to ultraviolet radiation in 2012.* Int J Cancer, 2018. **143**(6): p. 1305-1314.
2. Wikonkal, N.M. and D.E. Brash, *Ultraviolet radiation induced signature mutations in photocarcinogenesis.* J Investig Dermatol Symp Proc, 1999. **4**(1): p. 6-10.
3. Loeb, L.A., et al., *Smoking and lung cancer: an overview.* Cancer Res, 1984. **44**(12 Pt 1): p. 5940-58.
4. Raj, T., et al., *Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility.* Nat Genet, 2018. **50**(11): p. 1584-1592.
5. WYNDER, E.L. and E.A. GRAHAM, *Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases.* J Am Med Assoc, 1950. **143**(4): p. 329-36.
6. Rünger, T.M., B. Epe, and K. Möller, *Processing of directly and indirectly ultraviolet-induced DNA damage in human cells.* Recent Results Cancer Res, 1995. **139**: p. 31-42.
7. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.
8. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333-339.
9. Gibbons, D.L., L.A. Byers, and J.M. Kurie, *Smoking, p53 mutation, and lung cancer.* Mol Cancer Res, 2014. **12**(1): p. 3-13.
10. Easton, D.F. and R.A. Eeles, *Genome-wide association studies in cancer.* Hum Mol Genet, 2008. **17**(R2): p. R109-15.
11. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nat Rev Genet, 2016. **17**(2): p. 93-108.
12. Peto, J. and R.S. Houlston, *Genetics and the common cancers.* Eur J Cancer, 2001. **37 Suppl 8**: p. S88-96.
13. Wiemels, J.L., et al., *Prenatal origin of acute lymphoblastic leukaemia in children.* Lancet, 1999. **354**(9189): p. 1499-503.
14. Goodspeed, A., et al., *Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics.* Mol Cancer Res, 2016. **14**(1): p. 3-13.
15. Geraghty, R.J., et al., *Guidelines for the use of cell lines in biomedical research.* Br J Cancer, 2014. **111**(6): p. 1021-46.
16. Kaur, G. and J.M. Dufour, *Cell lines: Valuable tools or useless artifacts.* Spermatogenesis, 2012. **2**(1): p. 1-5.
17. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
18. Dermitzakis, E.T., *From gene expression to disease risk.* Nat Genet, 2008. **40**(5): p. 492-3.
19. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-8.
20. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.
21. Amoli, M.M., et al., *EBV Immortalization of human B lymphocytes separated from small volumes of cryo-preserved whole blood.* Int J Epidemiol, 2008. **37 Suppl 1**: p. i41-5.
22. Young, L.S. and A.B. Rickinson, *Epstein-Barr virus: 40 years on.* Nat Rev Cancer, 2004. **4**(10): p. 757-68.
23. Luskin, R. and H. Nathan, *Eligible Death Statistic: Not a True Measure of OPO Performance nor the Potential to Increase Transplantation.* Am J Transplant, 2015. **15**(8): p. 2019-20.
24. Hecht, J.L. and J.C. Aster, *Molecular biology of Burkitt's lymphoma.* J Clin Oncol, 2000. **18**(21): p. 3707-21.
25. Waszak, S.M., et al., *Population Variation and Genetic Control of Modular Chromatin Architecture in Humans.* Cell, 2015. **162**(5): p. 1039-50.

26. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
27. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
28. Garieri, M., et al., *The effect of genetic variation on promoter usage and enhancer activity.* Nat Commun, 2017. **8**(1): 1358. doi: 10.1038/s41467-017-01467-7.
29. Delaneau, O., et al., *Chromatin three-dimensional interactions mediate genetic effects on gene expression.* Science, 2019. **364**(6439). doi: 10.1126/science.aat8266.
30. Clottu, A.S., et al., *EBI2 Expression and Function: Robust in Memory Lymphocytes and Increased by Natalizumab in Multiple Sclerosis.* Cell Rep, 2017. **18**(1): p. 213-224.
31. Wei, T., & Simko, V., *R package "corrplot": Visualization of a Correlation Matrix.* (Version 0.92), *https://github.com/taiyun/corrplot*. 2021.
32. Delaneau, O., et al., *A complete tool set for molecular QTL discovery and analysis.* Nat Commun, 2017. **8**: 15452. doi: 10.1038/ncomms15452.
33. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
34. Aguet, F., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
35. Storey, J.D., et al., *qvalue: Q-value estimation for false discovery rate control.* R package version, 2015. **2**(0): p. 10.18129.
36. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
37. Kanehisa, M., *Toward understanding the origin and evolution of cellular organisms.* Protein Sci, 2019. **28**(11): p. 1947-1951.
38. Kanehisa, M., et al., *KEGG: integrating viruses and cellular organisms.* Nucleic Acids Res, 2021. **49**(D1): p. D545-D551.
39. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
40. Consortium, G.O., *The Gene Ontology resource: enriching a GOld mine.* Nucleic Acids Res, 2021. **49**(D1): p. D325-D334.
41. Mi, H., et al., *Large-scale gene function analysis with the PANTHER classification system.* Nat Protoc, 2013. **8**(8): p. 1551-66.
42. Peter, M.E., et al., *The role of CD95 and CD95 ligand in cancer.* Cell Death Differ, 2015. **22**(5): p. 885-6.
43. Pijuan, J., et al., *Cell Migration, Invasion, and Adhesion Assays: From Cell Imaging to Data Analysis.* Front Cell Dev Biol, 2019. **7**: p. 107. doi: 10.3389/fcell.2019.00107. eCollection 2019.
44. Bagati, A., et al., *A Modified In vitro Invasion Assay to Determine the Potential Role of Hormones, Cytokines and/or Growth Factors in Mediating Cancer Cell Invasion.* J Vis Exp, 2015(98):51480. doi: 10.3791/51480.
45. Wilson, M.A. and K.H. Buetow, *Novel Mechanisms of Cancer Emerge When Accounting for Sex as a Biological Variable.* Cancer Res, 2020. **80**(1): p. 27-29.
46. Zengin, T. and T. Önal-Süzek, *Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma.* J Pers Med, 2021. **11**(2):154. doi: 10.3390/jpm11020154.
47. Hu, R., et al., *SKA3 promotes cell proliferation and migration in cervical cancer by activating the PI3K/Akt signaling pathway.* Cancer Cell Int, 2018. **18**: 183. doi: 10.1186/s12935-018-0670-4. eCollection 2018.
48. Lin, Y., et al., *Integrative Multi-Omics Analysis of Identified SKA3 as a Candidate Oncogene Correlates with Poor Prognosis and Immune Infiltration in Lung Adenocarcinoma.* Int J Gen Med, 2022. **15**: p. 4635-4647.

49. Wang, C., et al., *SKA3 is a prognostic biomarker and associated with immune infiltration in bladder cancer.* Hereditas, 2022. **159**(1): 20. doi: 10.1186/s41065-022-00234-z.
50. Guicciardi, M.E., M. Leist, and G.J. Gores, *Lysosomes in cell death.* Oncogene, 2004. **23**(16): p. 2881-90.
51. Ivanova, S., et al., *Lysosomes in apoptosis.* Methods Enzymol, 2008. **442**: p. 183-99.
52. Mukherjee, M. and S. Goswami, *Identification of Key Deregulated RNA-Binding Proteins in Pancreatic Cancer by Meta-Analysis and Prediction of Their Role as Modulators of Oncogenesis.* Front Cell Dev Biol, 2021. **9**: 713852. doi: 10.3389/fcell.2021.713852. eCollection 2021.
53. Rorbach, J., P.A. Gammage, and M. Minczuk, *C7orf30 is necessary for biogenesis of the large subunit of the mitochondrial ribosome.* Nucleic Acids Res, 2016. **44**(2): 992. doi: 10.1093/nar/gkv1125. Epub 2015 Oct 19.
54. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
55. Juul, M., et al., *Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate.* Elife, 2017. **6**:e21778. doi: 10.7554/eLife.21778.
56. Osman, N., A.E. Shawky, and M. Brylinski, *Exploring the effects of genetic variation on gene regulation in cancer in the context of 3D genome structure.* BMC Genom Data, 2022. **23**(1): 13. doi: 10.1186/s12863-021-01021-x.
57. Oliva, M., et al., *The impact of sex on gene expression across human tissues.* Science, 2020. **369**(6509):eaba3066. doi: 10.1126/science.aba3066.
58. Haupt, S., et al., *Sex disparities matter in cancer development and therapy.* Nat Rev Cancer, 2021. **21**(6): p. 393-407.
59. Rubin, J.B., et al., *Sex differences in cancer mechanisms.* Biol Sex Differ, 2020. **11**(1): 17. doi: 10.1186/s13293-020-00291-x.
60. Cano-Gamez, E. and G. Trynka, *From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases.* Front Genet, 2020. **11**: 424. doi: 10.3389/fgene.2020.00424.
61. Consortium, G., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.
62. Viñuela, A., *Genetic analysis of blood molecular phenotypes reveals regulatory networks affecting complex traits: a DIRECT study*, V.O.P.A.A. Brown, Editor. 2021. doi: https://doi.org/10.1101/2021.03.26.21254347.

Figure legends

Figure 1. Gating strategy used in the phenotypic analysis.

Proliferation, apoptosis and chemotaxis phenotypes were analyzed by flow cytometry using the same three initial gating strategies, followed by specific ones according to each phenotype. **(A)** The gating strategy of LCLs for proliferation and chemotaxis included the CountBright™ beads for absolute cell count identified in dot-plots in the channels for yellow and red laser (YL1-A and RL1-A, respectively). Then, the cellular debris was excluded by using forward scatter-area (FSC-A) *versus* size scatter-area (SSC-A) dot-plots, beads (in red, back-gated from the precedent dot-plot) and LCLs (in black). The exclusion of duplets and aggregates was followed by the consecutively plotting SSC-with (SSC-W) *versus* SSC-high (SSC-H) and FCS-H *versus* FCS-W. **(B)** Proliferation and chemotaxis were obtained from gating into live cells in FCS-H versus 7-AAD (right). Proliferation was followed by histograms for the CTV channel (left). **(C)** Apoptosis required setting quadrants in Annexin V-Pacific blue (PB) *versus* 7AAD dot plots. Here, the cells are divided into dead (upper-left), late apoptotic (upper-right), early apoptotic (lower-right) and living LCLs (lower-left). Numbers in each plot correspond to the percentage of cells located in each quadrant.

Figure 2. Correlations among all the observed phenotypes.

(A) The figure summarizes the Spearman correlation (ρ) among the cancer-like phenotypes and sex computed by the *corrplot* package for the 87 samples analyzed. Positive and negative correlations are indicated in blue and red color shades, respectively. A strong correlation exists between the Apoptotic index with and without human Fas ligand-induced (hFAS) induction ($\rho = 0.82$, deep blue). Good positive correlations are between the three Proliferation phenotypes: Replication index, Proliferation index and Fold-dye dilution with a ρ of 0.6; 0.51; and 0.65 (dark blue). Proliferation index and Fold-dye dilution show a slight positive correlation with the biological covariate sex with ρ of 0.25 and 0.28, respectively (light blue). Opposite, a negative correlation exists between the Proliferation phenotypes and those related to Apoptosis (pale red). **(B-E**) Scatter-plots of representative examples from the correlation analysis in **(A)**.

Figure 3. Sex as the biological covariate shows a correlation between the cancer-like phenotypes.

Violin plots show the different correlations from a linear regression model between sex and **(A)** the proliferation index and **(B)** the fold dye dilution phenotypes. The information about sex obtained from Delaneau et al. [29] is displayed as 0 and 1, where 0 corresponds to males and 1 to females. Statistical significance, * *p*-values < 0.05.

Figure 4. GWAS for the six phenotypes reveals associations of single nucleotide variation with the Replication index.

The Manhattan plots represent the *p*-values, expressed as $-\log_{10}(p)$, along the chromosome for the single nucleotide polymorphism (SNP) represented as single dots resulting from GWAS. Alternated blue and yellow colors visually identify the different chromosomes. The GWAS threshold of 5e-8 is shown by the red lines. The phenotypes are **(A)** Replication index; **(B)** Proliferation index; **(C)** Fold-dye dilution; **(D)** Chemotaxis; **(E)** Apoptotic index without induction; and **(F)** Apoptotic index with human Fas ligand-induced (hFAS) induction

Figure 5. Magnification of the 1Mb window around the GWAS leading target rs12865307 SNP for replication index on chromosome 13.

(A) The Locus Zoom-plot shows enrichment of SNPs (red, green, yellow and light blue) in the 1MB window around the rs128654307 GWAS hits. These SNPs are correlated ($r^2 > 0.4$) with the leading SNP represented in purple. The color grading bar is for the r^2 values. **(B)** Representation of the genes located in genomic region surrounding the GWAS SNP rs128654307.

Figure 6. Representation of the *SKA3* gene as a putative target of trQTL effects.

(A) Graphical representation of the *SKA3* gene structure (grey) and the four different transcripts (yellow, green, blue and red). Each box depicts a different exon and the blue curved lines between exons represent all the possible splicing events between exons. (**B-E)** Box-plots showing the effect on the different transcript's expression of the rs12865307 SNP (GWAS – hit for replication index). The *x*-axis corresponds to the three different genotypes (0 for the homozygous for the reference allele,

1 for the heterozygous and 3 for the homozygous for the alternative allele of the SNP). The *y*-axis represents the transcript expression in TPM (Transcript per million units).

Figure 7. Representation of the GWAS – eQTLs gene overlaps between the six phenotypes.
The Upset-plot provide a visualization of the intersections of the genes linked to GWAS – eQTLs signal between the six different phenotypes listed on the left of the bottom part of the plot. In the upper part, the histogram shows the number of genes in each intersection between the different phenotypes represented below. On the left side are represented with different colors the number of genes detected in each phenotypes group.

Figure 8. Distribution of the *p*-values for each gene expression associated with phenotypes.
Histograms represent the *p*-value distribution for gene expression associated with each of the six phenotypes **(A)** Proliferation index, **(B)** Replication index, **(C)** Fold-dye dilution, **(D)** Apoptosis index, **(E)** Apoptosis index hFAS, and **(F)** Chemotaxis index. Significant *p*-value < 0.05.

Figures

Figure 1

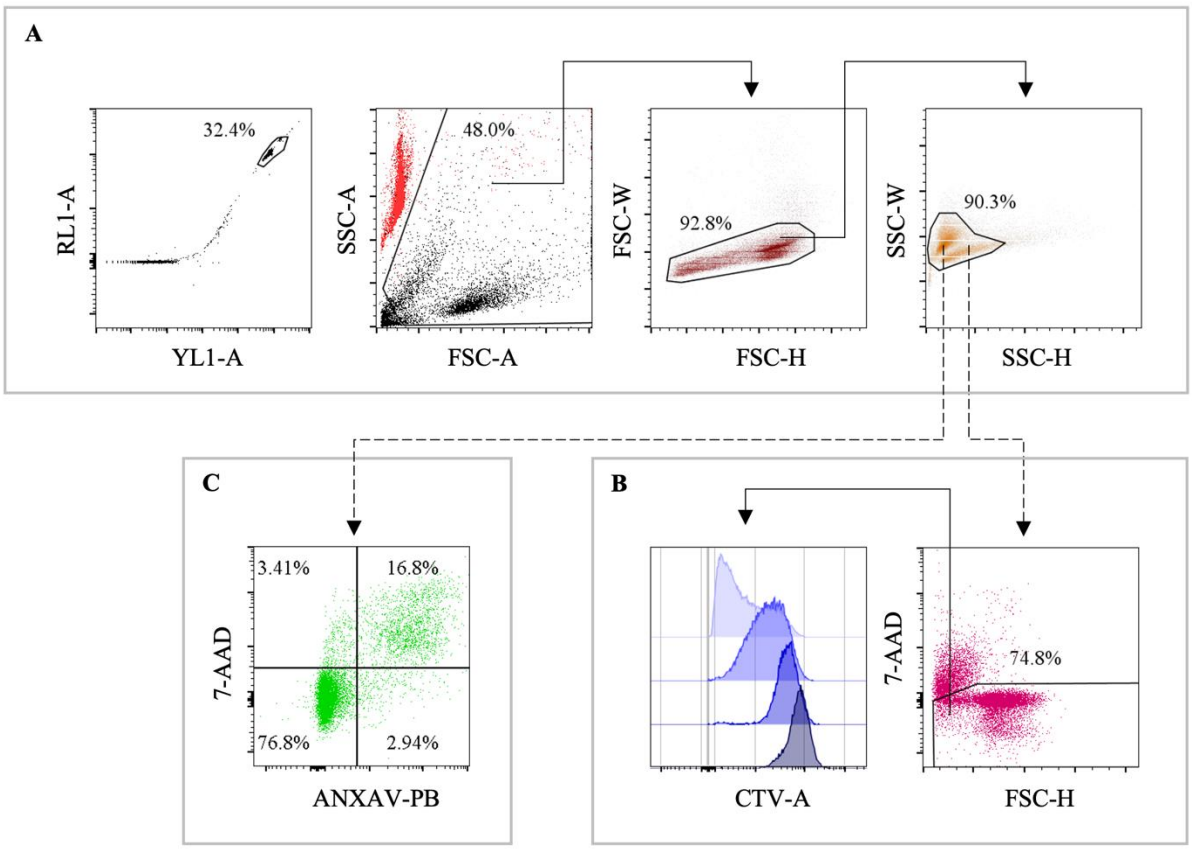


Figure 2

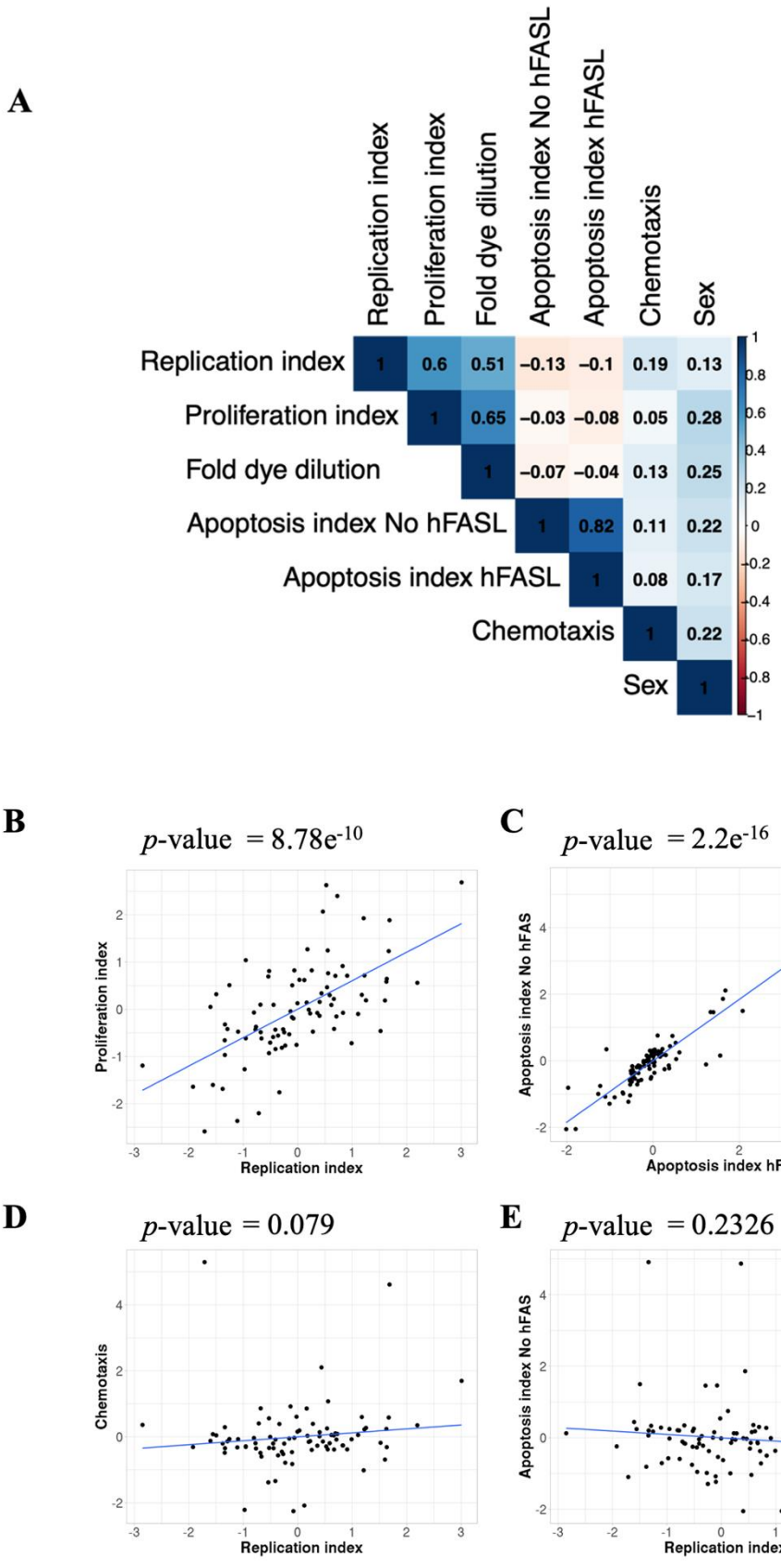

A
Replication index
Proliferation index
Fold dye dilution
Apoptosis index No hFASL
Apoptosis index hFASL
Chemotaxis
Sex
1 0.6 0.51 −0.13 −0.1 0.19 0.13
1 0.65 −0.03 −0.08 0.05 0.28
1 −0.07 −0.04 0.13 0.25
1 0.82 0.11 0.22
1 0.08 0.17
1 0.22
1
B
p-value = 8.78e-10
Proliferation index
Replication index
C
p-value = 2.2e-16
Apoptosis index No hFAS
Apoptosis index hFAS
D
p-value = 0.079
Chemotaxis
Replication index
E
p-value = 0.2326
Apoptosis index No hFAS
Replication index

Figure 3

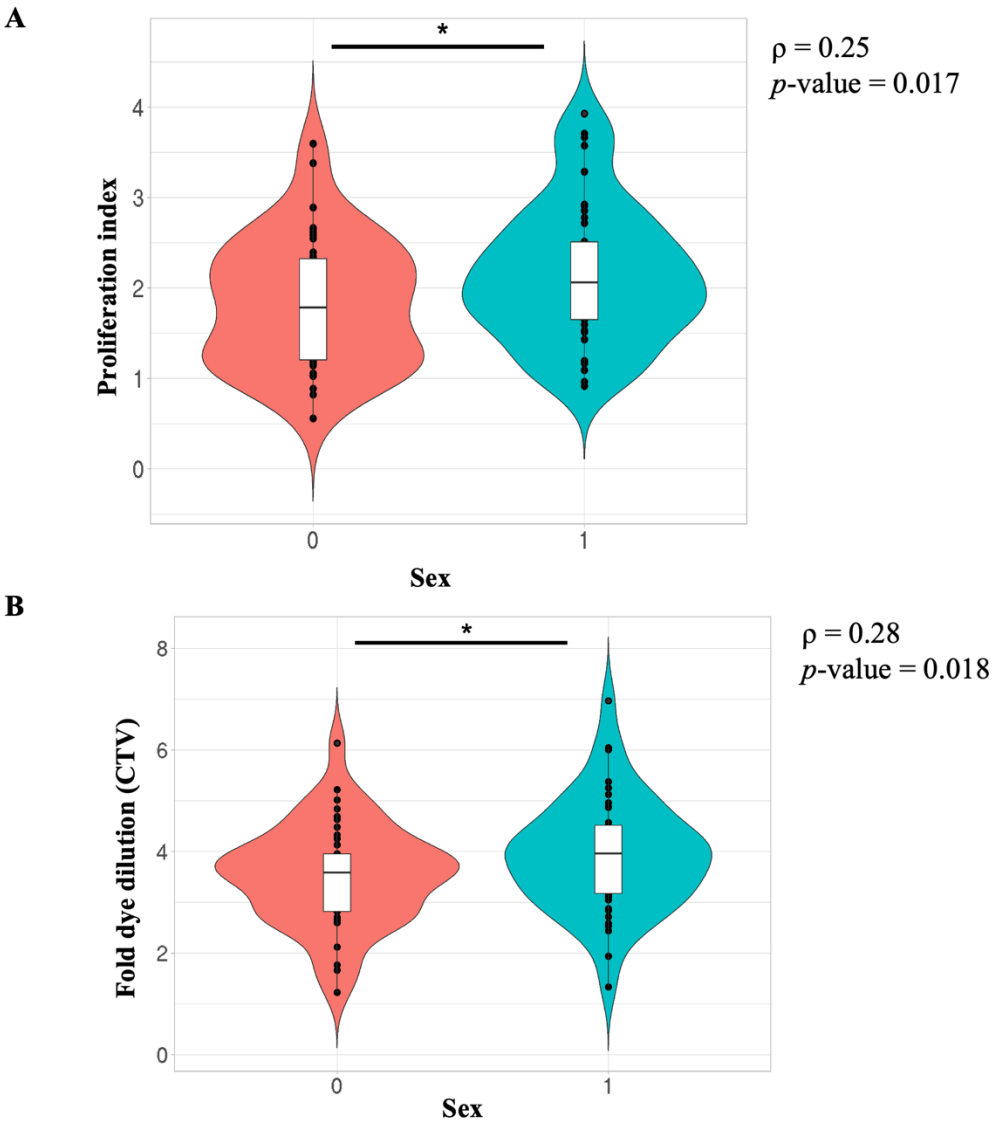
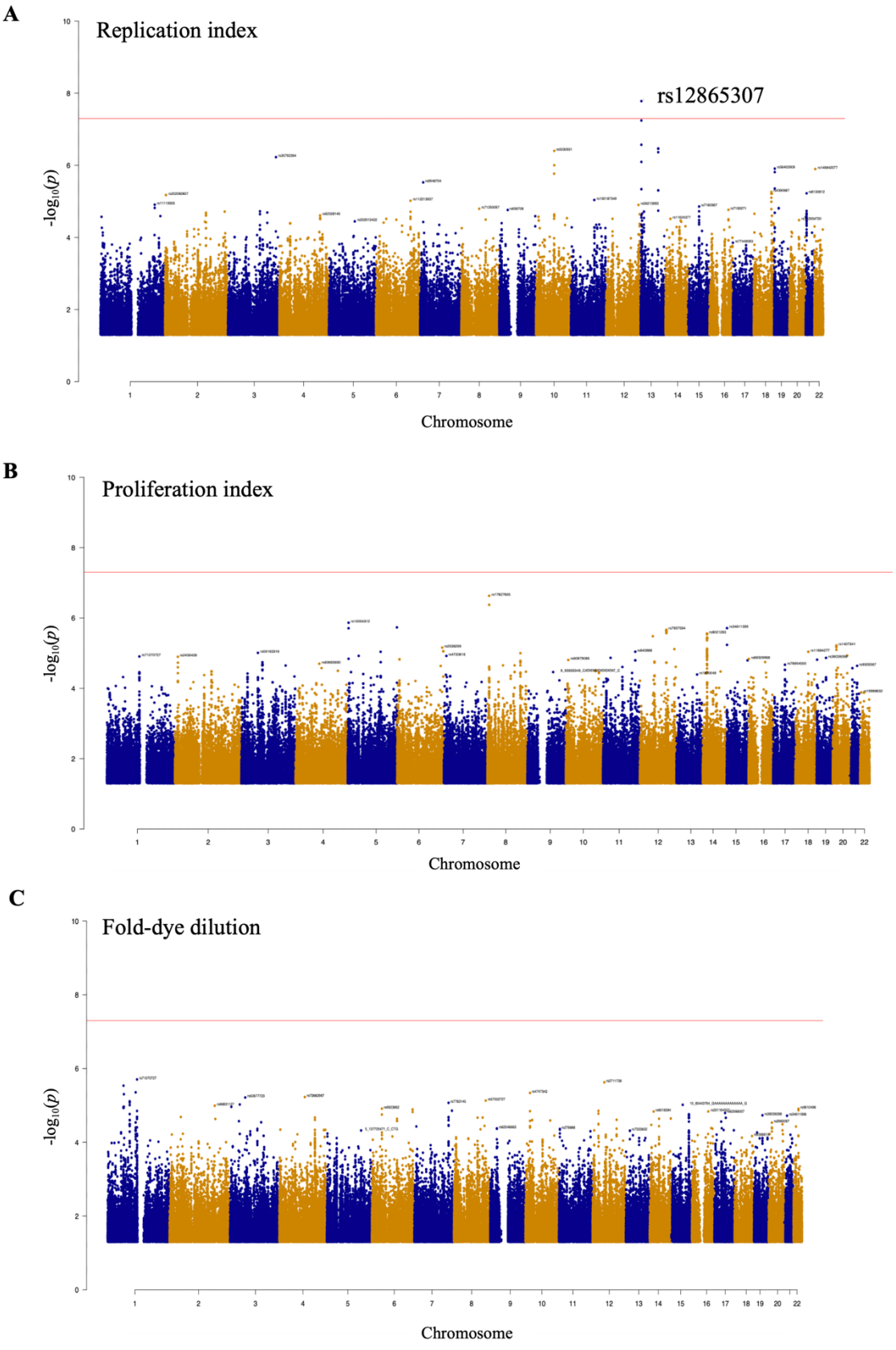
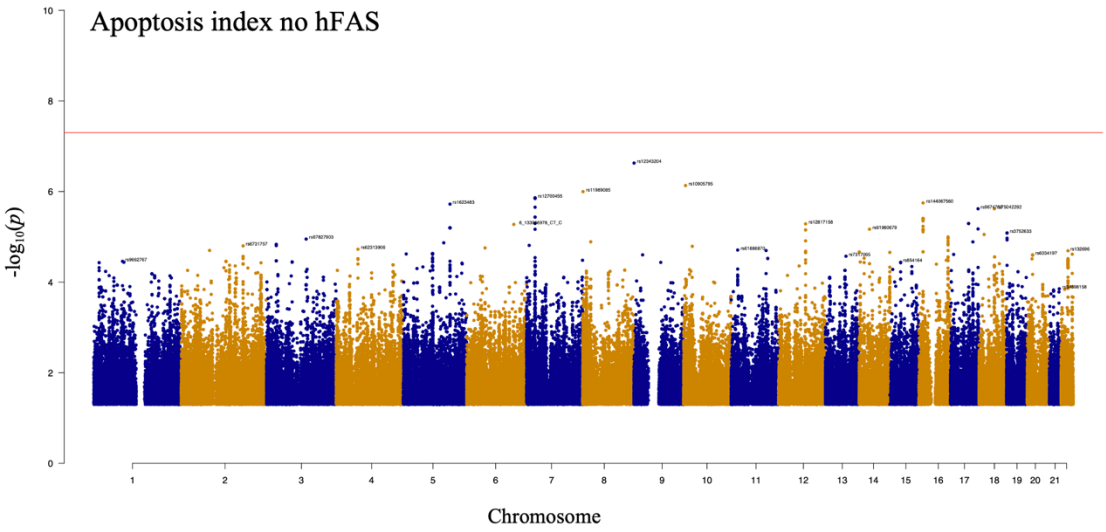

