Development of Bioinformatics Tools and Workflows for the Analysis of Cell Line Data

Robin, Thibault

UNIVERSITÉ DE GENÈVE

Département d'informatique

FACULTÉ DES SCIENCES
Docteur Frédérique Lisacek

Département de microbiologie
et médecine moléculaire

FACULTÉ DE MÉDECINE
Professeur Amos Bairoch

Development of Bioinformatics Tools and Workflows for the Analysis of Cell Line Data

THÈSE

Présentée à la Faculté des sciences de l'Université de Genève pour obtenir le grade de Docteur ès sciences, mention bioinformatique

Par

Thibault ROBIN

de

Lancy (GE)

Thèse n° 5471

GENÈVE

Atelier d'impression ReproMail

2020

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION BIOINFORMATIQUE

Thèse de Monsieur Thibault ROBIN

intitulée :

«Development of Bioinformatics Tools and Workflows for the Analysis of Cell Line Data»

La Faculté des sciences, sur le préavis de Monsieur A. BAIROCH, professeur ordinaire et codirecteur de thèse (Faculté de médecine, Département de science des protéines humaines), Madame F. LISACEK, docteure et codirectrice de thèse (Département d'informatique), Monsieur B. WOLLSCHEID, professeur (Gesundheitswissenschaften und Technologie, Eidgenössische Technische Hochschule Zürich, Schweiz), Madame B. PARODI, professeur (Insituto di Ricovero e Cura a Carettere Scientifico, Ospedale Policlinico San Martino, Gênes, Italie), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 3 août 2020

Thèse - 5471 -

Le Décanat

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".

Dedicated to my family, to my friends, and to my cat Sasaki who passed away the day I started writing this thesis

Acknowledgements

I would like to start by thanking Dr. Frédérique Lisacek and Prof. Amos Bairoch for being the co-directors of my PhD thesis and supporting me all the way through. They allowed me to undertake this ambitious work that ended up covering several biomolecular disciplines, requiring the development of distinct software solutions. As a result, I could further improve my bioinformatics training while learning how to write computer software following the best coding practices.

I also want to offer my special thanks to Dr. Barbara Parodi and Prof. Bernd Wollscheid for accepting being part of my PhD jury.

I am grateful to Dr. Lydie Lane and Dr. Markus Müller for their guidance during the mass spectrometry-based proteomics projects. Their experience and knowledge allowed me to complete insightful studies that taught me a lot about the large-scale study of proteomics data and data analysis in general.

I was very privileged to collaborate with Dr. Amanda Capes-Davis, Dr. Richard M. Neve, Dr. Christopher Korch, and Gregory Sykes on the topic of cell line authentication by STR profiling. They are passionate about the topic and took a lot of time to share their knowledge with me. I hope that the work that I achieved with them will have a long-term impact, and end up helping to curb the spread of misidentified and contaminated cell lines in the scientific literature.

I am also very thankful to my PhD colleagues, especially Julien Mariethoz, Thomas Stricker and Emma Ricart Altimiras, for being by my side during all these years. I wish them good fortune in the completion of their own thesis. I would like to end by thanking my family and friends for supporting me through this long and challenging journey, I hope that I made you all proud.

Thibault Robin

Table of Contents

List of Figures

List of Tables

Résumé en Français

Les lignées cellulaires sont devenues un atout essentiel pour la recherche biomédicale de par les nombreux avantages qu'elles procurent par rapport aux autres types de cultures cellulaires. Elles sont utilisées dans un grand nombre d'expériences dans les différents domaines *omiques*, qui sont rentrés dans une ère à haut débit ces dernières années avec l'arrivée de nouvelles technologies et instruments. La contamination-croisée et les erreurs d'identification des lignées cellulaires sont cependant reconnues être responsables d'affecter la fiabilité et reproductibilité des résultats expérimentaux. La génération à grande échelle de données biologiques soulève de nombreux défis bioinformatiques couvrant principalement la question de leurs formats, stockage, représentation et interprétation. Cette thèse se concentre sur le développement de logiciels bioinformatiques pour répondre à ces problématiques, offrant des outils pour l'analyse et l'interprétation de données omiques à la communauté scientifique. Trois applications distinctes furent développées au cours de cette thèse, ce qui amena à la publication de quatre articles scientifiques correspondants.

CLASTR est une application web qui fournit aux chercheurs un service en ligne fiable pour authentifier les lignées cellulaires avec lesquelles ils travaillent. Il permet d'effectuer des recherches de similarités de profils de séquences courtes répétées en tandem (STR) contenues dans la ressource en ligne Cellosaurus. Il possède à la fois une interface utilisateur web intuitive et une interface de programmation d'application efficace. De nombreux paramètres de recherche sont disponibles, pour lesquels un article de recherche spécifique fut écrit afin de détailler les conséquences de leur choix sur les identifications résultantes. CLASTR représente une contribution importante pour faciliter et démocratiser l'authentification de lignées cellulaires par profilage STR dans le but de freiner la propagation des lignées cellulaires avec des erreurs d'identification ou contaminations-croisées dans la littérature scientifique.

MzVar est une application de bureau qui a été conçue pour la compilation de base de données customisées contenant des protéines ou peptides variants. Dans l'approche de recherche de base de données, seules les séquences qui sont incluses dans la

base de données peuvent être par la suite identifiées. Des variants de séquence qui pourraient avoir une pertinence biologique pourraient être manqués s'ils ne sont pas inclus dans la base de données qui est recherchée. Cet outil fut utilisé dans une étude ultérieure pour identifier des variants de séquence dans des données protéomiques de spectrométrie de masse en tandem provenant de la lignée cellulaire HeLa, tout en essayant d'évaluer leur influence sur l'expression et la stabilité des protéines.

GlyConnect Compozitor est une application web qui permet aux chercheurs d'explorer le contenu de la base de données GlyConnect sous la forme de graphiques de compositions de glycanes. GlyConnect contient une richesse d'information à propos de la glycosylation et des glycoprotéines. Avec la transition récente vers les expériences de glycoprotéomique, une plus grande proportion de publications comporte une information détaillée concernant les sites de glycosylation au détriment des structures des glycanes qui ne sont pas complètement résolues. GlyConnect Compozitor a pour but de combler ces lacunes en permettant la visualisation et la comparaison de compositions de glycanes en relation avec une gamme d'autres entités biologiques. Les compositions de glycanes d'intérêt peuvent être exportées en différents formats pour être utilisées comme fichiers de compositions de glycanes pour des recherches glycoprotéomiques ultérieures.

Abstract

Cell lines became an essential asset for biomedical research through the numerous practical advantages they offer over other types of cell cultures. They are used in a wide range of experiments in the different *omics* fields, which entered in a high-throughput era in recent years with the emergence of new technology and instrumentation. Cell line cross-contamination and misidentification is however known to impede the reliability and reproducibility of experimental results. The large-scale generation of biological data also raises many bioinformatics challenges mainly spanning issues of formats, storage, representation, and interpretation. This thesis focuses on the development of bioinformatics software to address these problems, providing tools for the analysis and interpretation of omics data to the scientific community. Three distinct applications were developed in the course of this thesis, which led to the publication of four related scientific articles.

CLASTR is a web application that provides researchers with a reliable online service to authenticate the cell lines they are working with. It allows performing similarity searches on the short tandem repeat (STR) profiles stored in the Cellosaurus online resource. It has both an intuitive web user interface and an efficient application programming interface. Numerous search parameters are available, for which a specific research article was written in order to detail the impact of their choice on the resulting identifications. CLASTR represents a significant effort in facilitating and democratizing cell authentication by STR profiling with the goal of curbing the spread of misidentified and cross-contaminated cell lines in the scientific literature.

MzVar is a desktop application that was designed for the compilation of customized variant protein and peptide databases. In the database search approach, only the sequences that are included in the database can be subsequently identified. Sequence variants that may have a biological relevance may consequently be missed if they are not included in the searched database. This tool was used in a subsequent study to identify sequence variants in proteomics tandem mass spectrometry data from the HeLa cell line, while trying to assess their influence on protein expression and stability.

GlyConnect Compozitor is a web application that enables researchers to explore the content of the GlyConnect database in the form of glycan composition graphs. GlyConnect contains a wealth of information about glycoproteins and glycosylation. With the recent shift towards glycoproteomics experiments, a larger proportion of publications features detailed glycosylation site information but lacks fully resolved glycan structures. GlyConnect Compozitor aims to bridge the gap by allowing the visualization and comparison of glycan compositions in relation with a range of other biological entities. Glycan compositions of interest can be exported in different formats to be used as glycan composition files for subsequent glycoproteomics searches.

Abbreviations

ANSI	American National Standards Institute
API	Application Programming Interface
AQUA	Absolute Quantification
ATCC	American Type Culture Collection
CID	Collision-Induced Dissociation
CLDB	Cell Line Database
CODIS	Combined DNA Index System
DOM	Document Object Model
ESI	Electrospray Ionization
ETD	Electron-Transfer Dissociation
FDR	False Discovery Rate
FTICR	Fourier-Transform Ion Cyclotron Resonance
GBP	Glycan-Binding Protein
GUI	Graphical User Interface
HCD	Higher-energy Collisional Dissociation
HPLC	High Performance Liquid Chromatography
ICAT	Isotope-Coded Affinity Tag
ICLAC	International Cell Line Authentication Committee
iPS	Induced Pluripotent Stem (cells)
LC	Liquid Chromatography
LFQ	Label-Free Quantification
MALDI	Matrix-Assisted Laser Desorption/Ionization
MMR	Mismatch Repair
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
m/z	Mass-to-Charge
NMR	Nuclear Magnetic Resonance
PCR	Polymerase Chain Reaction
PSM	Peptide-Spectrum Match
PTM	Post-Translational Modification
SNP	Single Nucleotide Polymorphism

SIB	Swiss Institute of Bioinformatics
SILAC	Stable Isotope Labeling by Amino acids in Cell Culture
SPA	Single Page Application
SRM	Selected Reaction Monitoring
STR	Short Tandem Repeat
VCF	Variant Call Format
TMT	Tandem Mass Tag
VNTR	Variable Number Tandem Repeat

Chapter 1

Introduction

The last decades have known a paradigm shift in biomedical research. Following the successful genome sequencing of several viruses and bacteria in the late 1990s, the Human Genome Project was initiated. This international collaborative effort aimed to provide the complete human genome sequence to the scientific community. It resulted in a first draft of the human genome in 2001 [1], which was subsequently fully completed in 2003. The whole genomes of an increasing number of species were sequenced in the following years. With the availability of entire genome sequences, experiments that target every gene, transcript, or protein could be designed. New experimental technologies and approaches emerged in parallel in the different *omics* disciplines, bringing biomolecular sciences into a high-throughput era. This led to a significant increase in the rate at which biological data is being generated, a single experiment producing large amounts of raw and processed data. Biomedical research became as a result more data-driven and computational, requiring the use of bioinformatics to manage and interpret the vast amount of data produced.

1.1. Bioinformatics

Bioinformatics can be defined as an interdisciplinary field that uses computational approaches as well as statistical methods to store, analyze, and visualize biological data [2]. It lies at the intersection of molecular biology, computer science, information science, and mathematics. The use of bioinformatics resources and tools became a necessity for omics research, whether in the fields of genomics, transcriptomics, proteomics, or glycomics. To be understandable and exploitable, the ever-growing amount of biological data (Figure 1.1) generated by high-throughput technologies requires suitable analytical resources and computer software. The scope of bioinformatics notably includes: data storage, annotation and integration

in databases (i); data sharing and dissemination between scientists and resources (ii); and data processing, interpretation, and visualization by computer software and workflows (iii). In this section, we present some of the core bioinformatics concepts that were directly or indirectly used in this thesis, while mentioning related challenges that are faced nowadays.

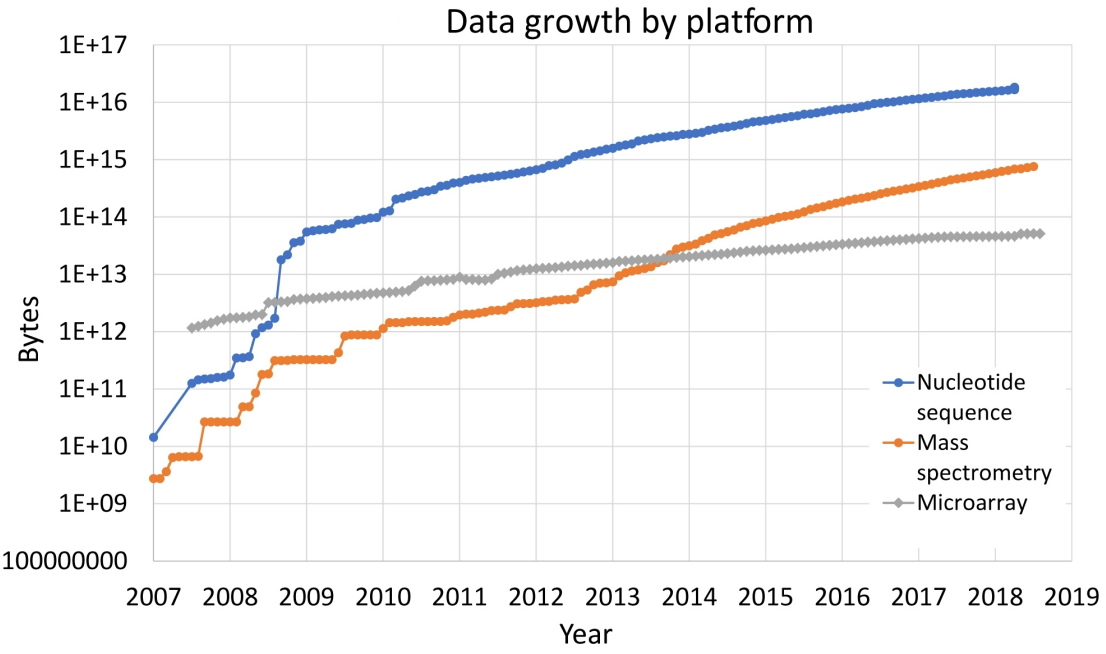


Figure 1.1: Data growth at EMBL-EBI by experimental platform. Figure courtesy of C.E. Cook *et al.* [3].

1.1.1 Data and Databases

As a direct consequence of being generated by a multitude of experimental methods and instruments, biological data distinguishes itself by its high heterogeneity [4]. Distinct high-throughput technologies emerged as references in the different omics fields, such as next-generation sequencing in genomics or mass spectrometry in proteomics and glycomics. To be able to reliably store the raw data generated by biological experiments, distinct data formats were designed. While all formats for a given experimental approach tend to share the same core information, they can significantly differ in terms of metadata. Metadata is the supplementary information that describes the data, encapsulating all the relevant attributes about an experiment. It thus describes the materials, methods, instrumentation, and additional information that were used to generated the data. To define the minimum information that data formats should contain, collaborative information standards were successively established for each experimental approach. These specifications

ensure experimental result interpretability by regulating the mandatory metadata information and its formatting. Following the successful example of a guideline regulating the minimum information about a microarray experiment (MIAME) [5] released in 2001, numerous other minimum information standards were introduced to the scientific community. Nowadays, all main high-throughput omics techniques have their own standard, which are coordinated together by the minimum information about a biomedical or biological investigation (MIBBI) [6] project since 2008. The minimum information about a proteomics experiment (MIAPE) [7] was a notably important initiative, as there is a high heterogeneity and complexity in the field of mass spectrometry for proteomics.

To be easily accessible and exploitable by researchers, the knowledge extracted from biological data requires to be organized and stored in databases. According to their content, biological databases have been historically classified as either *primary* or *secondary* databases. Primary databases, also referred to as archival databases (or repositories), are constituted of sequence or structural data directly extracted from experimental results. Secondary databases, also referred to as curated databases, are constituted of data derived from the processing of the information stored in primary databases [8]. These databases rely on manual annotation by experts and computational algorithms to reach a high level of data curation. The information they provide is often derived from multiple sources while using ontologies as regulatory frameworks for data representation. Some databases can also be hybrid and present characteristics of both a primary and secondary database at the same time. Biological databases are also frequently classified based on their scope. A database is thus deemed as *universal* when including many species, or as *specialized* when focusing on a single one. A multitude of databases emerged through the years to respond to the needs of modern biological research. While some act solely as repositories enabling the dissemination of raw experimental data, others have developed a panel of computer tools to access and further analyze their data. The 27$^{\text{th}}$ release of the NAR online molecular biology database collection [9] reported a total of 1,637 databases in 2019.

1.1.2 Experimental Reproducibility Crisis

In a 2016 survey from the *Nature* journal [10], a majority of the 1,576 interviewed researchers (52%) agreed that a significant reproducibility crisis is currently undergoing in science. Almost all of the respondents (90%) admitted that at least a slight crisis is occurring. An alarming number of experiments fail to be reproduced, and scientists can even struggle to replicate their own results. While the existence of a crisis is almost unanimous and sparked a lot of attention in recent years, scientists

are more divided on the factors that led to its inception. The reproducibility crisis is a complex multifaceted problem that is influenced by many causes, some of which are specific to each field of application. Among the many factors involved, selective reporting and the pressure to publish are deemed to be the biggest crisis contributors [10]. More technical causes are implicated, including bad experimental design and the unavailability of raw data, methods, or software. Poor statistical analysis is also a key factor, which worsened with the significant decrease in the production cost of data and its consequent large-scale generation [11]. Furthermore, the use of misidentified or contaminated cell lines is known to additionally hinder the reproducibility of experimental results [12].

1.2. Cell Lines

Since their first establishments in the middle of the last century, cell lines have grown to become an indispensable biotechnological tool for biomedical research. The numerous practical advantages they offer in comparison to traditional cell cultures led to their rapid adoption by laboratories. Cell lines are widely used as *in vitro* model systems for the study of numerous biological processes and diseases. In the pharmaceutical industry, cell lines play a major role in their ability to efficiently synthesize a large variety of drugs, hormones, and antibodies. In this section, we present cell lines and their defining traits, while discussing the spread of their misidentification and contamination and detailing the corresponding available genomics approaches to perform their authentication.

1.2.1 Characteristics

Primary cultures are cell cultures resulting from the growth of cells directly extracted from a tissue sample. When the primary cells constituting these cultures are passaged to a new culture vessel, they are referred to as *cell lines* [13]. Both cell lines and primary cells are valid *in vitro* models, bypassing the ethical and technical constraints of performing experiments on animals. In comparison to cell lines, primary cells represent model systems that are closer to their *in vivo* source and tend to produce more relevant biological results. Nonetheless, cell lines possess numerous technical advantages that led to their wide adoption in research. They proliferate faster and require less strict culture conditions, making them easier and less expensive to maintain. Since cell lines are derived from a single cell, they present the distinction of being uniform cell populations having in theory the same genetic makeup. Consequently, the experimental results produced on cell lines are prone to be more consistent and reproducible [14].

There are two main categories of cell lines based on their proliferation characteristics: *finite* and *continuous* cell lines. Finite cell lines, as well as primary cells, have a limited number of population doublings before reaching the *Hayflick limit* (Figure 1.2) and entering senescence [15]. This limit varies based on the cell type and species of origin, but it is usually comprised of between 40-60 cellular divisions [16]. Once the cells enter in senescence, they stop to proliferate despite remaining metabolically active [17]. Cells in this phase may survive for at least one year before dying [18]. Continuous cell lines, also referred to as *immortalized* cell lines, have or acquired the capability to proliferate indefinitely [15]. Thus, they provide an unlimited supply of cells having similar genotypes and phenotypes. In the case of stem cells, the immortality trait is innate and required for multicellular

organisms to function properly. Other continuous cell lines became immortal either spontaneously, in the case of cancer cells, or artificially through genetic engineering. Different approaches can be used to immortalize a cell line, namely irradiation, chemical mutagens, infection by transforming viruses, or recombinant DNA vectors expressing oncogenes [15]. Although not all cell lines are able to divide indefinitely, the term *cell line* is frequently used as a synonym to continuous cell lines.