A
*
Proliferation index
0
1
2
3
4
0
1
Sex
ρ = 0.25
p-value = 0.017
B
*
Fold dye dilution (CTV)
0
2
4
6
8
0
1
Sex
ρ = 0.28
p-value = 0.018

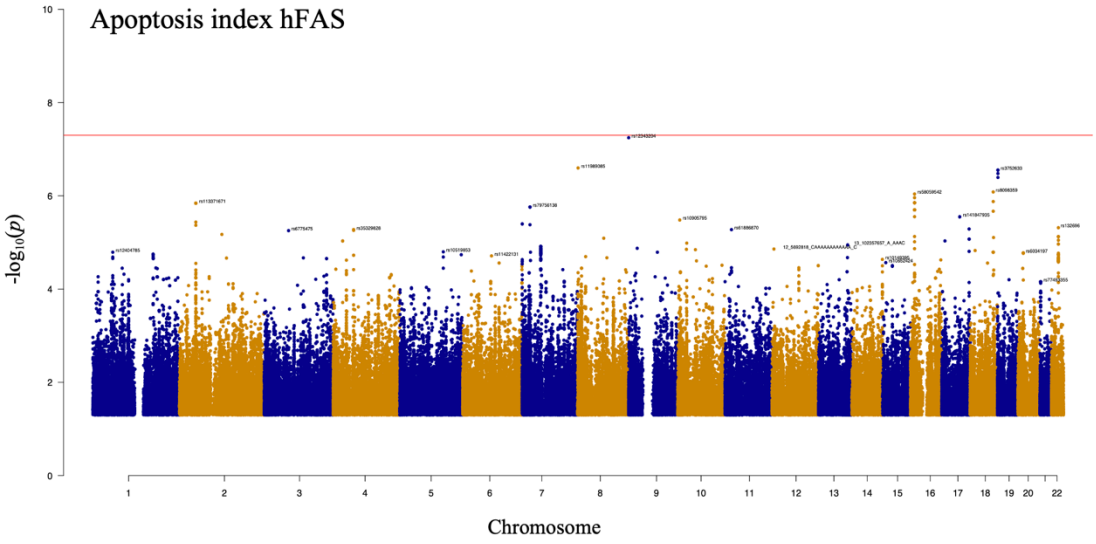
Figure 4


A
Replication index
rs12865307
-log10(p)
Chromosome
B
Proliferation index
-log10(p)
Chromosome
C
Fold-dye dilution
-log10(p)
Chromosome

D


Apoptosis index no hFAS
-log10(p)
Chromosome

E


Apoptosis index hFAS
-log10(p)
Chromosome

F

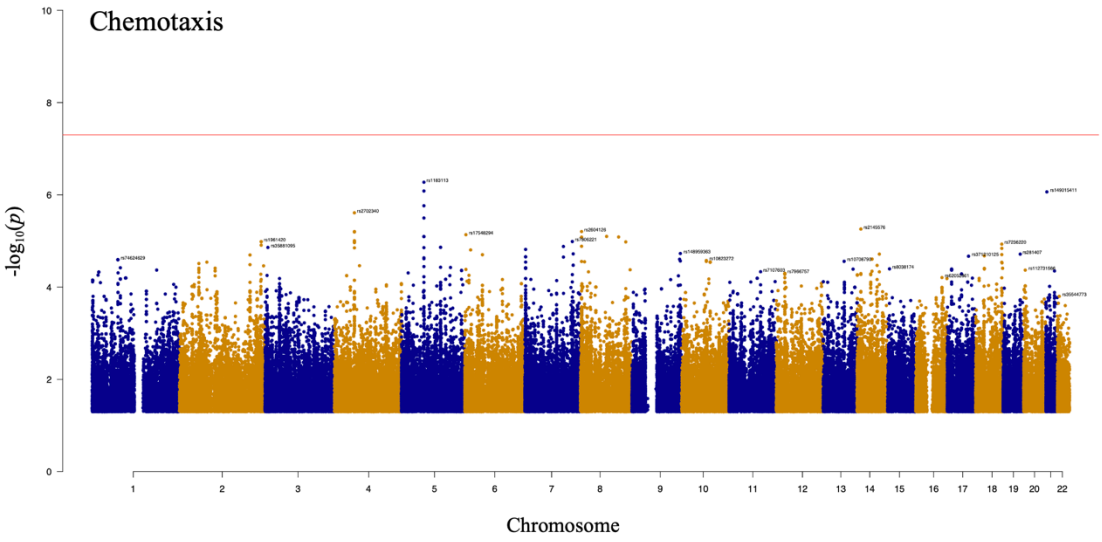

Chemotaxis
-log10(p)
Chromosome

Figure 5

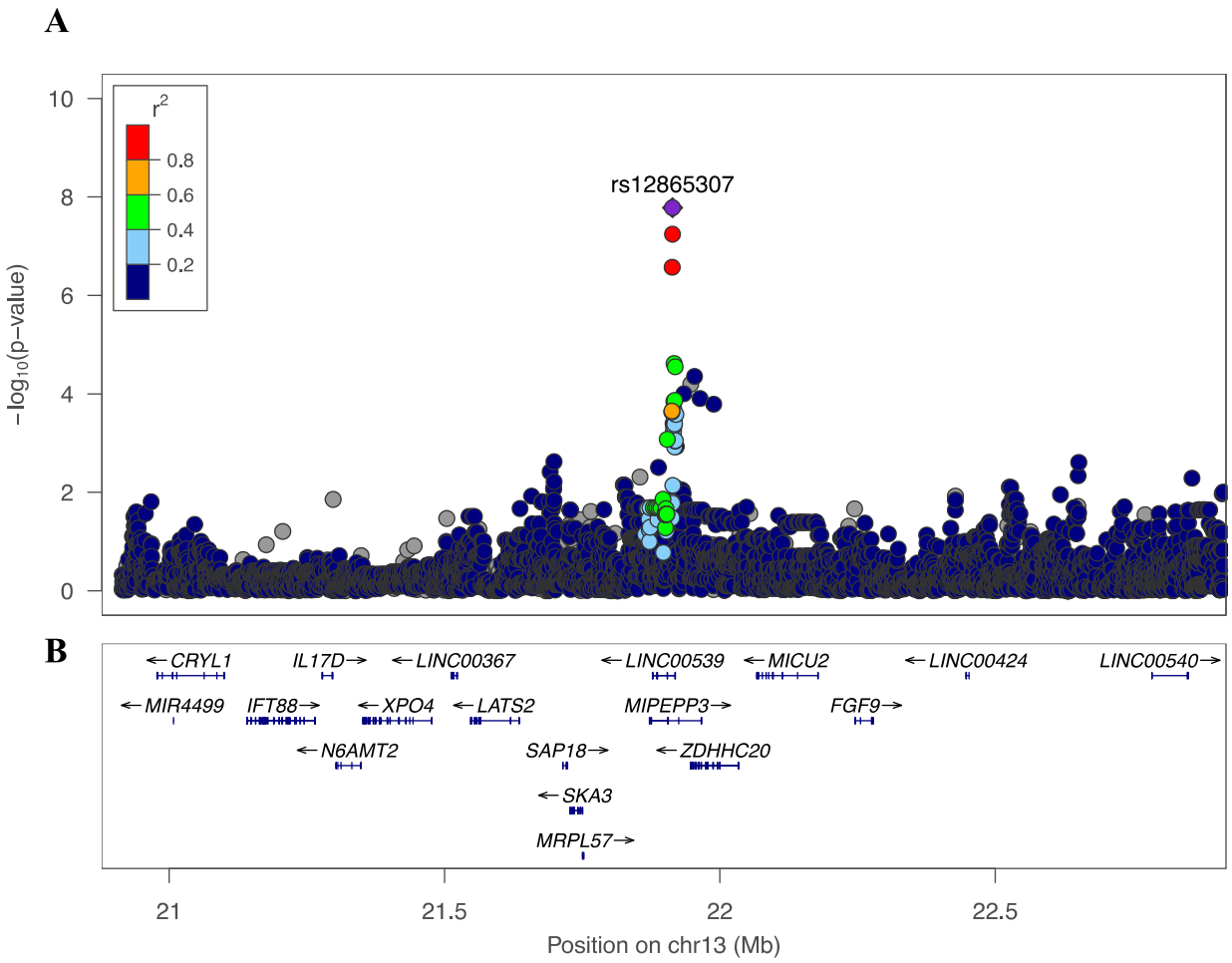

A
B
r2
0.8
0.6
0.4
0.2
rs12865307
−log10(p−value)
10
8
6
4
2
0
CRYL1
IL17D
LINC00367
LINC00539
MICU2
LINC00424
LINC00540
MIR4499
IFT88
XPO4
LATS2
MIPEPP3
FGF9
N6AMT2
SAP18
ZDHHC20
SKA3
MRPL57
21
21.5
22
22.5
Position on chr13 (Mb)

Figure 6

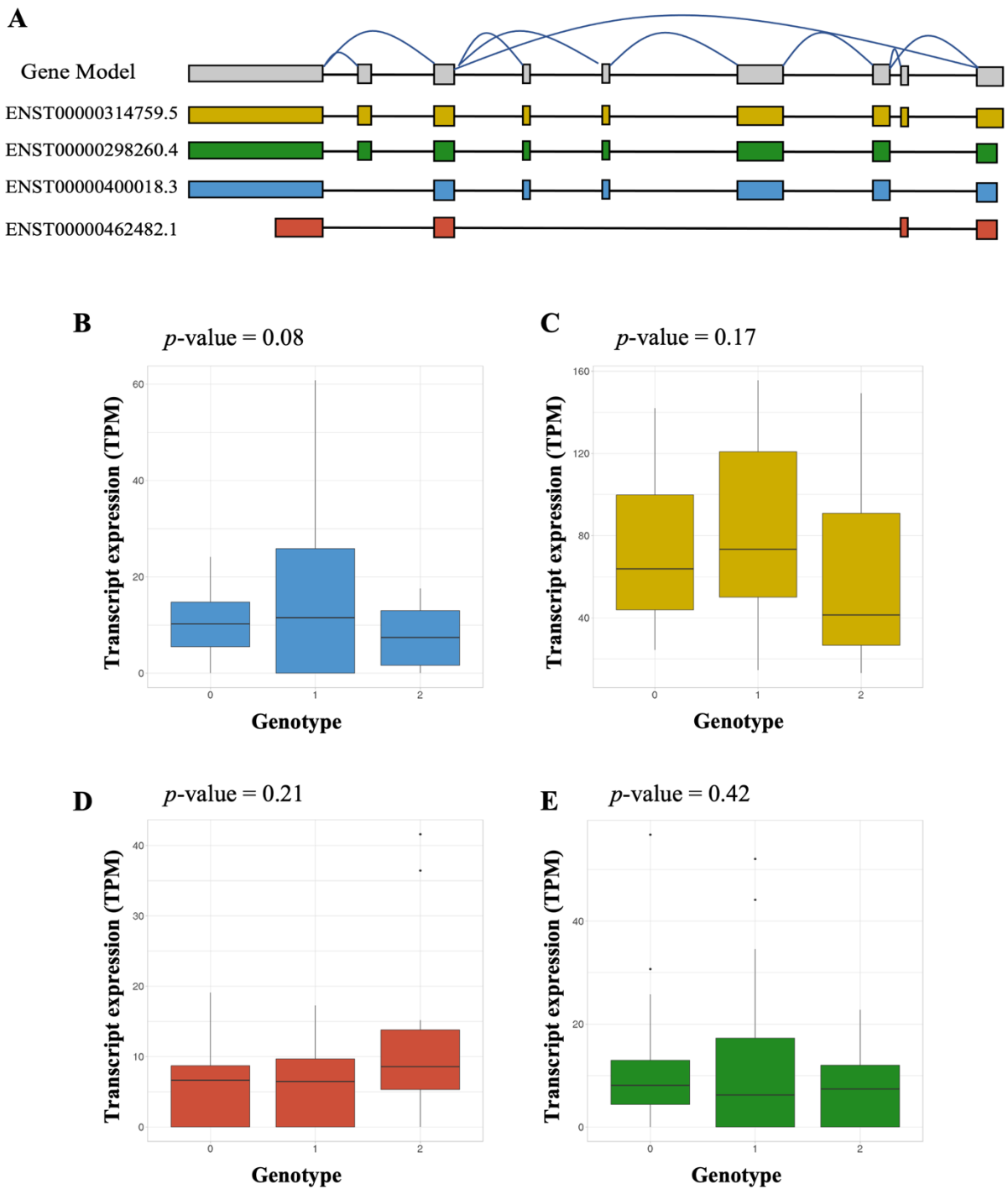

A
Gene Model
ENST00000314759.5
ENST00000298260.4
ENST00000400018.3
ENST00000462482.1
B
p-value = 0.08
Transcript expression (TPM)
Genotype
C
p-value = 0.17
Transcript expression (TPM)
Genotype
D
p-value = 0.21
Transcript expression (TPM)
Genotype
E
p-value = 0.42
Transcript expression (TPM)
Genotype

Figure 7

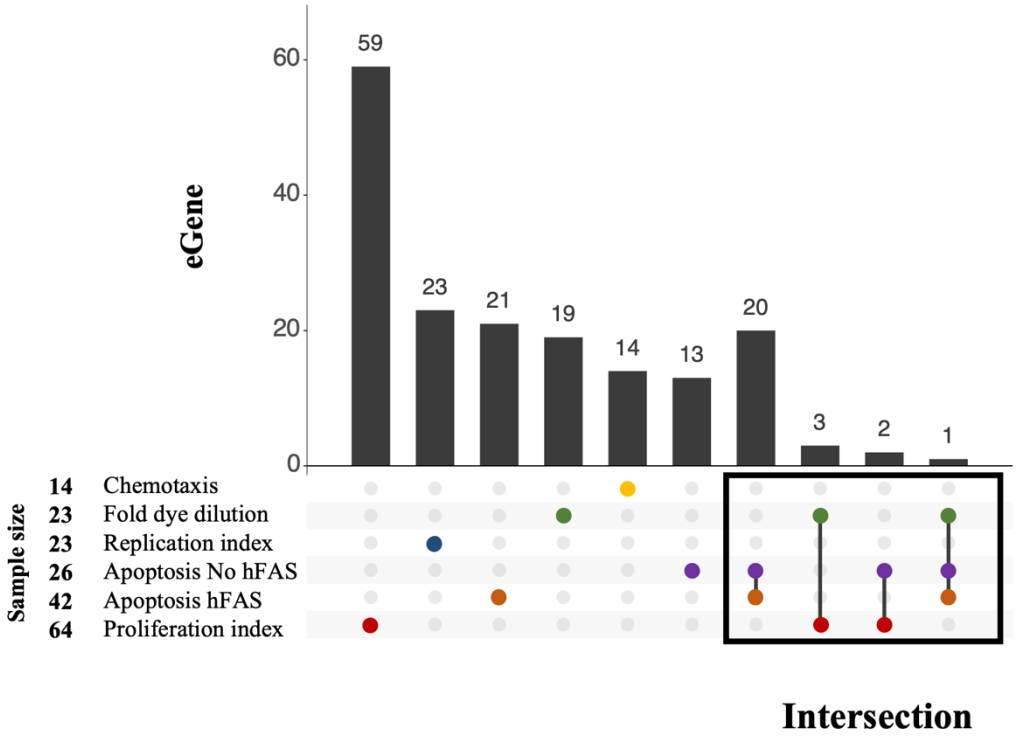
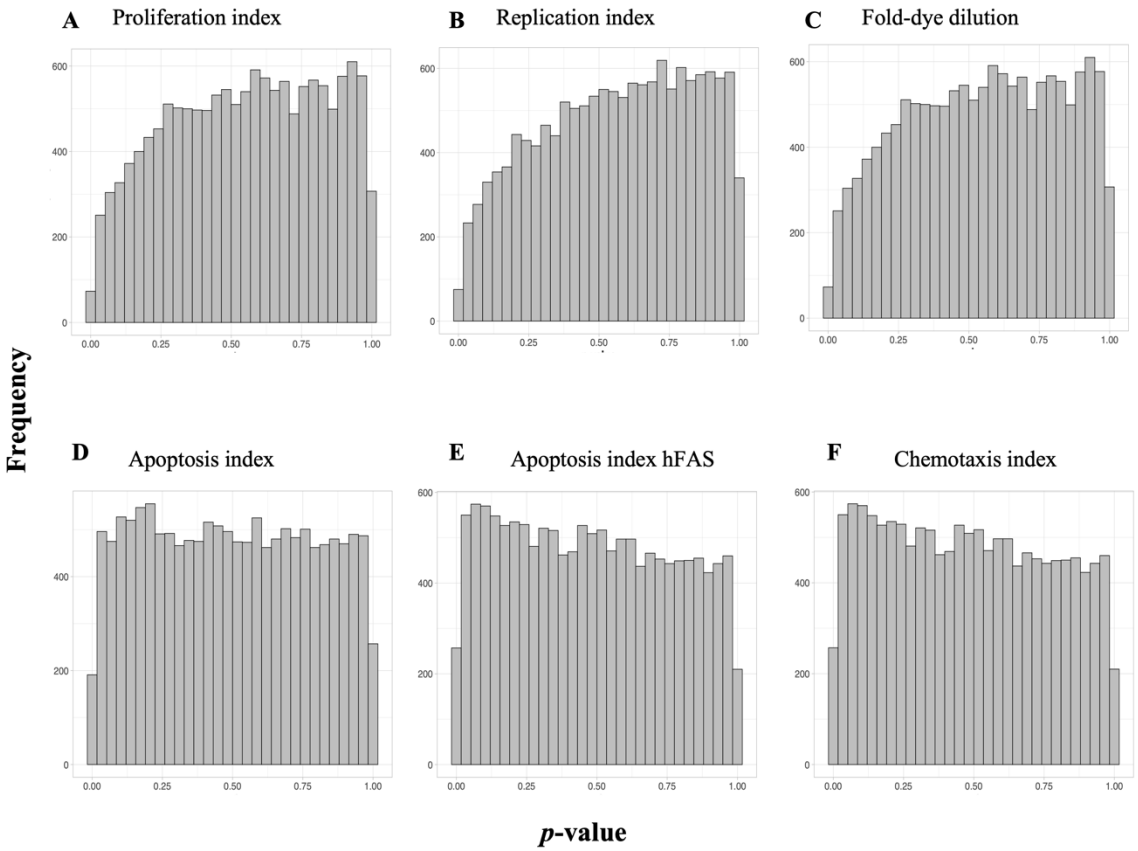

eGene
60
40
20
0
59
23
21
19
14
13
20
3
2
1
Sample size
14 Chemotaxis
23 Fold dye dilution
23 Replication index
26 Apoptosis No hFAS
42 Apoptosis hFAS
64 Proliferation index
Intersection

Figure 8


A Proliferation index
B Replication index
C Fold-dye dilution
D Apoptosis index
E Apoptosis index hFAS
F Chemotaxis index
Frequency
p-value

Supplementary tables

S4. SNP strongly correlated to the GWAS – hit rs12865307 with a GWAS – association p-value > − log10(2).

List of the 66 SNP located in the chr13 in the 1Mb window around the GWAS – hit SNP rs12865307 and strongly correlated to that SNP.

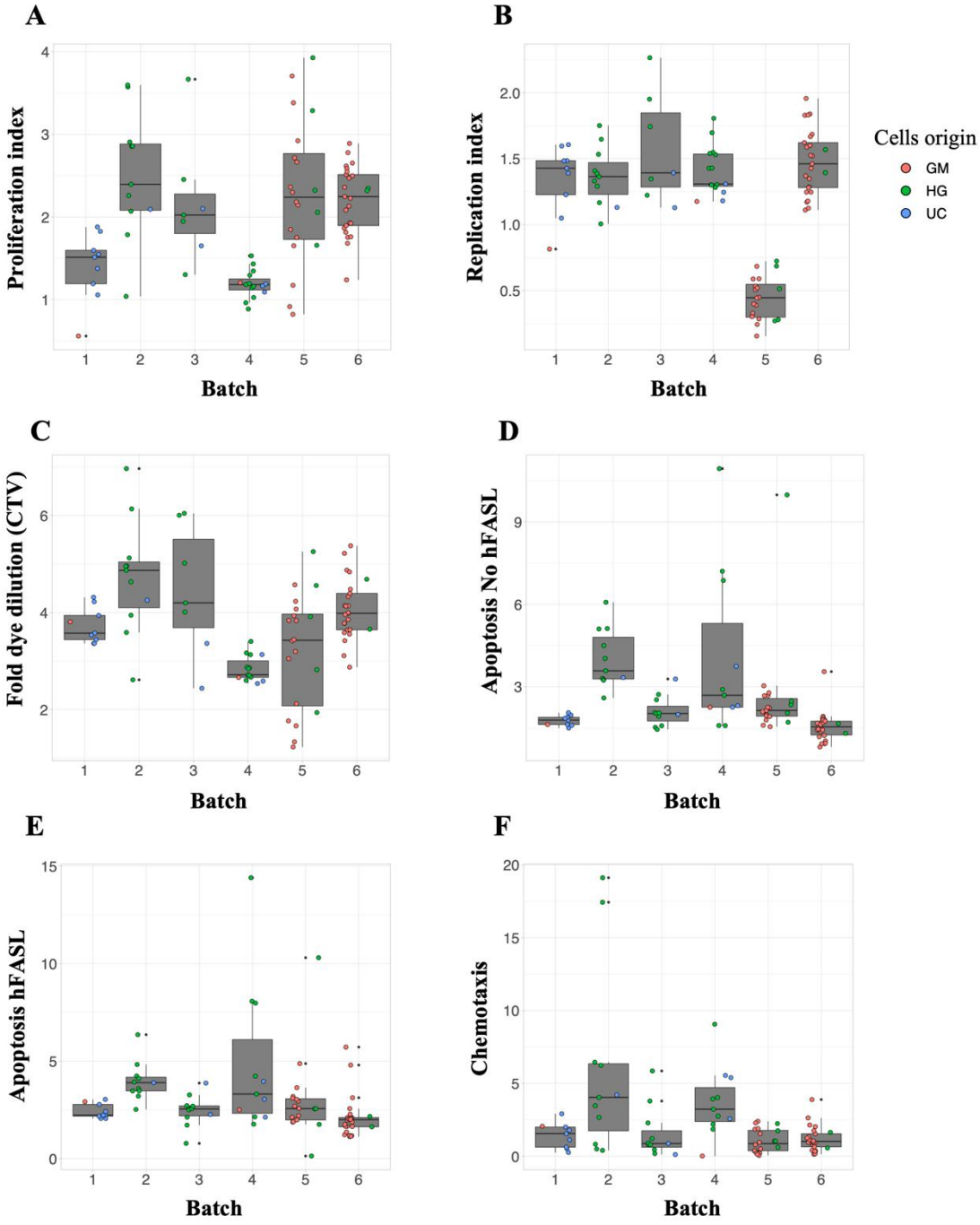
SNP ID	*p*-value	SNP ID	*p*-value
rs12865307	**1.66E-08**	rs12584855	9.16E-03
rs7333934	5.71E-08	rs9552414	9.17E-03
rs9578371	2.68E-07	rs9552415	9.18E-03
rs2315540	2.42E-05	rs4770140	1.03E-02
rs7337694	2.81E-05	rs75501492	1.21E-02
rs9316295	1.35E-04	rs35408239	1.27E-02
rs67864731	1.42E-04	rs35114912	1.28E-02
rs71202412	1.94E-04	rs1987501	1.34E-02
rs7320279	2.23E-04	rs7320093	1.52E-02
rs7327138	2.33E-04	rs7326077	1.52E-02
rs9552408	2.61E-04	rs9506652	1.66E-02
rs9509655	3.19E-04	rs4770139	1.84E-02
rs17356983	3.99E-04	rs261726	2.35E-02
rs9509656	4.10E-04	rs17065460	2.86E-02
rs17065406	4.50E-04	rs12853474	3.30E-02
rs9506654	5.52E-04	rs1963728	3.95E-02
rs9506653	6.49E-04	rs9316310	4.02E-02
rs73153907	8.01E-04	rs4770142	4.05E-02
rs2027154	8.37E-04	rs9316312	4.07E-02
rs7318257	8.89E-04	rs149443099	4.07E-02
rs7336525	9.11E-04	rs111360725	4.44E-02
rs7328466	1.17E-03	rs75361506	4.46E-02
rs9509657	1.20E-03	rs7336043	4.54E-02
rs11619856	7.25E-03	rs261727	4.55E-02
rs4770141	8.27E-03	rs9550727	4.82E-02
rs3066482	9.08E-03	rs7997365	5.32E-02
rs7983368	9.09E-03	rs7996457	5.81E-02
rs9552410	9.09E-03	rs57700509	6.13E-02
rs9552411	9.11E-03	rs9580087	6.88E-02
rs9552413	9.14E-03	rs9285163	8.20E-02

S5. List of the 20 eGenes in common between the Apoptosis index No hFAS and the Apoptosis index hFAS.