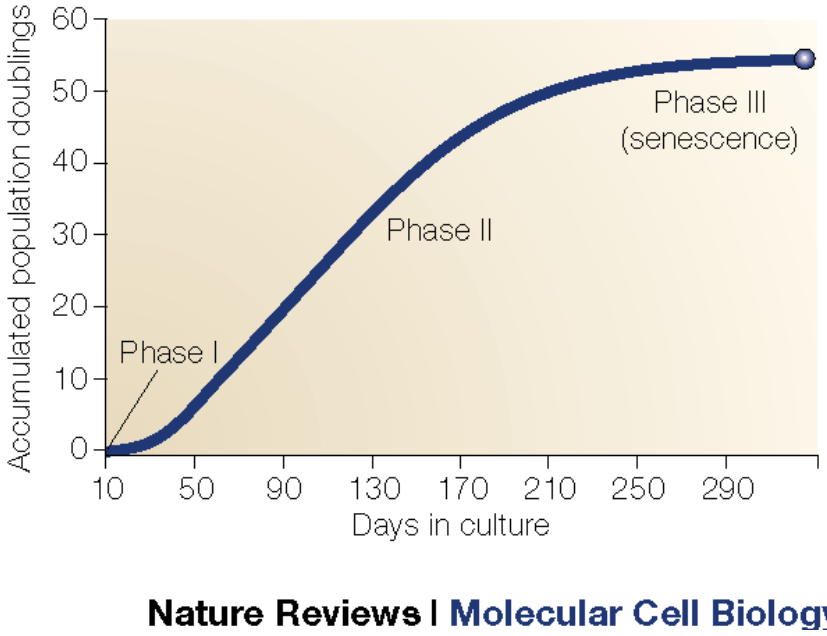


Figure 1.2: The three phases of a cell culture as described by L. Hayflick in 1961. They consist in the following: primary culture (phase I); exponential growth (phase II); and senescence (phase III). Figure courtesy of J.W. Shay *et al.* [18].

A large majority of the cell lines being used in research are cancer cells [19], whose most famous representative is the HeLa cell line. HeLa was the first human cell line to be established in 1951 by G.O. Gey [20] and quickly became used by laboratories all over the world. This cell line originated from the cervical cancer cells of Henrietta Lacks, who was a female patient at Johns Hopkins Hospital. It remains nowadays the most used cell line in terms of published scientific articles. Nonetheless, more specialized cell lines were also engineered through the years, such as hybridomas [21] for the production of monoclonal antibodies or induced pluripotent stem (iPS) cells [22, 23] which enables stem cell therapies. As a downside of their aggressive growth characteristics and resilience, cancer cell lines have been shown to be frequently involved in the contamination of other cell cultures.

1.2.2 Contamination and Misidentification

Cell line contamination is a long-lasting problem that appeared shortly after cell lines started to be used in research. It arises when a biological contaminant is accidentally introduced in a cell culture [17]. The contaminant can be either another cell line (cross-contamination) or a microorganism (microbial contamination). In the case of microbial contamination, the introduced microorganism (usually bacteria, fungi, or viruses) can greatly compromise experiments and the validity of their results [24]. The most frequent microbial contaminant is the *Mycoplasma* bacteria [24], which was shown to significantly alter cell physiology and metabolism [25]. Various techniques can be used to detect a Mycoplasma contamination, such as detection of prokaryote 16S rRNA by polymerase chain reaction (PCR) amplification, selective growth on a broth/agar culture, or fluorescent staining [26]. Mycoplasma testing is increasingly being mandated in order to be able to publish results when an experiment was performed on cell lines.

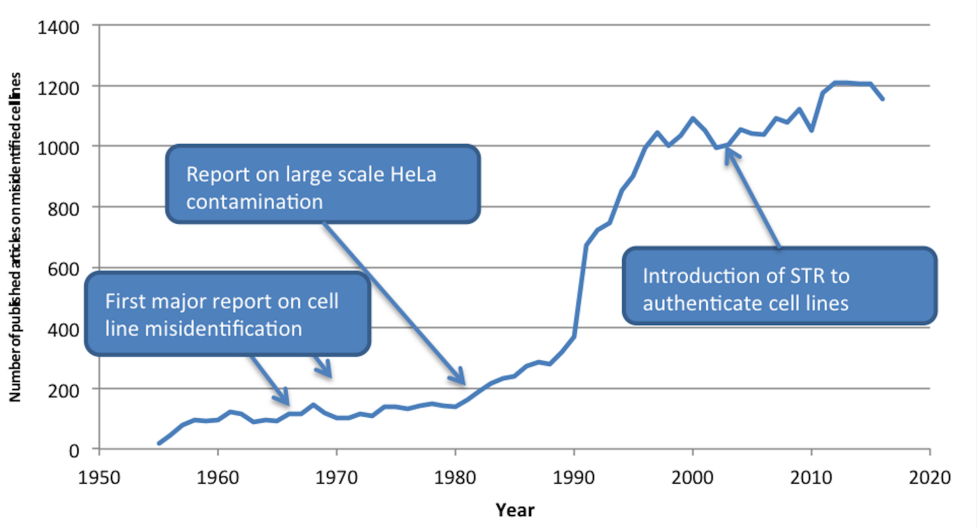


Figure 1.3: Contaminated literature over the years. Figure courtesy of S.P. Horbach *et al.* [27].

Cross-contamination occurs when a foreign cell line is introduced in a cell culture without the user's knowledge [17]. As the contaminant grows, the cell culture will contain a mixture of the two distinct cell lines. When the contaminant is a more aggressive cancer cell line, the authentic culture can get fully outgrown and replaced after several passages [28]. Cell lines in such cases are misidentified since they are thought to be authentic while having a completely different origin, no longer corresponding to their original donor or species. Cell line misidentification can also occur as the result of mislabeling, unwitting swap during manipulations, and improper freezer inventory [17]. Cell lines having errors in their tissue type, cell type,

or disease are referred to as misclassified cell lines. Several causes were singled out as potential sources of cross-contamination. For instance, improper technique from laboratory staff and sharing of reagents or culture medium between cell lines can introduce a contaminating cell line [17]. Surprisingly, cross-contamination is often an early event occurring in the original laboratory where the cell line was established [17]. This implies that if the cell line was already contaminated when it was transferred to other laboratories, no intact stock of the cell line exists anymore. In contrast, if the contamination occurred later, authentic stocks of the cell line can be retrieved by going back to the source.

Contaminating cell line	Cellosaurus AC	Cell description	Contaminated cell line count
HeLa	CVCL_0030	human cervical adenocarcinoma	121
T24	CVCL_0554	human bladder carcinoma	20
M14	CVCL_1395	human melanoma	18
HT-29	CVCL_0320	human colon carcinoma	16
U-937	CVCL_0007	human lymphoma, histiocytic	12
K-562	CVCL_0004	human leukemia, chronic myeloid, blast crisis	10
OCI-AML-2	CVCL_1619	human leukemia, acute myeloid, M4	8
PC-3	CVCL_0035	human prostate carcinoma, acute myeloid, M4	8
CCRF-CEM	CVCL_0207	human leukemia, acute lymphoblastic, T cell	7
JURKAT	CVCL_0065	human leukemia, acute lymphoblastic, T cell	7
HCu-10[1]	CVCL_M850	human esophageal squamous cell carcinoma	7
SW480[2]	CVCL_0546	human colon carcinoma	7
SW620[2]	CVCL_0547	human colon carcinoma	7
Hep-G2	CVCL_0027	human liver, hepatoblastoma	5
TPC-1	CVCL_6298	human thyroid, papillary carcinoma	5
U-251MG	CVCL_0021	human glioblastoma	5
UM-SCC-1	CVCL_7707	human oral squamous cell carcinoma	5

[1] Could also be HCu-18, HCu-22, HCu-27, HCu-33, HCu-37, or HCu-39 as they all share the same genetic identity

[2] SW480 and SW620 come from the same donor and share the same genetic identity

Table 1.1: Main contaminating cell lines sorted by the number of cell lines they affected. Only the cell lines with at least five cell lines contaminated are reported. Data courtesy of the ICLAC register of misidentified cell lines [17].

Cell line misidentification was pointed out as a major contributor to the inability to reproduce research results [12]. A recent study [27] estimated that over 30,000 scientific publications are based on data produced using misidentified cell lines. This number even increases to half a million when taking into account the publications citing those studies. The register of misidentified cell lines of the International Cell Line Authentication Committee (ICLAC) [17] reports 529 known misidentified cell

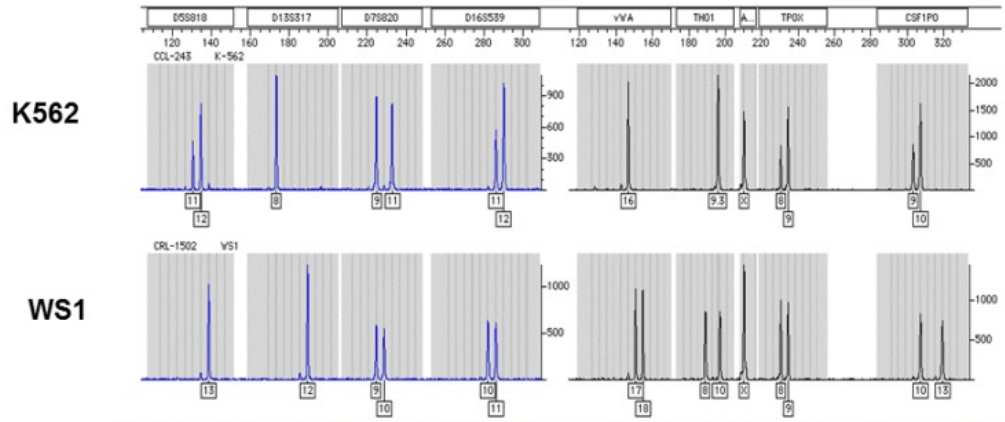
lines (version 9, October 2018). Human cancer cell lines are the most frequent contaminants with HeLa being, by far, the biggest contributor to this problem (Table 1.1). Although it means that only a very small fraction of the established cell lines is misidentified (~0.5%), they are widely used in research. Older estimations reported that between 18% and 36% of cell lines used in research may be misidentified [29]. Unfortunately, the awareness of the widespread of cell line contamination did little to prevent the publication of studies based on such cell lines (Figure 1.3). Even worse, there never has been so much contaminated literature submitted than in the past few years, despite the call for action over the years by the scientific community [30, 31]. To tackle the issue, an increasing number of journals are recommending, or even mandating, to test for cell line authenticity before being able to submit research results [32].

1.2.3 Authentication

The history of cell line authentication is closely related to that of cell line contamination. In the early 1960s, karyotyping and immunological techniques were used to detect inter-species contamination cases [33]. These approaches were, however, lacking the power of discrimination necessary to reliably identify contamination between cell lines from the same species [34]. In 1966, S. Gartler was able to distinguish human cell lines by applying a method based on isozyme expression [35]. He demonstrated that 18 human cell lines of distinct origins had been contaminated by HeLa cells. With the advent of molecular biology, new techniques based on DNA fingerprinting were developed. Short tandem repeat (STR) profiling arose as the recommended method to authenticate human cell lines. It was notably the subject of an American National Standards Institute standard (ANSI/ATCC ASN-0002-2011) entitled "Authentication of Human Cell Lines: Standardization of STR Profiling" [36], established by the American Type Culture Collection (ATCC) Standards Development Organization (SDO) workgroup. Single nucleotide polymorphism (SNP) profiling is an available alternative to STR profiling. SNPs being the most frequent sequence variations found in the human genome, they form collectively a unique combination specific to an individual [37]. This approach can prove to be particularly useful to identify cell lines having microsatellite instability as a consequence of mutations in the DNA mismatch repair (MMR) system. SNP profiling is however lacking a standard or guideline regulating the approach, and no database enabling the search of its profiles is currently available. With the development of next-generation sequencing approaches, partial or complete genome sequencing can also be used to authenticate cell lines. However, genome sequencing remains nowadays too time and cost expensive to be deployed on a large scale as an authentication methodology.

1.2.4 STR Profiling

The inception of the STR profiling method is linked to the discovery of a hypervariable locus in the human genome by A. Wyman and R. White in 1980 [38]. It had the peculiarity of having numerous distinct alleles that were inherited in a Mendelian fashion. This highly polymorphic locus would turn out to be the fist minisatellite to be reported. Minisatellites belong to the category of variable number tandem repeats (VNTR), as they are composed of a variable number of DNA motifs (6-100 base pairs long) that are repeated in tandem. In 1985, A.J. Jeffreys *et al.* demonstrated that these minisatellite regions are capable of hybridizing to numerous loci throughout the human genome to produce DNA fingerprints [39, 40]. One year earlier, P. Weller and A.J. Jeffrey described a short tandem repetitive sequence upstream of the myoglobin gene [41]. These shorter VNTR (2-6 base pairs long) would later on be referred to as microsatellites [42], but are also known under the name of short tandem repeats (STR). Of note, there is no consensus on the exact length of minisatellites and microsatellites, and the repeat length difference is their main distinction.



	D5S818	D13S317	D7S820	D16S539	vWA	THO1	Amel.	TPOX	CSF1PO
K562	11, 12	8	9, 11	11, 12	16	9.3	X	8, 9	9, 10
WS1	13	12	9, 10	10, 11	17, 18	8, 10	X	8, 9	10, 13

Figure 1.4: Electropherogram of the K562 (CVCL_0004) and WS1 (CVCL_2766) cell lines. The STR analysis was performed using the PowerPlex® 1.2 STR kit. Figure courtesy of Y. Reid *et al.* [43].

In parallel, the polymerase chain reaction (PCR) technique was developed by K. Mullis *et al.* in 1983 [44], enabling the amplification of specific DNA sequences using oligonucleotide primers. Multiple variants of the method emerged in the following years, including the Multiplex PCR in which multiple distinct sequences can be synchronously amplified. The STR profiling method consists in the application

of Multiplex PCR on selected microsatellite loci. The original workflow described by A. Edwards *et al.* in 1991 [45] did not change much in its core concept over time. In this approach, DNA is first extracted from a biological sample. Fluorescent-labeled PCR primer pairs targeting specific STR loci are then added. The corresponding DNA sequences are amplified in parallel using Multiplex PCR, before being separated using gel electrophoresis. Of note, capillary electrophoresis has nowadays replaced gel electrophoresis because of its higher resolution, enabling the separation of the bands produced by modern STR kits. The generated fluorescence is recorded to produce an electropherogram (Figure 1.4), which is then interpreted as an STR profile by translating the length of the amplified DNA sequences to the number of repeats at a given locus. This step is commonly performed using dedicated software followed by manual validation. Finally, the resulting STR profile is identified by its comparison to reference STR profiles.

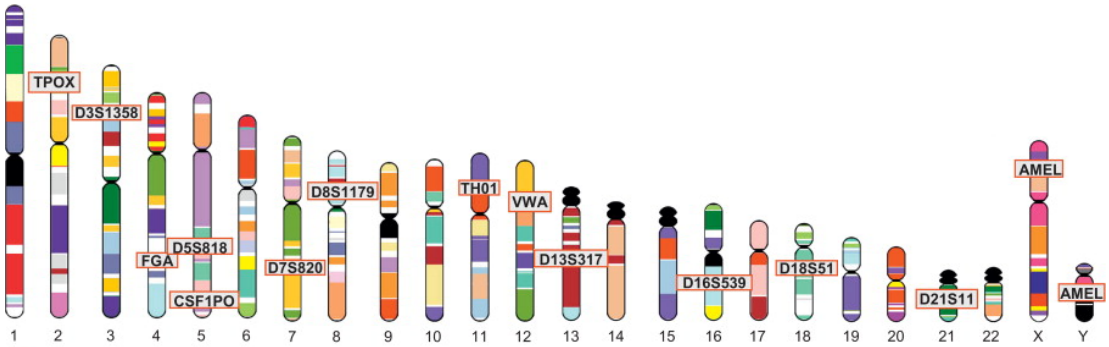


Figure 1.5: The 13 core CODIS STR loci and their chromosomal position. Figure adapted from W.T. Godbey [46].

At its beginnings, STR profiling was exclusively used for forensic applications and paternity testing. The approach was commonly used to identify victims in mass disasters [47, 48] or to resolve crime cases [49]. In 1997, the Federal Bureau of Investigation (FBI) established the combined DNA index system (CODIS) database, which consisted of a validated set of thirteen core STR loci (CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, TH01, TPOX, vWA, plus amelogenin for gender identification; Figure 1.5) [50]. It is important to note that not all STR markers are equal in terms of power of discrimination, as they can significantly vary in terms of allele diversity, enotype diversity, and heterozygosity (Table 1.2). Some have consequently a lower probability of identity ($P_{(ID)}$) than others, representing the probability that two given individuals have the same genotype at this locus, making them more capable to distinguish samples from one another.

The application of STR profiling to human cell line authentication was proposed by J.R. Masters *et al.* in 2001 [51]. In an effort to tackle cell line cross-contamination,

STR locus	Alleles	Genotypes	Heterozygosity	Probability of identity
SE33	52	304	0.9353	0.0066
Penta E	23	138	0.8996	0.0147
D2S1338	13	68	0.8793	0.0220
D1S1656	15	93	0.8890	0.0224
D18S51	22	93	0.8687	0.0258
D12S391	24	113	0.8813	0.0271
FGA	27	96	0.8745	0.0308
D6S1043	27	109	0.8494	0.0321
Penta D	16	74	0.8552	0.0382
D21S11	27	86	0.8330	0.0403
D8S1179	11	46	0.7992	0.0558
D19S433	16	78	0.8118	0.0559
vWA	11	39	0.806	0.0611
F13A01	16	56	0.7809	0.0678
D7S820	11	32	0.7944	0.0726
D16S539	9	28	0.7761	0.0749
D13S317	8	29	0.7674	0.0765
TH01	8	24	0.7471	0.0766
Penta C	12	49	0.7732	0.0769
D2S441	15	43	0.7828	0.0841
D10S1248	12	39	0.7819	0.0845
D3S1358	11	30	0.7519	0.0915
D22S1045	11	44	0.7606	0.0921
F13B	7	20	0.6911	0.0973
CSF1PO	9	31	0.7558	0.1054
D5S818	9	34	0.7297	0.1104
FESFPS	12	36	0.7230	0.1128
LPL	9	27	0.7027	0.1336
TPOX	9	28	0.6902	0.1358
DYS391	7	7	NA	0.4758

Table 1.2: Variation observed across 1036 U.S. population samples with 29 autosomal STR loci and the Y-STR locus DYS391. The CODIS 13 core STR markers are shown in purple. Data courtesy of Promega, available online at: https://ch.promega.com/products/pm/genetic-identity/population-statistics/power-of-discrimination/#table1 (accessed the 11th of November 2019).

the authors demonstrated that cell lines could be authenticated using a set of six core CODIS STR loci (D8S1179, D18S51, D21S11, FGA, TH01, vWA, plus amelogenin). The number of STR markers was later on increased to eight to improve the discrimination power, and discussions are currently held in the cell line community to further increase it to the thirteen CODIS loci. STR profiling was chosen as the international recommended method for cell line authentication based on several reasons. Authenticating a cell line using STR profiling is a cheap and fast process, and the corresponding workflow is simple to set up. The produced STR profiles are also easy to interpret and analyze. Finally, the STR kits have been used in forensic sciences for a long time, and they are consequently well-validated and commercially available.

1.2.5 Data and Databases

There are two core online resources gathering information about cell lines: the Cell Line Data Base (CLDB) [52] developed by B. Parodi and her group from the IRCCS Ospedale Policlinico San Martino, and the Cellosaurus [53] developed by A. Bairoch from the SIB CALIPHO group. Both databases share the same purpose of characterizing the cell lines that are, or were, commonly used in biomedical research. While CLDB tends to provide more detailed information about the cell line biochemical properties and growth conditions, the Cellosaurus encompasses a wider range of data categories [53]. Their main difference lies in the scope, CLDB storing information about 6,643 cell lines from 205 distinct species (version 4.0.201701) while the Cellosaurus contains information about 117,636 cell lines (87,495 human, 20,817 mouse and 2,131 rat cell lines) from 669 distinct species (version 33, December 2019). Unfortunately, no new update was released to the CLDB since January 2017. More specialized cell line databases are also available, although some have been discontinued, as discussed in [53].

The experimental data resulting from STR profiling consists essentially in a set of STR loci associated to their corresponding detected allele values. This simple key-value pair format led to the common use of tabular or PDF files to report STR profiling data, in which the columns represent the STR loci and each row represents the alleles of distinct samples (or reversely). The need for a dedicated standard format was thus less pressing compared to the high-throughput omics fields, explaining mainly its current absence. Tabular files are however notably unpractical to store metadata information in a standardized manner. STR profiles are publicly accessible from cell line collections, such as the American type culture collection (ATCC) or the German collection of microorganisms and cell cultures (DSMZ), and from the results of individual scientific publications. These public STR profiles

are additionally gathered and exposed by cell lines databases. Both the Cell Line Integrated Molecular Authentication (CLIMA) database of the CLDB and the Cellosaurus store STR profiles for researchers to access.

1.3. Proteomics

Proteomics is a multidisciplinary field that consists in the comprehensive study of the *proteome*, which is defined as the complete repertoire of proteins expressed by a cell, tissue, or organism at a given time [54]. Research in proteomics relies on high-throughput technologies and strategies to identify and quantify proteins from complex biological samples, while characterizing their functions, structures, expression patterns, interactions, and modifications. In this section, we present the core characteristics of proteins and the main mass spectrometry-based proteomics approaches for their identification and quantification.

1.3.1 Protein Complexity

Proteins are large linear biomolecules constituted of amino acid residues linked together by peptide bonds. They are biosynthesized by ribosomes in the rough endoplasmic reticulum through the translation of the sequence information conveyed by messenger RNA (mRNA) molecules. The mRNA sequences are transcripted from genes contained in DNA, and a single gene can produce multiple transcripts from the alternative splicing of exons in eukaryotes. This leads to distinct protein isoforms that differ in terms of sequence and structure, while sometimes having new biological functions, which drastically increase protein diversity [55]. There are 20 common amino acids that can be classified based on the physico-chemical properties of their side chains. Amino acids can thus be charged (positively or negatively), polar or hydrophobic. Consequently, the amino acid sequences directly affect the secondary and tertiary structures of proteins, influencing in turn their biological functions. The side chains of amino acids can also undergo post-translational modifications (PTMs), further increasing proteome complexity by their modulation of protein activity, turnover, localization, and interactions [56]. More than 300 eukaryotic PTMs have been observed experimentally [57], some of the most common being phosphorylation, glycosylation, acetylation, and methylation. The emergence of genetic variation, the most common form of which is single nucleotide polymorphism (SNP), can also deeply affect the properties and structural conformations of proteins when their sequence is altered. Additionally, proteins rarely act alone and often operate in complex plastic network of interactions regulating their functions [58]. All these factors together make the analysis of protein notoriously challenging, requiring the development and refinement of specialized experimental techniques and efficient computer software.

1.3.2 Mass Spectrometry

Following improvements in the instrumentation, mass spectrometry (MS) progressively arose as the reference proteomics approach for the analysis of complex protein samples [59]. The core principle of MS is based on the measurement of the mass-to-charge (m/z) ratio of ionized analytes in a gas phase. This analysis is performed by a mass spectrometer instrument, which is composed of three main components: an ion source that volatilize and ionize the analytes (i); a mass analyzer that separates the ionized analytes based on their measured m/z ration (ii); and a detector that records the number of ions for each detected m/z value (iii) [59]. The generated data consists of mass spectra, which are two-dimensional plots representing the ion intensity depending on the m/z ratio. The MS analysis can be performed either on intact proteins following *top-down* strategies or on digested peptides following *bottom-up* strategies (also known as *shotgun* proteomics) (Figure 1.6). Peptide MS is usually preferred since protein MS is less sensitive and the mass of intact proteins does not provide enough information for their reliable identification [59]. After or during protein isolation by biochemical fractionation and chromatography, a digestion step is thus frequently performed through the use of enzymes such as trypsin. The resulting peptides have a positively charged C-terminus, which is helping their subsequent sequencing. The *middle-down* hybrid strategy consisting in the MS analysis of large polypeptides is also available, providing more PTM coexistence information than bottom-up proteomics while being more accessible than top-down proteomics [60].

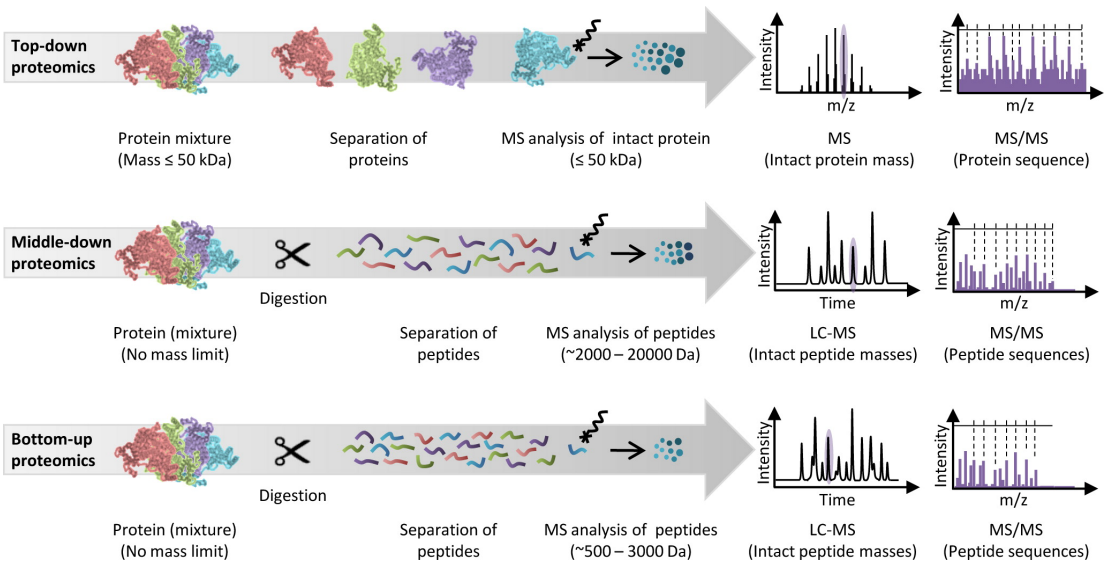


Figure 1.6: Overview of the main proteomics strategies. Figure courtesy of L. Switzar *et al.* [61].

The advent of mass spectrometry in proteomics is closely related to the emergence of new soft ionization techniques that enabled the analysis of biomolecules such as proteins in the late 1980s [62]. Electrospray ionization (ESI) [63] and Matrix-assisted laser desorption/ionization (MALDI) [64] successively arose as methods for volatilizing and ionizing biomolecules [65]. The main differences between the two techniques lie in the state of the sample and the source of the ionization, which in turn influence their applications. In the ESI technique, an analyte solution flowing through thin capillaries is subjected to a strong electric field, under atmospheric pressure, to generate an electrospray. In the MALDI technique, the analytes are embedded in a dry crystalline matrix and get sublimated and ionized by short laser pulses. Both analytical techniques possess high sensitivity and can thus be applied to low concentration samples [66]. Since the sample is a solution in ESI, it can easily be coupled with liquid-based separation techniques such as liquid chromatography (LC) or more recently high-performance liquid chromatography (HPLC). This property makes ESI the preferred approach for quantification experiments [66]. Another distinction is that MALDI tends to produce singly charged ions, while ESI tends to produce multiply charged ions and has consequently an increased mass range. Overall, ESI LC/MS has emerged as the reference for the analysis of complex protein samples in shotgun proteomics experiments, while MALDI is being used in more specific applications such as protein imaging [66].

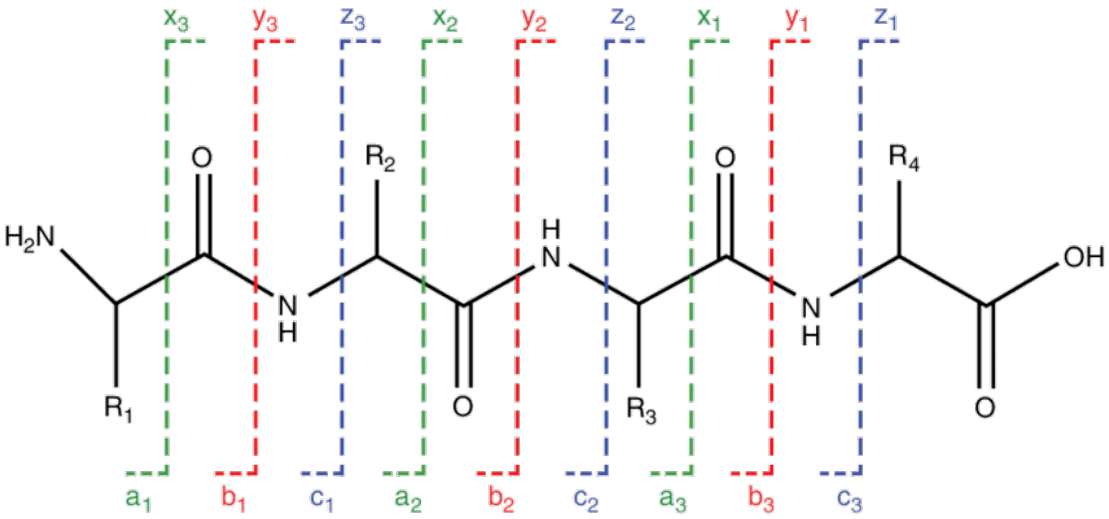


Figure 1.7: Ions produced from peptide fragmentation. Figure courtesy of Z. Hao *et al.* [67].

Further increasing the diversity of MS instrumentation, four main distinct types of mass analyzers were engineered through the years: the time-of-flight (ToF),

Fourier-transform ion cyclotron resonance (FTICR), quadrupole, and ion trap instruments. Since each mass analyzer is based on different physical principles, they vary in terms of their analytical performance [68]. Multiple mass analyzers can however be combined to take advantage of the strengths of each technology. While MALDI is mainly paired with ToF analyzers for the measurement of the mass of intact peptides [59], ESI has been commonly used with a wider variety of instruments. Single-stage MS is however insufficient to fully resolve peptide sequences. In tandem mass spectrometry (MS/MS) approaches, the ions isolated by the first mass analyzer (MS1), referred to as parent or precursor ions, are selected for a fragmentation step before being analyzed by a second mass analyzer (MS2). The exact location at which a peptide bond linking two amino acids will break depends on the fragmentation technique in use, thus determining the types of ions produced (Figure 1.7). Collision-induced dissociation (CID) has historically been the preferred fragmentation method in MS proteomics. In CID, the precursor ions are fragmented through their collision with an inert gas. More fragmentation techniques based on different principles appeared in recent years, with the electron-capture dissociation (ECD) in 1998 [69] and electron-transfer dissociation (ETD) in 2004 [70].

1.3.3 Data and Databases

MS instruments generate raw data in the form of large binary files that are often encoded in proprietary formats. In addition to the recorded MS/MS spectra, these formats also contain a wide range of metadata information describing the data acquisition process. To be exploitable by most bioinformatics software, raw data requires to be converted into open formats using tools such as MSConvert from the ProteoWizard toolkit [71]. The generated files are usually formatted either in text-based peak list formats (e.g., MGF, PKL and DTA) or in XML-based formats (e.g., mzData, mzXML, and more recently mzML). The tools for analyzing raw data produce processed data, which can also be formatted in various data formats. Most of these formats are XML-based, and are used to store either the identification results (e.g., mzIdentML) or quantification results (e.g., mzQuantML) from proteomics experiments along with the relevant metadata information.

Following the example of other omics fields such as genomics, research in proteomics progressively became more open and community-centered. Despite initial reluctance, the public sharing of data generated by MS-based proteomics experiments, whether raw data or processed data, became the norm over time. It is even starting to be mandatory by an increasing number of journals and editors to be able to publish an article, in order to ensure the reproducibility of the experimental results presented. Several repositories were established through the years to store

such MS data, such as PRIDE [72], MassIVE and jPOST [73]. These projects which were formerly completely independent are now coordinated in the framework of the ProteomeXchange consortium [74], ensuring of the proper and efficient sharing and dissemination of proteomics data between resources and users.

For a more in-depth review of the proteomics databases and data formats, please refer to the chapter I wrote for the *Elsevier Encyclopedia of Bioinformatics and Computational Biology* entitled "Proteomics Data Representation and Databases" (see the Section A.1 of the Appendix).

1.3.4 Search Strategies

An MS/MS spectrum represents the fragmentation pattern of a peptide precursor ion into product ions, obtained from the successive cleavage of its backbone. To be able to interpret and resolve the corresponding peptide sequence, the peaks in the spectra require to be annotated as ions. This process can be performed either manually or through the use of bioinformatics tools, but the large amount of MS/MS spectra generated by a single high-throughput experiment makes manual interpretation impractical. Two main strategies are available: database search (i) and spectral library search (ii). In database searches, the experimental MS/MS spectra are compared against theoretical ones that were predicted from a protein (or more rarely peptide) sequence database [75]. In the case of protein databases, the sequences are first digested *in silico*, following the cleavage pattern of specific digestion enzymes. Digestion enzymes being known to sometimes miss cleavage sites they were expected to cut, peptides with missed cleavages also require to be predicted at this step in order to be observable. Numerous bioinformatics tools were developed through the years to perform database searches, such as Mascot [76], Andromeda [77], and X!Tandem [78]. Search engines are based on more or less sophisticated scoring algorithms (usually normalized dot product or variations of it) and vary in terms of performance. However, they all share the same core principles. Search engines thus report the best peptide-spectrum matches (PSMs) for a given experimental MS/MS spectrum along with various scores and statistical parameters describing the quality of the spectral comparison. In spectral library searches, the experimental MS/MS spectra are compared against consensus spectra that were compiled using previously identified spectra. Some examples of spectral library search engines are SpectraST [79] and X!Hunter [80].

One of the main shortcomings of MS-based protein and peptide identification lies in the fact that a large proportion of matches are spurious and occurred by chance. To limit the presence of these false positive identifications in the results, a false discovery rate (FDR) threshold is usually applied following the target-decoy strat-

egy. In the case of database searches, a decoy version of each target protein stored in the database is usually computed by reversing the amino acid sequence. In the case of spectral library searches, the decoys are commonly generated by shuffling the peaks of the consensus MS/MS spectra. After the search, algorithms based on the score distribution of the target and decoy identifications are applied to compute the FDR at the protein, peptide and PSM level. Based on these FDR values, thresholds are applied so as to limit the proportion of false identifications to an expected percentage.

1.3.5 Quantitative Proteomics

In addition to the precise identification of proteins from complex biological samples, MS is abundantly used to perform the absolute and relative quantification of proteins and peptides. These analyses reveal a supplementary layer of information about the quantitative state of a proteome, providing insights about protein differential expression [81]. Nowadays, the quantification of proteins through untargeted approaches is commonly performed using isotopic labeling. Since the added labels introduce a known mass difference, mass spectrometers are able to distinguish the proteins of a labeled sample from those of an unlabeled sample. This thus enables to perform the relative quantification of proteins by comparing their abundance between distinct states or conditions. Numerous labeling techniques are available, including metabolic labeling (e.g. stable isotope labeling by amino acids in cell culture (SILAC)), isotopic labeling (e.g., isotope-coded affinity tag (ICAT)) and isobaric tagging (e.g. tandem mass tags (TMT)). With the recent improvements in instrumentation, the popularity of label-free quantification (LFQ) techniques increased. At the expense of lower quantitative accuracy, these techniques are cost and time effective by having simplified experimental designs and sample preparation steps. They also depend more on computer software and statistical validation to produce accurate results. Two distinct LFQ approaches are available: spectral counting and ion peak intensity quantification. Spectral counting is based on the rationale that an increase in protein abundance results in an increase in the number of corresponding peptides [82]. Consequently, a protein with a higher expression will tend to have better peptide coverage and a larger total number of PSMs. Quantification by ion peak intensity relies on the signal intensities of ions, which was shown to be correlated to the ion abundance. In comparison to spectral counting, ion peak intensity is only compatible with MS1. Absolute quantification (AQUA) of proteins and peptides can be achieved with the use of selected reaction monitoring (SRM) targeted approaches [83]. In this technique, labeled synthetic copies of peptides of interest are introduced at a known concentration during the digestion step to determine their abundance.

1.3.6 Proteoform Analysis

Proteoforms are defined as all the protein products that are derived from a unique gene, encompassing splice variants (isoforms), sequence variants and PTMs [84]. To be able to observe any alteration to the canonical protein sequences using MS, these changes to the amino acid sequences require to be included in the databases used by search engines. Protein isoforms are already searched in a large majority of proteomics experiments, and can easily be retrieved in the FASTA format along with canonical sequences from knowledge databases such as UniProtKB/Swiss-Prot [85] (many species) or neXtProt [86] (human only). In the case of SNPs and other small sequence variants such as indels, the altered protein sequences are not directly available. Protein databases annotate the changes at the protein level without including them in sequences. Of note, neXtProt enables to retrieve the protein sequences in the extended FASTA format (PEFF) [87] where the SNPs (and/or PTMs) are detailed in the FASTA header. While this new enhancement of the FASTA format allows including the known SNPs and PTMs in searches, only a limited number of search engines such as Comet [88] currently support this format. In most experiments focusing on SNPs, additional bioinformatics tools have thus to be used to generate the altered protein sequences. With the democratization of whole genome sequencing (WGS) and whole exome sequencing (WES), new proteogenomics approaches can also be used. The sequence variants observed at the gene level allow the prediction of their effect on protein sequences using specialized tools. This notably enables to generate customized protein sequence databases in which variants specific to a disease, tissue or cell line are included.

Two distinct approaches are commonly used in proteomics to detect PTMs and determine their location on proteins: targeted modification searches and open modification searches. In targeted approaches, the PTMs of interest are specified to the search engine prior to the database search step. The selection consists in specifying which amino acid can be modified, the mass of the modification, and if the modification is ubiquitous or not (fixed or variable modification). Targeted approaches are efficient but require prior knowledge about the PTMs expected to be present in the sample. Furthermore, selecting too many variable modifications drastically increase the search space, which in turn affects the FDR and the peptide and protein identifications. In open searches, the modifications are inferred from the mass difference between two spectra using spectral library searches. This enables the identification of both novel and known PTMs, but requires the use of specialized algorithms and software. Sample preparation is also known to introduce chemical artifacts that have no biological relevance. These chemical modifications further increase the complexity of PTM analysis. Unlike most PTMs, glycans are however unable to be fully identified using conventional search strategies. Because of their

great mass and structure diversity, specific MS approaches had thus to be developed to resolve the complexity of their structures.

1.4. Glycomics

Glycomics is an emerging discipline that focuses on the study of the *glycome*, which is defined as the complete repertoire of glycans expressed by a cell, tissue, or organism at a given time [89]. Glycans are large branched biomolecules essential to the normal function of cells while being directly involved in numerous diseases. They have important structural functions, such as the formation of the glycocalyx coat of eukaryote cells and the capsule of bacteria [90]. The ubiquitous presence of glycans on the cell surface gives them an important role in host-pathogen interactions and in the viral infection process. As nearly all cell surface receptors are glycosylated proteins, glycans are essential for the immune system and modulate the function of antibodies [91]. They are for instance well-known in the determination of the blood types by their presence on the surface of red blood cells. Glycans also assist the folding of newly synthesized proteins in the endoplasmic reticulum and regulate their trafficking. Through their recognition by specialized glycan-binding proteins (GBPs), glycans regulate many signaling and secretory pathways [90]. Altered glycosylation is additionally a known hallmark of cancer, allowing some glycans to be used as biomarkers of tumor progression [92]. In this section, we present the inherent diversity of glycans and the challenges that emerge in their analysis by mass spectrometry approaches.

1.4.1 Glycan Diversity

Glycans, also frequently referred to as carbohydrates or saccharides, are large linear or branched compounds constituted of multiple monosaccharides that are linked by glycosidic bonds. Since monosaccharides cannot be hydrolyzed into simpler forms [93], they represent the monomer units enabling the assembly of oligosaccharides (fewer than a dozen units) or polysaccharides (more than a dozen units) [94]. They are therefore the equivalent of nucleotides for nucleic acids or amino acids for proteins. Monosaccharides can be in linear or cyclic form and are generally constituted by a backbone of three to nine carbon atoms. Hexoses, such as glucose or galactose, are composed of six carbon atoms and represent the most common monosaccharides [95]. The hydroxyl and amino groups of monosaccharides can be further modified by enzymes, leading to an increase in structure diversity [96]. Two distinct stereoisomers (D and L) of the same monosaccharide can coexist based on the three-dimensional orientation of the atoms, but the D configuration is more frequent [96]. This stereoisomerism can modulate their biological properties. Different sites are available on monosaccharides to form glycosidic bonds, and the linkage itself can take two distinct orientations (α or β) depending on the relative stereochemistry of

the C1 carbon compared with the plane of the ring [97]. There is formation of an α-glycosidic bond when the two linking carbons have the same anomeric configuration and of a β-glycosidic bond when they have different anomeric configurations. The orientation of glycosidic bonds directly affects the structure of glycans.

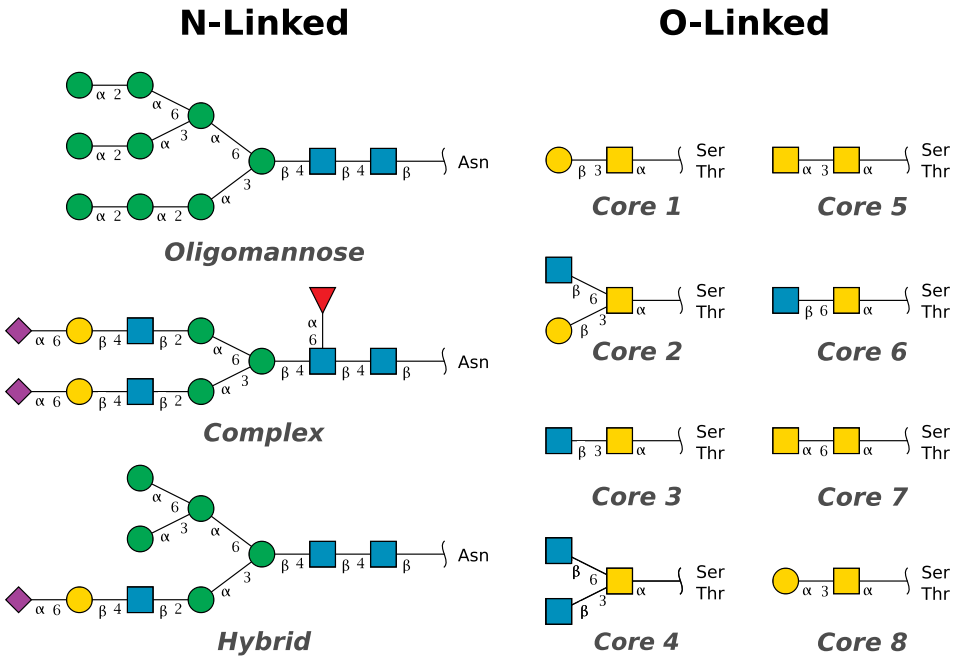


Figure 1.8: N-linked and O-linked glycans in the SNFG format.