Gene ID	Phenotypes
ENSG00000206530.4	**Apoptosis No hFAS and Apoptosis hFAS**
ENSG00000169193.7	
ENSG00000269918.1	
ENSG00000255310.2	
ENSG00000104643.5	
ENSG00000246477.2	
ENSG00000154319.10	
ENSG00000136573.8	
ENSG00000255518.1	
ENSG00000255354.1	
ENSG00000177570.9	
ENSG00000172785.14	
ENSG00000166800.5	
ENSG00000074319.8	
ENSG00000257941.1	
ENSG00000211935.2	
ENSG00000259715.1	
ENSG00000141543.5	
ENSG00000181523.8	
ENSG00000198683.2	

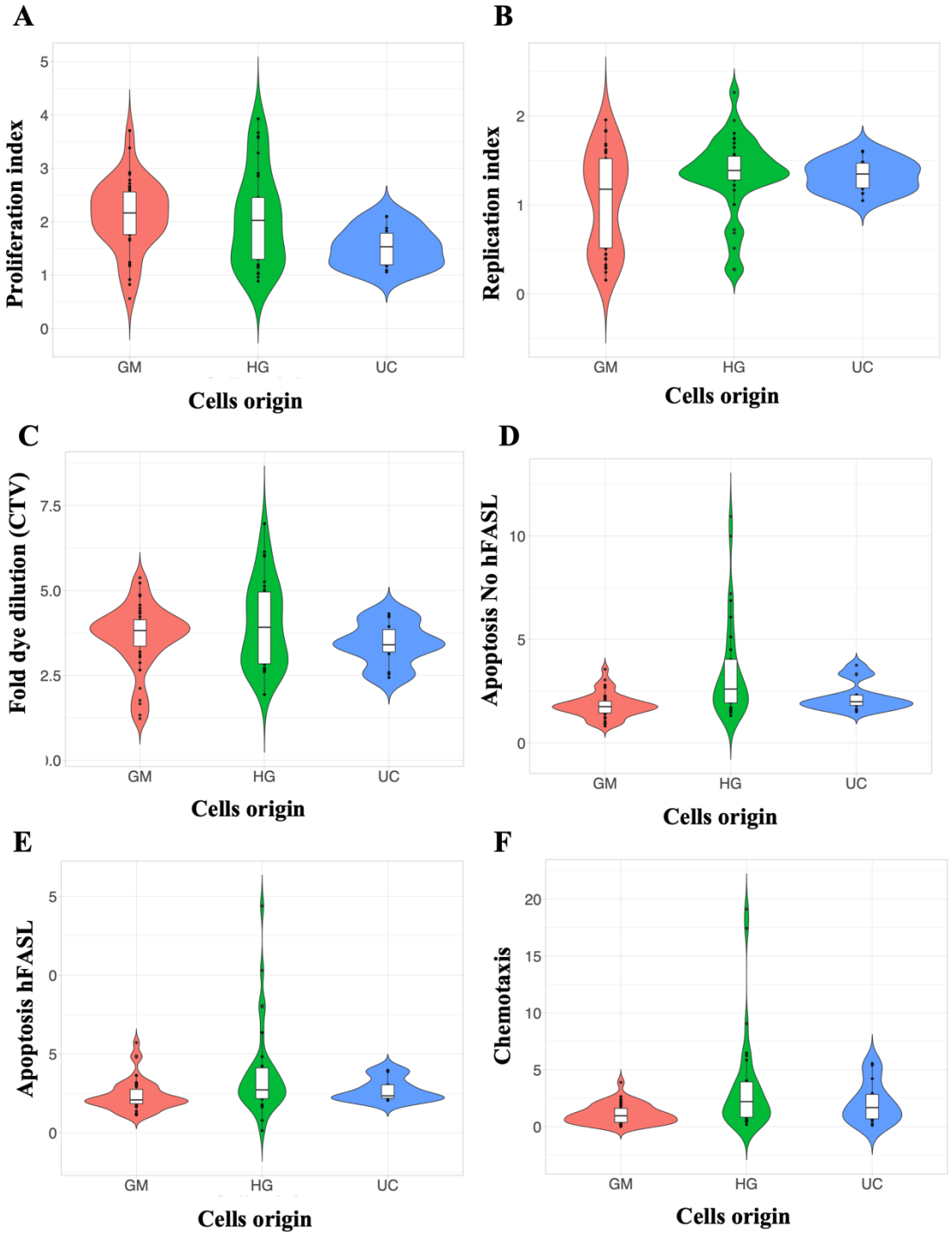
Supplementary figures

S1



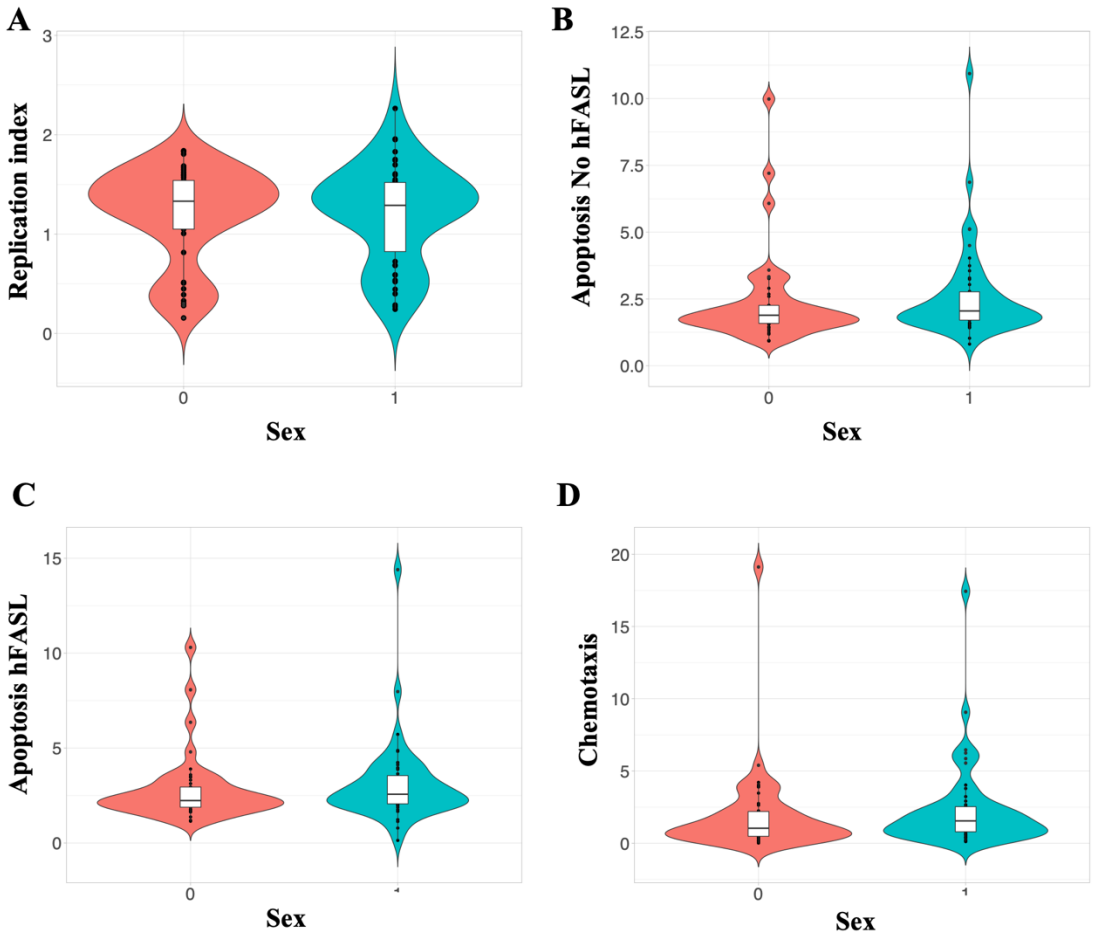
S1. Box-plots representing the batch effect for each of the six phenotypes analyzed. (A-F) Box-plots show the different correlations from a linear regression model between batch and **(A)** proliferation index, **(B)** replication index, **(C)** fold dye dilution, **(D)** apoptotic index no hFASL, **(E)** apoptotic index hFASL and **(F)** chemotaxis, respectively. We observed the presence of a different batch effect in all the six phenotypes produced that we take in consideration and correct for it all the analysis of this study.

S2



S2. Violin plots of the correlation between the cancer-like phenotypes and the technical covariate cells origin. **(A-F)** Violin plots show the different correlations from a linear regression model between cells origin and **(A)** proliferation index, **(B)** replication index, **(C)** fold dye dilution, **(D)** apoptotic index no hFASL, **(E)** apoptotic index hFASL and **(F)** chemotaxis, respectively. No significant correlation was observed.

S3



S3. Violin plots of the correlation between the cancer-like phenotypes and the biological covariate sex. (A-D) Violin plots show the different correlations from a linear regression model between sex and the **(A)** replication index, **(B)** apoptotic index no hFASL, **(C)** apoptotic index hFASL and **(D)** chemotaxis, respectively. The information about sex obtained from Delaneau et al. Science 2019 is displayed as 0 and 1, where 0 corresponds to male individuals and 1 to females. No strong correlation was observed between sex and these phenotypes.

General Discussion

The era of genomics has created unprecedented opportunities to discover and access a wide range of molecular data as a result of the development of new technologies. However, the availability of such a vast amount of data has also brought a new set of challenges. In particular, different high-throughput laboratory techniques have generated a significant amount of data that needs to be properly integrated and undergo the appropriate quantitative analysis, all of which remain a computational challenge requiring new tools and more powerful computers. The several different techniques used today to generate these data have their own their characteristics, biases, and artefacts that also need consideration when deriving biological meaning from them. Nevertheless, despite the challenges, this new wave of large dataset production provides opportunities to gain a deeper understanding of how organisms function, how tissues develop, and what causes disease and complex traits.

The field of human genetics shifted focus from studying the inheritance of acquired characteristics to identifying thousands of genomic variants impacting hundreds of diseases, complex traits, and thousands of molecular phenotypes [69, 129-131]. In parallel, sequencing transcriptomic studies (or RNA-seq studies) lead to a better understanding of how genome differences across individuals affect gene expression across various tissues, organs, and cell types, giving us a tool to understand how genetics can cause disease [132]. However, to date, most common sequencing approaches require laboratory manipulations before sequencing such as sequence amplification, reverse transcription and fragmentation of the molecules for sequencing (short reads) [133]. These manipulations preclude rapid, comprehensive and unbiased RNA molecule sequencing, preventing a more accurate measure of gene expression. At least partially, long-read sequencing technologies resolved one of these challenges by providing single-molecule sequencing resolution able to directly measure transcripts, the essential biological unit of the transcriptome [25]. We see the gains possible with long-read sequencing on our understanding of the human genome as well, making possible the sequencing of the remaining 8% of the genome previously missing from human reference genome on account of

technological limitations [134]. Specifically, the combination of PacBio and Nanopore long-read sequencing together with more established short-read sequencing removed the technological barriers that were present for 20 years, enabling the generation of a most complete human reference genome, which undoubtedly will result in advances in the field of genomic health and diseases [134].

All the current advances move us towards answering a fundamental question – How does a particular genotype result in a given phenotype? More specifically, – How does gene expression mediate the effect of genetic variants on phenotypes? In this discussion, I will examine the central role of splicing and whole transcript expression in understanding the functioning of the human genome and its link to disease. I will also discuss the importance of genetic variants influencing complex traits, and how cancer-related phenotypes can be studied using standard tools for the genetics of complex traits. My last goal is to discuss the connections between SNPs, genes and phenotypes with a specific focus on cancer-like phenotypes, and how challenging this can be using an *in vitro* model. Finally, I will expose limitations and lessons learned from this thesis, with a guide to fill out the gaps remaining in these two complex above-mentioned questions.

The central role of splicing in complex traits

Throughout this thesis, I aimed to answer questions related to the interconnection between genetic variants and changes in gene and transcript expression that may result in phenotypic differences. In the past, many genetic loci associated with diseases were identified through GWAS. However, we still lack clear insight into how these variants exert their effects on human traits [56]. Previous studies observed that the effects of GWAS variants are often mediated by variations in the expression of genes and also by RNA splicing events [85, 135]. In **Article 1**, we took advantage of the new ONT long-read direct RNA-seq technology to directly sequence full-length transcripts, thus improving the detection of structural differences between alternative transcripts produced by the same gene. The generation of multiple mRNAs from one gene is the result of alternative splicing events that are affected by genetic variants [85, 86, 129]. Therefore, by studying the influence of genetic variation

on the abundance of different transcripts coming from the same gene, we gain insight into how these variants exert their effect on human traits and shed light on some of the mechanisms behind this genetic regulation.

Genetic variants affecting transcript structure and abundance can be identified using quantitative trait loci (QTL) analyses of transcripts (trQTL) and splice events (sQTL). Specifically, trQTLs describe the effect of SNPs on the expression of a specific transcript among the other transcripts generated by the same gene; while sQTLs test for the effects of SNPs on splicing events, resulting in the generation of different mRNA molecules. Both types of phenotypes are highly interconnected, with splicing phenotypes being a proxy for transcript production as they identify the splice events that may produce particular transcripts [135]. It is important to highlight that previous studies have struggled to discriminate between sQTL and trQTL effects; only identifying eQTLs as SNPs affecting gene expression. In many of these cases, the detected eQTLs could be the result of effects on splicing or transcripts producing longer transcripts and thus more mRNA to sequence, or SNPs which caused one transcript to be more expressed. An eQTL will only inform of the presence of a genetic effect on the gene expression without elucidating the nature of this effect and therefore we expect that many eQTLs are also sQTLs and trQTLs [85, 86, 129]. Nevertheless, studies identifying sQTLs using proxy phenotypes and eQTLs in multiple human tissues concluded that genetic variations affecting gene expression levels and splicing were distinct [135, 136]. Furthermore, these studies highlighted that sQTL and eQTLs were independent, and the consequences of sQTLs on GWAS traits seemed larger than the consequences of eQTLs on GWAS traits [135, 136]. Using long-reads RNA-seq approaches we can direct measuring the full-length of the transcript, removing the need for complex transcripts structure reconstruction after sequencing or the development of proxy splice phenotypes, and precisely address the nature of a variants effect on expression [25, 26, 48, 137, 138].

In our study, I addressed the limitations of short-reads approaches by directly measuring the transcript expression, which in turn allowed us to perform trQTLs analysis not possible using short-reads

technologies. We observed substantial differences between eQTLs discovered with short and long-read technologies. Disregarding our limited sample size (60 LCLs), we identified more trQTLs than eQTLs as shown in **Figure 6** of **Article 1**. I discovered 72 trQTLs, of which 60 did not overlap with the significant eQTLs. Among the trQTLs, only one was also an eQTL previously discovered using Illumina short-reads [139]. This result strongly supports the notion that the effects of eQTLs and trQTLs are independent. Furthermore, of the SNP-gene associations found using both Illumina and Nanopore (4,157), only 5% were trQTLs in the ONT Nanopore dataset. The power of the analysis could explain this low number, an aspect discussed in detail in the following heading. However, the low number of trQTLs may also be the result of the independent effects of eQTLs and trQTLs, as has previously been shown in other studies [85, 86, 129]. Overall, these findings shows that multiple genetic mechanisms are relevant for changes in gene expression, and that eQTLs do not necessary recapitulate all the genetic effects on gene expression.

Splicing was also examined in **Article 2** in the context of cancer-like phenotypes. Many studies have already underlined the association between alternative splicing events and cancer [140, 141]. In particular, mutations affecting the splicing machinery have been reported in human hematological malignancies [142-144]. Moreover, it is common for tumor cells to present an abnormal splicing pattern with the production of oncogenic transcripts due to alternative splicing events [145, 146]. For instance, oncogenic transcripts were shown to play a role in the maintenance of the anomalous proliferative and apoptotic rhythm of cancer cells [145, 146]. Having large-scale genome, transcriptome and functional genomic data to study abnormal splicing events is essential to unveil the mechanism of gene and transcript expression dysregulation in cancer. In my study, I analyzed the relationship between genetic variants, gene expression and cancer-like phenotypes using long-read sequencing data. In **Article 2** I proposed the *SKA3* as a candidate gene to mediate the effect of the significant GWAS SNP rs12665307, associated with the replication index. In the ONT dataset (**Article 1**) I observed that *SKA3* produces four different transcripts. Even if the strength of the trQTL

effect I observed was not significant, I could detect a genetic effect on the expression of two of the isoforms of the *SKA3* gene (**Article 2**, **Figure 6B-C**).). In addition, *SKA3* has already been described as an important player in increased cell proliferation and migration in cervical cancer [147], a candidate oncogene associated with poor outcomes in lung adenocarcinomas [148], and a biomarker for immune infiltration in bladder cancers [149]. These findings highlight the relevance of designing population-based functional studies that could integrate not only transcriptomic and genomic data but also *in vitro* generated phenotypic data. With this approach, we are one step closer of narrowing the gap between the different molecular steps that link genotype to phenotype.

Finally, using allele-specific transcript structure analysis (ASTS), in addition to allele-specific expression analysis (ASE), could provide new information about how rare and common variants affect transcript structure and disease risk [150]. Long-read sequencing allows us to map allelic effects on transcripts instead of solely on gene expression [48]. In ASTS analysis, the inheritance of a transcript is deduced based on the allele present at heterozygous locations. In a previous study, Glinos *et al* highlighted that there was a widespread coexistence of ASTS with ASE and showed eQTLs manifesting as ASTS [48]. This contradicts evidence that distinct regulatory variants and processes regulate expression and splicing [77, 85, 129]. Glinos *et al* showed that with ASTS data it is possible to isolate allele-specific differences in 5' mRNA structure as the cause of eQTLs manifesting in transcript structure changes [48]. In light of this evidence, I strongly believe that it is important to analyze the transcriptome at the level of individual transcripts and the combinations of them, and not only focus on the gene level, which is now possible thanks to long-reads. Genetic variants that affect transcript structure are known to play a significant role in disease risk [77, 151, 152]. Therefore, high-resolution characterization of the transcriptome with long-read data will prove useful for investigating disease-associated regulatory mechanisms.

Long-reads direct RNA – sequencing limitations

The results from **Article 1** described in the previous heading stressed the advantages of using long-read direct RNA-seq for the detection of splicing events and transcripts abundance. However, this novel technology also has technical limitations that deserve further discussion. For instance, short-read RNA-seq often produces a very large number of reads per sample, comprehensively describing the transcriptome [153]. However, sequencing of direct RNA, currently only possible with ONT, has an upper limit of sequencing depth due to saturation of the nanopores and a limited amount of starting biological material. When this project was designed, with the state of the technology available, it was not possible to achieve high sequencing depth unless the same sample was sequenced multiple times. In the past few months, the technology for direct RNA-seq has particularly improved, increasing the RNA-sequencing depth and the reads coverage (nanoporetech.com). In the current study, limited sequencing depth was probably one reason we were unable to discover the vast majority of the significant eQTLs identified with Illumina short-reads. Since the eQTLs obtained by Illumina technology were already confirmed by many independent studies using the short-reads approach [129, 132, 139], they were proven to be robust genetic effects and our inability to capture them underline a possible limitation of the ONT direct RNA-seq. While sequencing the same sample multiple times would increase the sequencing depth, this is complicated by the amount of starting material required for direct RNA-seq long-read and increased cost. For each sequencing sample, 500ng of PolyA$^+$ enriched mRNA was required for the library preparation (nanoporetech.com). Since PolyA$^+$ mRNA constitutes only 2-3% of the total RNA obtained per sample, this means that, with the current state of the technology, a larger amount of initial biological material (> 40×10^6 LCLs) would be required to achieve higher sequencing depth. Thus, to achieve a coverage comparable to Illumina short reads (~30×10^6, approximately equivalent to dRNA 3×10^6 reads per sample for ONT), one would need to sequence each LCL sample at least three times, tripling the cost. At the time of the study, the

sequencing of a sample cost nearly CHF 900 (all material included); multiplying that by three the sequencing of 60 samples would have cost at least an additional CHF 108,000.

Other groups have attempted to overcome these limitations by using cDNA as the starting material for sequencing. While this approach partially solves the limitation posed by sequencing depth and cost, it maintains the biases represented by the retro-transcription process. In addition, RNA-seq using cDNA precludes the study of mRNA molecules in their native form, with all their modifications. All in all, the direct RNA long-read sequencing technology is a promising tool still under development. However, I believe that addressing these limitations and being aware of the caveats of the technology will assist researchers in better designing their own studies and will pave the way for its use in a clinical setting.

Challenges in genetic – phenotypic associations in cancer-like phenotypes

Another aspect to consider is the complexity that we faced when we investigated the effects of genetic variants on gene expression and ultimately on cancer-like phenotypes produced *in vitro*. As previously mentioned, GWAS indicates that non-coding regulatory variants play a significant role in cancer development and progression, which expands the traditional view of cancer as a disease caused by somatic mutations [154-156]. To understand the basic biological mechanisms of disease initiation and progression, it is essential to gain a thorough understanding of the genes that influence these processes. In the study, **Article 2**, we investigated the link between genetic variants and cancer-like phenotypes in an attempt to identify genes mediating the effect of germinal variations. I faced difficult challenges regarding the small sample size (n = 87) and the complexity of the cancer-like phenotypes "mimicked" *in vitro*. I assumed that reproducing cancer-like phenotypes *in vitro* will reduce the degree of complexity observed *in vivo*, and that finding associations between genetic variants and phenotypes would be a straightforward endeavor. Although I performed the functional assays in a controlled experimental environment, under specific conditions, and measured subtle differences in the LCL population, the link between genetic variation and changes in cancer phenotypes was

difficult to establish. Certainly, the small sample size played a key role in limiting the statistical power. But our study design also reduces complex phenotypes such as cell proliferation, apoptosis and chemotaxis to a single number and loses the global perspective of physiological cell interactions for in cancer *in vivo*.

Regardless of the limitations of the experimental system, I discovered the association between SNP rs12865307 and the replication index in the GWAS analysis (**Article 2**, **Figure 4A**). Furthermore, after identifying the putative causal variants with colocalization in both GWAS and eQTL studies, I could determine whether a single variant was responsible for both GWAS and eQTL signals at the locus to identify putative genes mediating the genetic effects. The SNP rs12865307 is known to be a significant LCL eQTL (https://gtexportal.org) affecting the expression of the Long Intergenic Non-Protein Coding RNA 539 (*LINC00539*) gene. *LINC00539* was already described as an important player in the tumor immune response in lung adenocarcinoma and lung squamous cell carcinoma [157] and associated with the overall survival in Acute myeloid leukemia [158]. These findings underline the potential of using this approach to identify germline variants influencing cancer-related phenotypes and find genes involved in such effects. Increasing the sample size will augment the statistical power to discover significant genome-wide associations and produce more putative cancer driver genes.

Concluding remarks

This thesis has attempted to answer two important questions in the field of genetics of human complex traits: How genetic variation affects gene expression? And what are the mechanisms that determine complex phenotypes and diseases such as cancer? Since the advent of high-throughput sequencing methods, several large-scale efforts have been pursued to establish resources of gene expression data, with the aim to build a comprehensive understanding of gene expression specificity and variability across human tissues. One critical challenge remaining is how to use these resources to better understand complex traits and diseases like cancer, with the ultimate goal to move from bench science

to clinical practice. One aspect that needs more focus is the role of non-coding variants which are associated with splicing in the development of diseases, including cancer. With this work, I contributed to this with a novel layer of information coming from new sequencing technology. I have also attempted to integrate this technology into *in vitro* models of disease.

In summary, to understand human genetics from genotype to disease, it is necessary to study gene expression, genetic variant, and disease together within the same framework. It was my intention in this thesis to shed light on the effects of genetic variation on gene expression and phenotypes in humans and the mechanisms that mediate these effects, a complex problem with real potential for improving human health.

General references

1. https://www.genome.gov/. 01/07/2022; Available from: https://www.genome.gov/human-genome-project/Completion-FAQ.
2. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
3. Clancy, S.B., W., *Translation: DNA to mRNA to Protein*. 2008, Nature Education 1(1):101, Corpus ID: 89529269
4. Proudfoot, N.J., A. Furger, and M.J. Dye, *Integrating mRNA processing with transcription.* Cell, 2002. **108**(4): p. 501-12.
5. Sims, R.J., S.S. Mandal, and D. Reinberg, *Recent highlights of RNA-polymerase-II-mediated transcription.* Curr Opin Cell Biol, 2004. **16**(3): p. 263-71.
6. Sharp, P.A., *Split genes and RNA splicing.* Cell, 1994. **77**(6): p. 805-15.
7. Kornblihtt, A.R., et al., *Multiple links between transcription and splicing.* RNA, 2004. **10**(10): p. 1489-98.
8. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
9. Matlin, A.J., F. Clark, and C.W. Smith, *Understanding alternative splicing: towards a cellular code.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-98.
10. Park, E., et al., *The Expanding Landscape of Alternative Splicing Variation in Human Populations.* Am J Hum Genet, 2018. **102**(1): p. 11-26.
11. Kelemen, O., et al., *Function of alternative splicing.* Gene, 2013. **514**(1): p. 1-30.
12. Scheper, G.C., M.S. van der Knaap, and C.G. Proud, *Translation matters: protein synthesis defects in inherited disease.* Nat Rev Genet, 2007. **8**(9): p. 711-23.
13. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
14. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
15. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.
16. Ferreira, P.G., et al., *Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia.* Genome Res, 2014. **24**(2): p. 212-26.
17. Consortium, S.M.-I., *A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium.* Nat Biotechnol, 2014. **32**(9): p. 903-14.