Glycans can be either covalently bonded to proteins (glycoproteins) or lipids (glycolipids) to form glycoconjugates [93], or free when released in body fluids by those glycoconjugates [98]. In the case of glycoproteins, there are two main categories of glycans based on the type of linkage: N-linked and O-linked glycans. N-linked glycans consist in the attachment of an oligosaccharide to the amine nitrogen of the side chain of an asparagine (Asn) residue [99]. The N-glycosylation sites usually follow a Asn-X-Ser/Thr pattern, where X is any amino acid with the exception of proline [99]. In eukaryote organisms, N-glycans all share a common core sequence defined as Manα1-3(Manα1-6)Manβ1-4GlcNAcβ1–4GlcNAcβ1 [99]. They can be further subdivided into three categories based on their structure (Figure 1.8): oligomannoses (i) when the core is solely extended by mannose (Man) residues; complexes when the core is extended by antennas starting with a N-Acetylglucosamine (GlcNAc) residue (ii); and hybrid when the Manα1-3 arm of the core is extended by one or two GlcNAc residues and the Manα1-6 arm is extended by Man residues (iii) [99]. O-linked glycans consist in the attachment of an oligosaccharide to the oxygen atom of the side chain of generally a serine (Ser) or threonine (Thr) residue [100]. O-glycans all start with a N-Acetylgalactosamine (GalNAc) monosaccharide that is mostly elongated into eight different core structures (Figure 1.8), which can be in turn further elongated and modified [101].

1.4.2 Mass Spectrometry

In comparison to the characterization of proteins, additional challenges are encountered in glycomics when aiming to characterize glycan structures using MS-based approaches. A given precursor mass can only provide reliable information about the composition of a glycan, failing to fully resolve the corresponding structure [102]. As they have the same mass, isomeric structures and isomeric monosaccharides cannot be directly distinguished. Through the use of MS/MS approaches, the fragmentation of glycans can provide supplementary information about their primary sequences. The additional use of orthogonal techniques, such as nuclear magnetic resonance (NMR), is however often required to reduce the number of possible glycan structures through the determination of glycan configurations and linkage positions [102]. In comparison to the traditional CID or HCD fragmentation, electron activation fragmentation techniques, such as ECD or ETD, were shown to result in more detailed glycan structures [102]. In a typical protein-based glycomics MS experiment, the glycans are first released from their proteins either enzymatically for N-glycans using peptide N-glycosidase F (PNGase F) or chemically for O-glycans using β-elimination [103]. In glycoproteomics approaches, this step is not performed in order to be able to identify glycosylation sites inside the amino acid sequences of intact glycoprotein and glycopeptides. Site localization is obtained at the cost of precise structural information, as only the glycan composition can be fully resolved [102]. As for glycomics, ECD and ETD fragmentation techniques nowadays tend to be preferred in glycoproteomics experiments as they provide overall more information about glycan structures. They also mainly fragment the peptide backbone instead of glycosidic bonds, preventing the complete loss of glycans and their corresponding site information [103].

1.4.3 Data and Databases

Once fully resolved by glycomics approaches, the glycan structures require to be encoded into standardized data formats to reliably store the corresponding sequence information. As no clear community standard was originally available, numerous glycan databases and initiatives ended up developing their own internal format. This led to the coexistence of a multitude of glycan structure formats (Table 1.3), which can vary regarding the level of detail they encapsulate. They are also differentiated by the way the sequence information is encoded. Some formats like GLYDE [104] are XML-based, while others like KCF [105] are formatted as indented tables. The majority of formats, such as LINUCS [106] or Linear Code [107], consists however in a one-line string. While those string formats tend to produce very condensed

results, they can prove to be difficult to read. This is especially true in the case of branched structures. Another issue is that each format uses a distinct notation for the encoded monosaccharides. The MonosaccharideDB [108] database notably addresses this problematic by listing the existing notations and providing translation tools. While tools enabling the conversion of a given glycan format to another exist, they usually only support a restricted number of input and output formats. Recently, GlycoCT [109] was selected by several databases as a default format in an effort to improve the information exchange and cross-references between glycoinformatics resources. This however did not prevent new formats, such as WURCS [110], from being developed.

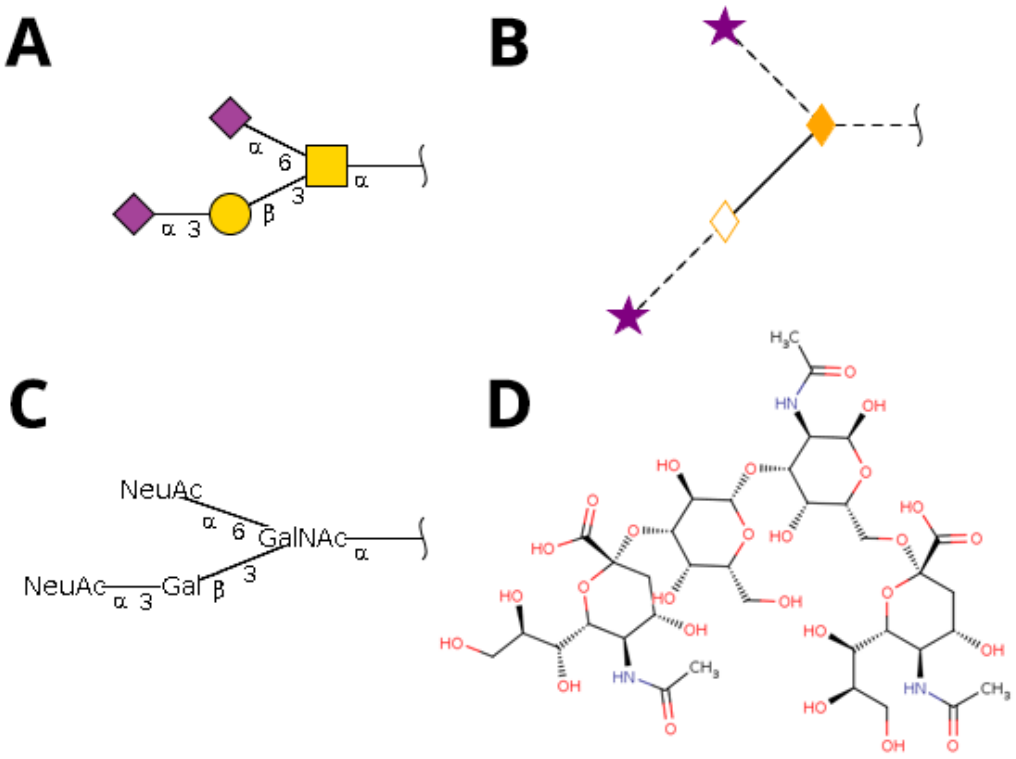


Figure 1.9: Graphical representation of glycans. Example of a glycan structure (GlyConnect: 2641; GlyTouCan: G01614ZM) represented in four distinct formats: SNFG (a); Oxford (b); text (c); and as its chemical structure depiction (d).

With the advent of MS-based glycoproteomics approaches, there has been a consequential increase in the number of observed glycans that lack a fully resolved structure. For these cases, the structure formats detailed above cannot be used since solely the monosaccharide composition is fully known. Once more, no consensus was originally achieved and distinct glycan composition formats were developed by databases and tools working with such data. Glycan composition formats all consist in describing the observed monosaccharides and their number of occurrences

Format	Value
BCSDB	Ac(1-5)aXNeup(2-3)bDGalp(1-3)[Ac(1-2)aXNeup(2-6),Ac(1-2)]aDGalpN
GlycoCT	RES 1b:a-dgal-HEX-1:5 2s:n-acetyl 3b:b-dgal-HEX-1:5 4b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d 5s:n-acetyl 6b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d 7s:n-acetyl LIN 1:1d(2+1)2n 2:1o(3+1)3d 3:3o(3+2)4d 4:4d(5+1)5n 5:1o(6+2)6d 6:6d(5+1)7n
GLYDE	<?xml version="1.0" encoding="UTF-8?> <GlydeII><molecule subtype="glycan" id="From_GlycoCT_Translation"> <residue subtype="base_type" partid="1" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=a-dgal-HEX-1:5"> <residue subtype="substituent" partid="2" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=n-acetyl"> <residue subtype="base_type" partid="3" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=b-dgal-HEX-1:5"> <residue subtype="base_type" partid="4" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=a-dgro-dgal-NON-2:6|1:a|2:keto|3:d"> <residue subtype="substituent" partid="5" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=n-acetyl"> <residue subtype="base_type" partid="6" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=a-dgro-dgal-NON-2:6|1:a|2:keto|3:d"> <residue subtype="substituent" partid="7" ref="http://www.monosaccharideDB.org/GLYDE-II.jsp?G=n-acetyl"> <residue_link from="2" to="1" <atom_link from="N1H" to="C2" to_replace="O2" bond_order="1"></residue_link> <residue_link from="3" to="1" <atom_link from="C1" to="O3" from_replace="O1" bond_order="1"></residue_link> <residue_link from="4" to="3" <atom_link from="C2" to="O3" from_replace="O2" bond_order="1"></residue_link> <residue_link from="5" to="4" <atom_link from="N1H" to="C5" to_replace="O5" bond_order="1"></residue_link> <residue_link from="6" to="1" <atom_link from="C2" to="O6" from_replace="O2" bond_order="1"></residue_link> <residue_link from="7" to="6" <atom_link from="N1H" to="C5" to_replace="O5" bond_order="1"></residue_link> </molecule></GlydeII>
IUPAC	NeuAc(a2-3)Gal(b1-3)[NeuAc(a2-6)]GalNAc
KCF	ENTRY G00027 Glycan NODE 5 1 *Ser/Thr* 13 0 2 *GalNAc* 3 0 3 *Gal* -5 -4 4 *Neu5Ac* -6 4 5 *Neu5Ac* -13 -4 EDGE 4 1 *2:a1* 1 2 *3:b1* 2:3 3 *4:a2* 2:6 4 *5:a2* 3:3 ///
Linear Code	NNa3Ab3(NNa6)AN
LINUCS	[]\[a-D-GalpNAc\]{[(3+1)]\[b-D-Galp\]{[(3+2)]\[a-D-Neup5Ac\]{}}[(6+2)]\[a-D-Neup5Ac\]{}}
WURCS	2.0/3,4,3/[a2112h-1a_1-5_2*NCC/3=O][a2112h-1b_1-5][Aad21122h-2a_2-6_5*NCC/3=O]/1-2-3-3/a3-b1_a6-d2_b3-c2

Table 1.3: Common glycan sequence formats. Example of a glycan structure (GlyConnect: 2641; GlyTouCan: G01614ZM) represented in different data format that can be found in glycomics.

as key-value pairs (e.g., *Hex:5 HexNAc:4 NeuAc:2*). While these formats are inherently less complex than glycan structure formats, they share the same issue of having distinct notations for monosaccharides.

The graphical representation of glycan structures also required the design of specialized formats (Figure 1.9). In addition to the classical chemical structure depiction commonly preferred by chemists, new formats emerged using specific encoding for the monosaccharide appearance. While the most basic one uses text abbreviations, the Oxford [111] and SNFG [112] formats opted for symbols varying in shape and color to represent the various monosaccharide classes. In contrast to the glycan sequences formats, the SNFG format was rapidly adopted by a large number of databases and related resources and is notably recommended by the *Essentials of Glycobiology* [113].

Database	Content	URL
CFG	Glycan array data	http://www.functionalglycomics.org
CSDB	Plant, fungal and bacterial glycan structures	http://csdb.glycoscience.ru
GlycoEpitope	Glycan epitopes	https://www.glycoepitope.jp
GlyConnect	Glycan structures, glycoproteins, glycosylation sites	https://glyconnect.expasy.org
GLYCOSCIENCES.de	3D glycan structures, NMR data	http://www.glycosciences.de
GlyTouCan	Glycan structures	https://glytoucan.org
KEGG	Glycan structures, pathways	https://www.genome.jp/kegg/glycan/
MonosaccharideDB	Monosaccharide data	http://www.monosaccharidedb.org
SugarBind	Pathogen-glycan binding data	https://sugarbind.expasy.org
UniCarb-DB	Glycan MS/MS data	https://unicarb-db.expasy.org
UniCarbKB	Glycan structures, glycoproteins, glycosylation sites	http://www.unicarbkb.org

Table 1.4: Main glycan databases and repositories. The resources that were discontinued or focus on topics that are too specifc (e.g., enzymes) are not listed.

Databases in glycomics differ significantly in terms of goals and content (Table 1.4). While some focus solely on describing the structures and compositions of glycan, others aim to characterize their interactions with proteins or pathogens. GlyConnect [114] is a database aiming to characterize the different molecular actors of protein glycosylation. It was developed at the SIB Proteome Informatics Group (PIG) and is openly accessible on the ExPASy server (https://glyconnect.expasy.org). The database specifically covers the N- and O-linked glycans and the glycoproteins that harbor them. The data stored in GlyConnect can be accessed through any of the nine available categories: structures (i); compositions (ii); proteins (iii); peptides (iv); sites (v); taxonomies (vi); tissues (vii); diseases (viii); and references (ix). This enables users to perform multi-criteria queries answering specific questions, such

as finding all the glycan structures that have been reported in a given disease. GlyConnect is built upon the content of the GlycoSuiteDB database [115], following the transfer of the license to the SIB. The database is consequently constituted of a manually curated data core containing numerous glycan structures. With the rise of high-throughput glycoproteomics approaches, a large amount of data was subsequently added to GlyConnect. While a lot of information was gained about the location of glycosylation sites in protein sequences, the proportion of glycan for which only the monosaccharide composition is known drastically increased. As a result, there is a growing need to develop new computer software to help to make sense of this data.

1.5. Objectives and Thesis Overview

The work presented in this thesis was carried out from November 2015 to March 2020. This PhD was co-supervised by the Proteome Informatics Group (PIG) and the Computer Analysis and Laboratory Investigation of Proteins of Human Origin (CALIPHO) group, both part of the University of Geneva (UNIGE) and the Swiss Institute of Bioinformatics (SIB). It started with the development of new bioinformatics tools and workflows for the mass spectrometry analysis of protein post-translational modifications and sequence variants. The goal of the PhD later on expanded to the development of bioinformatics tools for the scientific community to make sense of omics data, with a focus on data related to cell lines. Reflecting the corresponding fields of research of the two groups, the type of data analyzed ranged mainly from DNA fingerprinting data, to mass spectrometry proteomics and glycoproteomics data. While the concerned fields vary significantly in terms of purpose and scope, the underlying bioinformatics challenges regarding biological data format, storage, and subsequent computer-aided processing and interpretation are similar. As summarized in Figure 1.10, three distinct bioinformatics tools (MzVar, CLASTR, and GlyConnect Compozitor) were developed in the course of this thesis, two of which have been published in scientific journals. Two additional publications were written based on studies performed with the help of two of these tools.

1.5.1 Cell Line Authentication using STR

Chapter 2 presents the Cellosaurus STR similarity search tool named CLASTR (Cell Line Authentication using STR). While the experimental protocols and methods to generate STR profiles from a cell line sample are well defined and available, the bioinformatics tools to interpret such data were lacking. An STR profile can only allow the authentication of a cell line through its comparison to a reference. Over the years, the Cellosaurus grew to become the largest database in terms of human publicly available human STR profiles but was lacking a dedicated tool to fully exploit this data. CLASTR is an important component in the international efforts targeted toward curbing the spread of contaminated cell lines in scientific research.

This tool resulted in the publication of an article entitled ”CLASTR: the Cellosaurus STR Similarity Search Tool - A Precious Help for Cell Line Authentication” in the Cancer Genetics and Epigenetics section of the *International Journal of Cancer*, which was generously put in open access for free.

1.5.2 STR Profiling Search Parameters

Chapter 3 presents an extensive analysis of the influence of the different parameters available for STR similarity searches on the identification of related cell lines. This includes the search algorithm, the count of homozygous loci as one or two alleles, the inclusion of amelogenin in the scoring computation, and the number of STR markers compared. A good set of parameters is expected to lead to high scores for related and low scores for unrelated cell lines, allowing reliably distinguishing them for authentication purposes. Scientific journals are unanimously recommending to ensure cell line authenticity throughout experimental procedures. This measure requires the establishment of guidelines to be able to reach its full potential, enabling non-experts to access more easily to STR profiling.

The results of this analysis are detailed in an article draft entitled ”Evaluation of the Effect of STR Profiling Search Parameters on Cell Line Authentication”.

1.5.3 Reanalysis of HeLa Proteomics Data

Chapter 4 presents a large-scale reanalysis of 40 tandem mass spectrometry data sets from the HeLa cell line. This study aimed to identify SNPs inferred from genomic sequencing data at the protein level, while trying to further assess their impact on protein expression and stability through changes in relative protein abundance. The project included the development and use of a tool named MzVar, enabling the generation of customized variant protein and peptide databases using variant calling data. In the scope of the Human Protein Project, special attention was paid to the identification of peptides from *missing proteins*. Overall, this study showcases the benefits, as well as the many challenges and technical difficulties, of reanalyzing publicly available proteomics data at a large scale.

The results of this analysis were published in an article entitled ”Large-Scale Reanalysis of Publicly Available HeLa Cell Proteomics Data in the Context of the Human Proteome Project” in the *Journal of Proteome Research* as part of the Human Proteome Project 2018 special issue.

1.5.4 GlyConnect Compozitor

Chapter 5 presents Glyconnect Compozitor, the latest addition to the visual toolkit of the Glyconnect database. With the advent of high-throughput glycoproteomics studies focusing on the precise mapping of glycosylation sites, a growing proportion

of experiments does not fully resolve glycan structural information. There is consequently an accumulation of glycan data for which only the composition is known. As a result, there is a critical need to develop new bioinformatics tools and approaches to exploit such data. Glyconnect Compozitor generates glycan composition graphs from various available search criteria, enabling to detect and export the compositions that are in common between several biological entities or specific to a single one.

This tool and its usage are detailed in an article entitled "Examining and Fine-Tuning the Selection of Glycan Compositions with GlyConnect Compozitor", which was published in the *Molecular & Cellular Proteomics* journal.

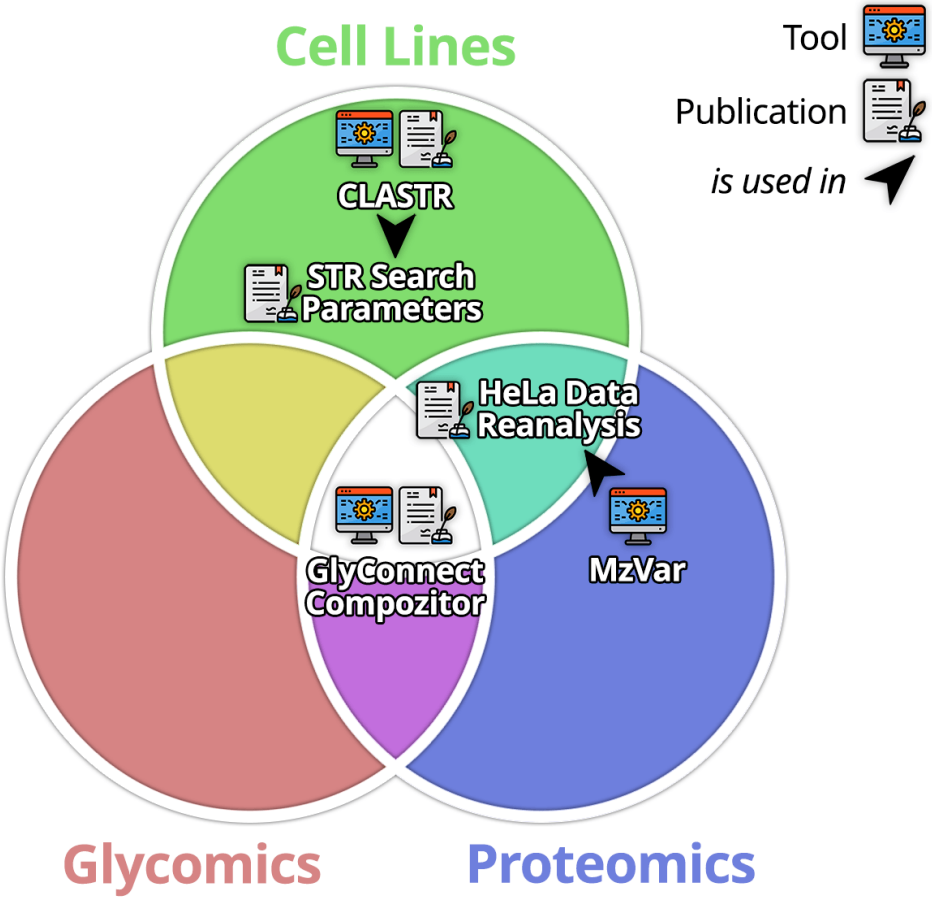


Figure 1.10: Overview of the work achieved during the thesis. Three distinct software applications were developed, two of which having been described in a dedicated publication. Two additional studies could be published through their use.

Chapter 2

Cell Line Authentication using STR

2.1. Overview

In this chapter, we introduce CLASTR (Cell Line Authentication using STR), a new web application that performs similarity searches on the STR profiles stored in the Cellosaurus. While similar tools have been available for a while, they are fairly limited with respect to the features and search parameters they offer. Additionally, these tools are based on their own STR profile databases, which are often too restrictive and failing to include STR profiles available in the literature from individual publications. The Cellosaurus compiles all publicly available STR profiles and as such, offers the largest library of reference STR profiles to the scientific community. CLASTR is an important step in making cell line authentication accessible for everyone, providing both an intuitive web interface and an efficient API. It directly contributes to curbing the propagation of unreliable experimental results produced with contaminated or misidentified cell lines. Interestingly, the *International Journal of Cancer* generously waived the open access fees for the publication describing CLASTR in recognition of the definite advantage of such a tool for the whole community.

IJC
International Journal of Cancer

CLASTR: The Cellosaurus STR similarity search tool - A precious help for cell line authentication

Thibault Robin [1,2,3,4], Amanda Capes-Davis [5] and Amos Bairoch [1,2]

[1]CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva, Switzerland
[2]Microbiology and Molecular Medicine Department, Faculty of Medicine, University of Geneva, Geneva, Switzerland
[3]Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva, Switzerland
[4]Computer Science Department, Faculty of Sciences, University of Geneva, Geneva, Switzerland
[5]CellBank Australia, Children's Medical Research Institute, The University of Sydney, Westmead, NSW, Australia

Despite an increased awareness of the problematic of cell line cross-contamination and misidentification, it remains nowadays a major source of erroneous experimental results in biomedical research. To prevent it, researchers are expected to frequently test the authenticity of the cell lines they are working on. STR profiling was selected as the international reference method to perform cell line authentication. While the experimental protocols and manipulations for generating a STR profile are well described, the available tools and workflows to analyze such data are lacking. The Cellosaurus knowledge resource aimed to improve the situation by compiling all the publicly available STR profiles from the literature and other databases. As a result, it grew to become the largest database in terms of human STR profiles, with 6,474 distinct cell lines having an associated STR profile (release July 31, 2019). Here we present CLASTR, the Cellosaurus STR similarity search tool enabling users to compare one or more STR profiles with those available in the Cellosaurus cell line knowledge resource. It aims to help researchers in the process of cell line authentication by providing numerous functionalities. The tool is publicly accessible on the SIB ExPASy server (https://web.expasy.org/cellosaurus-str-search) and its source code is available on GitHub under the GPL-3.0 license.

Introduction

The Cellosaurus[1] is a knowledge resource on cell lines. It aims to describe all cell lines used in biomedical research. It currently contains more than 113,000 cell line entries from 625 species. The majority of the cell lines originate from two species, namely, human (75%) and mouse (17%). For each cell line, a wealth of information is provided, allowing researchers to easily get an idea of how the cell line was generated and its main characteristics. It provides cross-links to 88 different external resources (cell repository catalogs, ontologies, databases) and provides more than 105,000 references to 17,700 publications. The Cellosaurus is updated regularly and is available on the Swiss Institute of Bioinformatics (SIB) ExPASy web server[2] (https://web.expasy.org/cellosaurus) where users can browse the database or download the full set of data in three different formats (structured text, OBO and XLM).

One of the many purposes of the Cellosaurus is to make users aware of cell lines that are known or suspected to be cross-contaminated, misidentified or misclassified. Contaminated cell lines arise from the accidental introduction of foreign cell lines (cross-contamination) or microorganisms (microbial contamination). Misidentified cell lines are the result of errors in their gender or species, while cell lines are defined as misclassified when the tissue type, cell type or disease is incorrect. Cell line cross-contamination was singled out as a major contributor to the reproducibility crisis that has been recently highlighted in life sciences.[3] A recent study[4] estimated that over 30,000 scientific publications were based on data produced using misidentified cell lines. As a consequence, the results of such experiments are partially or totally unreliable. This situation has led journal editors and publishers to ask authors to authenticate the cell lines that they have used prior to publication. The preferred experimental approach to authenticate a cell line is the short-tandem repeat (STR) profiling method (Fig. 1) that had already proven its effectiveness in forensic applications.[5,6] Once the STR profile of a given cell line sample is obtained, it must be compared against a database of reference STR profiles to verify that it does not have an unexpectedly high similarity with an unrelated cell line. If that is the case, it could potentially mean that the cell line sample is either partially or completely contaminated or misidentified.

Author contributions: CLASTR concept: Robin T, Bairoch A and Capes-Davis A. Code writing: Robin T. Project supervision and context development of the Cellosaurus collected the STR profiles used by CLASTR: Bairoch A. Manuscript writing with support from Capes-Davis A: Robin T and Bairoch A.
Key words: authentication, cell lines, cell culture, contamination, misidentification, STR profiling
Conflict of interest: None declared.
DOI: 10.1002/ijc.32639
History: Received 9 Jul 2019; Accepted 14 Aug 2019; Online 23 Aug 2019
Correspondence to: Thibault Robin, Microbiology and Molecular Medicine Department, Faculty of Medicine, University of Geneva, Switzerland, Tel.: +41-22-379-02-40, Fax: +41-22-379-11-34, E-mail: thibault.robin@unige.ch

What's new?

Despite increased awareness, cell line cross-contamination and misidentification remain a major source of erroneous experimental results in biomedical research. Nowadays, researchers performing experiments on cell lines are thus expected to ensure their authenticity using short-tandem repeat (STR) profiling. The Cellosaurus, which compiles all publicly available STR profiles, has become a valuable knowledge resource for this purpose. However, the database lacked a dedicated tool allowing a similarity search for a query STR profile. Here, the authors present CLASTR, the Cellosaurus STR similarity search tool that aims to facilitate the authentication process and the detection of potentially cross-contaminated or misidentified cell lines.

All currently available cell line STR search tools are restricted in the number and scope of the STR profiles that are available in their database (see Table 1, next section), and a given sample would have to be tested against multiple data sets to ensure its authenticity. By compiling all publicly available STR profiles, the Cellosaurus provides a remedy for this issue. However, the database lacked a dedicated tool allowing a similarity search for a query STR profile until recently. Here we present CLASTR (Cell Line Authentication using STR), the Cellosaurus STR similarity search tool, which aims to provide a large panel of functionalities to facilitate the similarity search process.

Materials and Methods

STR profiles

The current Cellosaurus release (release July 31, 2019) contains STR profiles for 6,556 distinct cell lines (6,474 human, 46 mouse and 36 dog cell lines) from 444 different sources providing, by far, the largest publicly available data set in terms of the number of STR profiles (Table 1). More than two-thirds of these human cell lines have only a single source for their STR profile (Fig. 2). For cell lines with more than one source, these frequently disagree on the exact allele value of a given STR marker. Such "conflicting" markers have their different alleles and corresponding sources clearly labeled on the STR profile section of a Cellosaurus entry. Collectively, individual scientific publications constitute the largest source for the Cellosaurus STR profiles. Cell line collections such as ATCC, CLS, DSMZ, ECACC, JCRB, KCLB, RCB and TKG, and initiatives such as the COSMIC cell line project[7] are also major contributors (Fig. 3).

STR markers

The first multiplex STR amplification kits were limited in terms of the number of amplified STR markers, usually containing

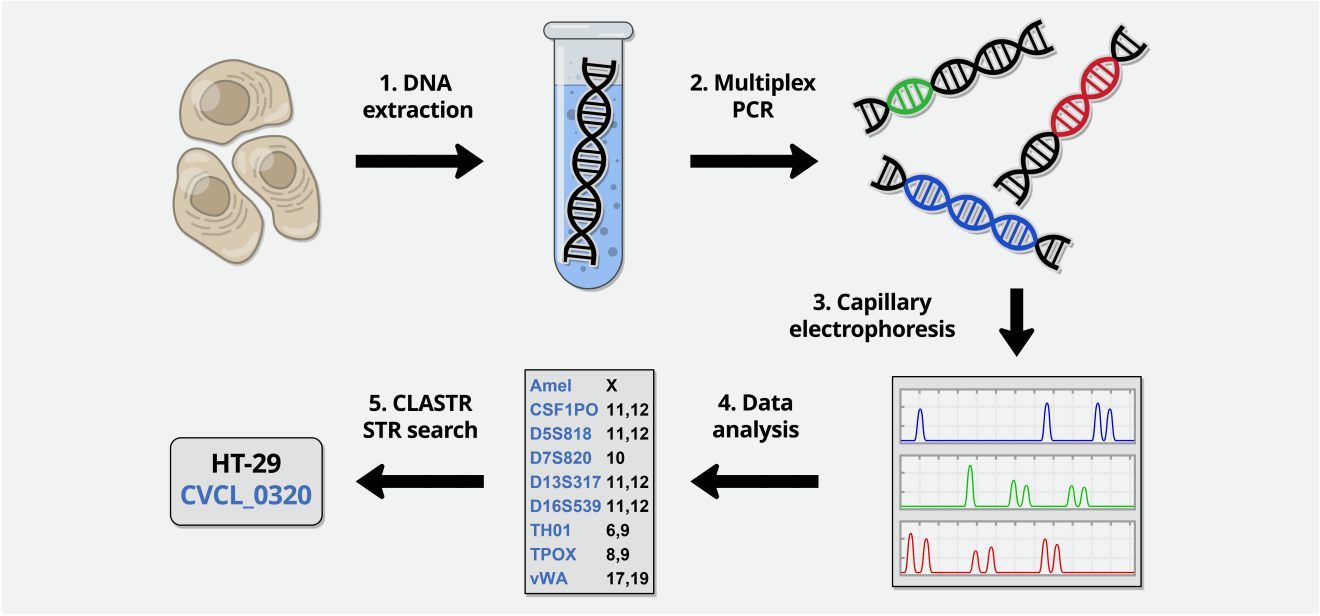


Figure 1. Workflow of cell line authentication by STR profiling. The process of the authentication of a cell line using STR profiling can be summarized by the following steps: (1) DNA is extracted from a cell sample; (2) fluorescent-labeled primers targeting specific STR loci are added and the corresponding DNA sequences are amplified simultaneously though Multiplex polymerase chain reaction (Multiplex PCR); (3) capillary electrophoresis is used to separate the amplified DNA fragments and the fluorescence is recorded to produce an electropherogram; (4) the electropherogram is interpreted as a STR profile by converting the size of each amplified fragment to the number of repetitions at each locus using specific software and controls, with manual validation as required; (5) the STR profile is searched against the Cellosaurus using CLASTR, allowing to know the identity of the cell sample and detect a potential cross-contamination. [Color figure can be viewed at wileyonlinelibrary.com]

Table 1. Comparison of the publicly available human STR profile data sets

Database	ATCC STR profile database	Cellosaurus	CLIMA	COG single record STR database	DSMZ STR profile database	NCBI BioSample
Number of human cell line with a STR profile	1,626	6,474	4,354	3,380	2,455	3,083

only four markers plus amelogenin used for gender determination. Although these studies continue to be informative, the number of STR loci used was insufficient to discriminate between cell lines from different donors. Over time, their limitations became clear, leading to the publication of an American National Standards Institute standard (ANSI ASN-0002-2011) for the authentication of human cell lines by STR profiling[8] in 2011. This standard requires the use of eight STR markers (CSF1PO, D13S317, D16S539, D5S818, D7S820, TH01, TPOX and vWA) plus amelogenin. In recent years, a large panel of multiplex STR systems has become commercially available, with a great variability in the STR markers and their numbers. The largest kits can amplify up to 27 STR markers at a time.

As a result, the STR markers constituting the STR profiles contained in the Cellosaurus can vary drastically between two entries. Currently, 32 distinct human STR markers (including amelogenin) are allowed in the database, although only 17 of them are commonly used in cell line STR genotyping.

It should also be noted that there have been some efforts recently to develop sets of STR markers for nonhuman species, specifically for dog[9,10] and mouse.[11,12]

Results

CLASTR, the Cellosaurus STR similarity search tool provides an intuitive and reliable platform to perform similarity searches on the human STR profiles contained in the Cellosaurus resource. It allows the efficient identification of cell lines and thus helps to detect contaminated and/or misidentified samples. The tool provides a wide range of search options for the users to choose from while implementing many useful functionalities. CLASTR is freely accessible on the ExPASy web server (https://web.expasy.org/cellosaurus-str-search). As the current number of mouse (44) and dog (28) cell lines with a STR profile is very limited we did not implement the option to search for similarity across samples from these species, but this may change if, as expected, the number of such authenticated nonhuman cell lines significantly increases over time.

Comparison to similar tools

Five bioinformatics tools enabling the pairwise comparison of STR profiles have been developed over time. These are the ATCC STR Profiling Analysis (https://www.atcc.org/STR%20Database.aspx), CLIMA,[13] the COG Single Record STR Database Search (https://strdb.cccells.org), the DSMZ Online STR matching analysis (OSTRA)[14] and the Search Program for the STR profile database of the JCRB Cell Bank (https://cellbank.nibiohn.go.jp/legacy/str2/top.html). All these tools were designed to support a specific set of STR profiles and, consequently, cannot be applied on other data sets. Furthermore, their source code is not publicly available. As the Cellosaurus has

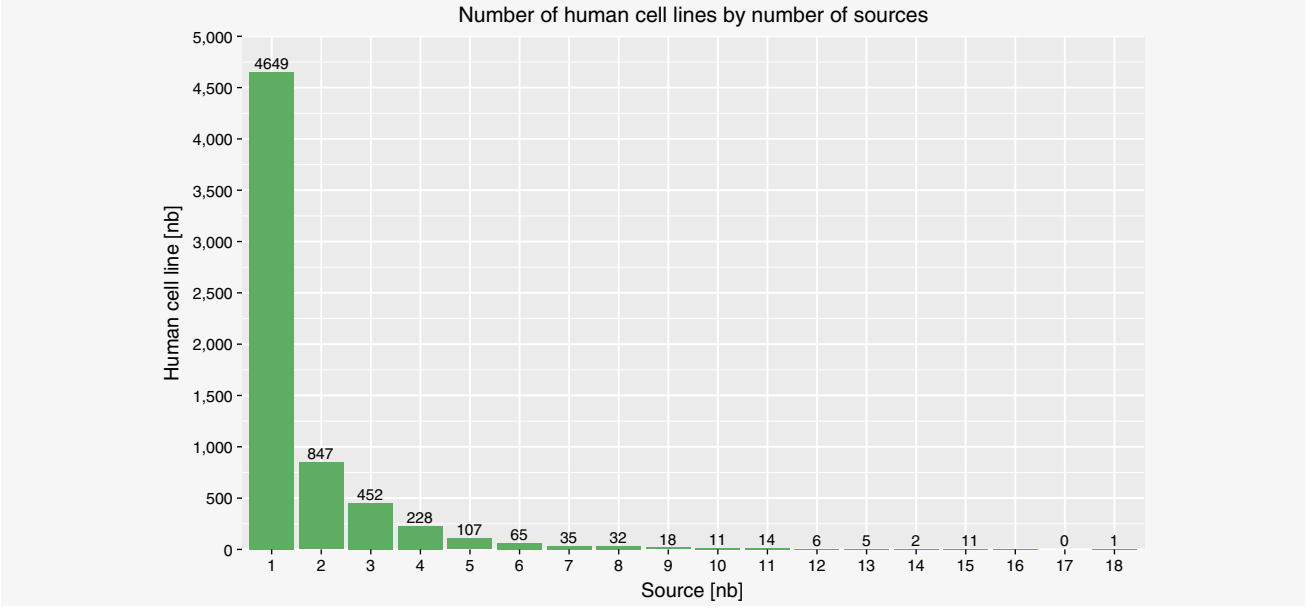


Figure 2. Number of human cell lines by number of sources. [Color figure can be viewed at wileyonlinelibrary.com]

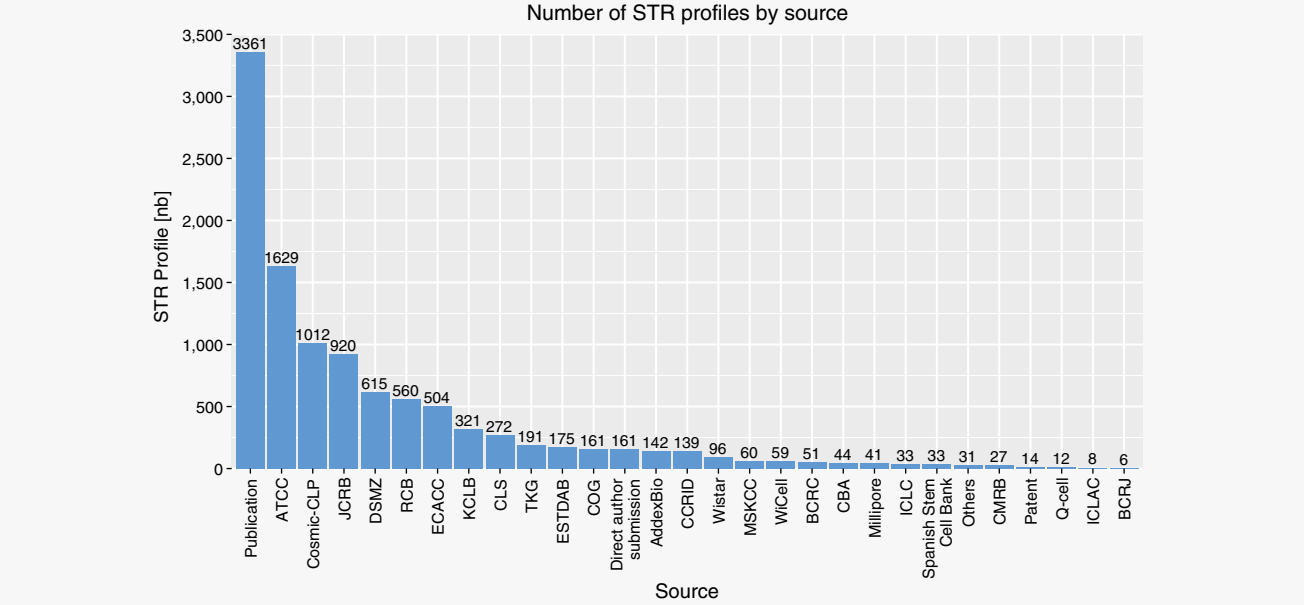


Figure 3. Number of STR profiles by source. [Color figure can be viewed at wileyonlinelibrary.com]

grown to become the largest public repository of human cell line STR profiles, the need for a tool enabling searching for similar STR profiles became more pressing and many users requested this feature. This prompted the development of CLASTR, which implements numerous features to facilitate and automate the search process.

While all previously available STR similarity search tools are based on the same core principle and share similar scoring algorithms, they differ from CLASTR in terms of user experience and interactivity (Table 2). The large majority of these tools do not allow the generated result table to be exported directly, preventing the storage of matching cell line information. In all these tools, allele information can only be entered manually, which can be tedious and prone to errors. Moreover, the absence of a public API or any other means to perform batch searches makes the process challenging and time-consuming when analyzing more than a few STR profiles. CLASTR is currently the software with the most features available and the added advantage of accessing the largest data set of human cell line STR profiles. The tool is designed to provide an easy, intuitive experience for experimental researchers and bioinformaticians.