18. Stark, R., M. Grzelak, and J. Hadfield, *RNA sequencing: the teenage years.* Nat Rev Genet, 2019. **20**(11): p. 631-656.
19. Corchete, L.A., et al., *Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis.* Sci Rep, 2020. **10**(1): p. 19737.
20. Inc., I. *Illumina. For all you seq. Illumina* 2014; Available from: https://emea. illumina.com/techniques/sequencing/ngs-library-prep/ library-prep-methods.html.
21. Leinonen, R., et al., *The sequence read archive.* Nucleic Acids Res, 2011. **39**(Database issue): p. D19-21.
22. Hollox, E.J., L.W. Zuccherato, and S. Tucci, *Genome structural variation in human evolution.* Trends Genet, 2022. **38**(1): p. 45-58.
23. Salzberg, S.L. and J.A. Yorke, *Beware of mis-assembled genomes.* Bioinformatics, 2005. **21**(24): p. 4320-1.
24. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat Rev Genet, 2011. **13**(1): p. 36-46.
25. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome Biol, 2020. **21**(1): p. 30.
26. Tilgner, H., et al., *Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events.* Nat Biotechnol, 2015. **33**(7): p. 736-42.
27. Tilgner, H., et al., *Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome.* Genome Res, 2018. **28**(2): p. 231-242.
28. Wright, D.J., et al., *Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes.* BMC Genomics, 2022. **23**(1): p. 42.
29. Wang, Y., et al., *Nanopore sequencing technology, bioinformatics and applications.* Nat Biotechnol, 2021. **39**(11): p. 1348-1365.
30. Logsdon, G.A., M.R. Vollger, and E.E. Eichler, *Long-read human genome sequencing and its applications.* Nat Rev Genet, 2020. **21**(10): p. 597-614.
31. Chaisson, M.J., R.K. Wilson, and E.E. Eichler, *Genetic variation and the de novo assembly of human genomes.* Nat Rev Genet, 2015. **16**(11): p. 627-40.
32. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads.* Nat Biotechnol, 2018. **36**(4): p. 338-345.
33. Miga, K.H., et al., *Telomere-to-telomere assembly of a complete human X chromosome.* Nature, 2020. **585**(7823): p. 79-84.

34. Rang, F.J., W.P. Kloosterman, and J. de Ridder, *From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy.* Genome Biol, 2018. **19**(1): p. 90.
35. Wick, R.R., L.M. Judd, and K.E. Holt, *Performance of neural network basecalling tools for Oxford Nanopore sequencing.* Genome Biol, 2019. **20**(1): p. 129.
36. Mantere, T., S. Kersten, and A. Hoischen, *Long-Read Sequencing Emerging in Medical Genetics.* Front Genet, 2019. **10**: p. 426.
37. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
38. Mikheyev, A.S. and M.M. Tin, *A first look at the Oxford Nanopore MinION sequencer.* Mol Ecol Resour, 2014. **14**(6): p. 1097-102.
39. Salk, J.J., M.W. Schmitt, and L.A. Loeb, *Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations.* Nat Rev Genet, 2018. **19**(5): p. 269-285.
40. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores.* Nat Methods, 2018. **15**(3): p. 201-206.
41. Liu, H., et al., *Accurate detection of m^6A modifications in native RNA sequences.* Nat Commun, 2019. **10**(1): p. 4079.
42. Oikonomopoulos, S., et al., *Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations.* Sci Rep, 2016. **6**: p. 31602.
43. Simpson, J.T., et al., *Detecting DNA cytosine methylation using nanopore sequencing.* Nat Methods, 2017. **14**(4): p. 407-410.
44. Keller, M.W., et al., *Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome.* Sci Rep, 2018. **8**(1): p. 14408.
45. nanoporetech.com. *Nanopore accuracy*. 3rd December 2021; Available from: https://nanoporetech.com/.
46. Byrne, A., et al., *Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells.* Nat Commun, 2017. **8**: p. 16027.
47. Tang, A.D., et al., *Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.* Nat Commun, 2020. **11**(1): p. 1438.
48. Glinos, D.A., Garborcauskas, G., Hoffman, P. et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature (2022). doi: 10.1038/s41586-022-05035-y.
49. Consortium, I.H., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

50. Thorpe, J., et al., *Mosaicism in Human Health and Disease.* Annu Rev Genet, 2020. **54**: p. 487-510.
51. Thorisson, G.A., et al., *The International HapMap Project Web site.* Genome Res, 2005. **15**(11): p. 1592-3.
52. Altshuler, D.M., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.
53. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
54. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
55. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
56. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.
57. Harold, D., et al., *Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease.* Nat Genet, 2009. **41**(10): p. 1088-93.
58. Simón-Sánchez, J., et al., *Genome-wide association study reveals genetic risk underlying Parkinson's disease.* Nat Genet, 2009. **41**(12): p. 1308-12.
59. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. **17**(9): p. 502-10.
60. Hardy, J. and A. Singleton, *Genomewide association studies and human disease.* N Engl J Med, 2009. **360**(17): p. 1759-68.
61. Pe'er, I., et al., *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genet Epidemiol, 2008. **32**(4): p. 381-5.
62. Li, M.J., et al., *GWASdb: a database for human genetic variants identified by genome-wide association studies.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1047-54.
63. Dermitzakis, E.T., *From gene expression to disease risk.* Nat Genet, 2008. **40**(5): p. 492-3.
64. Nica, A.C. and E.T. Dermitzakis, *Expression quantitative trait loci: present and future.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120362.
65. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743-7.
66. Bryois, J., et al., *Cis and trans effects of human genomic variants on gene expression.* PLoS Genet, 2014. **10**(7): p. e1004461.

67. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.
68. Fu, J., et al., *Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression.* PLoS Genet, 2012. **8**(1): p. e1002431.
69. Grundberg, E., et al., *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nat Genet, 2012. **44**(10): p. 1084-9.
70. Montgomery, S.B., et al., *Rare and common regulatory variation in population-scale sequenced human genomes.* PLoS Genet, 2011. **7**(7): p. e1002144.
71. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population.* Nature, 2010. **464**(7289): p. 773-7.
72. Nica, A.C., et al., *The architecture of gene regulatory variation across multiple human tissues: the MuTHER study.* PLoS Genet, 2011. **7**(2): p. e1002003.
73. Pickrell, J.K., et al., *Understanding mechanisms underlying human gene expression variation with RNA sequencing.* Nature, 2010. **464**(7289): p. 768-72.
74. Stranger, B.E., et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes.* Science, 2007. **315**(5813): p. 848-53.
75. Stranger, B.E., et al., *Patterns of cis regulatory variation in diverse human populations.* PLoS Genet, 2012. **8**(4): p. e1002639.
76. Ongen, H., et al., *Putative cis-regulatory drivers in colorectal cancer.* Nature, 2014. **512**(7512): p. 87-90.
77. Consortium, G., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.
78. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease.* Nat Rev Genet, 2015. **16**(4): p. 197-212.
79. Cano-Gamez, E. and G. Trynka, *From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases.* Front Genet, 2020. **11**: p. 424.
80. Ongen, H., et al., *Estimating the causal tissues for complex traits and diseases.* Nat Genet, 2017. **49**(12): p. 1676-1683.
81. Raj, T., et al., *Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility.* Nat Genet, 2018. **50**(11): p. 1584-1592.
82. Takata, A., N. Matsumoto, and T. Kato, *Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci.* Nat Commun, 2017. **8**: p. 14519.

83. Tian, J., et al., *CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer.* Nucleic Acids Res, 2019. **47**(D1): p. D909-D916.
84. Caswell, J.L., et al., *Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors.* Hum Mol Genet, 2015. **24**(25): p. 7421-31.
85. Li, Y.I., et al., *RNA splicing is a primary link between genetic variation and disease.* Science, 2016. **352**(6285): p. 600-4.
86. Battle, A., et al., *Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.* Genome Res, 2014. **24**(1): p. 14-24.
87. Monlong, J., et al., *Identification of genetic variants associated with alternative splicing using sQTLseekeR.* Nat Commun, 2014. **5**: p. 4698.
88. Zhao, K., et al., *GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data.* Genome Biol, 2013. **14**(7): p. R74.
89. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
90. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
91. Dagogo-Jack, I. and A.T. Shaw, *Tumour heterogeneity and resistance to cancer therapies.* Nat Rev Clin Oncol, 2018. **15**(2): p. 81-94.
92. Levine, A.J., *p53: 800 million years of evolution and 40 years of discovery.* Nat Rev Cancer, 2020. **20**(8): p. 471-480.
93. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
94. Nakagawa, H., et al., *Role of cancer-associated stromal fibroblasts in metastatic colon cancer to the liver and their expression profiles.* Oncogene, 2004. **23**(44): p. 7366-77.
95. McPherson, K., C.M. Steel, and J.M. Dixon, *ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics.* BMJ, 2000. **321**(7261): p. 624-8.
96. Bignell, G.R., et al., *Signatures of mutation and selection in the cancer genome.* Nature, 2010. **463**(7283): p. 893-8.
97. Rahman, N., *Realizing the promise of cancer predisposition genes.* Nature, 2014. **505**(7483): p. 302-8.
98. Vasen, H.F., et al., *New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC.* Gastroenterology, 1999. **116**(6): p. 1453-6.

99. Antoniou, A., et al., *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies.* Am J Hum Genet, 2003. **72**(5): p. 1117-30.
100. Maxwell, K.N., et al., *Prevalence of mutations in a panel of breast cancer susceptibility genes in BRCA1/2-negative patients with early-onset breast cancer.* Genet Med, 2015. **17**(8): p. 630-8.
101. Consortium, B.C.L., *Cancer risks in BRCA2 mutation carriers.* J Natl Cancer Inst, 1999. **91**(15): p. 1310-6.
102. van Asperen, C.J., et al., *Cancer risks in BRCA2 families: estimates for sites other than breast and ovary.* J Med Genet, 2005. **42**(9): p. 711-9.
103. Braithwaite, D., et al., *Screening outcomes in older US women undergoing multiple mammograms in community practice: does interval, age, or comorbidity score affect tumor characteristics or false positive rates?* J Natl Cancer Inst, 2013. **105**(5): p. 334-41.
104. Southey, M.C., et al., *Morphological predictors of BRCA1 germline mutations in young women with breast cancer.* Br J Cancer, 2011. **104**(6): p. 903-9.
105. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.
106. Chen, C.Y., et al., *On the identification of potential regulatory variants within genome wide association candidate SNP sets.* BMC Med Genomics, 2014. **7**: p. 34.
107. Sud, A., B. Kinnersley, and R.S. Houlston, *Genome-wide association studies of cancer: current insights and future perspectives.* Nat Rev Cancer, 2017. **17**(11): p. 692-704.
108. Horn, S., et al., *TERT promoter mutations in familial and sporadic melanoma.* Science, 2013. **339**(6122): p. 959-61.
109. Easton, D.F. and R.A. Eeles, *Genome-wide association studies in cancer.* Hum Mol Genet, 2008. **17**(R2): p. R109-15.
110. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-8.
111. Amoli, M.M., et al., *EBV Immortalization of human B lymphocytes separated from small volumes of cryo-preserved whole blood.* Int J Epidemiol, 2008. **37 Suppl 1**: p. i41-5.
112. Young, L.S. and A.B. Rickinson, *Epstein-Barr virus: 40 years on.* Nat Rev Cancer, 2004. **4**(10): p. 757-68.
113. Luskin, R. and H. Nathan, *Eligible Death Statistic: Not a True Measure of OPO Performance nor the Potential to Increase Transplantation.* Am J Transplant, 2015. **15**(8): p. 2019-20.

114. Hecht, J.L. and J.C. Aster, *Molecular biology of Burkitt's lymphoma.* J Clin Oncol, 2000. **18**(21): p. 3707-21.
115. Bhowmick, N.A., E.G. Neilson, and H.L. Moses, *Stromal fibroblasts in cancer initiation and progression.* Nature, 2004. **432**(7015): p. 332-7.
116. Cheng, M., et al., *Diagnostic utility of LunX mRNA in peripheral blood and pleural fluid in patients with primary non-small cell lung cancer.* BMC Cancer, 2008. **8**: p. 156.
117. Wong, R.S., *Apoptosis in cancer: from pathogenesis to treatment.* J Exp Clin Cancer Res, 2011. **30**: p. 87.
118. Carneiro, B.A. and W.S. El-Deiry, *Targeting apoptosis in cancer therapy.* Nat Rev Clin Oncol, 2020. **17**(7): p. 395-417.
119. Galluzzi, L. and I. Vitale, *Oncogene-induced senescence and tumour control in complex biological systems.* Cell Death Differ, 2018. **25**(6): p. 1005-1006.
120. Wajant, H., *The Fas signaling pathway: more than a paradigm.* Science, 2002. **296**(5573): p. 1635-6.
121. Gibert, B. and P. Mehlen, *Dependence Receptors and Cancer: Addiction to Trophic Ligands.* Cancer Res, 2015. **75**(24): p. 5171-5.
122. Bird, C. and S. Kirstein, *Real-time, label-free monitoring of cellular invasion and migration with the xCELLigence system.* Nature Methods, 2009. **6**(8): p. v-vi.
123. Roussos, E.T., J.S. Condeelis, and A. Patsialou, *Chemotaxis in cancer.* Nat Rev Cancer, 2011. **11**(8): p. 573-87.
124. McSherry, E.A., et al., *Molecular basis of invasion in breast cancer.* Cell Mol Life Sci, 2007. **64**(24): p. 3201-18.
125. Balkwill, F., *Cancer and the chemokine network.* Nat Rev Cancer, 2004. **4**(7): p. 540-50.
126. Lazennec, G. and A. Richmond, *Chemokines and chemokine receptors: new insights into cancer-related inflammation.* Trends Mol Med, 2010. **16**(3): p. 133-44.
127. Müller, A., et al., *Involvement of chemokine receptors in breast cancer metastasis.* Nature, 2001. **410**(6824): p. 50-6.
128. Murphy, P.M., *Chemokines and the molecular basis of cancer metastasis.* N Engl J Med, 2001. **345**(11): p. 833-5.
129. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
130. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

131. Westra, H.J., et al., *Systematic identification of trans eQTLs as putative drivers of known disease associations.* Nat Genet, 2013. **45**(10): p. 1238-1243.
132. Battle, A., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
133. Hrdlickova, R., M. Toloue, and B. Tian, *RNA-Seq methods for transcriptome analysis.* Wiley Interdiscip Rev RNA, 2017. **8**(1).
134. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**(6588): p. 44-53.
135. Garrido-Martín, D., et al., *Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome.* Nat Commun, 2021. **12**(1): p. 727.
136. Li, Y.I., et al., *Annotation-free quantification of RNA splicing using LeafCutter.* Nat Genet, 2018. **50**(1): p. 151-158.
137. Sedlazeck, F.J., et al., *Piercing the dark matter: bioinformatics of long-range sequencing and mapping.* Nat Rev Genet, 2018. **19**(6): p. 329-346.
138. Weirather, J.L., et al., *Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis.* F1000Res, 2017. **6**: 100. doi: 10.12688/f1000research.10571.2.eCollection 2017.
139. Delaneau, O., et al., *Chromatin three-dimensional interactions mediate genetic effects on gene expression.* Science, 2019. **364**(6439). doi: 10.1126/science.aat8266.
140. Dvinge, H., et al., *RNA splicing factors as oncoproteins and tumour suppressors.* Nat Rev Cancer, 2016. **16**(7): p. 413-30.
141. Singh, B. and E. Eyras, *The role of alternative splicing in cancer.* Transcription, 2017. **8**(2): p. 91-98.
142. Liu, J., et al., *Aberrant expression of splicing factors in newly diagnosed acute myeloid leukemia.* Onkologie, 2012. **35**(6): p. 335-40.
143. Kanagal-Shamanna, R., et al., *Myeloid neoplasms with isolated isochromosome 17q demonstrate a high frequency of mutations in SETBP1, SRSF2, ASXL1 and NRAS.* Oncotarget, 2016. **7**(12): p. 14251-8.
144. Itzykson, R., et al., *Prognostic score including gene mutations in chronic myelomonocytic leukemia.* J Clin Oncol, 2013. **31**(19): p. 2428-36.
145. Oltean, S. and D.O. Bates, *Hallmarks of alternative splicing in cancer.* Oncogene, 2014. **33**(46): p. 5311-8.
146. Belluti, S., G. Rigillo, and C. Imbriano, *Transcription Factors in Cancer: When Alternative Splicing Determines Opposite Cell Fates.* Cells, 2020. **9**(3): 760. doi: 10.3390/cells9030760.

147. Hu, R., et al., *SKA3 promotes cell proliferation and migration in cervical cancer by activating the PI3K/Akt signaling pathway.* Cancer Cell Int, 2018. **18**: p. 183.
148. Lin, Y., et al., *Integrative Multi-Omics Analysis of Identified SKA3 as a Candidate Oncogene Correlates with Poor Prognosis and Immune Infiltration in Lung Adenocarcinoma.* Int J Gen Med, 2022. **15**: p. 4635-4647.
149. Wang, C., et al., *SKA3 is a prognostic biomarker and associated with immune infiltration in bladder cancer.* Hereditas, 2022. **159**(1): p. 20.
150. Castel, S.E., et al., *Tools and best practices for data processing in allelic expression analysis.* Genome Biol, 2015. **16**, 195. doi: 10.1186/s13059-015-0762-6.
151. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
152. Gandal, M.J., et al., *Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder.* Science, 2018. **362**(6420). doi: 10.1126/science.aat8127.
153. Nellore, A., et al., *Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive.* Genome Biol, 2016. **17**(1): 266.
154. Xu, X., et al., *Germline genomic patterns are associated with cancer risk, oncogenic pathways, and clinical outcomes.* Sci Adv, 2020. **6**(48). doi: 10.1126/sciadv.aba4905.
155. Lichtenstein, P., et al., *Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland.* N Engl J Med, 2000. **343**(2): p. 78-85.
156. Phelan, C.M., et al., *Mutation analysis of the BRCA2 gene in 49 site-specific breast cancer families.* Nat Genet, 1996. **13**(1): p. 120-2.
157. Zengin, T. and T. Önal-Süzek, *Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma.* J Pers Med, 2021. **11**(2): 154. doi: 10.3390/jpm11020154.
158. Wang, S., et al., *A Novel Immune-Related Competing Endogenous RNA Network Predicts Prognosis of Acute Myeloid Leukemia.* Front Oncol, 2020. **10**: 1579. doi: 10.3389/fonc.2020.01579. eCollection 2020.

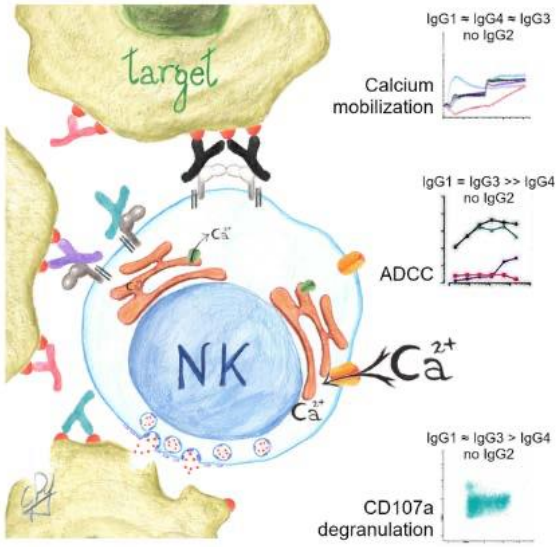
Annex 1

Anti-CD20 rituximab IgG1, IgG3 and IgG4 but not IgG2 subclass trigger Ca^{2+} mobilization and cytotoxicity in human NK cells

Marta Freitas Monteiro, Maria Papaserafeim, **Aline Réal**, Gisella L. Puga Yung, Jörg D. Seebach

Published on *Journal Leukocyte Biology*

DOI: 10.1002/JLB.5MA0620-039R. Epub 2020 Jul 3.



Goal: The current study aimed to analyze *in vitro* the role of IgG subclasses and *FCGR3A* polymorphism on the cytotoxic function of effector immune cells. For that, we performed calcium mobilization assays and cytotoxic assays using NK cells from healthy human donors (effectors cells), Daudi cell line (target cells), and variable concentration of anti-CD20 monoclonal antibody subclasses. We found out that Ca^{2+} signaling and Ab-dependent cell cytotoxicity induced by anti-CD20 rituximab in NK cells depend on the IgG subclass and weakly correlate with *FCGR3A* polymorphism.

Personal contribution: My involvement in this paper was in performing the molecular biology experiments that included the DNA extraction, the RT-PCR and the production of the **Supplementary figure 4**. I also took part of the cytotoxicity assays.

DOI: 10.1002/JLB.5MA0620-039R

JLB JOURNAL OF LEUKOCYTE BIOLOGY

ARTICLE

Anti-CD20 rituximab IgG1, IgG3, and IgG4 but not IgG2 subclass trigger Ca^{2+} mobilization and cytotoxicity in human NK cells

Marta Freitas Monteiro | Maria Papaserafeim | Aline Réal | Gisella L. Puga Yung | Jörg D. Seebach

Division of Immunology and Allergology, Department of Medicine, University Hospitals and Medical Faculty, Geneva, Switzerland

Correspondence
Prof. Jörg D. Seebach, University Hospitals Geneva, Division of Clinical Immunology and Allergy, Rue Gabrielle-Perret-Gentil 4, CH-1205 Geneva, Switzerland
E-mail: joerg.seebach@hcuge.ch

Abstract

NK cell-mediated Ab-dependent cellular cytotoxicity (ADCC) is increasingly recognized to play an important role in cancer immunotherapy, transplant rejection, and autoimmunity. However, several aspects of the molecular interactions of IgG subclasses with the Fc-gamma receptor IIIA (FcγRIIIA)/CD16a expressed on NK cells remain unknown. The aim of the current study was to further analyze the role of IgG subclasses and *FCGR3A* V158F single nucleotide polymorphism (SNP) on Ca^{2+} signaling and NK cell-mediated ADCC against Daudi target cells in vitro. NK cells were isolated from donors with different *FCGR3A* SNP. The affinity of rituximab IgG subclasses to CD20 expressed on Daudi cells showed similar dissociation constant as tested by flow cytometry. Induction of Ca^{2+} signaling, degranulation, intracellular cytokine production, and ADCC was demonstrated for IgG1 and IgG3, to a lesser degree also for IgG4, but not for IgG2. Compared to NK cells carrying the low-affinity (FF) variant for the *FCGR3A* V158F SNP, binding of IgG1 and IgG3 to NK cells carrying the high-affinity (VV) and VF SNP variants was two- to threefold higher. Variations of *FCGR3A* SNP among the eight tested donors (1 VV, 3FF, and 4VF) revealed no significant differences of Ca^{2+} signaling and degranulation; however, ADCC was somewhat weaker in donors with the low-affinity FF variation. In conclusion, this is the first study correlating Ca^{2+} signaling and NK cell-mediated ADCC triggered by the four IgG subclasses with the *FCGR3A* V158F SNP. Our findings indicate important differences in the interactions of IgG subclasses with FcγRIIIA/CD16a but no major impact of *FCGR3A* SNP and may therefore help to better correlate the functional properties of particular engineered therapeutic antibodies in vitro with individual differences of their clinical efficacy.

KEYWORDS

ADCC, anti-CD20 antibodies, calcium flux, degranulation, FCGR3A gene polymorphism, IgG subclasses, NK cells, rituximab

1 | INTRODUCTION

NK cell-mediated Ab-dependent cellular cytotoxicity (ADCC) plays a key role in the lysis of IgG-coated/targeted cells and transplantation immunology.[1,2] Lately, ADCC has come under the spotlight because immunotherapies using a variety of mAb are increasingly used to target cancer cells as well as auto- or alloreactive immune cells.[3–5] Recognition of IgG-coated/targeted cells is mainly mediated through the expression of CD16a, the activating Fc-gamma receptor IIIA (FcγRIIIA), on NK cells and to a smaller extent via activating FcγRIIC/CD32c or inhibitory FcγRIIB/CD32b.[6,7] Copy number variations and several single nucleotide polymorphisms (SNP) have been described for the *FCGR3A* gene,[7] the most broadly studied SNP by far is V158F

Abbreviations: ADCC, Ab-dependent cell cytotoxicity; anti-hCD20-IgG, chimeric IgG subclasses antibodies against human CD20; AUC, area under the curve; B_{max}, total number of receptors expressed in the same units; FcγR, Fc-gamma receptor; K_d, dissociation constant; MFIR, geometric mean fluorescence intensity ratio; SNP, single nucleotide polymorphism; SPR, surface plasmon resonance.

Received: 12 February 2020 | Revised: 11 June 2020 | Accepted: 16 June 2020

determining a high-affinity variant of the FcγRIIIA receptor (VV). From the therapeutic perspective, the association between the V158F SNP and the efficacy of mAb therapy, however, remains contradictory.[3,5,8–16]

As a result of the specific characteristics of IgG subclasses including effector functions and half-life,[17,18] most cell-depleting mAb therapies, such as anti-CD20 rituximab, and cytokine-neutralizing mAb, such as anti-TNF infliximab, are IgG1 antibodies.[4,19] For some other indications of immunomodulatory mAb therapy, IgG2 or IgG4 are the subclasses of choice in order to avoid complement activation or FcγRIIIA engagement.[20] Nevertheless, the field of Ab engineering is quickly evolving always aiming to identify the most effective mAb for each indication.[18] The affinity of the different IgG subclasses for FcγRs has been well characterized by previous studies using ELISA and surface plasmon resonance (SPR). These techniques require the attachment of FcγRs or IgG to rigid matrixes.[21–23] In contrast, to our knowledge, only one other study compared IgG subclass-FcγRIIIA/CD16a interactions when the FcγRs were immersed in a lipid bilayer or expressed on cellular membranes.[24]

The increase in intracellular calcium (Ca^{2+}) is often used as a marker for FcγR activation following the binding of antibodies or more efficiently of immune complexes. Ca^{2+} flux has been implicated in the regulation of cell differentiation, gene transcription, and effector functions.[25] In NK cells, engagement of FcγRIIIA/CD16a induces phosphorylation of ITAM motifs in the adaptor chain CD3ζ, which in turn activates tyrosine kinases downstream leading to activation of phospholipase C gamma. Consequently, intracellular Ca^{2+} storages are mobilized from the endoplasmic reticulum (ER) into the cytosol (Ca^{2+} efflux), followed by extracellular Ca^{2+} influx.[26,27] This later step turned out to be fundamental for NK cell functions and plays a critical role in NK cytotoxicity against cancer cells.[27] Rare patients carrying mutations in two different store-operated Ca^{2+} influx channels exhibit impaired target cell-induced lytic granule exocytosis and thus NK cytotoxicity, but also reduced cytokine production.[28] Furthermore, as demonstrated more than 20 yr ago, engagement of the high-affinity VV variant of the FcγRIIIA/CD16a by IgG aggregates leads to a larger increase in Ca^{2+}.[29]

The strength of ADCC mediated by NK cells is the resultant of several factors including Ab subclass and FcγR affinity. Nevertheless, it has been difficult to judge the relevance of each factor when it comes to explaining heterogeneous individual responses to mAb treatment. NK cell-mediated ADCC depends on: (i) receptor-ligand interactions, which are defined by the number/density of antigens present on the target cells, the affinity of IgG for the antigen and FcγR (defined by the IgG subclass and the glycosylation of FcγR and IgG), and the genetic variants of the *FCGR* expressed by NK cells; (ii) triggering of the signaling cascade, given by the strength of IgG-FcγR interactions, the accessibility and availability of adaptor molecules, Ca^{2+} storage release, and intake, and degradation/recycling of IgG-FcγRIIIA/CD16a complexes; (iii) formation of the immune synapses, here the clustering of the FcγR given by the membrane malleability, adhesion molecules, and extracellular Ca^{2+} are important; (iv) the NK cell killing machinery, characterized by the proteolytic enzymes contained in the lytic granules, capability of NK cells to degranulate, which depends on the cytoskeleton, microtubule-organizing center formation, and granule polarization, the fusion of cytolytic granules to cellular membrane, pH, and exhaustion of cytotoxic mechanisms; and, finally, (v) the environmental conditions, including pharmacologic agents such as immunosuppressants, cytokines, and neighboring cells contribute to the outcome.[27,30–35]

In conclusion, the molecular mechanisms regulating NK cell responses upon IgG-FcγRIIIA/CD16a interactions need to be analyzed at different levels. The current study addressed the role of all four IgG subclasses and the V158F gene polymorphism on the binding of the chimeric anti-CD20 mAb rituximab to FcγRIIIA/CD16a on NK cells, the triggering of Ca^{2+} mobilization, degranulation, ADCC, and intracellular cytokine production.