Scoring algorithms

Two similar algorithms are frequently used to perform the pairwise comparison between a "query" and a "reference" STR profile: the Masters algorithm[15] and the Tanabe algorithm[16] (also known as the Sørensen–Dice coefficient). Both algorithms are based on the same concept, where a ratio is calculated based on the total number of alleles in each sample and the number of alleles that are shared by both samples. The Masters algorithm consists of the ratio between the number of shared alleles and the total number of alleles in the query (or in the reference for its modified version), while the Tanabe algorithm consists of the ratio between twice the number of shared alleles and the sum of the alleles in the query and reference. While these algorithms are fairly simple, they have been shown to be sufficiently

Table 2. Comparison of the features of the previously existing STR similarity search tools with CLASTR

Software	Available STR markers	Algorithms	Filters	Import file	Batch queries	Export formats	API
CLASTR	31	Tanabe and Masters	Score, min markers and max results	Yes	Yes	Excel (XLSX), CSV and JSON	REST
ATCC STR Profiling Analysis	8	Masters[1]	Score	No	No	CSV	No
CLIMA	8	Masters[1]	Score	No	No	None	No
COG single record STR database search	15	Masters[1]	Score	No	No	None	No
DSMZ Online STR matching analysis	8	Tanabe	Score	No	No	None	No
Search Program for the STR profile database of the JCRB Cell Bank	8	?[2]	?[2]	No	No	?[2]	No

[1]Modified version of the Masters algorithm only.
[2]This tool is no longer maintained and is nonfunctional, thus we could not check the status of these features.

effective to discriminate related from unrelated cell lines, provided that enough STR markers were compared.[17] Both the Masters and Tanabe algorithms are implemented in CLASTR and can be selected when performing a search. As the Tanabe algorithm is symmetrical and produces the same score if the query and reference cell lines are swapped, we have selected it as the default algorithm.

Scoring modes

Although the scoring algorithms are precisely described in terms of the score computation itself, they do not define a default behavior in the case of missing allele data for one of the two STR profiles to be compared. Such a problem is particularly relevant to the Cellosaurus data set because it contains STR profiles originating from many different sources that vary in the number and extent of the analyzed STR markers. To address this problem, we implemented different scoring modes. By default, only the STR markers for which both the query and reference have allele data are included in the score computation. However, as an option, the user can choose to compute the score based on all the query markers, even if the reference is lacking allele data for some of the markers. The reverse option is also available.

It is important to note that both algorithms do not define if the homozygous STR loci should count as one or two alleles in the score computation. Based on feedback from members of the International Cell Line Authentication Committee (ICLAC) and from experts in cell line STR profiling we decided to count homozygous STR loci as one allele. This choice is motivated by the fact that many cell lines present abnormal karyotypes and are no longer diploid. The chromosome counts can vary even between cells of the same culture. By counting homozygous loci as one, we do not falsely imply a specific number of chromosomes that could turn out to be often erroneous.

Conflicting STR profiles

Another consequence of the great diversity of sources of STR profiles in the Cellosaurus is that they may disagree on the exact allele value of a given STR marker. Since the similarity search is performed as a pairwise comparison, only one version of a given conflicting STR marker can be searched at a time. Cell lines that contain one or more conflicting STR marker values need to be handled differently than those without any conflicts. By default, our tool will try to resolve the conflicts by grouping the alleles of conflicted STR markers in distinct STR profiles based on their common sources. If this step cannot be properly completed because the set of sources differ between the STR markers, all the possible combinations of the alleles (up to a maximum of 150) of conflicted STR marker are then computed and stored as "virtual" STR profiles. Since it would not be manageable to report all these STR profiles in the results, only those with the best and worst scores are displayed so as to represent the extremes of the range of computed STR profiles.

Problematic cell lines

As noted in the Introduction, the contamination and misidentification of cell lines is a critical issue directly affecting the reproducibility of scientific research. Hence, one of the top priorities for the Cellosaurus is to clearly list and label all cell lines deemed to be problematic. The development of CLASTR was initiated so as to further endorse the objective of warning investigators about potential problems regarding the cell lines they are using in their research. As the identification of problematic cell lines is one of the key features of the tool, special care was taken to ensure that these cell lines would be properly flagged in both the web interface and export formats.

Web interface

The CLASTR web interface was designed as a single-page application composed of two key components: the input form and the result table. The input form consists of a main panel containing the STR marker input fields and a side panel containing the different parameters and possible actions (Fig. 4). The STR markers on the main panel are divided into two columns: the left one representing Amelogenin and the most common STR markers (CSF1PO, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA) and the right one representing the less common STR markers (D1S1656, D2S441, D6S1043, D10S1248, D12S391, D22S1045, DXS101, DYS391, F13A01, F13B, FESFPS, LPL, Penta C and SE33). A cross-reference to the STRBase database[18] is also provided, which acts as a source of further information for all STR loci having a corresponding entry page in this resource. By default, only the most common STR markers are included in the result table. The less common STR markers need to be checked prior to the search so as to be displayed in the results. At the right of the markers input section, there is a panel that contains (from top to bottom): (*i*) choices for scoring algorithms and modes along with the possibility to include amelogenin in the score computation; (*ii*) the option to filter the search results by a minimum score, by a minimum number of common STR markers and by a maximum number of Cellosaurus entries returned; and (*iii*) buttons to initiate the main actions available and to be directed to the Help and About pages.

A useful feature of the web interface is the ability to load STR profile data from a file into the input form for subsequent searches. The input table file can either be a plain text file (CSV, TSV, TXT) format, a Microsoft Excel file (XLS, XLSX) format or a file produced by the GeneMapper ID-X software (https://www.thermofisher.com/ch/en/home/industrial/human-identification/genemapper-id-x-software.html). The table needs to include a column representing the sample name (labeled as "Name," "Sample" or "Sample Name") and a range of columns representing the STR markers. Note that the order of the columns does not matter and additional columns will be ignored. As an option, a "batch query" can be performed, performing iteratively the similarity search on all the samples contained in the

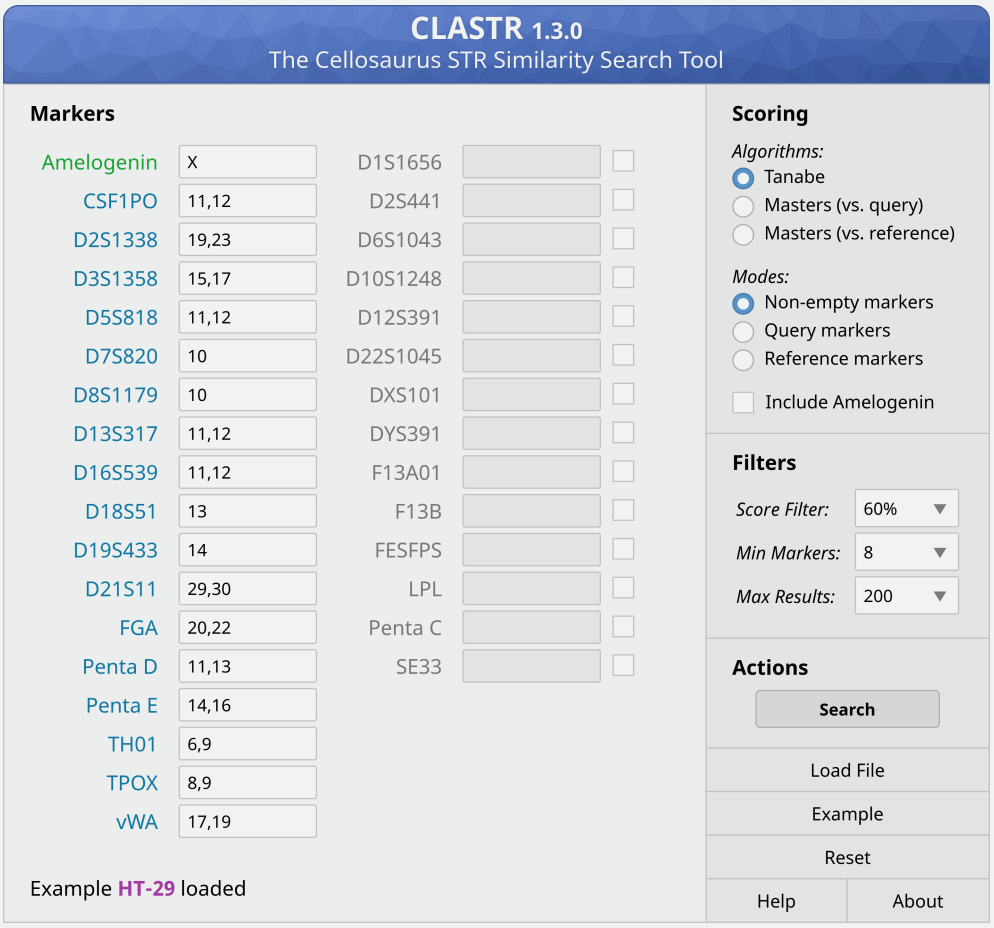


Figure 4. Screenshot of the input form. [Color figure can be viewed at wileyonlinelibrary.com]

table input file and directly returning all the results in XSLX, JSON or CSV formats. Note that in the case of the CSV format, one file per submitted sample is generated; these are returned as a single compressed archive in ZIP format.

The result table is a dynamic and sortable HTML table that is displayed once the similarity search is complete. The first row of the table always represents the query that was submitted to be searched against the STR profiles in the Cellosaurus. The first column labeled "Accession" provides the Cellosaurus accession number of the cell line along with a link toward its corresponding Cellosaurus entry page on ExPASy. In the case of problematic cell lines, the accession number is displayed in red and is associated with a tooltip describing the problem. In the case of a cell line with one or more STR allele value conflicts, an additional keyword is inserted to specify if the tested STR profile corresponds to the "best" or "worst" result (see section on conflicting STR profiles). The second column labeled "Name" provides the name of the cell line. The third column labeled "N° Markers" provides the number of STR markers that were used in the score computation, which depends on the scoring mode selected (see section on scoring modes). The fourth column labeled "Score" provides the computed score from the pairwise comparison based on the selected scoring algorithm and related parameters. The left border of the table is color coded so as to conveniently indicate if the cell line is highly related with the query or not. The following columns provide the alleles of the STR markers. The alleles that do not match the ones of the query are displayed in red. In the case of conflicted STR markers, the alleles are underlined and tooltips indicating the corresponding sources are provided.

An export table button located at the top left of the table provides the ability to directly export the generated table in XSLX, CSV or JSON formats. Note that all the files generated by the tool contain metadata, making it possible to keep track of the relevant search information (version of the Cellosaurus, version of CLASTR, run date and parameters) regardless of the selected format.

Additionally, each Cellosaurus entry page associated with a human cell line STR profile is directly linked to the CLASTR home page. The link encapsulates the STR profile information as URL parameters thus allowing the tool to load the corresponding allele data automatically. This is particularly useful to search a specific cell line and identify the cell lines that have similar STR profiles.

RESTful API

The CLASTR RESTful API allows STR profile searches to be performed without needing to use the web interface. Two main distinct public API resources are available: "single entry mode query" and "batch mode query." The single entry mode query provides the ability to search a single STR profile using a GET or POST HTTP method and retrieve the response content in XLSX, CSV or JSON formats. The batch mode query provides the ability to search several STR profiles at the same time using a POST HTTP method and retrieve the response content in XLSX, CSV or JSON formats. More information about the RESTful API is available in the online help page (https://web.expasy.org/cellosaurus-str-search/help.html).

Source code

The CLASTR source code is composed of three main parts: the front end, the back end and the web app. The front end handling the web interface is written in HTML/CSS and JavaScript with the jQuery and jQuery UI libraries. The back end performing the similarity search is written in Java 8. The web app integrating the back end and linking it to the front end, while also managing the RESTful API, is written in Java 8 and is deployed as a web application using the Apache Tomcat application server. All source code is publicly available on GitHub (https://github.com/calipho-sib/cellosaurus-STR-similarity-search-tool) under the GPL-3.0 license. Python 3 scripts showcasing the use of the RESTful API are also available.

Data privacy

Data privacy is an important concern when it comes to genomic data.[19,20] However, this has not hindered the growth in the number of publicly available DNA sequences in recent years.[21] Although a STR profile represents only a fraction of the total genomic information, it still has the potential to identify the individual from which it originated. To address this issue, we made sure that no query data is kept on the server once the similarity search is completed. Moreover, all connections are encrypted using the HTTPS security protocol in order to ensure data confidentiality during transfers.

Discussion

Cell line authentication remains the best approach to address the problem of cross-contamination and misidentification. Although it does not directly prevent contamination cases, it allows researchers to verify the authenticity of a cell line before performing any experiments on it. Because contamination has numerous causes, including poor technique (e.g., sharing media between cell lines),[22] we expect that new cases will continually arise despite the implementation of stricter requirements. Consequently, contamination is expected to be an enduring issue in biomedical research, and bioinformatics will need to play an important role to limit its prevalence. Increased knowledge about contaminated cell lines needs to be gathered and the authentication process needs to be facilitated for all actors involved with cell lines, aims which are being addressed by the Cellosaurus and by CLASTR, respectively.

A big advantage of using the Cellosaurus STR data set is that its extensiveness increases the probability of detecting potential contamination cases. As all other available STR similarity search tools are based on their own specific data sets, which are restricted in scope, a number of STR profiles are never compared against each other and some problematic cases can be missed. This is especially relevant since the majority of the STR profiles contained in the Cellosaurus come from individual publications, as mentioned in the data section.

For each cell line entry, the Cellosaurus reports its hierarchy (i.e., if it has parents or children) and if other cell lines originate from the same individual. In the case of a contaminated cell line entry, its hierarchy is modified to indicate the contaminating cell line as parent. This type of information allows one to know if two given cell lines are annotated as related and are thus expected to have similar STR profiles. If two unrelated cell lines turn out to have a high similarity, further investigations are required to determine if they are actually related or if contamination is involved. To automate this auto-validation process of the Cellosaurus, CLASTR was adapted to run as a procedure in which a search is performed against all STR profiles. This workflow enables the investigation of any cell line pair that has an unexpectedly high similarity. Over time, new problematic cases will regularly be reported to ICLAC (https://iclac.org) to be investigated and the corresponding Cellosaurus entries will be updated accordingly to reflect their problematic status.

Originally, eight core STR markers were believed to be sufficient to be able to discriminate related from unrelated cell lines[17] fully. However, at the time, studies were already pointing out that this number may be too limited and that at least 13 core STR markers should be compared to identify cell lines with a high confidence.[23] While the cell line community seems at present to agree that using only eight STR markers is too limiting, few recent studies explore in detail how beneficial it is to include more STR markers and what is the preferred number of markers to be used. With the large amount of STR profiles that the Cellosaurus provides and the flexibility brought by CLASTR, we expect that it could also help in the process of establishing new standards and guidelines.

Variations in STR profiles obtained from the same cell line are an important issue when interpreting data from CLASTR searches. These variations arise due to biological and technical factors. Many cell lines were established from malignant tissue, which is inherently heterogeneous. Clonal populations are present that evolve as the culture is passaged and when derivatives are established. Laboratories with expertise in cell line authentication have developed match criteria for interpretation of variable STR profiles, which are discussed elsewhere.[15–17,23] However, it is important to minimize such variation by "banking" cell lines at early passage. This is an

important part of good cell culture practice for all laboratories, along with the need to perform authentication testing before further work commences.[24] Technical factors may also arise, due to variations in STR profiling procedures and data interpretation. All laboratories that perform STR profiling must do so using a standardized approach. Standards have been developed for authentication of human and nonhuman cell lines that set out consensus requirements.[8] Adherence to these technical standards, and the use of early passage material for testing, will ensure that published STR profiles are fit for use in CLASTR searches.

In its initial testing phase, CLASTR received an overwhelmingly positive response from beta-testers with expertise in cell lines and STR profiling; it was significantly improved thanks to their feedback. CLASTR was made publicly available on ExPASy on March 22, 2019. Since then it has been used ~3,500 times. In approximately half of the cases, it has been accessed from its home (input) page while the second half consist of users accessing it from a Cellosaurus cell line entry. With the progressive increase in the number of STR profiles stored in the Cellosaurus, we expect that the popularity of CLASTR will grow concomitantly over time.

Acknowledgements

We are very grateful to Richard Neves and Gregory Sykes for actively beta-testing CLASTR and for providing extremely useful feedback. We thank Elisabeth Gasteiger for installing the tool on the ExPASy server and for enabling a Cellosaurus entry with a STR profile to be directly linked to CLASTR. We are also thankful to Monique Zahn for careful proofreading of this manuscript.

References

1. Bairoch A. The Cellosaurus, a cell-line knowledge resource. *J Biomol Tech* 2018;29:25–38.
2. Artimo P, Jonnalagedda M, Arnold K, et al. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* 2012;40:W597–603.
3. Freedman LP, Gibson MC, Ethier SP, et al. Reproducibility: changing the policies and culture of cell line authentication. *Nat Methods* 2015;12:493–7.
4. Horbach SPJM, Halffman W. The ghosts of HeLa: how cell line misidentification contaminates the scientific literature. *PLoS One* 2017;12:e0186281.
5. Clayton TM, Whitaker JP, Maguire CN. Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Sci Int* 1995;76:7–15.
6. Holt CL, Stauffer C, Wallin JM, et al. Practical applications of genotypic surveys for forensic STR testing. *Forensic Sci Int* 2000;112:91–109.
7. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;45:D777–83.
8. Almeida JL, Cole KD, Plant AL. Standards for cell line authentication and beyond. *PLoS Biol* 2016;14:e1002476.
9. Fowles JS, Dailey DD, Gustafson DL, et al. The Flint animal cancer center (FACC) canine tumour cell line panel: a resource for veterinary drug discovery, comparative oncology and translational medicine. *Vet Comp Oncol* 2017;15:481–92.
10. Berger B, Berger C, Hecht W, et al. Validation of two canine STR multiplex-assays following the ISFG recommendations for non-human DNA analysis. *Forensic Sci Int Genet* 2014;8:90–100.
11. Almeida JL, Hill CR, Cole KD. Mouse cell line authentication. *Cytotechnology* 2014;66:133–47.
12. Almeida JL, Dakic A, Kindig K, et al. Interlaboratory study to validate a STR profiling method for intraspecies identification of mouse cell lines. *PLoS One* 2019;14:e0218412.
13. Romano P, Manniello A, Aresu O, et al. Cell line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res* 2009;37:D925–32.
14. Dirks WG, MacLeod RAF, Nakamura Y, et al. Cell line cross-contamination initiative: an interactive reference database of STR profiles covering common cancer cell lines. *Int J Cancer* 2010;126:303–4.
15. Masters JR, Thomson JA, Daly-Burns B, et al. Short tandem repeat profiling provides an international reference standard for human cell lines. *Proc Natl Acad Sci USA* 2001;98:8012–7.
16. Tanabe H, Takada Y, Minegishi D, et al. Cell line individualization by STR multiplex system in the cell bank found cross-contamination between ECV304 and EJ-1/T24. *Tissue Cult Res Commun* 1999;18:329–38.
17. Capes-Davis A, Reid YA, Kline MC, et al. Match criteria for human cell line authentication: where do we draw the line? *Int J Cancer* 2013;132:2510–9.
18. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res* 2001;29:320–2.
19. Naveed M, Ayday E, Clayton EW, et al. Privacy in the genomic era. *ACM Comput Surv* 2015;48:1–44.
20. Kaye J. The tension between data sharing and the protection of privacy in genomics research. *Annu Rev Genomics Hum Genet* 2012;13:415–31.
21. Poon H, Quirk C, DeZiel C, et al. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics* 2014;30:2840–2.
22. Capes-Davis A, Theodosopoulos G, Atkin I, et al. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer* 2010;127:1–8.
23. Bady P, Diserens A-C, Castella V, et al. DNA fingerprinting of glioma cell lines and considerations on similarity measurements. *Neuro-oncology* 2012;14:701–11.
24. Geraghty RJ, Capes-Davis A, Davis JM, et al. Guidelines for the use of cell lines in biomedical research. *Br J Cancer* 2014;111:1021–46.

2.2. Concluding Remarks

CLASTR received an overwhelmingly positive reception from experts in cell authentication and users. Since it was made available on the SIB ExPASy server [116] exactly a year ago in March 2019, it accumulated a total of about 12,000 unique page views. It is also directly referenced in the ”Cell line authentication and mycoplasma testing” section of the author guidelines from the *International Journal of Cancer*.

The project is open-source (GPL-3.0 license) and hosted on the GitHub platform. Thus, it can be adapted with little effort to other STR profile databases while benefiting from all the features implemented in the tool. The code meets high-quality criteria having good test coverage and being well documented. The code was purposefully written to be easily maintained and extended. This was imperative as the implementation of new features may be required in the future. Since the version presented in the article (1.3.0), the tool was already adapted to support the similarity search of STR profiles from mouse and dog cell lines, which is a feature no other existing STR similarity search tool currently offers.

Chapter 3

STR Profiling Search Parameters

3.1. Overview

Several search parameters can influence the computed similarity between two STR profiles under comparison. In most of the tools performing STR similarity searches, the large majority of these parameters are fixed to an internal default that was pre-selected by the developers. In CLASTR, we opted to let researchers choose from a wide range of available search options. Our aim was not to exclude anyone who would have different work habits concerning STR profiling, as CLASTR is the only existing application allowing searching the wealth of STR profiles stored in the Cellosaurus. However, this diversity of options can make users confused about which parameter combination to pick, and whether the different options can significantly affect the resulting cell line identifications. In this study, we provide a detailed analysis assessing the precise effect of each parameter choice on the STR similarity scores based on the processing of a large data set. This endeavor provides guidelines regarding the optimal search parameter options to pick while justifying their selection with data evidence.

Evaluation of the Effect of STR Profiling Search Parameters on Cell Line Authentication

Thibault Robin[1,2,3,4], Amos Bairoch[1,2,5], Richard M. Neve[5,6], and Christopher Korch[5,7]

DRAFT

***Abstract*—Cell line misidentification and cross-contamination constitute a major source of erroneous experimental results in biomedical research. Consequently, researchers working with cell lines are expected to ensure their authenticity. Different genomic testing approaches are available for cell line authentication, but only STR profiling was subject to a standard. In this method, several search parameters can affect the computed similarity between STR profiles, including the scoring algorithm, the count of homozygous loci as one or two alleles, the inclusion of the amelogenin marker, and the number of STR loci constituting the profiles. There is currently a lack of recommendations concerning which parameter choices should be selected to uniquely identify cell lines. The effects of these parameters were analyzed using a large STR profile data set consisting of 2,284 cell lines extracted from the Cellosaurus knowledge resource. The STR profiles of all cell lines were searched against each other using the different parameter options and the resulting scores were analyzed based on the cell line pair relationships. The results presented here show a clear advantage to increasing the number of loci reported in STR profiles from 8 to 13. The Tanabe algorithm presents the best overall performance, while the inclusion of amelogenin in the scoring decreases the power to discriminate related from unrelated cell lines. Counting homozygous loci as two alleles does decrease, on average, the scores of unrelated cell lines, but it makes incorrect biological assumptions about their ploidy.**

***Abbreviations*—*ANSI**: American National Standards Institute; **ATCC SDO**: American Tissue Culture Collection Standards Development Organization; **CODIS**: Combined DNA Index System; **LOH**: Loss of Heterozygosity; **MMR**: Mismatch Repair; **MSI**: Microsatellite Instability; **MSS**: Microsatellite Stability; **PCR**: Polymerase Chain Reaction; **SNP**: Single Nucleotide Polymorphism; **STR**: Short Tandem Repeat; **WGS**: Whole Genome Sequencing.*

Introduction

THE use of misidentified and contaminated cell lines have been widely reported for many years as being the source of spurious experimental results [1–3]. Despite the call for action by members of the research community [4, 5] and the establishment of stricter submission guidelines in journals [6], it still remains a major concern in biomedical research. Regular cell line authentication represents the best solution to detect the emergence of such cases and curb the spread of unreliable results in the literature [7]. Several genomic approaches are available to evaluate the authenticity of a cell line, such as short tandem repeat (STR) profiling, single nucleotide polymorphism (SNP) profiling, and whole genome sequencing (WGS). STR profiling was however selected as the preferred technique as it had been extensively used in forensic and paternity applications, in addition to having a relatively low cost and high discrimination power. In 2011, an American National Standards Institute standard (ANSI/ATCC ASN-0002-2011) for the authentication of human cell lines through STR profiling was established by the American Tissue Culture Collection (ATCC) Standards Development Organization (SDO) workgroup [8]. ANSI/ATCC ASN-0002-2011 [9, 10] thoroughly describes the materials and methods required for the generation of an STR profile and its subsequent interpretation, and recommends the use of a minimum of eight core STR loci (CSF1PO, D5S818, D7S820, D13S317, D16S539, TH01, TPOX, and vWA) to uniquely identify human cell lines. Nonetheless, several additional factors should be taken into account when performing STR similarity searches.

There are four main search parameters than can influence the computed similarity between two STR profiles under investigation: (i) the search algorithm; (ii) the count of homozygous loci as one or two alleles; (iii) the inclusion of amelogenin in the scoring (when testing human cell lines); and (iv) the number of STR loci compared. A good set of parameters is expected to give high similarity scores to cell lines that are related, enabling the confident identification of such cell samples and the detection of potential cross-contamination and misidentification cases, where a cell line culture has been taken over by a different cell line. This diversity of options can however make the interpretation of STR profiles confusing for researchers with little experience in STR profiling. As cell line authentication is increasingly being mandated, it is consequently essential that clear instructions and recommendations are available. Few studies focusing on analyzing the precise effect of each parameter on STR similarity scores can be found in the literature, and the ANSI/ATCC ASN-0002-2011 standard did not cover these specific issues in detail. In this work, we have analysed a large test set of cell line STR profiles to study the impact of each parameter on STR matching, thereby identifying the optimal parameter settings.

Materials and Methods

Test Data Set

The test data set is composed of a total of 2,284 human cell lines extracted from the Cellosaurus database [11] (version 33, 19th of December 2019), whose STR profiles all share a minimum of the 13 commonly used STR loci (CSF1PO,

[1]Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, CH-1211 Geneva, Switzerland
[2]Computer Science Department, Faculty of Science, University of Geneva, Switzerland
[3]CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, CH-1211 Geneva, Switzerlan
[4]Microbiology and Molecular Medicine Department, Faculty of Medicine, University of Geneva, Switzerland
[5]Member, International Cell Line Authentication Committee
[6]Department of Disease Biology, Frontier Medicines, South San Francisco, CA, USA
[7]Division of Medical Oncology, University of Colorado School of Medicine, Aurora, CO, USA

D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, TH01, TPOX, vWA, plus amelogenin, which is used for gender determination). They represent the 13 core loci that were originally selected for the CODIS (Combined DNA Index System) database by the Federal Bureau of Investigation (FBI) to be used in forensic applications. In addition, for 1,196 of these cell lines we also possess the values for the Penta D and Penta E loci (total of 15 STR markers), and for a further 532 cell lines the values for the D2S1338 and D19S433 loci (total of 17 STR markers). Among the 2,284 cell lines, 453 have discordant allele data (i.e., different allele values and/or different numbers of alleles) for at least one STR locus resulting from the various sources that have performed the STR profiling. Based on the annotation information stored in the Cellosaurus, a given pair of cell lines was labeled either as "related" when they had parent-child (original cell line or patient sample versus a subline derived from the original sample), sister (two different sublines), or autologous (i.e., originating from the same donor) relationships, or as "unrelated" when it was not the case. This distinction is important since it is used to determine whether the STR profiles of two given cell lines are expected to have high similarity.

The complete list of selected cell lines and their corresponding information (Cellosaurus accession number, name, category, etc.) is available in supplementary material Table S1.

Methodology

Each of the 2,284 cell lines had its STR profile searched against that of all others for every set of search parameters studied. The STR similarity scores were computed once per pairwise comparison. A given cell line was also never tested against itself to avoid introducing bias into the results. As some cell lines had discordant allele data from distinct sources, they consequently had more than one distinct STR profile. In order to not give too much weight to these cell lines, only the highest STR similarity score was reported for a given pair under comparison when one of the two cell lines had more than one STR profile. In total, each run for a set of parameters is composed of 2,607,186 ((2,284−1)×2,284/2) pairwise comparisons, comprised of 799 related and 2,606,387 unrelated cell line pairs. The results were subsequently plotted to show the influence of each parameter on the STR similarity score distributions based on the cell line relationships. In each plot, one parameter was evaluated for both 8 and 13 STR markers while the other parameters were set to their default in CLASTR (Tanabe algorithm, amelogenin not included, homozygous loci counted as one). These default parameters were originally selected based on feedback from experts in cell line authentication. In order to additionally provide more additional specific examples, each presented plot also features subsets of the global results for the M14 (CVCL_1395) and Jurkat (CVCL_0065) cancer cell lines. These filtered results for each of the two cell lines are thus composed of 2,283 pairwise comparisons, as they were each tested once against all others. The two cell lines were selected on the basis that they differ in terms of microsatellite stability, besides having many relatives in the test data set (8 for M14 and 9 for Jurkat). Based on their respective annotations from the Cellosaurus, M14 is microsatellite stable (MSS) while Jurkat presents microsatellite instability (MSI-high). This enabled us to observe in detail whether this stability criterion leads to different observations and conclusions concerning the choice of the evaluated parameters.

Software

All the STR similarity score computations were performed on a local machine using CLASTR [12], the Cellosaurus STR similarity search tool. The different parameter choices under evaluation were already implemented in the tool, with the exception of counting homozygous alleles as more than one. This functionality was not part of the original design and had to be consequently implemented specifically for the purpose of this study.

The source code of CLASTR, including all the code used to generate the results of this study, is publicly available on GitHub (https://github.com/calipho-sib/cellosaurus-STR-similarity-search-tool/) under the GPL-3.0 license. The whole analysis workflow can be launched with a single command line using Docker so as to easily regenerate the result files.

Results and Discussion

The plots were generated using the R programming language and the ggplot2 library (https://ggplot2.tidyverse.org). The results are presented as violin plots, which corresponds to box plots (in white) surrounded by kernel density plots (colored based on the cell line pair relationship). As per the library default options, the box plot whiskers are placed on the most extreme data points that are within 1.5 times the interquartile range (IQR; defined as the difference between the 3rd and 1st quantile, which corresponds to the length of the box and contains 50% of the values). A cell line pair score is thus considered as an outlier (represented as a black dot) when it is located outside of the whiskers. Additionally, the bold black bar inside the box plots represents the median (2nd quantile), while the colored square represents the mean.The main result values (mean, median, and outliers) presented in each plot are summarized in Table 1. A black dashed line is also placed on the plots at 80% STR similarity, representing the recommended score threshold used to discriminate related from unrelated cell lines [13]. The outlier values that exceed this limit, either by being smaller or equal to it for related cell line pairs or greater or equal to it for unrelated cell line pairs, are considered as spurious. The possible origin of such cases is addressed in Table 2.

Search Algorithms

The purpose of search algorithms is to score the similarity between two STR profiles under comparison. The resulting score allows determining if the corresponding cell lines are related by applying a determined similarity threshold. Different search algorithms with varying levels of complexity were developed over the years, but only two are commonly used in cell line authentication: the Masters [14] and Tanabe [15] algorithms. The Masters algorithm is defined as the number of alleles in common divided by the total number of query alleles (or reference alleles in its modified reverse version) at shared loci. It has been historically used as the default algorithm to authenticate cell lines, and it is notably described

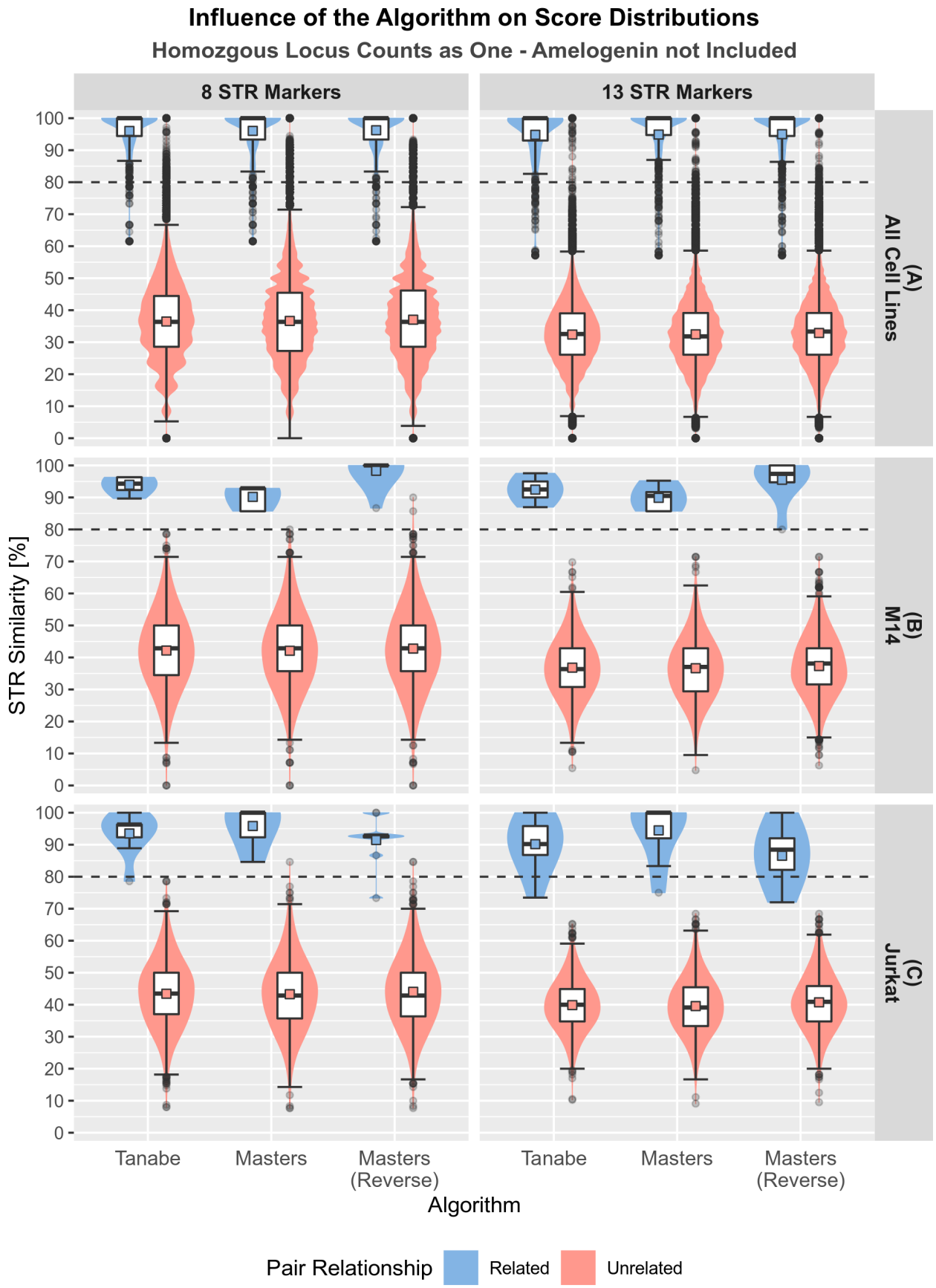


Figure 1: *Influence of the algorithm on the score distributions of related and unrelated cell line pairs*. The main results are presented on the first row (A), while the second (B) and third (C) rows present the specific cases of the M14 (MSS) and Jurkat (MSI-high) cell lines respectively.

Parameters				Related Cell Lines				Unrelated Cell Lines			
STR markers	Algorithm	Homo-zygote counts	Amelo-genin included	Mean	Median	Lower outliers	Spurious outliers	Mean	Median	Upper outliers	Spurious outliers
8	Tanabe	as one	no	96.00	100.00	73	40	36.47	36.36	11096	373
	Masters	as one	no	96.03	100.00	57	49	36.65	36.36	11463	2307
	Masters (Reverse)	as one	no	96.23	100.00	52	45	37.03	36.36	13378	2851
	Tanabe	as two	no	94.8	100.00	92	50	31.96	31.25	28992	173
	Tanabe	as one	yes	96.20	100.00	63	36	41.00	41.38	9576	540
13	Tanabe	as one	no	94.81	100.00	67	60	32.39	32.56	7264	43
	Masters	as one	no	94.89	100.00	111	66	32.46	31.82	17132	90
	Masters (Reverse)	as one	no	95.01	100.00	107	61	32.88	33.33	18813	83
	Tanabe	as two	no	93.69	100.00	92	71	28.41	26.92	13447	41
	Tanabe	as one	yes	95.00	100.00	67	56	35.51	35.56	6930	47

Table 1: *Summary of the effect of search parameters on the STR similarity scores.* In addition to the mean and median values for related and unrelated cell line pairs, the number of outliers values is also indicated. Of note, only the lower outliers (smaller than the lower whisker) and upper outliers (greater than the upper whisker) are respectively indicated for related and unrelated cell line pairs. The spurious outliers represent either the number of outliers that are smaller or equal to 80% for related pairs or the number of outliers that are greater or equal to 80% for unrelated pairs. They constitute the spurious cases that have unexpected similarities based on their described relationship. The causes of these spurious results are addressed in the Spurious Outlier Cases subsection.

in the ANSI/ATCC ASN-0002-2011 standard. The Tanabe algorithm (also known as Sørensen–Dice coefficient or Sørensen similarity index) is defined as twice the number of alleles in common divided by the number of query and reference alleles at shared loci. Both of the algorithms were shown in a previous study [13] to be capable of discriminating between related and unrelated cell lines using an 80% score threshold. The main distinction between them is that the Masters algorithm is asymmetrical, which can lead to different scores for a given cell line pair based on which STR profile is used as the denominator in the scoring equation (query or reference profile).

As illustrated in Figure 1, the two search algorithms present overall similar performances, the mean and median scores of related and unrelated cell line pairs being almost identical for a given number of STR markers (see Table 1). One major difference that can be noted is that the two versions of the Masters algorithm produce more spurious outliers (greater or equal to the 80% score threshold) for unrelated cell lines in comparison to the Tanabe algorithm (373 for Tanabe versus 2,307 and 2,851 for Masters and its reverse version respectively with 8 STR markers, and 43 for Tanabe versus 90 and 83 for Masters and its reverse version respectively with 13 STR markers). This is also observed with the M14 and Jurkat cell lines with both having additional spurious outliers (greater than 80% cut-off) when using either version of the Masters algorithm. However, these spurious cases disappear when the number of compared STR markers is increased to 13.

The differences in the number of outliers between algorithms can be partially explained by the fact that using the Masters algorithm, additional allele values in the reference profile (or query profile in the reversed version) would not decrease the computed score, because they are not included in the denominator. In contrast, any change to the allele data of one of the two STR profiles under comparison will modify the final score using the Tanabe algorithm. Another shortcoming of the Masters algorithm consists in its asymmetry, leading to the two possible implementations. This can prove confusing for researchers to determine which algorithm version to choose and that a given pair of STR profiles can produce two distinct STR similarity scores depending on which one is considered as the query. Consequently, we recommended using the Tanabe algorithm for cell line authentication by STR profiling and will use only it for the other analyses below.

Homozygous Locus Counts

A given STR locus can be either homozygous or hemizygous (only one distinct allele), or heterozygous (more than one distinct allele). As the search algorithms are based on allele counts to compute an STR similarity score, it is important to define how these homozygous loci are counted. The two conventional methods in use are to either count homozygous loci as one or as two alleles. The first approach is based on only counting the unique set of allelic possibilities that are detected by STR profiling, while the second is based on the assumption that the cells are diploid and contain two copies of a given homozygous allele.

Although expecting the cell lines to be diploid is valid for normal human cells in forensic applications and paternity testing, most immortalized human cell lines are aneuploid (tumor cells, transformed cells, etc.) and can exhibit a large number of chromosomal rearrangements and copy number variations [16]. Interestingly, the two SKY karyotypes of the human glioblastoma U-251MG cell line in this publication shows only a single copy of chromosome 10, i.e, it is haploid for this chromosome. For example, KBM-7 (CVCL_A426) from the test data set is haploid for all chromosomes except chromosome 8 and this is manifested in its STR profile where all STR loci are haploid except STR D8S1179 which has two alleles [17, 18]. SK-OV-3 (CVCL_0532) [19] and ML14 (CVCL_DI50) [20] from the test data set are polyploid from both their karyotype and their STR profiles. It can thus be biologically incorrect to assume an arbitrary chromosomal

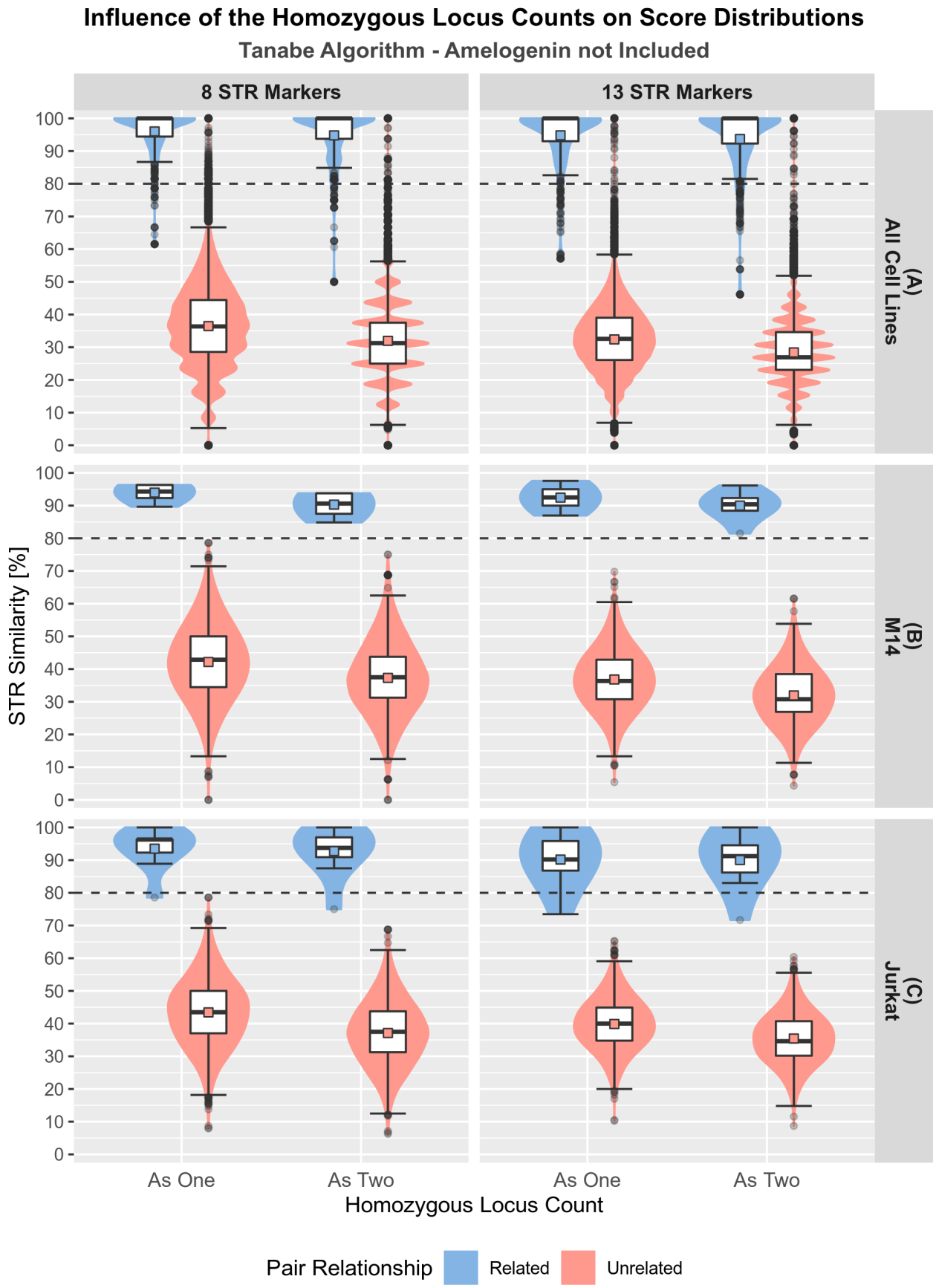


Figure 2: *Influence of counting homozygous loci as one or two on the score distributions of related and unrelated cell line pairs*. The main results are presented on the first row (A), while the second (B) and third (C) rows present the specific cases of the M14 (MSS) and Jurkat (MSI-high) cell lines respectively.