2 | MATERIAL AND METHODS

2.1 | Reagents and antibodies

The complete list of antibodies with their characteristics is given in Table 1. BSA, FBS, ionomycin, EGTA, probenecid, and thapsigargin were from Sigma-Aldrich (St. Louis, MO, USA). The tissue culture media AIM-V and RPMI 1640, the buffers Dulbecco's phosphate-buffered saline (DPBS), HBSS with and without Ca^{2+} and Mg^{2+} and HEPES, and penicillin/streptomycin were all from Gibco (Grand Island, NY, USA); 2 mM L-Alanyl-L-Glutamine was from Bioswisstec (Schaffhausen, Switzerland); and Ficoll-Hypaque Plus was from GE Healthcare (Uppsala, Sweden).

2.2 | NK cell isolation and cell line

Human PBMCs were isolated from buffy coats or peripheral blood obtained from the Blood Transfusion Center of the University Hospitals Geneva (CTS) or healthy volunteers after informed consent according to the local ethical committee (CER13-149 and CER18-00552). For donors' details, check Supporting Information Table S1. PBMCs were isolated by centrifugation using Ficoll-Hypaque Plus followed by negative magnetic purification using the MACS NK isolation kit (Miltenyi) according to the manufacturer's instructions. Purity was 90–95% and was analyzed by flow cytometry (Supporting Information Fig. S1).

The Daudi cell line was purchased from American Type Culture Collection (Manassas, VA, USA) and cultured in RPMI 1640 supplemented with 10% heat-inactivated FBS, 2 mM L-Alanyl-L-Glutamine, 100 U/ml penicillin, and 100 μg/ml streptomycin at 37°C, 5% CO_2.

2.3 | Binding of chimeric anti-human CD20 rituximab Ab subclasses to Daudi cells

BCRs on Daudi cells were blocked by incubation of 1×10^6 cells with blocking anti-h-F(ab')$_2$ fragments at a final concentration of 0.4 mg/ml for 30 min at 4°C. After washing with DPBS-0.1% BSA at

TABLE 1 List of antibodies used in the current study

Ab	Antigen	Host species, isotype	Clone	Fluorochrome	Company
Anti-hCD20-IgG1	h CD20	Chimeric h IgG1, m	NS	–	InVivoGen[a]
Anti-hCD20-IgG2	h CD20	Chimeric h IgG2, m	NS	–	InVivoGen
Anti-hCD20-IgG3	h CD20	Chimeric h IgG3, m	NS	–	InVivoGen
Anti-hCD20-IgG4	h CD20	Chimeric h IgG4, m	NS	–	InVivoGen
Detection anti-h-F(ab′)$_2$	h (Fab′)$_2$ κ	Goat, polyclonal	–	PE	Southern Biotech[b]
Blocking anti-h-F(ab′)$_2$, h-cross-linker Ab	h (Fab′)$_2$	Goat, F(ab′)$_2$ fragment	–	–	Jackson ImmunoResearch[c]
m-cross-linker Ab	m IgG	Goat, polyclonal	polyclonal	–	Southern Biotech
CD3	h CD3	m, IgG1	UCHT1	FITC	Biolegend[d]
CD14	h CD14	m, IgG1	HCD14	FITC	Biolegend
CD16	h CD16	m, IgG1	B73.1	BV605	Biolegend
		m, IgG1	3G8	–	Biolegend
				BV421	Biolegend
				BV605	Biolegend
CD19	h CD19	m, IgG1	HIB19	PE	Biolegend
CD33	h CD33	m, IgG1	HIM3-4	FITC	Biolegend
CD45	h CD45	m, IgG1	5B1	VioGreen	Miltenyi[e]
CD56	h CD56	m, IgG1	HCD56	BV421	Biolegend
				BV605	BioLegend
CD107a	h CD107a	m, IgG1	HA43	PE	BD Pharmingen[f]
IFNγ	hIFNγ	m, IgG1	4S.B3	FITC	Biolegend
GM-CSF	hGM-CSF	r, IgG2a	BVD2-21C11	PerCPCy5.5	Biolegend
TNF	hTNF	m, IgG1	MAb11	BV421	Biolegend
Isotype control	–	m, IgG1	MOPC-21	FITC	BD Pharmingen
				PE	BD Pharmingen
				BV421	BD Pharmingen
				BV605	BD Pharmingen
Isotype control	–	r, IgG2a	RTK2758	PerCPCy5.5	BD Pharmingen
Isotype control	–	Goat polyclonal	–	PE	Southern Biotech

Abbreviations: BV421, brilliant violet 421; BV605, brilliant violet 605; FITC, fluorescein isothiocyanate; h, human; m, mouse; NS, not specified; and r, rat.
[a]San Diego, CA, USA; [b]Birmingham, AL, USA; [c]West Groove, PA, USA; [d]San Diego, CA, USA; [e]Miltenyi, Bergisch Gladbach, Germany; [f]San Diego, CA, USA.

4°C, 5×10^4 cells were incubated with anti-hCD20-IgG Abs of each subclass (InVivoGen) followed by the detection by polyclonal goat anti-h-(Fab′)$_2$ PE, both incubations were performed for 30 min at 4°C. Cells were then analyzed using CytoFLEX (Beckman Coulter, Brea, CA, USA) and a minimum of 20,000 events was acquired. To compare the levels of surface expression, the geometric mean fluorescence intensity ratios (MFIR) were calculated as described earlier[36] and was used to estimate total number of receptors expressed in the same units (B_{max}) and dissociation constant (K_d) using a nonlinear fit for one binding site with GraphPad Prism, version 8.2.0.

2.4 | Binding of chimeric anti-human CD20 Ab subclasses to FcγRIIIA/CD16a

To release naturally bound IgG, NK cells were kept overnight in AIM-V medium supplemented with 2% HEPES at 37°C, 5% CO_2. The release of naturally bound IgG was confirmed by flow cytometry before continuing with the binding assays. Thereafter, NK cells (5×10^4) were incubated with anti-hCD20-IgG Abs of each subclass, washed twice with DPBS-0.1% BSA at 4°C, and labeled with detection anti-h-(Fab′)$_2$ PE as described earlier for Daudi cells. Cells were analyzed using an Attune flow cytometer (Life Technologies, Eugene, OR, USA) and a minimum of 20,000 events was acquired. Binding of anti-hCD20-IgG subclasses to NK cells was analyzed by gating on the CD56$^+$ live cell population (Supporting Information Fig. S2A).

2.5 | CD107a degranulation assay and intracellular cytokine detection

Degranulation and intracellular cytokines were analyzed by CD107a surface expression in 6 h assays using CD56$^+$ NK cells according to Bryceson et al.[37] with some modifications. Different concentrations

of anti-hCD20-IgG subclasses (0.125; 0.5 and 2 μg/ml) were added to Daudi cells in AIM-V + 2% HEPES; 2×10^5 cells/well were seeded in 96-well plates, followed by the addition of freshly purified NK cells and anti-CD107a mAb or isotype-matched control Ab at an effector to target ratio of 1:1. After 1 h, GolgiPLUG-Brefeldin (BD Biosciences, La Jolla, CA, USA) were added and cocultures maintained for an additional 5 h. For intracellular cytokine determination, cells were first stained for anti-CD56 followed by anti-IFNγ, anti-TNF, and anti-GM-CSF antibodies using the fixation/permeabilization solution kit (BD Biosciences) according to the manufacturer's instructions. Before the addition of anti-cytokine antibodies, cells were blocked with 2% mouse immunoglobulins. Cells were analyzed using Attune flow cytometer. Cytotoxicity controls consisted of Daudi cells without anti-hCD20-IgG subclasses and mAb isotypes. The gating strategy is depicted in Supporting Information Figure S3.

2.6 | Ca^{2+} mobilization assay

Freshly isolated NK cells (4.8×10^6/ml) were incubated with anti-hCD20-IgG subclasses and anti-CD16 mAb at 100 and 10 μg/ml, respectively, in HBSS supplemented with 0.2% BSA and 20 mM HEPES, pH 7.4 for 45 min at 4°C. Next, NK cells were labeled by adding FLIPR calcium 6-QF kit (Molecular Probes, Eugene, OR, USA) supplemented with 2.5 mM probenecid; 25 μl of labeled cells (25,000 cells) were transferred to a 384-well flat clear bottom plate (Corning, Corning, NY, USA) in quadruplicate, and incubated for 1 h at 37°C and 5% CO_2. The concentrations of IgG subclasses before and after adding cross-linker were of 17.4 and 15 μg/ml, respectively. For each set of experiments, either HBSS plus 1.3 mM EGTA (no Ca^{2+} condition) or HBSS Ca^{2+}/Mg^{2+} buffer (Ca^{2+} condition) was assayed. The plates were centrifuged at 100 × *g* for 3 min to pull down the cells. For each experiment, two reagent-supplying plates were prepared for automatic loading by the FDSS/μCELL plate reader (Functional Drug Screening System/Microcell, Hamamatsu Photonics, Shizuoka, Japan). The first plate contained h-cross-linker Ab, m-cross-linker Ab, ionomycin, and thapsigargin to reach final concentrations of 20 μg/mL, 0.11 μg/mL, 2 μg/mL, and 2 μM, respectively. The second plate contained Ca^{2+} buffer to reach a 2 mM final concentration. After establishing a baseline for 30 s, reagents from the first plate were added to the correspondent wells and recorded for 10 min, followed by the addition of the Ca^{2+} buffer and recorded for another 10 min. Results are expressed as the ratio (F/F_0) vs. time, where F is the intensity of fluorescence emission recorded as the experiment runs and F_0 the fluorescence intensity at the beginning of the experiment. To combine several experiments, two different Ca^{2+} flux measurements were employed: (i) Ca^{2+} efflux from the ER (0–630 s), estimated by two parameters: maximum peak (Max peak, F/F_0 ratio) obtained after 30 s reagent addition, and the time when the Max peak was reached (s). (ii) Ca^{2+} influx to the cell (630–1230 s) was estimated by the Max peak measured after the addition of Ca^{2+}, and Ca^{2+} influx rate at 20 s measured after extra Ca^{2+} addition.

2.7 | Ab-dependent cellular cytotoxicity assay

Nonradioactive DELFIA Europium^{3+} chelate of 2,2′:6′2′′-terpyridine-6,6′′-dicarboxylic acid cytotoxicity reagents (PerkinElmer, Waltham, MA, USA) was used for NK cell-mediated cytotoxicity assays according to the manufacturer's instructions and as previously described.[32] Briefly, Daudi target cells were labeled with bis(acetoxymethyl)2,2′:6′2′′-terpyridine-6,6′′-dicarboxylate , washed with DPBS supplemented with 20 mM HEPES and 2.5 mM probenecid, suspended in RPMI 1640, and anti-hCD20-IgG subclasses were added at the indicated concentrations (0.0078-2 μg/ml). Immediately thereafter, freshly isolated NK cells (1×10^5 cells per well) were added, at an effector to target ratio of 5:1. The 100 μL coculture was performed at 37°C for 2 h in a 96-well U-bottom plate. Controls included direct cytotoxicity where Daudi cells and NK cells were cocultured in medium without antibodies. Maximum release was obtained using labeled Daudi cells lysed with either 10% Triton-X 100 (Sigma-Aldrich) or Lysis Buffer (PerkinElmer). Cytotoxicity was evaluated by the release of fluorescence of coculture supernatants and measure in a time-resolved fluorometer (EnVision 2014 Multilabel reader, PerkinElmer). Cytotoxicity was expressed as a percentage (%) of specific lysis.

2.8 | FCGR3A V158F polymorphism

DNA was extracted from total blood using QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany). The *FCGR3A* V158F allelic variations were determined by nested-PCR and restriction enzyme digestion assay using 2 ng of genomic DNA as previously described.[38] The products were then run in 10% acrylamide gels stained with GelRed (Biotium, Fremont, CA, USA) according to the manufacturer's instructions.

2.9 | Statistical analysis

For data consisting of multiple repeats, mean ± SD is shown. The comparison of differences among multiple groups was performed by 2-way ANOVA with Tukey posttest correction. t-student was used to compare the differences ADCC between donors grouped as VV + VF versus FF. Differences were considered statistically significant for *P*-values less than 0.05. Statistical values as well as area under the curve (AUC) and EC_{50} were calculated using GraphPad Prism.

3 | RESULTS

3.1 | Variable regions of rituximab anti-human CD20 IgG subclasses bind equally to the CD20 antigen expressed on Daudi cells

Before analyzing interactions between rituximab anti-hCD20 IgG subclasses and the IgG-FcγRIIIA/CD16a, binding to the antigen CD20 on Daudi cells was assayed in dose-response curves (0.0078–16 μg/ml) by flow cytometry. Bound anti-hCD20-IgG was measured

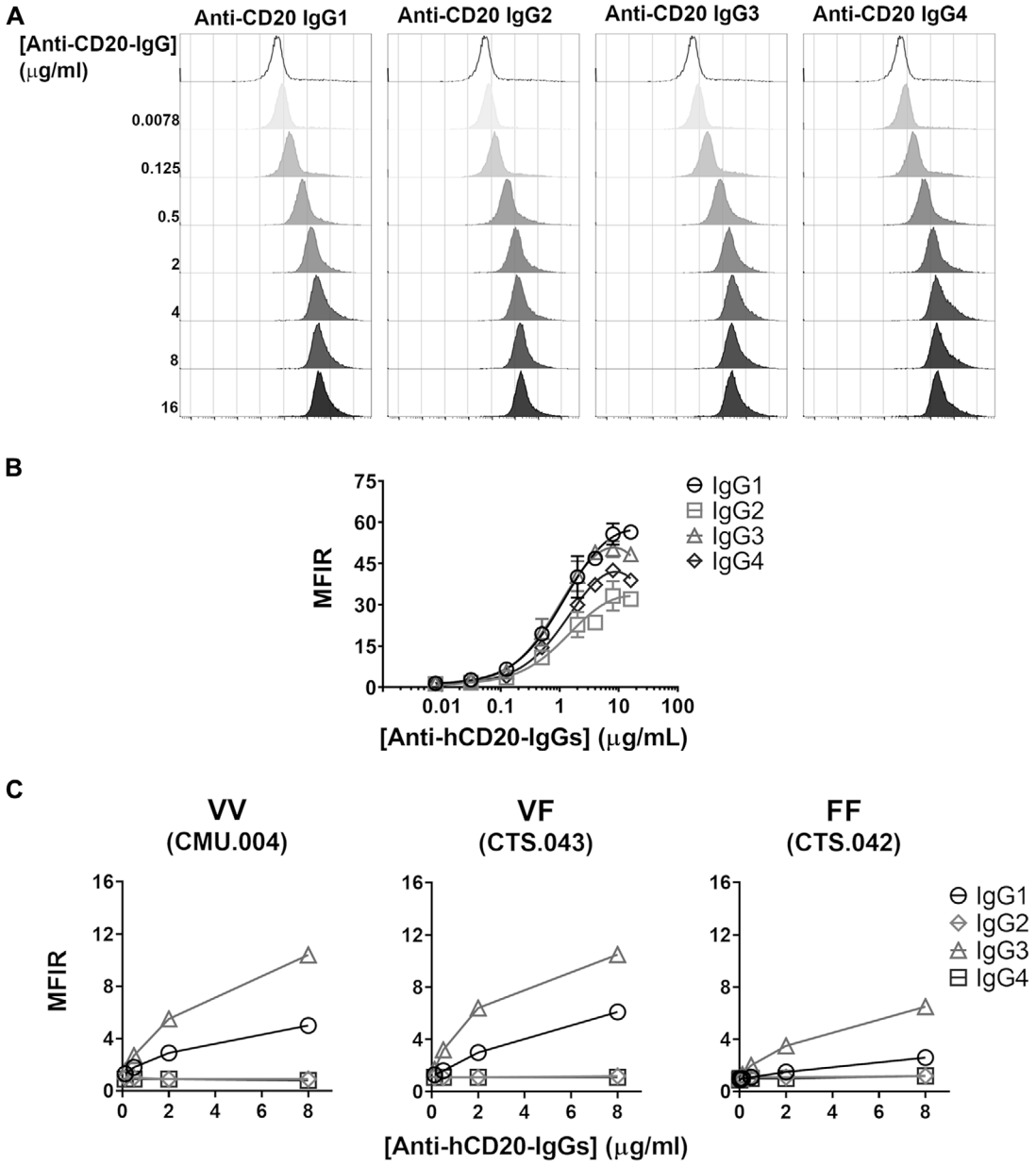


FIGURE 1 **Binding of anti-hCD20-IgG rituximab subclasses to CD20 and Fc-gamma receptor IIIA (FcγRIIIA)/CD16a. A** and **B**. Binding of anti-hCD20-IgG subclasses to CD20 expressed on Daudi cells. After blocking of BCRs with anti-h-F(ab′)$_2$fragments, Daudi cells were incubated with increasing concentrations of each subclass (0.0078–16 μg/ml) plus the detection anti-h-F(ab′)$_2$-PE polyclonal goat Ab and then analyzed by flow cytometry. **A**, Representative histogram showing the recognition of CD20 by anti-CD20-IgG1, -IgG2, -IgG3, and -IgG4 at different concentrations. **B**, Dose-response curve showing the binding of the different anti-hCD20 IgG subclasses to the CD20 epitope in flow cytometry. Mean ± SD of the mean intensity ratio (MFIR) from three independent experiments are plotted. This curve was used to determine the K_d and B_{max} with GraphPad Prism (not indicated). Anti-hCD20-IgG1, -IgG2, -IgG3, and -IgG4 are represented as circle, square, triangle, and diamond, respectively. **C** and **D**, Binding of anti-hCD20-IgG subclasses to FcγRIIIA/CD16a in donors with different V158F *FCGR3A* single nucleotide polymorphism variations. NK cells were cultured overnight in AIM-V HEPES to release naturally bound IgG. After washing, NK cells were incubated separately with anti-hCD20-IgG subclasses at increasing concentrations (0.0078–8 μg/ml), followed by detection anti-h-F(ab′)$_2$-PE polyclonal goat Ab staining, and analysis by flow cytometry. **C**, Dose-response curves showing the affinity of the four anti-hCD20 IgG subclasses to *FCGR3A* VV (left), VF (center), and FF (right) donors, ($n = 1$ donor for each genotype).

(Continued on next page)

using detection anti-h-F(ab′)$_2$ PE Ab and expressed as MFIR. The binding of anti-hCD20-IgG subclasses to the CD20 antigen was similar for IgG1, IgG2, IgG3, and IgG4 at concentrations lower than 1 μg/ml of Ab (Fig. 1A, B). The dose-binding curves were then used to determine the K_d that were in the micromolar range (8.7 ± 2.6, 10.6 ± 5.0, 6.9 ± 1.7, and 12.1 ± 1.8 μM, respectively), and the number of binding sites (B_{max}), which were also in the same order of magnitude of MFIR, 64.2 ± 8.2; 38.8 ± 8.4; 65.51 ± 6.45; and 59.0 ± 4.3, for

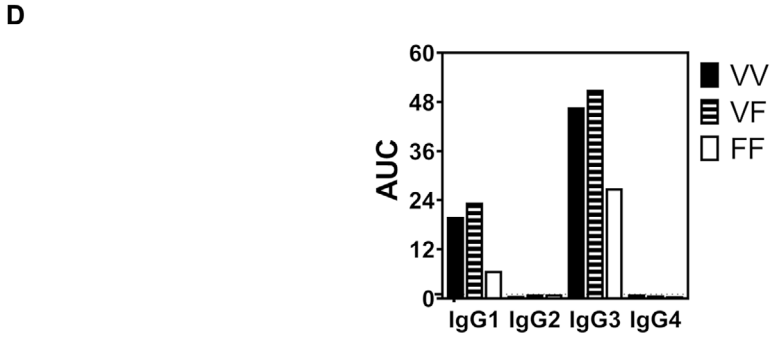


FIGURE 1 (Continued) **D**, The affinity of each rituximab anti-hCD20-IgG subclass was compared using the area under the curve (AUC) of the dose-response for each donor. Data of VV and FF homozygosity, and VF heterozygosity, are shown as black, white, and striped bars, respectively

anti-hCD20-IgG1, -IgG2, -IgG3 ,and -IgG4, respectively. In summary, rituximab anti-hCD20-IgG subclasses mAb exhibited similar binding to the CD20 antigen expressed on the surface of Daudi cells, and thus were considered suitable reagents for the following experiments.

3.2 | Differential binding of the constant region of rituximab anti-CD20 IgG subclasses to the FcγRIIIA/CD16a on NK cells

The binding of rituximab anti-hCD20-IgG subclasses to FcγRIIIA/CD16a was tested using NK cells isolated from donors carrying different V158F polymorphisms (Supporting Information Fig. S4). To avoid competition with naturally bound IgG to FcγRIIIA/CD16a during the assay,[39] NK cells were first cultured overnight in media deprived of any source of IgG to release naturally bound IgG. Minimal binding was identified the next day before running the binding experiments (Supporting Information Fig. S2B). Strong binding of anti-hCD20-IgG3 and -IgG1 to NK cells was found, whereas -IgG2 and -IgG4 were below the detection limit of the assay giving MFIR values of 1.0 irrespective of the *FCGR3A* genotype. Donors carrying the VV or VF variant of the *FCGR3A* SNP V158F showed two threefold higher binding of anti-hCD20-IgG1 and -IgG3 than the FF carriers as previously described (K_d of VV 19.9 and 46.8; VF 23.4 and 50.9 vs. FF 6.8; and 26.9 for IgG1 and IgG3, respectively; Fig. 1C, D). In summary, the binding of anti-hCD20-Ig subclasses to FcγRIIIA/CD16a on NK cells depends on the IgG subclass (IgG3 > IgG1, no binding of IgG2/4) and *FCGR3A* SNP V158F (VV and VF > FF).

3.3 | Binding of IgG1, IgG3, and IgG4, but not of IgG2 to FcγRIIIA/CD16a induces calcium mobilization in NK cells

The engagement of IgG FcγRIIIA/CD16a signals via Ca^{2+} dependent pathways. It was previously shown that IgG aggregates induce greater mobilization of Ca^{2+} in NK cells expressing high-affinity *FCGR3A* V158F SNP (VV).[29,40] However, the impact of IgG subclasses on Ca^{2+} signaling remains unknown. Thus, we analyzed Ca^{2+} mobilization in NK cells upon binding and crosslinking of all four anti-hCD20-IgG subclasses to FcγRIIIA/CD16a. NK cells were freshly isolated from VV, VF, and FF *FCGR3A* V158F SNP donors. Cytosolic Ca^{2+} flux was barely detectable following cross-linking of anti-hCD20-IgG2, independently of the donor genotype (Fig. 2, Supporting Information Fig. S5B). In the absence of extracellular Ca^{2+}, and compared to human cross-linker, a modest Ca^{2+} efflux from the ER was detected following cross-linking of anti-hCD20-IgG1 and -IgG4 (Fig. 2A). However, upon addition of extracellular Ca^{2+} a strong store-operated Ca^{2+} entry (Ca^{2+} influx) into NK cells was observed following cross-linking of anti-hCD20-IgG1, -IgG3, and to a lower extent of anti-hCD20-IgG4. Suitable experimental controls for each experimental condition including registrations in the presence of Ca^{2+} at physiologic concentrations are shown in Supporting Information Figure S5. When the results obtained from eight different donors were pooled (Fig. 2B), Ca^{2+} efflux was similar for anti-hCD20-IgG1, -IgG4, and the CD16 control; however, anti-hCD20-IgG2 and -IgG3 hardly mobilized Ca^{2+} (compared to human cross-linker). In fact, Ca^{2+} efflux was not detectable for most of the donors at any of the time points after the cross-linking of anti-hCD20-IgG2. Finally, the response induced by the cross-linking of CD16 (positive control) reached a close maximum response after 91.98 ± 11.20 s showing little deviation for the eight donors, thus supporting the notion that the experimental conditions were similar (Fig. 2B). On the other hand, in the presence of external Ca^{2+}, the peak of Ca^{2+} influx was the lowest for anti-hCD20-IgG2 (1.29 ± 0.15) compared to the other subclasses and the human cross linker (1.62 ± 0.19, 1.59 ± 0.17, 1.51 ± 0.16, 1.40 ± 0.15 for anti-hCD20-IgG1, -IgG3, -IgG4, and the human cross linker alone, respectively); and statistically significant differences were seen for anti-hCD20-IgG1, -IgG3, and -IgG4 ($P = 0.0238$, 0.0023 and 0.0455, respectively). Analyzing the Ca^{2+} rate 20 s after the addition of Ca^{2+}, in NK cells exposed to the anti-hCD20-IgG2 cross-linking the mean was lower than cross-linker alone (0.0556 ± 0.0067 vs. 0.0625 ± 0.0039, respectively) (Fig. 2C). Anti-hCD20-IgG3 was able to induce Ca^{2+} influx. The mean of maximum peak was similar to anti-hCD20-IgG4 (1.587 ± 0.1746 vs. 1.511 ± 0.1567 respectively) (Fig. 2C). In summary, anti-hCD20-IgG2 did not induce Ca^{2+} efflux from the ER or Ca^{2+} influx. IgG3 induced Ca^{2+} influx but no Ca^{2+} efflux, whereas IgG1 and IgG4 induced both significant Ca^{2+} efflux and Ca^{2+} influx in NK cells. As to the *FCGR3A* V158F SNP, there were no significant

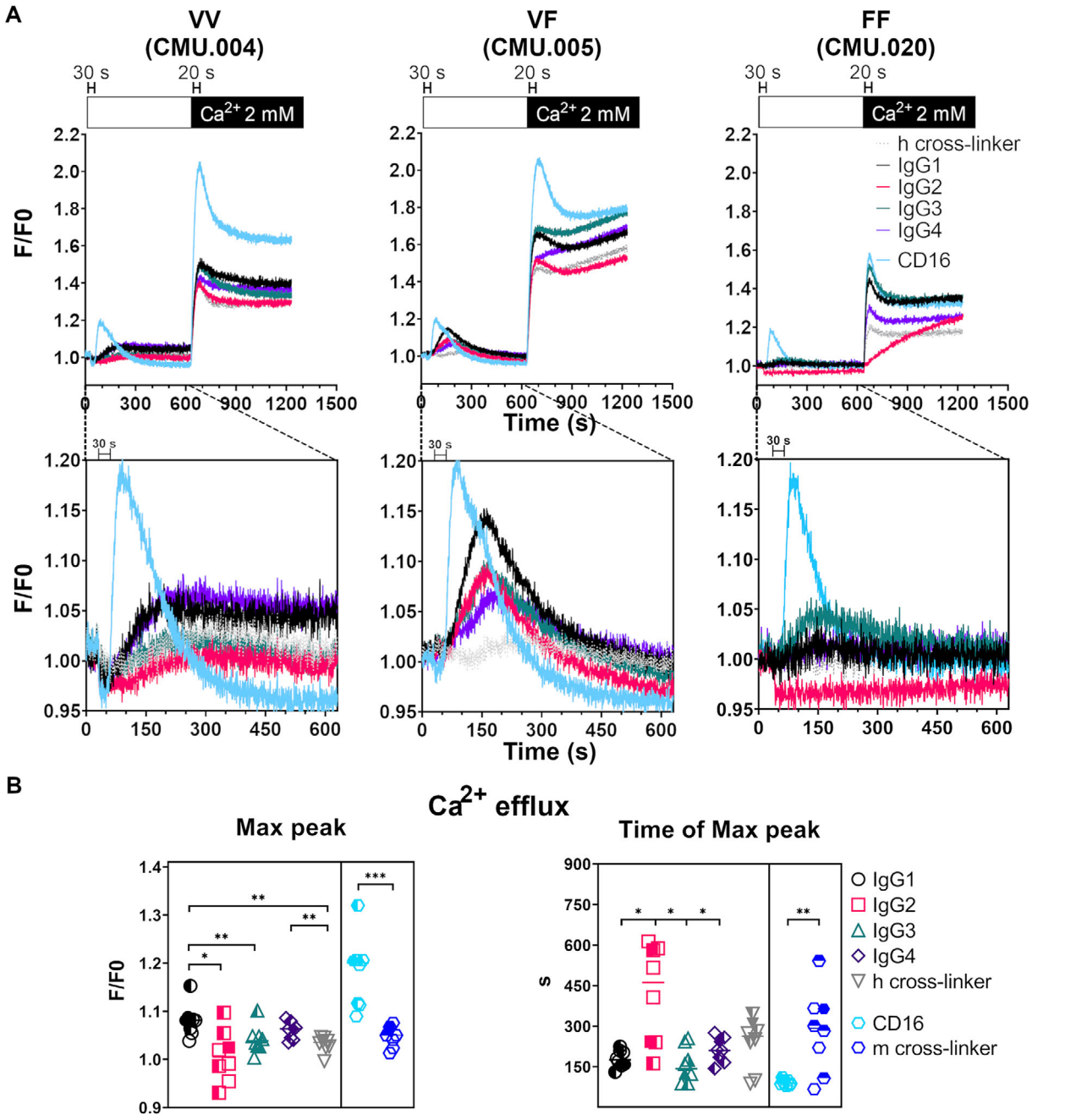


FIGURE 2 **Anti-hCD20-IgG rituximab subclasses induce Ca^{2+} mobilization in NK cells.** Calcium 6-QF-labeled NK cells were either pre-incubated with 1.5 μg/ml anti-CD16 mAb (positive control), 15 μg/ml anti-hCD20-IgG subclass, and loaded to the FDSS/μCELL plate reader. For each set of experiments HBSS plus 1.3 mM EGTA (no Ca^{2+} condition, white bar) was assayed; specific reagents were added at 60 or 630 s, respectively. After establishing a baseline for 30 s, h cross-linker Ab was added to reach a final concentration of 20 μg/mL. At 630 s, Ca^{2+} buffer was added to reach a final concentration of 2 mM (black bar) and Ca^{2+} mobilization was recorded for an additional 600 s. Results are expressed as the ratio of flux (F/F_0) vs. time (s), representing the average of quadruplicates obtained from eight different donors (one VV, four VF, and three FF). **A**, Traces show Ca^{2+} efflux from 30 to 630 s, indicated by a white bar and magnified in the lower part of the panel; and Ca^{2+} influx (630–1230 s, black bar) for all anti-hCD20-IgG subclasses (anti-CD20-IgG1, -IgG2, -IgG3, and -IgG4 in black, pink, green, and purple traces, respectively), anti-CD16 (light blue), and h cross-linker Ab (dashed gray traces). **B**, Plots show pooled data of the means of the maximum peak (Max peak) of Ca^{2+} efflux (F/F_0), measured between 60 and 630 s (left panel); and the time evolved to the Max peak value (s, right panel).

(Continued on next page)

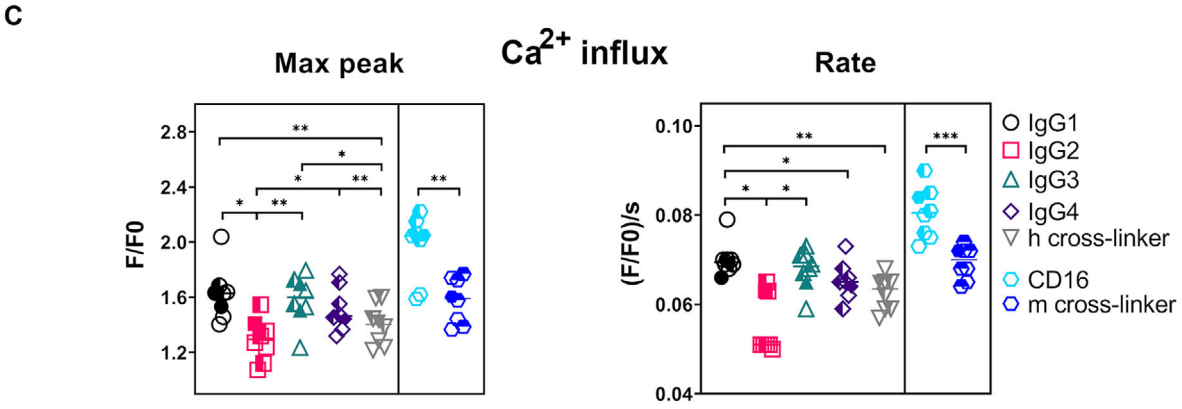


FIGURE 2 (Continued) **C**, Plots show pooled data of the means of the Max peak of Ca^{2+} influx between 630 and 1230 s (left) and the Ca^{2+} influx rate 20 s after the addition of Ca^{2+} (right). Results in **B** and **C** were plotted for each anti-hCD20-IgG subclass (-IgG1, -IgG2, -IgG3, and -IgG4 symbolized as a black circle, pink square, green triangle, and purple diamond, respectively), anti-CD16 (light blue hexagon) for the eight donors grouped by V158F *FCGR3A* polymorphism. VV and FF homozygosity and VF heterozygosity are shown as filled, open, and half-open symbols, respectively. The differences between the anti-hCD20-IgG subclasses were compared using Tukey's multiple comparisons test; differences between the IgG subclasses that were found statistically significant are presented in the figure. $^{*}P \leq 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$. No significant difference was found among donor's genotype

differences ($P = 0.1396$) among the eight (1 VV, 3 FF, and 4 VF) tested donors (filled, open, and half-filled symbols in Fig. 2B, C). However, the number of donors tested per subgroup was too low to draw firm conclusions.

3.4 | IgG1 and IgG3 subclasses and to a minor extent IgG4 trigger ADCC by degranulation in NK cells

Most reports showing that ADCC depends on the V158F polymorphism were using only one type of IgG subclass.[41–45] This study went further and investigated the differences between all four different subclasses directed against the same epitope and the effect of the most common SNP for *FCGR3A*, V158F.

First, we measured the degranulation marker CD107a after coculture of NK cells and Daudi cells coated with anti-hCD20-IgG subclasses. In the absence of anti-CD20, NK cell degranulation as the result of direct NK cytotoxicity against Daudi target cells was 8.7 ± 5.1%. Addition of increasing amounts of anti-hCD20-IgG1, -IgG3, -IgG4, but not -IgG2 led to increasing percentages of $CD56^{+}CD107a^{+}$ NK cells, as shown in the representative example in Figure 3A. Dose-response curves for the percentage of positive cells for four donors (2 VF and 2 FF genotype) are shown (Fig. 3B). For comparisons, the AUC for the dose-response plots was calculated for each donor. Pooled results of five individuals showed similar results, the higher the concentration of anti-hCD20-IgG1 -IgG3, and -IgG4, the higher the NK degranulation, whereas anti-hCD20-IgG2 did not induce degranulation (Fig. 3C, Supporting Information Fig. S6).

Complementarily, ADCC assays were performed at a fixed E:T ratio of 5:1 using the same donor and increasing concentrations of all four anti-hCD20-IgG subclasses simultaneously. In accordance with the findings from Ca^{2+} mobilization and degranulation, the dose-response curves demonstrate that ADCC was triggered by even very low concentrations of anti-hCD20-IgG1 and -IgG3. In contrast, anti-hCD20-IgG4 triggered ADCC only at the highest concentrations, and IgG2 was not able to induce ADCC at all (Fig. 4, Supporting Information Fig. S7). Remarkably, when donors were grouped by the presence of valine (VV + VF) and compared to FF donors, the VV + VF group displayed higher ADCC for anti-hCD20-IgG1 and -IgG3. In addition, anti-hCD20-IgG4 showed a trend to trigger ADCC at the highest concentration (2 μg/ml) in the group containing valine ($P = 0.0555$, *t*-test) (Fig. 4B). Again, to compare different donors, the AUC for the dose-response plots was calculated for each donor (Fig. 4C). The mean of the AUC for anti-hCD20-IgG3, -IgG1, -IgG4, and -IgG2 in the pooled population were 119.1 ± 31.6, 114.9 ± 36.5, 36.0.1 ± 16.6, and 6.6 ± 5.2, respectively. Intriguingly, VF donors overall showed somewhat higher ADCC than VV or FF individuals for anti-hCD20-IgG1 (AUC 156.8 ± 45.2 vs. 98.5 ± 35.8 and 89.5 ± 25.5; $P = 0.0553$ and 0.0023, respectively), and IgG3 (AUC 153.3 ± 32.8 vs. 113.0 ± 75.5 and 91.0 ± 13.0; $P = 0.2351$ and 0.0048). Taken together, the observed variations of the AUC were linked to both IgG subclasses and genotype ($P < 0.0001$ and 0.0484, respectively; subclass × genotype $P = 0.0355$).

To conclude, anti-hCD20-IgG1 and -IgG3 efficiently trigger ADCC and degranulation, whereas -IgG4 did it only at the highest micromolar concentration tested in individuals containing valine. Anti-hCD20-IgG2 did not induce ADCC.

3.5 | Intracellular cytokine production of IFNγ, TNF, and GM-CSF is triggered by anti-hCD20-IgG1 and -IgG3 subclasses during ADCC

Cytokine production by NK cells has been reported after activation of several receptors.[46] Thus, we interrogated whether anti-hCD20-IgG subclasses also differentially affect IFNγ, TNF, and GM-CSF production by NK cells. Altogether, TNF was the major intracellular cytokine

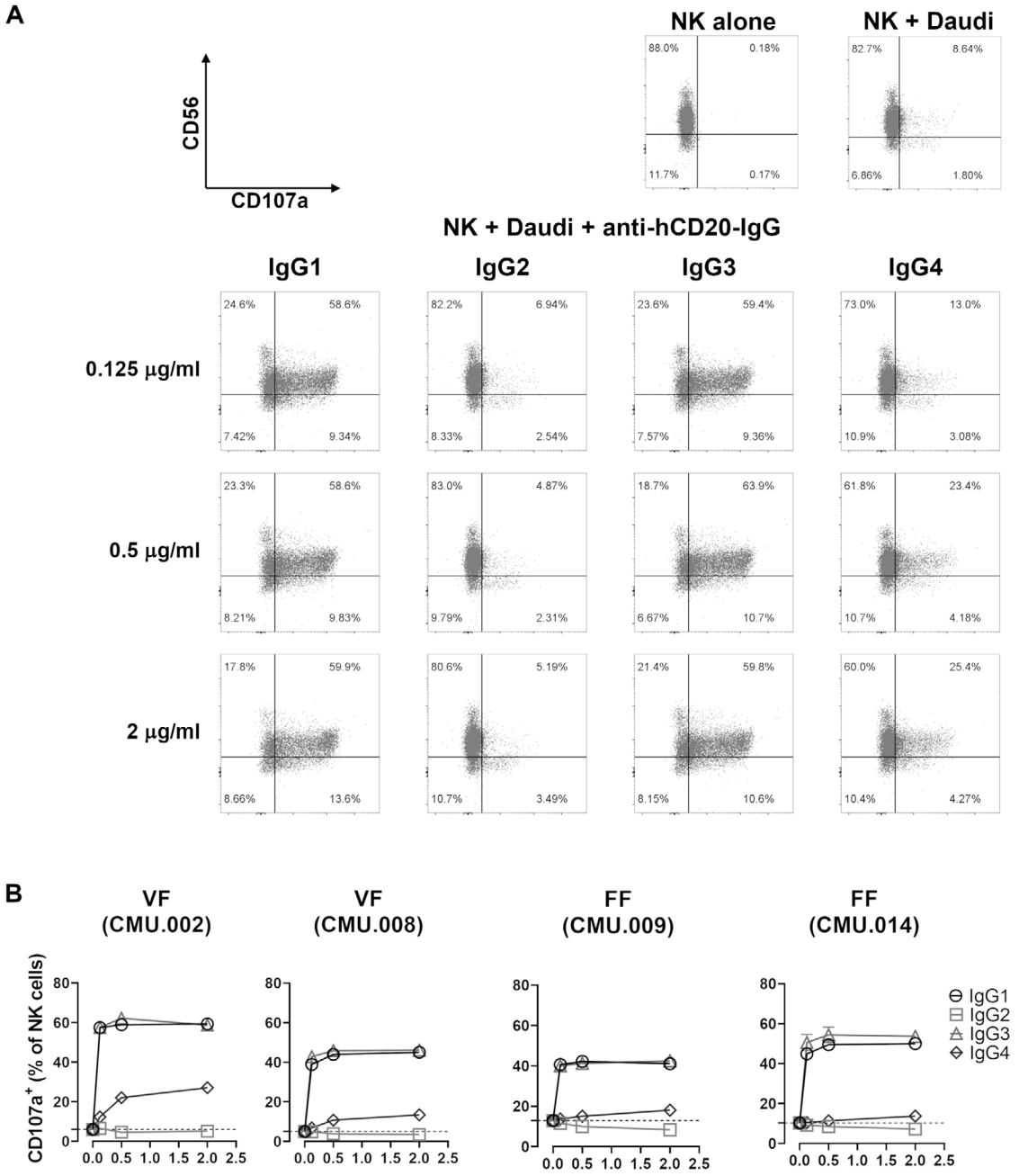


FIGURE 3 **Degranulation of NK cells is altered in the presence of the IgG subclasses.** Freshly isolated NK cells were incubated with Daudi target cells in the presence of anti-hCD20-IgG subclasses for 6 h and stained for CD56 and CD107a. **A**, Representative plots for a VF donor (CMU 0.002) are shown for three different concentrations of anti-hCD20-IgG subclasses. The two plots on the top show control for degranulation values when NK cells were cultured alone (top left), or with Daudi cells (top right) in the absence of anti-CD20-IgG. The remaining panel shows the percentage of $CD56^+CD107a^+$ NK cells (upper right quadrant) in the presence of increasing concentrations of anti-hCD20-IgG1, IgG2, IgG3, and IgG4 subclasses, respectively. **B**, Representative plots of dose response for CD107a degranulation using NK cells isolated from 2 healthy *FCGR3A* VF (left), and 2 FF (right) donors, and anti-CD20-coated Daudi cells as targets are shown. Daudi cells were coated with anti-hCD20-IgG1 (circle), IgG2 (square), IgG3 (triangle), or IgG4 (diamond) subclasses at increasing concentrations (0–2 μg/ml). Subsequently, anti-hCD20-coated Daudi cells were cocultured with freshly isolated NK cells for 6 h at an E:T ratio of 1:1. The dashed line shows degranulation induced by Daudi cells in the absence of anti-hCD20-IgG. The experiments were performed in duplicates and the data are shown as mean ± SD. **C**, Pooled data of NK cell degranulation comparing the AUC of the percentage of anti-hCD20-IgG subclass induced $CD56^+CD107a^+$ NK cells of each donor, using as baseline the percentage of $CD56^+CD107a^+$ NK cells coculture with Daudi cells alone. Results were plotted for each IgG subclass (IgG1, IgG2, IgG3, and IgG4 depicted as circle, square, triangle, and diamond, respectively). Donors with FF homozygosity ($n = 2$), and VF heterozygosity ($n = 3$), are shown as open and half-filled symbols, respectively.

(Continued on next page)

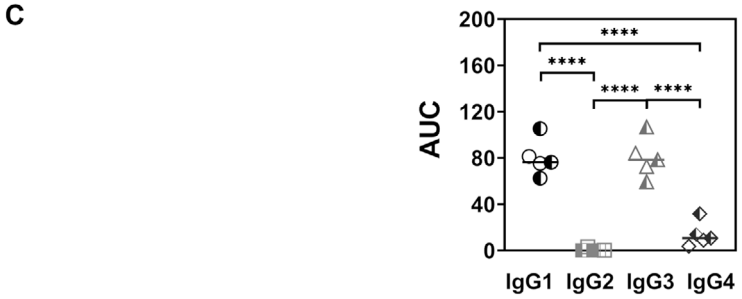


FIGURE 3 (Continued) The difference between the mean of anti-hCD20 IgG subclasses was compared using Tukey's multiple comparisons test; statistical significance was found between all IgG subclasses as detailed in the figure. $^{*}P \leq 0.05$, $^{**}P \leq 0.001$. No significant difference was found between the donors' genotype. Data of one experiment performed in duplicates for each donor are shown

detected, followed by IFNγ, and to a much lesser extent GM-CSF. Moreover, a large diversity of the cytokine response was observed among the four donors (Fig. 5). Finally, a dose-response effect for IgG1 and IgG3 anti-hCD20-IgG subclasses was found, whereas IgG2 and IgG4 induced no cytokine response when compared to the basal level present in NK cells stimulated only with Daudi cells.

4 | DISCUSSION

ADCC represents one of the major functions of NK cells that are increasingly explored for the treatment of autoimmune diseases and cancer using depleting therapeutic mAb. However, despite a plethora of studies and engineering to improve Ab-mediated cytotoxicity, several aspects of the interactions of IgG subclasses with FcγRIIIA/CD16a remain unknown. Here, we compared the ability of all four rituximab anti-hCD20 IgG subclasses, sharing a close affinity for CD20, to trigger Ca^{2+} signaling, cytokine production, and NK cell-mediated ADCC. In addition, we studied whether the *FCGR3A* V158F polymorphism plays a role in the interactions with the different IgG subclasses and consequently, NK cell effector functions. We found that IgG1 and IgG3, and to a lesser degree IgG4 at high concentrations, are able to induce NK cell effector functions, but not IgG2. In contrast, variations of *FCGR3A* SNP had no impact on Ca^{2+} signaling and degranulation; however, ADCC was somewhat weaker in donors with the low-affinity FF variation.

Flow cytometry binding assays showed a slightly lower affinity of anti-hCD20-IgG2 for the CD20 antigen expressed on Daudi cells, especially at higher concentrations. This finding might be explained by the rigidity of the IgG2 hinge region, which could interfere with the binding to the antigen, as discussed in detail by Roux et al.[47] Nevertheless, the affinity of all subclasses were in the same micromolar range and thus considered as similar. Binding of IgG subclasses to FcγRIIIA/CD16a has been analyzed by several previous studies using ELISA or SPR. In contrast, NK cells freshly isolated from healthy donors carrying different genotypes were used in the present study to measure the differential binding of IgG subclasses to FcγRIIIA/CD16a by flow cytometry and to test the impact of the V158F *FCGR3A* SNP. Our results confirm the relative strong binding of anti-hCD20-IgG3 and -IgG1, and that the binding was weaker on NK cells carrying the low-affinity *FCGR3A* variant (FF). However, using this approach we could not detect significant binding of IgG4 or IgG2, as previously reported by Armour et al. for anti-D antibodies and myeloma-derived IgG using CHO cells transfected with the VV and FF *FCGR3A* variants.[24] These differences might be explained by the different affinities of the mAb used and/or a larger range of concentrations evaluated. Furthermore, Repp et al. reported binding of IgG1 single-chain Fv-Fc-engineered fusion proteins to FcγR on NK cells by flow cytometry, protein- and glyco-engineering enhanced CD16a affinity but beyond a certain threshold did not result in a further increase of NK-cell mediated ADCC. However, the latter study did not test the other IgG subclasses.[48]

Until now, the majority of Ca^{2+} mobilization studies in lymphoid cells have focused on B and T lymphocytes. Nevertheless, cross-linking of FcγRIIIa/CD16a mobilizes large amounts of Ca^{2+} that initiate NK cell-mediated ADCC.[49–51] More importantly, recent findings showed the relevance of store-operated Ca^{2+} entry (Ca^{2+} influx) into NK cells in patients with an immunodeficiency syndrome associated with autoimmunity.[52] Patients with chronic fatigue syndrome/myalgic encephalomyelitis showed abnormalities in NK cells activity and a particular Ca^{2+} channel affecting the NK cells' activation threshold.[53,54] In the current study, Ca^{2+} mobilization in NK cells was explored for the first time by distinguishing between efflux and influx following IgG-FcγRIIIA/CD16a interaction. To detect the Ca^{2+} efflux from the ER we performed the determinations in the absence of extracellular Ca^{2+}. Controls using physiologic Ca^{2+} conditions were run in parallel and showed similar Ca^{2+} treats. Significant amounts of Ca^{2+} efflux were triggered by IgG1 and IgG4, whereas Ca^{2+} influx was triggered by anti-hCD20-IgG1, -IgG3, and -IgG4. On the other hand, IgG2 results were consistently negative in all the assays performed in this study, with the exception of the binding to CD20 on Daudi cells. We can only speculate why anti-hCD20-IgG3 did not induce Ca^{2+} efflux, but significant Ca^{2+} influx, similar as anti-hCD20-IgG4 (1.587 ± 0.1746 vs. 1.511 ± 0.1567, respectively), and strong ADCC. IgG3 might induce ADCC via a signaling cascades independent of Ca^{2+}; or alternatively the extremely flexible structure of IgG3 with the highest Fab-Fab and Fab-Fc bending[47]

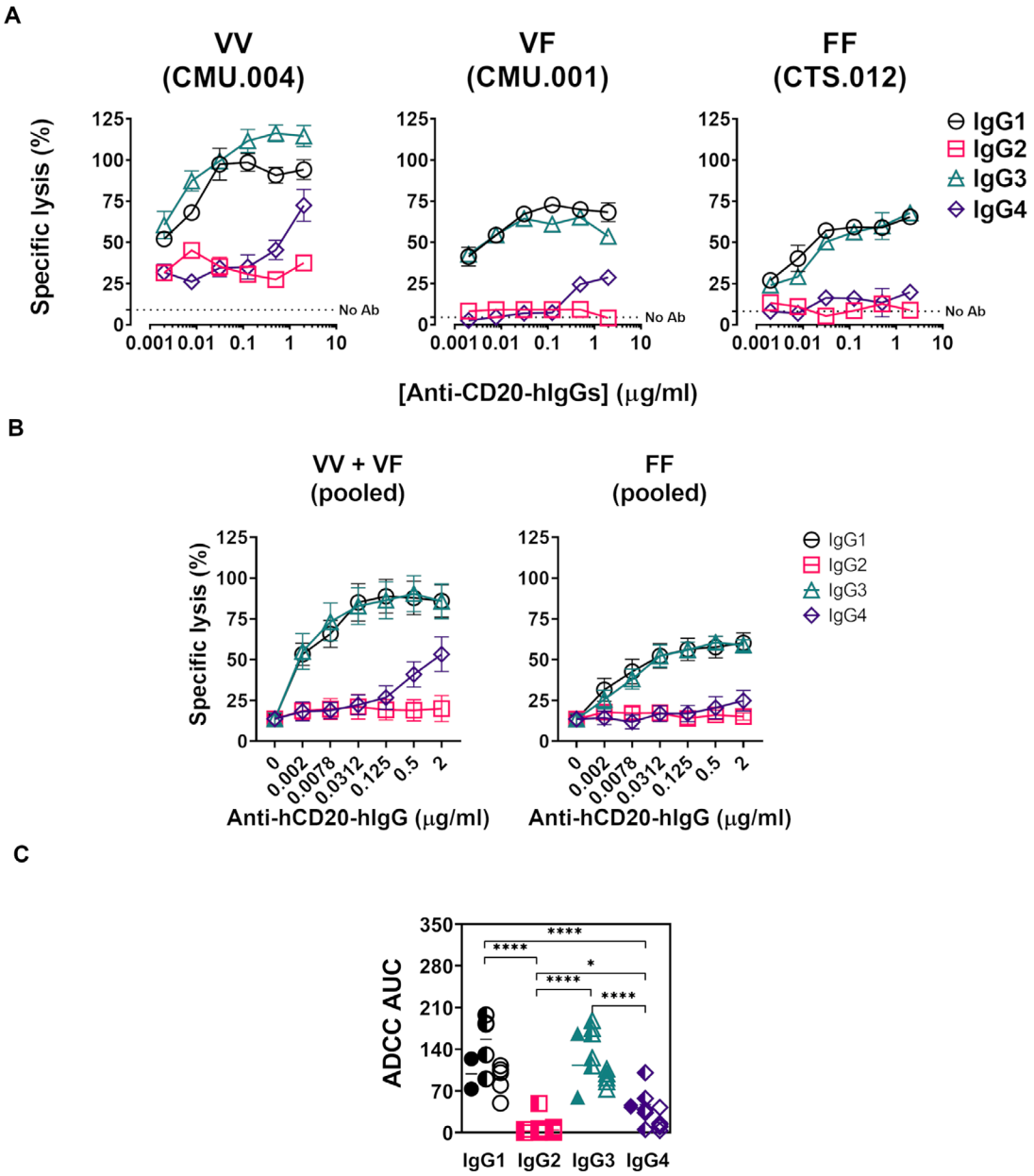


FIGURE 4 **Anti-hCD20-IgG rituximab subclasses and V158F *FCGR3A* polymorphism induce different levels of Ab-dependent cellular cytotoxicity (ADCC) in NK cells. A**, Representative plots of ADCC using NK cells isolated from healthy *FCGR3A* VV (left), VF (center), and FF (right) donors, and anti-CD20-coated Daudi cells as targets are shown. Nonradioactive labelled Daudi cells were coated with anti-hCD20-IgG1 (black circle), IgG2 (pink square), IgG3 (green triangle), or IgG4 (purple diamond) subclasses at increasing concentrations (0.0078–2 μg/ml). Subsequently, anti-hCD20-coated Daudi cells were cocultured with freshly isolated NK cells for 2 h at an E:T ratio of 5:1. The dashed line shows the direct cytotoxicity of Daudi cells in the absence of anti-hCD20-IgG. The experiments were performed in triplicates and the data are shown as mean ± SD of one representative experiment. **B**, ADCC plots showing pooled data grouped by the presence of at least one valine (VV+VF) vs. FF homozygosity. Data are shown as mean ± SEM for 2 and 5 VV or VF donors, respectively, and 5 FF healthy individuals. **C**, ADCC induced by anti-hCD20-IgG subclasses was compared using the AUC of the percentage of specific lysis vs. anti-hCD20-IgG concentration for each donor. The percentage of direct cytotoxicity was used as baseline. Results were plotted with all donors grouped by V158F *FCGR3A* polymorphism. Data obtained from VV ($n = 2$) vs. FF homozygous ($n = 5$), and VF heterozygous ($n = 5$) donors, are shown as filled, open, and half-filled symbols, respectively. The differences between the anti-hCD20-IgG subclasses were compared using Two-way ANOVA and Tukey's multiple comparisons test; $^{*}P \leq 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$

hinders cross-linking under the experimental conditions used in Ca^{2+} assays, but not when IgG3 is bound/fixed to cell surface antigen as in the ADCC assays. IgG4 induced Ca^{2+} flux although we could not detect binding to NK cells by flow cytometry. It is possible that the higher concentrations of IgG subclasses used in the Ca^{2+} mobilization assay were responsible for the discrepancy in the results (15 vs. 8 μg/mL). However, we think that Ca^{2+} mobilization is more sensitive than flow cytometry to detect binding of IgG4 to FcγRIIIa/CD16a, corroborating

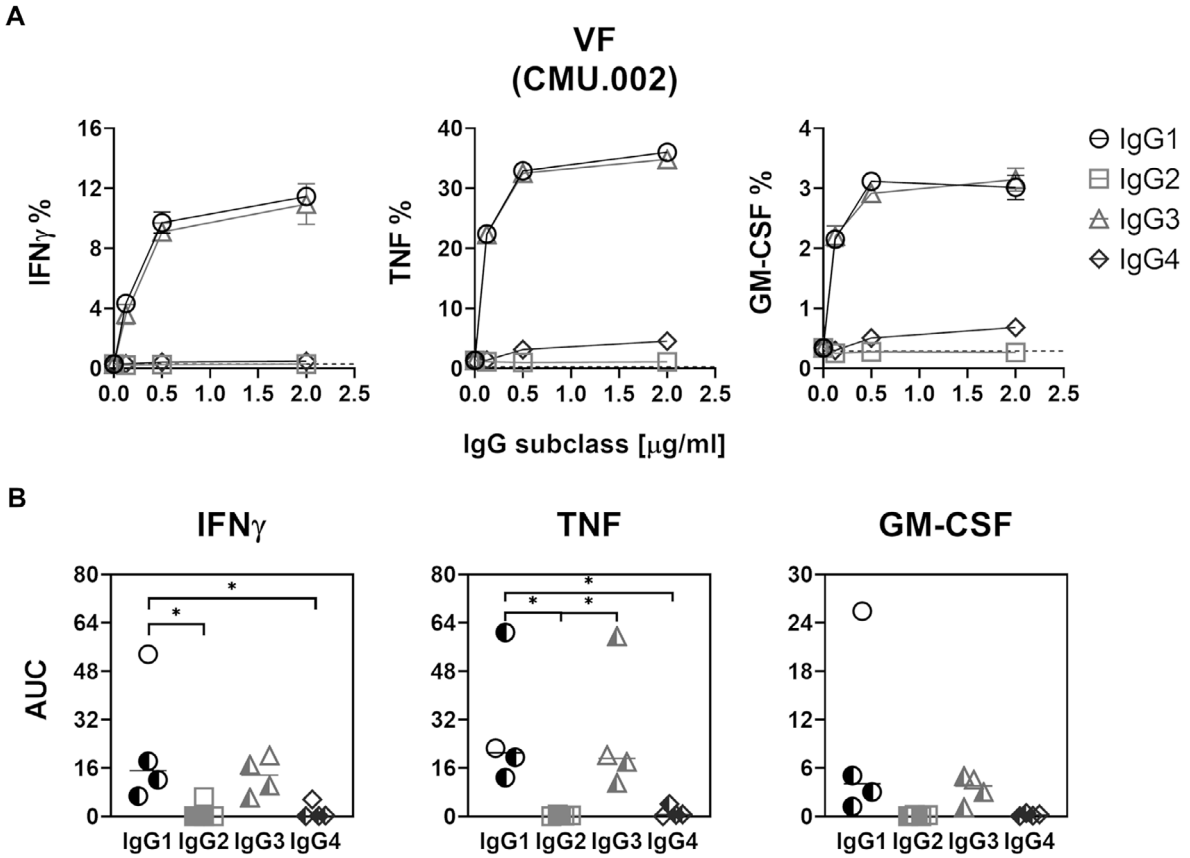


FIGURE 5 **Intracellular detection of cytokines produced by NK during/in Ab-dependent cellular cytotoxicity correlates with anti-CD20 IgG subclasses.** Freshly isolated NK cells were incubated with Daudi target cells in the presence of anti-hCD20-IgG subclasses for 6 h and stained for CD56 and the intracellular cytokines IFNγ, TNF, and GM-CSF. **A**, For a representative VF donor (CMU 0.002) is shown the dose-response curves for the percentage of NK cells producing each cytokine. Basal cytokine levels in the absence of anti- CD20-IgG rituximab are shown in dotted line. **B**, Pooled data of NK cell cytokine content comparing the AUC of the percentage of anti-hCD20-IgG subclass induced $CD56^+IFN\gamma^+$, $CD56^+TNF^+$, and $CD56^+GM\text{-}CSF^+$; NK cells of each donor, using as baseline the percentage of $CD56^+$cytokine$^+$ NK cells coculture with Daudi cells alone. Results were plotted for each IgG subclass (IgG1, IgG2, IgG3, and IgG4 depicted as circle, square, triangle, and diamond, respectively). Donors with FF homozygosity ($n = 2$), and VF heterozygosity ($n = 2$), are shown as open and half-filled symbols, respectively. The difference between the mean of anti-hCD20 IgG subclasses was compared using Tukey's multiple comparisons test; statistical significance was found between all IgG subclasses as detailed in the figure. $^*P \leq 0.05$. No significant difference was found between the donors' genotype due to the low the low number of cases analyzed did not allow. Data of one experiment performed in duplicates for each donor are shown

the results of our ADCC and degranulation assays as well as previous studies.[22,55,56]

The notion that IgG4 has lower ability to trigger effector functions should be interpreted with caution. Waldmann and colleagues reported that IgG4 anti-CD52 Campath-1 mAb induced poor ADCC in vitro in some individuals as expected, but in others it was as active as the potent IgG1. When they investigated the in vivo biologic effects in patients, IgG4 anti-CD52 depleted peripheral blood lymphocytes, albeit less efficiently than IgG1. Intriguingly in vitro assays did not predict the in vivo effector function.[22,55,56] Another previous study by Warncke et al. used total PBMCs of either human or cynomolgus monkey origin to study ADCC induced by different IgG subclasses in vitro and demonstrated that in contrast to human IgG2 and IgG4, cynomolgus IgG2 and IgG4 display strong ADCC.[22] Their findings are also partially in agreement with ours; IgG3 and IgG1, and to a lesser degree IgG4, triggered ADCC, but killing mediated by high concentrations of IgG2 subclass was also observed.[22] Because PBMCs were used instead of purified NK cells, monocytes might have been involved in the depletion of Ab-coated target cells by phagocytosis and ADCC.[57–59] These findings might partially explain the outcome of a trial in 2006, in which an anti-CD28 IgG4 agonist Ab (TGN1412) did not show severe adverse side effects in preclinical nonhuman primate models, but generated a dramatic cytokine storm in all six treated human subjects.[60]

However, it is still not clear why IgG4 at high concentrations is able to trigger ADCC in some individuals and not in others. One explanation might be the fact that IgG4 molecules have the unique property to exchange Fab arms becoming bivalent, consequently, losing the ability to cross-link antigen.[61] However, under the experimental conditions in the present study, no IgG4 with specificities for other antigens were present. Furthermore, other *FCGR3A* gene polymorphisms might structurally favor the interaction with IgG4 in

those individuals showing ADCC.[62] This question warrants further examination.

As to intracellular cytokine production during ADCC, we found the highest levels for TNF followed by IFNγ and only very low levels of GM-CSF. These results corroborate the findings of Fauriat et al. of who used reversed ADCC to measure cytokine secretion after different receptor engagements on NK cells.[46] Only anti-hCD20-IgG1 and -IgG3 lead to cytokine production by NK cells whereas to -IgG2 and -IgG4 NK cells remained irresponsive. To be certain that IgG4 does not induce cytokine production, however, a larger number than four donors should be analyzed.

Our study has several limitations. First, the results of ADCC assays are notoriously variable even in a single donor. Our analysis does not include repeated assays and reports only a limited number of donors of the different genotypes for the *FCGR3A* V158F polymorphism. The Ca^{2+} mobilization assays and ADCC could not be tested in parallel for the same donor, and the rarity of the high-affinity VV variant limited the genotype comparisons. Nevertheless, we were able to show a statistical impact of the *FCGR3A* V158F polymorphism on ADCC induced by the different IgG subclasses. Second, we did not analyze the conjugate and synapse formation between NK cells and anti-hCD20-IgG-coated Daudi target cells, or other signaling pathway besides Ca^{2+} mobilization, for instance CD3ξ, SYK, or p38/ERK. Moreover, neither the copy number variation of the *FCGR3A* gene nor the expression of the activating receptor FcγRIIC, which can be found on NK cells, was determined. Therefore, ADCC mediated by NK cells carrying FcγRIIC, as shown by others in vitro,[63] might have had an impact on our results. Finally, because we only tested anti-hCD20 IgG subclass mAb, the results of our study cannot be generalized to antibodies recognizing other antigens.

In conclusion, using a fully human system, we showed that anti-CD20 rituximab IgG1, IgG3, and to a lesser extent IgG4, but not IgG2 subclass mAb, trigger ADCC and degranulation in NK cells. Nonetheless, we could not provide evidence that *FCGR3A* V158F SNP has an impact on store-operated Ca^{2+} entry in NK cells. Our findings indicate important differences in the interactions of IgG subclasses with FcγRIIIA/CD16a and may therefore help to correlate better the functional properties of particular engineered therapeutic antibodies in vitro with individual differences of their clinical efficacy.

AUTHORSHIP

M.F.M. provided the experimental design, data analysis, and statistics and drafted approved the, final version. M.P. prepared the experimental design and carried out the revision and approval of the article. A.R. contributed the molecular biology experimentation and approval of the article. G.L.P.Y. conceived the concept and design, provided data analysis, and involved in interpretation and statistics; critical revision, approval of the article and final versions of the article. J.D.S. also conceived the concept and design, carried out interpretation of results and critical revision, secured the funding for the project, and provided approval of the article and final versions. G.L.P.Y. and J.D.S. contributed equally to this work.

DISCLOSURES

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

The authors thank R. Spirig (CSL Behring) for critical discussions; L. Gruaz for PBMC isolation and technical help; Y. Cambet from the READS platform for his assistance in calcium flux experiments; G. Schneiter and C. Gameiro from the flow cytometry facility for assistance with plate reading; O. Hartley for the use of the CytoFLEX cytometer; and finally H. Buvelot and H. Kutaish for blood sampling of healthy donors enrolled at CMU. This work was supported by a Private Foundation to J.D.S. and the Faculty of Medicine, University of Geneva to M.F.M.

ORCID

Marta Freitas Monteiro https://orcid.org/0000-0002-8613-9622
Aline Réal https://orcid.org/0000-0002-9437-4411
Gisella L. Puga Yung https://orcid.org/0000-0002-2283-7798
Jörg D. Seebach https://orcid.org/0000-0001-5748-4577

REFERENCES

1. Perussia B, Loza MJ. Assays for antibody-dependent cell-mediated cytotoxicity (ADCC) and reverse ADCC (redirected cytotoxicity) in human natural killer cells. *Methods Mol Biol*. 2000;121:179-192.
2. Loupy A, Lefaucheur C. Antibody-mediated rejection of solid-organ allografts. *N Engl J Med*. 2018;379:1150-1160.
3. Seidel UJ, Schlegel P, Lang P. Natural killer cell mediated antibody-dependent cellular cytotoxicity in tumor immunotherapy with therapeutic antibodies. *Front Immunol*. 2013;4:76.
4. Redman JM, Hill EM, AlDeghaither D, Weiner LM. Mechanisms of action of therapeutic antibodies for cancer. *Mol Immunol*. 2015;67:28-45.
5. Guillerey C, Huntington ND, Smyth MJ. Targeting natural killer cells in cancer immunotherapy. *Nat Immunol*. 2016;17:1025-1036.
6. Nagelkerke SQ, Schmidt DE, de Haas M, Kuijpers TW, Phenotypic variation in IgG receptors by nonclassical FCGR2C alleles. *J Immunol*. 2012;188:1318-1324.
7. Mahaweni NM, Olieslagers TI, Rivas IO, et al. *Sci Rep*. 2018;8:15983.
8. Lin TS, Flinn IW, Modali R, et al. FCGR3A and FCGR2A polymorphisms may not correlate with response to alemtuzumab in chronic lymphocytic leukemia. *Blood*. 2005;105:289-291.
9. Taylor RJ, Chan SL, Wood A, et al. FcgammaRIIIa polymorphisms and cetuximab induced cytotoxicity in squamous cell carcinoma of the head and neck. *Cancer Immunol Immunother*. 2009;58:997-1006.
10. Calemma R, Ottaiano A, Trotta AM, et al. Fc gamma receptor IIIa polymorphisms in advanced colorectal cancer patients correlated with response to anti-EGFR antibodies and clinical outcome. *J Transl Med*. 2012;10:232.
11. Park SJ, Hong YS, Lee JL, et al. Genetic polymorphisms of FcgammaRIIa and FcgammaRIIIa are not predictive of clinical outcomes after cetuximab plus irinotecan chemotherapy in patients with metastatic colorectal cancer. *Oncology*. 2012;82:83-89.
12. Mellor JD, Brown MP, Irving HR, Zalcberg JR, Dobrovic A. A critical review of the role of Fc gamma receptor polymorphisms in the response to monoclonal antibodies in cancer. *J Hematol Oncol*. 2013;6:1.

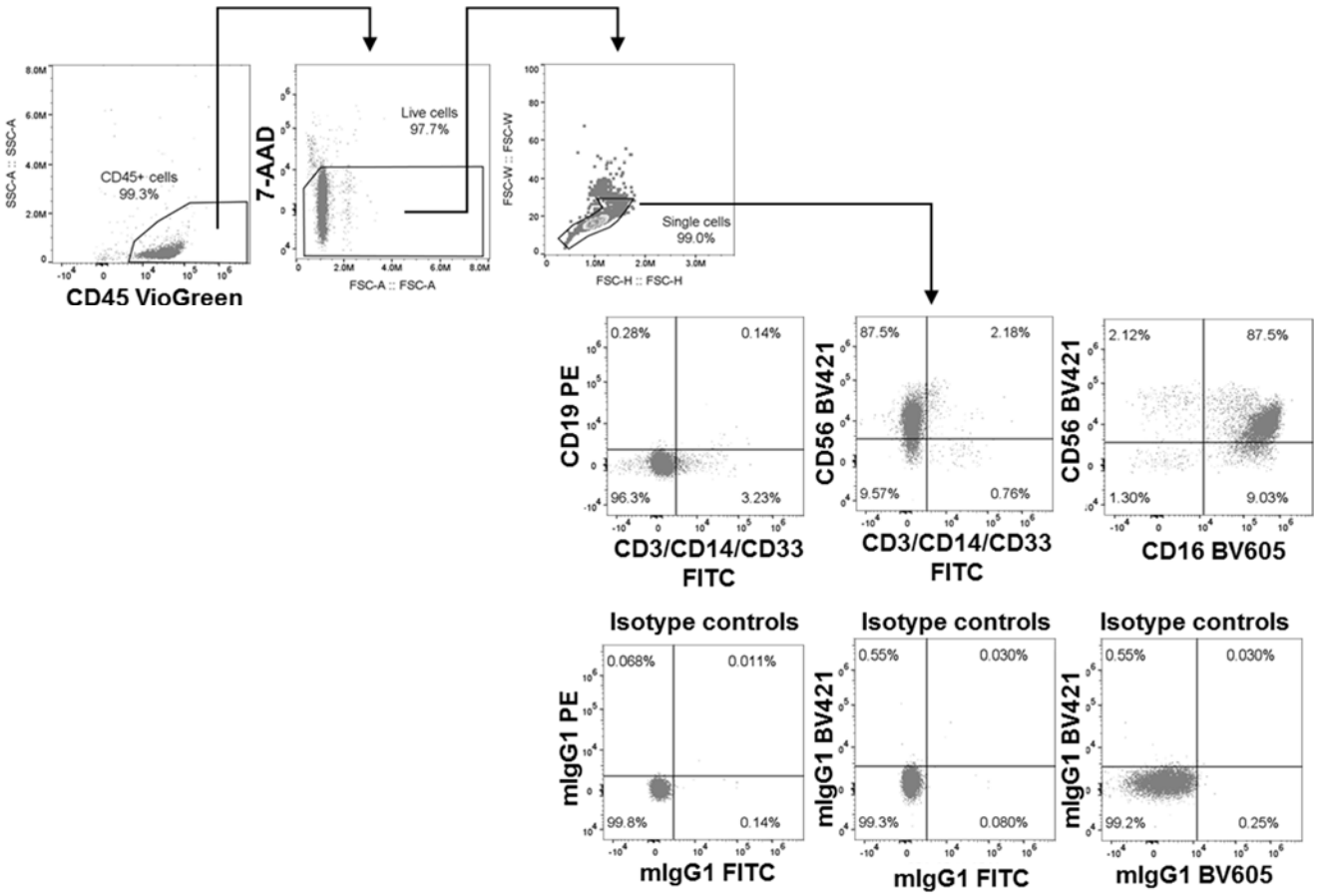
13. Stork AC, Notermans NC, van den Berg LH, et al. Fcgamma receptor IIIA genotype is associated with rituximab response in antimyelin-associated glycoprotein neuropathy. *J Neurol Neurosurg Psychiatry*. 2014;85:918-920.
14. Negri FV, Musolino A, Naldi N, et al. Role of immunoglobulin G fragment C receptor polymorphism-mediated antibody-dependant cellular cytotoxicity in colorectal cancer treated with cetuximab therapy. *Pharmacogenomics J*. 2014;14:14-19.
15. Boero S, Morabito A, Banelli B, et al. Analysis of in vitro ADCC and clinical response to trastuzumab: possible relevance of FcgammaRIIIA/FcgammaRIIA gene polymorphisms and HER-2 expression levels on breast cancer cell lines. *J Transl Med*. 2015;13:324.
16. Gavin PG, Song N, Kim SR, et al. Association of polymorphisms in FCGR2A and FCGR3A with degree of trastuzumab benefit in the adjuvant treatment of ERBB2/HER2-positive breast cancer: analysis of the NSABP B-31 Trial. *JAMA Oncol*. 2017;3:335-341.
17. Vidarsson G, Dekkers G, Rispens T. IgG subclasses and allotypes: from structure to effector functions. *Front Immunol*. 2014;5:520.
18. Goulet DR, Atkins WM. Considerations for the design of antibody-based therapeutics. *J Pharm Sci*. 2020;109:74-103.
19. Weiner LM, Surana R, Wang S. Monoclonal antibodies: versatile platforms for cancer immunotherapy. *Nat Rev Immunol*. 2010;10:317-327.
20. Stewart R, Hammond SA, Oberst M, Wilkinson RW. The role of Fc gamma receptors in the activity of immunomodulatory antibodies for cancer. *Journal for ImmunoTherapy of Cancer*. 2014;2:29.
21. Bruhns P, Iannascoli B, England P, et al. Specificity and affinity of human Fcgamma receptors and their polymorphic variants for human IgG subclasses. *Blood*. 2009;113:3716-3725.
22. Warncke M, Calzascia T, Coulot M, et al. Different adaptations of IgG effector function in human and nonhuman primates and implications for therapeutic antibody treatment. *J Immunol*. 2012;188:4405-4411.
23. Patel RJ, Johnson KK, Andrien BA, Tamburini PP. IgG subclass variation of a monoclonal antibody binding to human Fc-gamma receptors. *Am J Biochem Biotechnol*. 2013;9:12.
24. Armour KL, Smith CS, Clark MR. Expression of human FcgammaRIIIa as a GPI-linked molecule on CHO cells to enable measurement of human IgG binding. *J Immunol Methods*. 2010;354:20-33.
25. Feske S. Calcium signalling in lymphocyte activation and disease. *Nat Rev Immunol*. 2007;7:690-702.
26. Bergmeier W, Weidinger C, Zee I, Feske S. Emerging roles of store-operated Ca(2)(+) entry through STIM and ORAI proteins in immunity, hemostasis and cancer. *Channels (Austin)*. 2013;7:379-391.
27. Schwarz EC, Qu B, Hoth M. Calcium, cancer and killing: the role of calcium in killing cancer cells by cytotoxic T lymphocytes and natural killer cells. *Biochim Biophys Acta*. 2013;1833:1603-1611.
28. Maul-Pavicic A, Chiang SC, Rensing-Ehl A, et al. ORAI1-mediated calcium influx is required for human cytotoxic lymphocyte degranulation and target cell lysis. *Proc Natl Acad Sci U S A*. 2011;108:3324-3329.
29. Wu J, Edberg JC, Redecha PB, et al. A novel polymorphism of FcgammaRIIIa (CD16) alters receptor function and predisposes to autoimmune disease. *J Clin Invest*. 1997;100:1059-1070.
30. Brander C, Matter-Reissmann UB, Jones NG, Walker BD, Sachs DH, Seebach JD. Inhibition of human NK cell-mediated cytotoxicity by exposure to ammonium chloride. *J Immunol Methods*. 2001;252:1-14.
31. Patel KR, Roberts JT, Barb AW. Multiple variables at the leukocyte cell surface impact Fc gamma receptor-dependent mechanisms. *Front Immunol*. 2019;10:223.
32. Pradier A, Papaserafeim M, Li N, et al. Small-molecule immunosuppressive drugs and therapeutic immunoglobulins differentially inhibit NK cell effector functions in vitro. *Front Immunol*. 2019;10:556.
33. Long EO, Kim HS, Liu D, Peterson ME, Rajagopalan S. Controlling natural killer cell responses: integration of signals for activation and inhibition. *Annu Rev Immunol*. 2013;31:227-258.
34. Prager I, Liesche C, van Ooijen H, et al. NK cells switch from granzyme B to death receptor-mediated cytotoxicity during serial killing. *J Exp Med*. 2019;216:2113-2127.
35. Prager I, Watzl C. Mechanisms of natural killer cell-mediated cellular cytotoxicity. *J Leukoc Biol*. 2019;105:1319-1329.
36. Millard AL, Valli PV, Stussi G, Mueller NJ, Yung GP, Seebach JD. Brief exercise increases peripheral blood NK cell counts without immediate functional changes, but impairs their responses to ex vivo stimulation. *Front Immunol*. 2013;4:125.
37. Bryceson YT, Fauriat C, Nunes JM, et al. Functional analysis of human NK cells by flow cytometry. *Methods Mol Biol*. 2010;612:335-352.
38. Koene HR, Kleijer M, Algra J, Roos D, von dem Borne AE, de Haas M. Fc gammaRIIIa-158V/F polymorphism influences the binding of IgG by natural killer cell Fc gammaRIIIa, independently of the Fc gammaRIIIa-48L/R/H phenotype. *Blood*. 1997;90:1109-1114.
39. Mota G, Moldovan I, Calugaru A, et al. Interaction of human immunoglobulin G with CD16 on natural killer cells: ligand clearance, FcgammaRIIIA turnover and effects of metalloproteinases on FcgammaRIIIA-mediated binding, signal transduction and killing. *Scand J Immunol*. 2004;59:278-284.
40. Bryceson YT, March ME, Ljunggren HG, Long EO. Activation, coactivation, and costimulation of resting human natural killer cells. *Immunol Rev*. 2006;214:73-91.
41. Cartron G, Dacheux L, Salles G, et al. Therapeutic activity of humanized anti-CD20 monoclonal antibody and polymorphism in IgG Fc receptor FcgammaRIIIa gene. *Blood*. 2002;99:754-758.
42. Dall'Ozzo S, Tartas S, Paintaud G, et al. Rituximab-dependent cytotoxicity by natural killer cells: influence of FCGR3A polymorphism on the concentration-effect relationship. *Cancer Res*. 2004;64:4664-4669.
43. Li Y, Huang K, Liu L, et al. Effects of complement and serum IgG on rituximab-dependent natural killer cell-mediated cytotoxicity against Raji cells. *Oncol Lett*. 2019;17:339-347.
44. Musolino A, Naldi N, Bortesi B, et al. Immunoglobulin G fragment C receptor polymorphisms and clinical efficacy of trastuzumab-based therapy in patients with HER-2/neu-positive metastatic breast cancer. *J Clin Oncol*. 2008;26:1789-1796.
45. Weng WK, Levy R. Two immunoglobulin G fragment C receptor polymorphisms independently predict response to rituximab in patients with follicular lymphoma. *J Clin Oncol*. 2003;21:3940-3947.
46. Fauriat C, Long EO, Ljunggren HG, Bryceson YT. Regulation of human NK-cell cytokine and chemokine production by target cell recognition. *Blood*. 2010;115:2167-2176.
47. Roux KH, Strelets L, Michaelsen TE. Flexibility of human IgG subclasses. *J Immunol*. 1997;159:3372-3382.
48. Repp R, Kellner C, Muskulus A, et al. Combined Fc-protein- and Fc-glyco-engineering of scFv-Fc fusion proteins synergistically enhances CD16a binding but does not further enhance NK-cell mediated ADCC. *J Immunol Methods*. 2011;373:67-78.
49. Sanchez-Mejorada G, Rosales C. Signal transduction by immunoglobulin Fc receptors. *J Leukoc Biol*. 1998;63:521-533.
50. Getahun A, Cambier JC. Of ITIMs, ITAMs, and ITAMis: revisiting immunoglobulin Fc receptor signaling. *Immunol Rev*. 2015;268:66-73.
51. Theorell J, Bryceson YT. Analysis of intracellular Ca(2+) mobilization in human NK cell subsets by flow cytometry. *Methods Mol Biol*. 2016;1441:117-130.
52. Lian J, Cuk M, Kahlfuss S, et al. ORAI1 mutations abolishing store-operated Ca(2+) entry cause anhidrotic ectodermal dysplasia with immunodeficiency. *J Allergy Clin Immunol*. 2018;142:1297-1310 e11.
53. Nguyen T, Johnston S, Clarke L, Smith P, Staines D, Marshall-Gradisnik S. Impaired calcium mobilization in natural killer cells from chronic fatigue syndrome/myalgic encephalomyelitis patients is associated

with transient receptor potential melastatin 3 ion channels. *Clin Exp Immunol*. 2017;187:284-293.
54. Nguyen T, Staines D, Nilius B, Smith P, Marshall-Gradisnik S. Novel identification and characterisation of transient receptor potential melastatin 3 ion channels on natural killer cells and B lymphocytes: effects on cell signalling in chronic fatigue syndrome/myalgic encephalomyelitis patients. *Biol Res*. 2016;49:27.
55. Greenwood J, Clark M, Waldmann H. Structural motifs involved in human IgG antibody effector functions. *Eur J Immunol*. 1993;23:1098-1104.
56. Isaacs JD, Wing MG, Greenwood JD, Hazleman BL, Hale G, Waldmann H. A therapeutic human IgG4 monoclonal antibody that depletes target cells in humans. *Clin Exp Immunol*. 1996;106:427-433.
57. Yeap WH, Wong KL, Shimasaki N, et al. CD16 is indispensable for antibody-dependent cellular cytotoxicity by human monocytes. *Sci Rep*. 2016;6:34310.
58. Roda JM, Joshi T, Butchar JP, et al. The activation of natural killer cell effector functions by cetuximab-coated, epidermal growth factor receptor positive tumor cells is enhanced by cytokines. *Clin Cancer Res*. 2007;13:6419-6428.
59. Herter S, Birk MC, Klein C, Gerdes C, Umana P, Bacac M. Glycoengineering of therapeutic antibodies enhances monocyte/macrophage-mediated phagocytosis and cytotoxicity. *J Immunol*. 2014;192:2252-2260.
60. Suntharalingam G, Perry MR, Ward S, et al. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. *N Engl J Med*. 2006;355:1018-1028.
61. van der Neut Kolfschoten M, Schuurman J, Losen M, et al. Anti-inflammatory activity of human IgG4 antibodies by dynamic Fab arm exchange. *Science*. 2007;317:1554-1557.
62. Davies AM, Sutton BJ. Human IgG4: a structural perspective. *Immunol Rev*. 2015;268:139-159.
63. Metes D, Morel PA, Nellis J, Fung JJ, Rao AS. FcgammaRIIc 13Q/STP polymorphism influences the antibody-dependent cytotoxicity levels triggered by natural killer cells against pig aortic endothelial cells. *Transplant Proc*. 2001;33:333.

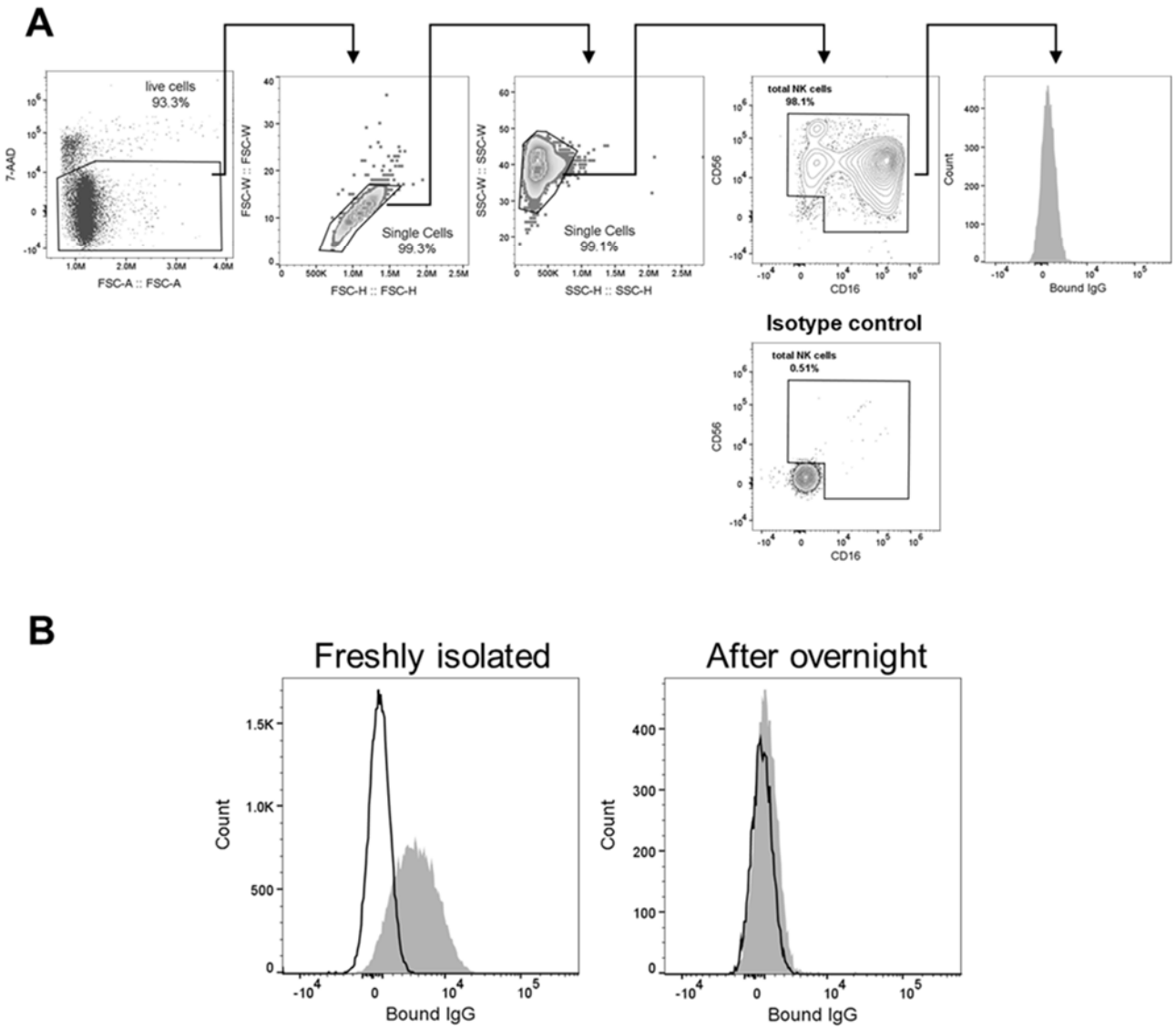
SUPPORTING INFORMATION

Additional information may be found online in the Supporting Information section at the end of the article.

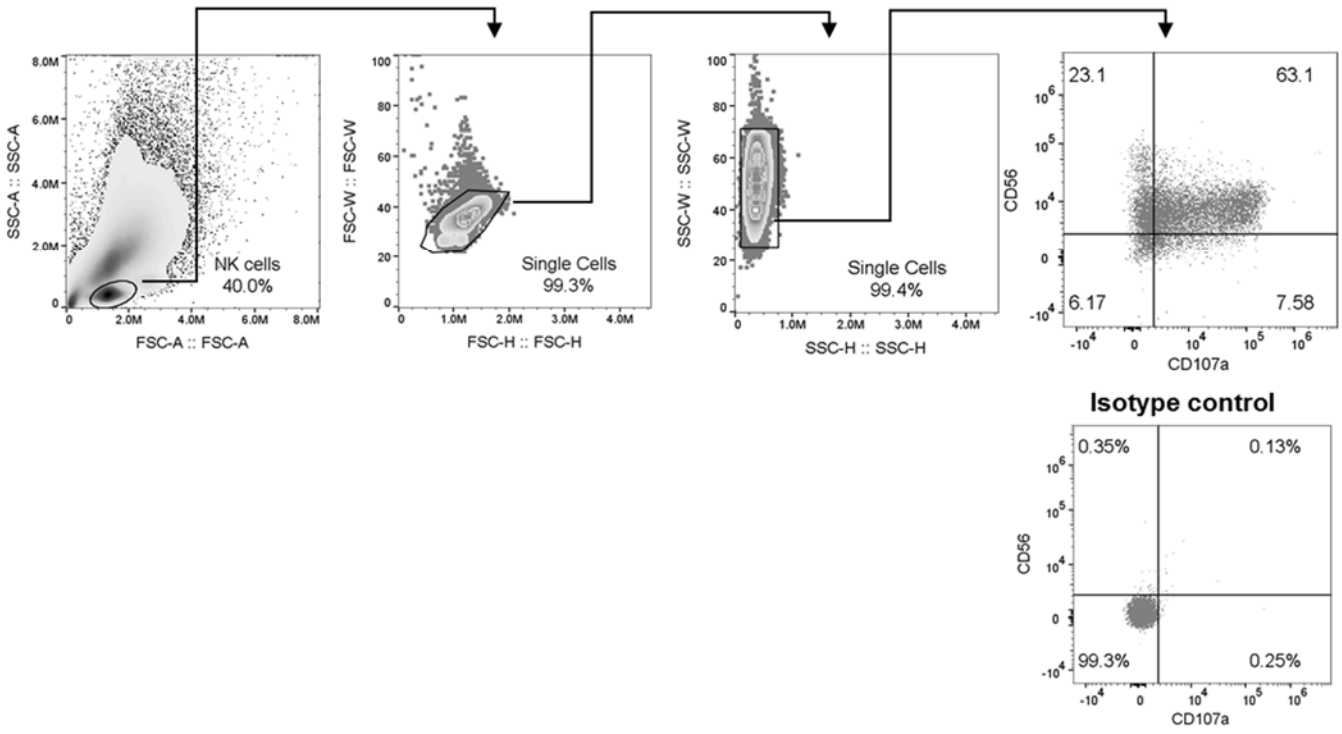
How to cite this article: Freitas Monteiro M, Papaserafeim M, Réal A, Puga Yung GL, Seebach JD. Anti-CD20 rituximab IgG1, IgG3, and IgG4 but not IgG2 subclass trigger Ca^{2+} mobilization and cytotoxicity in human NK cells. *J Leukoc Biol*. 2020;108:1409–1423. https://doi.org/10.1002/JLB.5MA0620-039R



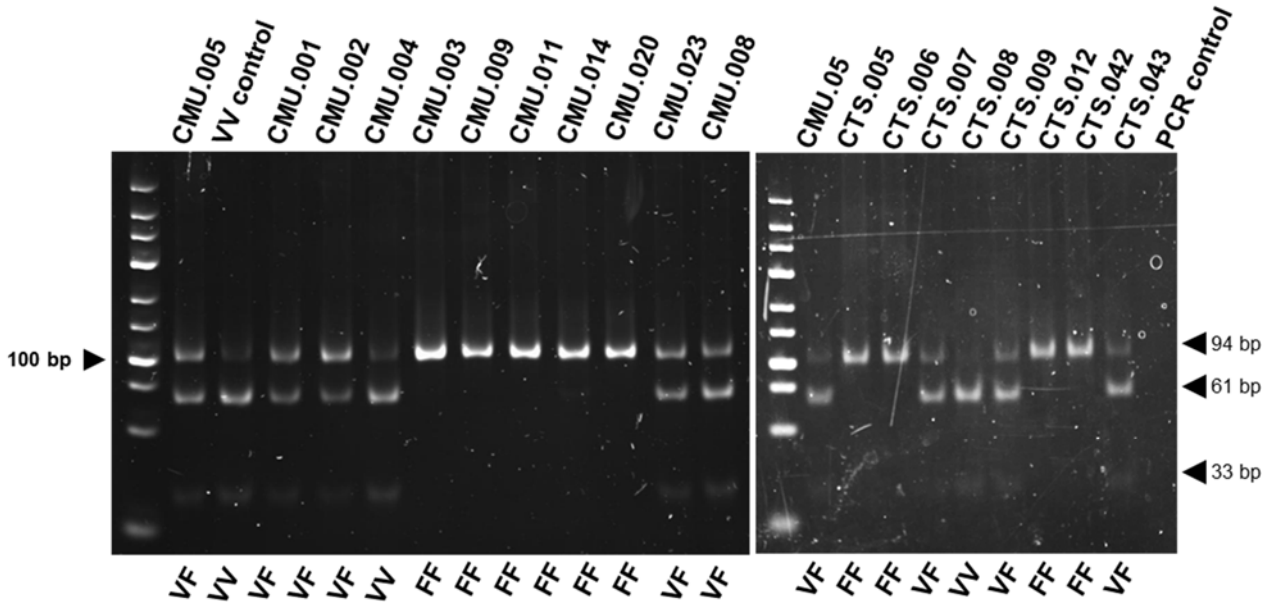
Supplementary Figure 1. NK cell purity after isolation. PBMC were isolated from human blood by Ficoll-gradient centrifugation, NK cells isolated by negative magnetic isolation using the MACS NK isolation kit (Miltenyi Biotec). NK cells were identified by sequential gating on $CD45^+$ cells, live cells ($7AAD^-$) followed by single cell gating using the FSC width, and finally by expressing CD56 and/or CD16. NK cell purity was verified by staining with anti-CD3/CD14/CD19/CD33 Abs and isotype matched control Abs and analyzed by flow cytometry. Dot plots of a representative donor are presented.



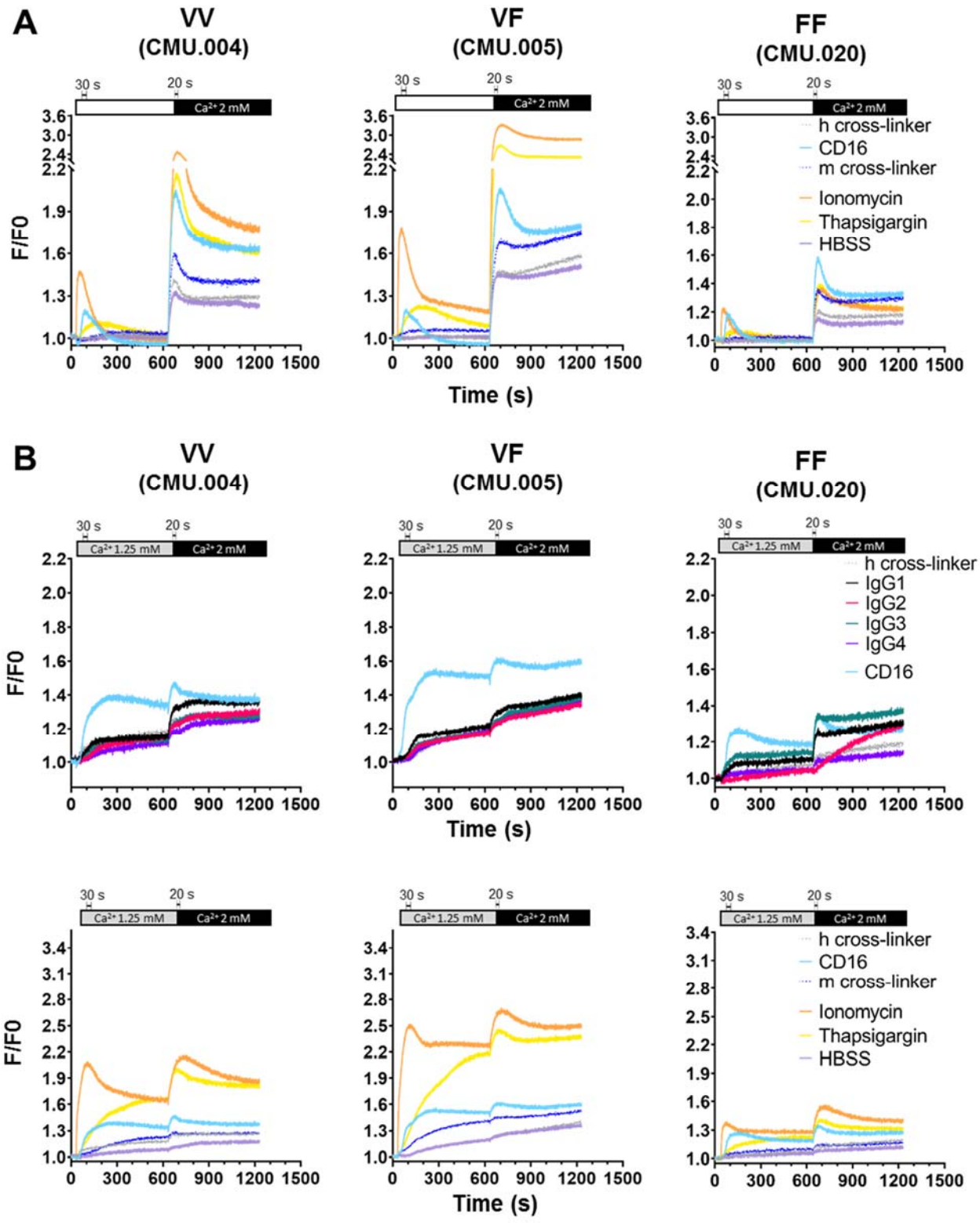
Supplementary Figure 2. Gating strategy for the detection of naturally bound IgG on the surface of freshly isolated NK cells. After purification, freshly isolated NK cells were kept overnight in AIM-V medium + 2% HEPES and analyzed by flow cytometry. **A**, NK cells were identified by sequential gating on $CD45^+$ cells, live cells ($7AAD^-$) followed by single cell gating using the FSC width, and finally by expressing CD56 and/or CD16. Naturally bound IgG were detected using detection goat anti-h-F(ab')2-PE polyclonal Ab (grey-filled histogram). **B**, Representative histogram overlays showing the naturally bound IgG (gray) on NK cells and correspondent isotype control (dark line) immediately after isolations (left) and after overnight culture (right).



Supplementary Figure 3. Gating strategy for the determination of NK cell degranulation by CD107a expression. After co-culture of NK cells and Daudi cells in the presence of anti-hCD20-IgG subclasses, the degranulation marker CD107a$^+$ on NK cells was quantified by flow cytometry as described in the Material and Methods section. Gating on NK cells is shown on Forward (FSC-A) *versus* Side Scatter (SSC-A) dot-plots, followed by gating on single cells in both forward and side scatter width plots, and finally by the expression of CD56 and CD107a. Representative dot plots of NK cells co-cultured with Daudi cells show the percentages of CD56$^+$ CD107a$^+$ NK cells (upper right quadrant) when stained with CD107a or isotype control.

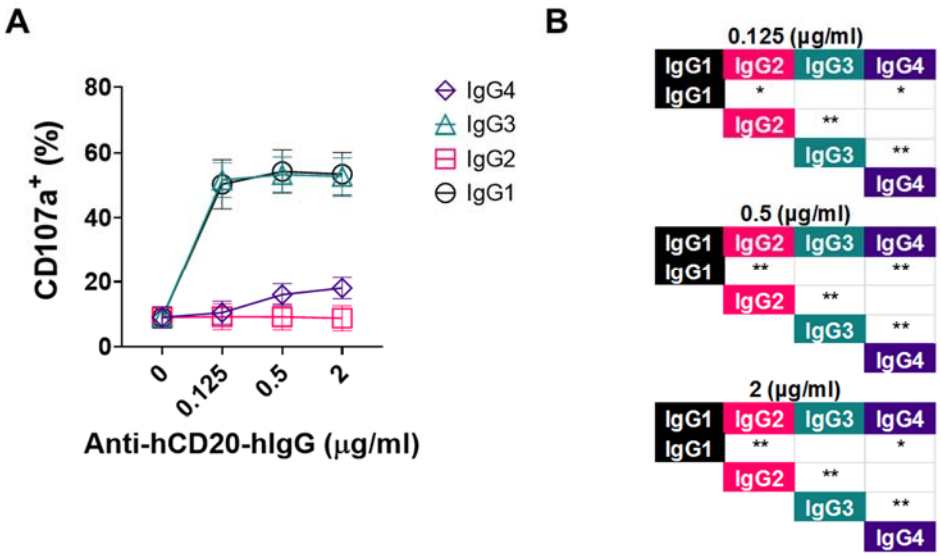


Supplementary Figure 4. Determination of V158F *FCGR3A* polymorphism. Genomic DNA was extracted from total blood and the *FCGR3A* V158F allelic variation of each NK cell donor was determined by nested-PCR followed by cutting of the second PCR product with the NlaIII restriction enzyme. Separation of DNA fragments was performed in 10% acrylamide gels stained with GelRed. Digested PCR products from donors homozygous for V158F polymorphism VV give two bands, a strong 61 bp and a faint of 33 bp band; whereas donors homozygous for FF are not cut, thus, only one band of 94 bp is seen; VF carries show three bands of 94, 61 bp of similar strong intensity and a faint band of 33 pb. Occasionally, after digesting VV shows a weak band at 94 bp. Five µl of digestion reaction were loaded in 6× TriTrack DNA Loading Dye buffer (Thermo Fisher, Vilnius, Lithuania), the size of the DNA fragments was estimated using the Gene Ruler Low Range DNA ladder (Thermo Fisher).

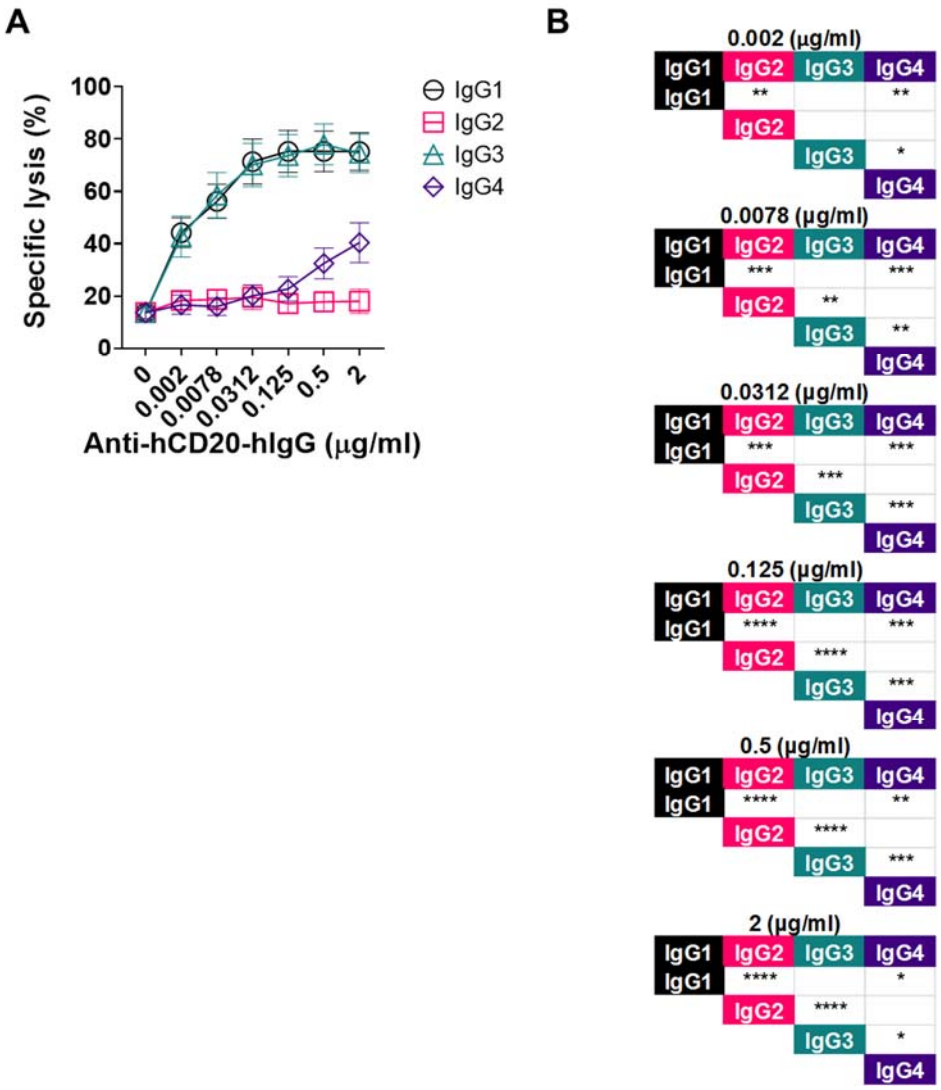


Supplementary Figure 5. Controls used in the Ca^{2+} mobilization assay. Freshly isolated NK cells were labeled with anti-hCD20-IgG subclasses and anti-CD16 mAb at 15µg/ml and 1.5 µg/ml final, respectively. For each set of experiments, either HBSS plus 1.3 mM EGTA

(no Ca^{2+} condition, white bar) or HBSS Ca^{2+}/Mg^{2+} buffer (Ca^{2+} condition, grey bar) were assayed. For each experiment, reagents were added at 60 or 630 seconds (s). After establishing a baseline for 30 s on the FDSS/µCELL plate reader, h cross-linker Ab, m cross-linker Ab, ionomycin and thapsigargin were added to reach the final concentrations of 20 µg/mL, 0.11 µg/mL, 2 µg/mL and 2 µM, respectively. At 630 s, Ca^{2+} buffer was added at a final concentration of 2 mM (black bar) and Ca^{2+} mobilization recorded for additional 600 s. Results are expressed as the ratio of Ca^{2+} flux (F/F0) versus time. **A**, Treats of Ca^{2+} efflux (first 630 s) and Ca^{2+} influx (last 600 s) for all controls are shown. **B**, Treats for Ca^{2+} mobilization at 1.25 mM Ca^{2+} (first 630 s) or 2 mM (last 600 s), for anti-hCD20-IgG subclasses (upper) and all controls (lower) are shown. Results represent the average of quadruplicates of three donors. Designation and genotype for each donor is presented in the panels.



Supplementary Figure 6. Pooled data of CD107a degranulation. A, Plot of percentage of $CD56^{+}CD107^{+}$ cells of pooled donors using NK cells, and anti-CD20-coated Daudi cells as targets are shown. Flow cytometry analysis using Daudi coated with anti-hCD20-IgG1 (circle), IgG2 (square), IgG3 (triangle) or IgG4 (diamond) subclasses at increasing concentrations (0.125 - 2 μg/ml). Subsequently, anti-hCD20-coated Daudi cells were co-cultured with freshly isolated NK cells for 6 hours at an E:T ratio of 1:1. The experiments were performed in duplicates, and the data are shown as mean ± SEM of 5 healthy individuals. **B**, Summary table showing the differences between the anti-hCD20-IgG subclasses at increasing concentrations obtained by two-way ANOVA and Tukey's multiple comparisons test; * $P \leq 0.05$ and ** $P < 0.01$.



Supplementary Figure 7. Pooled data of ADCC. A, ADCC plot of pooled donors using NK cells, and anti-CD20-coated Daudi cells as targets are shown. Non-radioactive labelled Daudi cells were coated with anti-hCD20-IgG1 (circle), IgG2 (square), IgG3 (triangle) or IgG4 (diamond) subclasses at increasing concentrations (0.0078-2 µg/ml). Subsequently, anti-hCD20-coated Daudi cells were co-cultured with freshly isolated NK cells for two hours at an E:T ratio of 5:1. The experiments were performed in triplicates, and the data are shown as mean ± SEM of 12 healthy individuals. **B**, Summary table showing the differences between the anti-hCD20-IgG subclasses at increasing concentrations obtained by two-way ANOVA and Tukey's multiple comparisons test; * $P \leq 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$.

Supplementary Table 1. General characteristics of healthy donors.

Identifier*	Source	Gender	Age (years)	V158F *FCGR3A*	Assay
CMU.001	Blood	M	29	VF	Ca^{2+} mobilization, ADCC
CMU.002	Blood	M	55	VF	Ca^{2+} mobilization, CD107a, ADCC, cytokines
CMU.003	Blood	M	28	FF	ADCC
CMU.004	Blood	F	33	VV	IgG affinity, Ca^{2+} mobilization, ADCC,
CMU.005	Blood	F	49	VF	Ca^{2+} mobilization, CD107a, ADCC, cytokines
CMU.008	Blood	F	26	VF	CD107a, cytokines
CMU.009	Blood	M	57	FF	Ca^{2+} mobilization; CD107a, cytokines
CMU.011	Blood	F	29	FF	IgG affinity, ADCC; Ca^{2+} mobilization
CMU.014	Blood	F	28	FF	CD107a
CMU.020	Blood	M	30	FF	Ca^{2+} mobilization
CTS.005	Buffy coat	ND	ND	FF	ADCC
CTS.006	Buffy coat	ND	ND	FF	ADCC
CTS.007	Buffy coat	ND	ND	VF	ADCC
CTS.008	Buffy coat	F	27	VV	ADCC
CTS.009	Buffy coat	M	61	VF	ADCC
CTS.012	Buffy coat	ND	ND	FF	ADCC
CTS.042	Blood	ND	63	FF	IgG affinity
CTS.043	Blood	ND	21	VF	IgG affinity

Abbreviations: ADCC, antibody-dependent cell cytotoxicity; ND, not disclosed