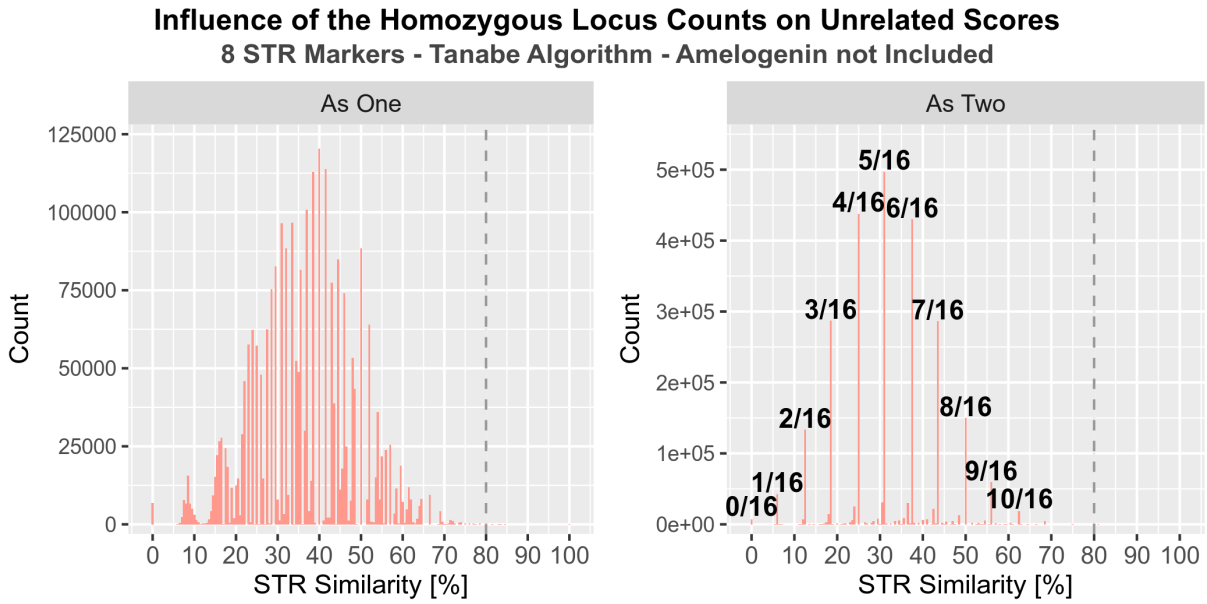


Figure 3: *Influence of counting homozygous loci as one or two on the scores of unrelated cell line pairs*. The numbers in the right graph (e.g., 5/16) indicate the matches divided by the total number of 16 possible alleles. Note the sharper discretization of the distribution of Tanabe match scores when counting single STR peaks twice, versus only once, which appear as wavy violin plots in Figure 2.

number when cell lines can be aneuploid. Furthermore, most cell lines are populations of cells with varying chromosome numbers per cell, that show aneuploidy [21].

As shown in Figure 2, counting homozygous loci as two alleles strongly decreases the scores of unrelated cell lines (mean decrease of 4.51% and 3.98% for 8 and 13 STR markers respectively) in contrast to a smaller decrease for related cell lines of 1.20% and 1.19% for 8 and 13 STR markers, respectively (Table 1). In contrast, the number of upper outlier pairs of unrelated cell lines (i.e., those with scores above 80%) is increase by factors of 2.61 for 8 STR loci and 1.85 for 13 STR loci (Table 1).

Moreover, counting single alleles twice significantly alters the shape of their score distributions, which can be explained by the fact that the total number of alleles of an STR profile will most of the time be twice the number of STR markers. As a result, the denominator part of the search algorithm equation is fixed (to either 16 or 26 alleles for 8 vs 13 STR markers, respectively), and only the numerator varies and influences the final scores. As reflected in Figure 3, this leads to a discretization of the results (i.e., visible as wavy violin plots in Figure 2) due to the more restrictive possibilities. An exception to this rule occurs when a given heterozygous locus has more than two unique allele values, producing the smaller peaks that can be observed.

Although the counting of single alleles twice seems to increase the differences between the overall mean scores of STR profiles between related vs unrelated cell lines, the distributions of the scores show increased overlap (compare all cell lines panel in Figure 2). This latter effect could be due to the addition of an allele that was not in the original sample being considered as a mismatch and thus reducing the percent match score. Overall, we recommend counting single peak (i.e., homozygous or hemizygous) alleles once rather than twice and not assume a specific number of chromosomes.

Inclusion of the Amelogenin Sex Typing Marker

Amelogenin is a human protein involved in the formation of dental enamel, which is encoded by genes on the X (AMELX) and Y (AMELY) chromosomes [22]. The sequence of these two versions of the amelogenin gene differ in sequence and overall length. The standard amelogenin allele test is based on a 6 base pair (bp) deletion in exon 3 of AMELX in a region conserved between the two genes [23]. PCR amplification of this region produces two amplicons that differ by 6 bp if the sample is from a male derived cell line and has both AMELX and AMELY intact. PCR from a female-derived cell line will only produce the shorter amplicon. Liang-Chu *et al.* [24] showed that as high as 45% of male-derived cell lines have lost the AMELY allele and are scored as female cell lines. Amelogenin testing has been extensively used in forensic applications, and the AMELX and AMELY genes are frequently co-amplified with STR markers by multiplex PCR in STR profiling. Despite not being an STR marker in itself, amelogenin has historically been included in the score computation of STR similarity searches.

Since amelogenin is not an STR marker, it does not possess the same allelic diversity as STR alleles. Amelogenin is thus limited to the X and Y alleles only, the presence of the Y allele indicating a male sample. Apart from some rare potential cases of X chromosome losses (two in the test data set, VMRC-RCW (CVCL_1790) and NCI-H810 (CVCL_1590)), the cell lines always share the X allele even when they have nothing else in common. As shown in Figure 4, the inclusion of the amelogenin allele data increases the mean Tanabe STR similarity scores for unrelated pairs by 4.53% and 3.12% for 8 and 13 STR markers, respectively. In contrast, the mean Tanabe scores for related pairs increased by only 0.2% for both 8 and 13 STR loci. As a consequence, the power to discriminate between cell lines is reduced (Table 1). This is further confirmed with the emergence of one spurious unrelated outlier (greater or equal to the 80% score threshold) for the

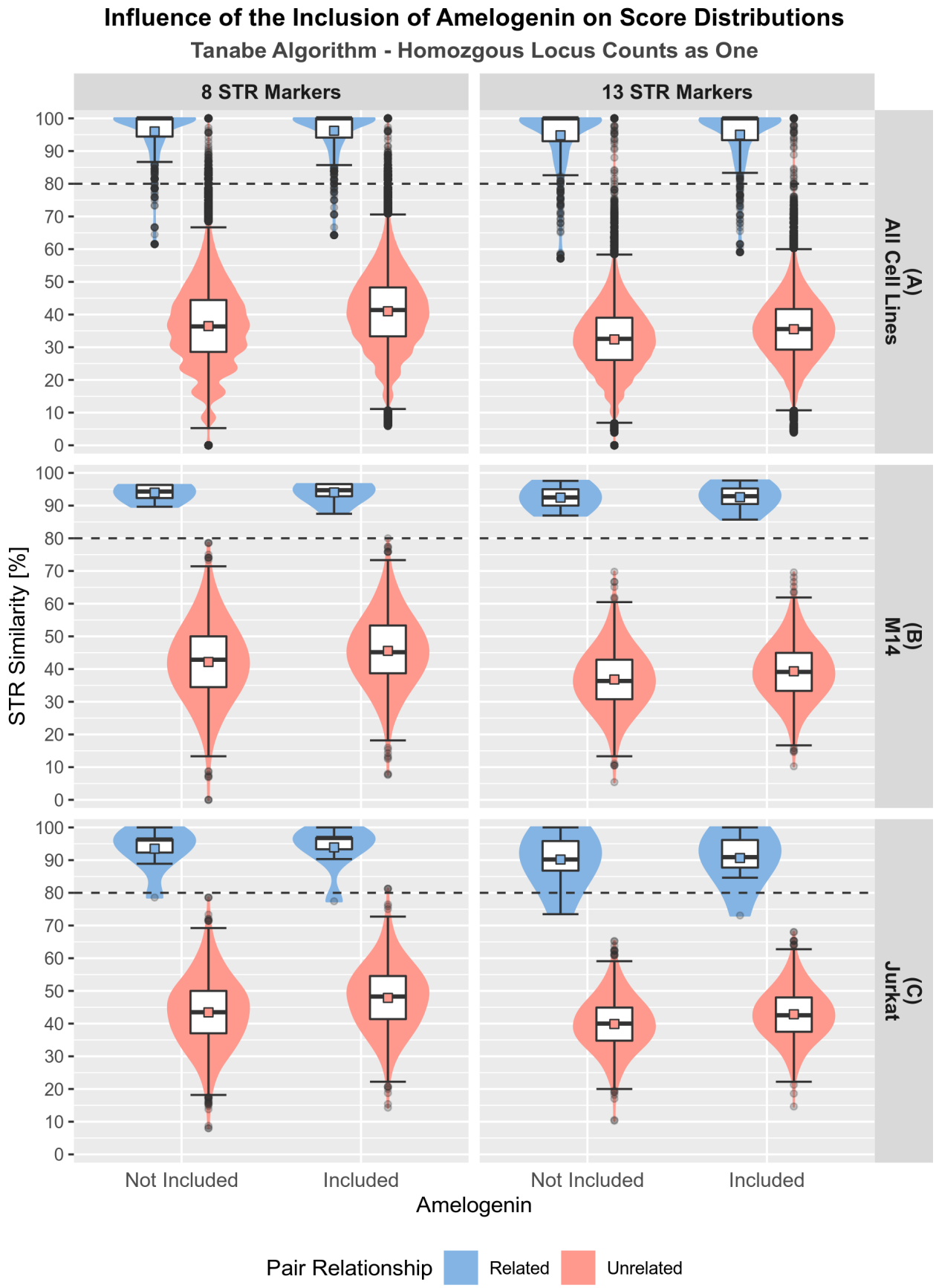


Figure 4: *Influence of the inclusion of amelogenin on the score distributions of related and unrelated cell line pairs*. The main results are presented on the first row (A), while the second (B) and third (C) rows present the specific cases of the M14 (MSS) and Jurkat (MSI-high) cell lines respectively.

Probable Causes of Spurious Cell Line Pairs		Number of Cell Line Pairs	
		Low Scoring Related	High Scoring Unrelated
1	Misidentified origin of cell line in Cellosaurus as related but are unrelated	10	
2	Misidentified origin of cell line in Cellosaurus as unrelated but are related		
2a	Incorrectly described as unrelated		3
2b	Cell lines established from identical monozygotic twins		1
2c	Cell lines established from sister cells of a blastocyst, i.e., identical genotype		1
2d	Cell lines established from same patient at different times or tissues		3
2e	Other error in Cellosaurus database		1
3	Defect in DNA mismatch repair causing MSI	40	
4	Genetic drift not due to MSI	1	
5a	Cross-contamination of a member of the cell line pairs	3	22
5b	Probable cross-contamination of a member of the cell line pairs		4
6	Cell fusion due to EBV transfection	1	2
7	Probable error in STR analysis	5	
8	Awaiting clarification from originators of cell lines		6
	Total Cell Line Pairs	**60**	**43**

Table 2: *Possible causes of spurious outliers.* Possible causes of Related cell line pairs with spurious Tanabe match scores of 80% or less and Unrelated cell line pairs with spurious Tanabe match scores of 80% or greater using STR profiles with 13 loci.

M14 cell line and three for the Jurkat cell line when using 8 STR loci when amelogenin is included in the scoring. As seen in the comparison of the different search algorithms (Figure 1), these spurious cases disappear when using 13 STR loci.

The influence of amelogenin inclusion is stronger with only 8 than with 13 STR markers as the contribution of a single allele to the global STR similarity score is greater when there are fewer loci. The Y chromosome does vary between cell lines, but it only indicates whether they share the same gender and provides little information concerning their relatedness. Interestingly, most of the score increase seems to be linked to the X allele, as the results of including solely the Y allele leads to results closer to the exclusion of amelogenin (mean for unrelated pairs with 8 STR markers of 36.5% when not including amelogenin, 37.7% when only counting the Y allele and 41.0% when counting both alleles ; data not shown). This lack of effect may reflect that many male-derived cell lines tend to lose the AMELY allele often [24].

Overall, we recommend not to include amelogenin results in the score computation of STR similarity searches, but rather to use it as a separate indicator of cell line provenance. For example, a cell line purportedly derived from a female donor, but which has an AMELY allele is an indication that the cell line is misidentified, possibly by a mix-up or cross-contamination. In contrast, the absence of AMELY allele for a purportedly male-derived cell line is not informative on the gender origin because of the frequent loss of this allele.

Number of STR Markers

Historically, STR profiles were composed of very few (four or six) STR markers when STR profiling was first introduced to cell line authentication [14]. This number was subsequently increased to eight (CSF1PO, D5S818, D7S820, D13S317, D16S539, TH01, TPOX, and vWA) following the recommendations of the ANSI/ATCC ASN-0002-2011 standard [8] and the analysis of Capes-Davis *et al.* [13]. However, recent studies [18, 25] reported that this number was still insufficient to reliably authenticate some cell lines having similar STR profiles. The revision of the ASN-0002 proposes that the number of STR loci in such profiles be further increased to the 13 core CODIS STR markers (addition of D3S1358, D8S1179, D18S51, D21S11, and FGA) to improve the power of discrimination. Also, the addition of these loci allowed comparisons between older results using kits with different sets of loci, some of which are shared. For example, Promega Powerplex 1.2 and Applied Biosystems Profiler Plus only shared four of the 13 core CODIS STR loci.

As illustrated in Figures 1, 2, and 4, increasing the number of markers in STR profiles from 8 to 13 systematically decreases the similarity scores and the ranges of scores of unrelated cell line pairs. The number of spurious outlier values (greater or equal to the 80% score threshold) for unrelated cell line pairs is also significantly reduced in all cases (see Table 1). As a consequence, using profiles with 13 STR loci, makes it much easier to distinguish between related and unrelated cell lines and it reduces the number of incorrectly categorized cell lines as related or unrelated. This is reflected when the number of STR markers was increased; the number of unrelated suspicious outliers decreased from 373 to 43 (Table 1).

Interestingly, some related cell line pairs can undergo a slight similarity score decrease when expanding the number of CODIS STR markers to 13. This is probably due to the fact that with an increased number of STR markers slight variations between two given STR profiles emerge. This is specifically the case for cell lines presenting MSI, such as the MSI-high Jurkat cell line whose plots highlight a consistent mean and median similarity score decrease when compared to its relatives from the test data set. Overall, our data supports the recommendation to increase the number of STR markers to a minimum of 13 to improve cell line authentication accuracy and reliability.

Spurious Outlier Cases

Table 2 summarizes the possible causes for spurious match scores for related and unrelated cell line pairs. The history, relationships, and STR data of the 58 cell lines in the 60 spurious related cell line pairings were reexamined and where possible the originators of the cell lines were contacted for clarification about these cell lines. The majority of these spurious cases (Cause 3, 40 pairs) are associated with both cell lines in the pair being defective in DNA mismatch repair (MMR), which manifests as microsatellite instability (MSI). Such cell lines tend to present more than two unique alleles at several loci, which can significantly decrease the computed similarity scores. Indeed, the supplementary alleles raise the total allelic count of the corresponding STR profiles, while not necessarily providing more alleles in common.

Ten purportedly *related* cell line pairs gave low scores because two of the lines in these 10 pairings were in fact not related to the others and therefore belong to the unrelated category (Cause 1). Extensive loss of heterozygosity (LOH) could also be involved, reducing the number of common alleles between two STR profiles, one such example is the cell line pair that showed genetic drift that did not present as being caused by MSI (Cause 4). Three pairs of lines appeared to be mixtures of two cell lines (Cause 5). One pair of cell lines were obtained from a EBV associated Burkitt lymphoma patient and probably arose through EBV-mediated cell fusion in vivo. EBV is often used to immoralize lymphocytes and to fuse cells into hybrids (Cause 6). In the case of five pairs of cell lines from the same lab showed consistent differences in allele calling with profiles of these same lines from other sources, suggesting an error in the calibration of the allele calling software (Cause 7).

As shown in Table 1, 373 unrelated cell line pairs had scores greater or equal to 80% using only 8 STR loci. This number decreased to 43 pairs of 71 cell lines after raising the number of STR loci to 13. In contrast to the erroneous annotations mentioned for related cell lines, lacking annotations tended by nature to be much more frequent among the purportedly unrelated cell lines. Thus, some unrelated cell line pairs could actually be related, explaining the high similarity of their STR profiles. The cross-contamination of a cell line strain by another would also lead to high STR similarity scores for cell lines being annotated as unrelated. Moreover, if a cell line is mixed or has MSI, the supplementary alleles could match by chance with an unrelated cell line and increase the computed similarity in their STR profiles.

We examined the descriptions of these 71 cell lines in Cellosaurus, their descriptions in the literature, and contacted the originators of the cell lines where possible to determine the potential causes of these purportedly unrelated cell line pairs showing Tanabe scores reater or equal to 80%. In Table 2, the majority (22/43, Cause 5a) of the originally unrelated cell line pairings were found to actually be related because of cross-contamination of one of the members of each of the 22 cell line pairs. Four additional pairs are suspected to be cross-contaminated cell cultures, but we have insufficient additional genetic data to support that conclusion (Cause 5b). Nine of these ”unrelated” cell line pairings were found actually to be related; e.g., they were established from monozygotic twins, sister cells in a blastocyst, or different tissues of the same patient, or some other labeling error (Causes 2 a-e). Two cell line pairs were probably due to EBV mediated cell fusion with HeLa cells (Cause 6). Finally, we were not able to explain why 6 pairs of purportedly unrelated cell lines showed Tanabe scores greater than 80% using 15 STR loci (Cause 8).

Conclusions

The comparisons of the different match algorithms and the influence of their parameters demonstrate the clear advantage of using additional loci for the comparison between samples, with a minimum of the 13 core CODIS STR loci. The Tanabe match algorithm is recommended for the authentication of cell lines because it is simple to use and produces consistent scores, which are more consistent when 13 STR loci are used instead of 8 loci. Inclusion of the amelogenin genotype degrades the power of discriminating between related and unrelated pairs of cell lines and should consequently be avoided.

Counting alleles from homozygous loci twice increases the discrimination capabilities by lowering the mean scores of the unrelated cell lines more than the scores of the related cell lines, but the range of values is also increased with increased overlap of scores (Figure 2). Furthermore, it makes biological assumptions about the ploidy of cell lines that are more often than not incorrect. Counting single alleles twice also causes confusion about which alleles might be present at a locus in the original patient sample and biases matches. For example, if a patient sample showed alleles A and B at a locus, but the cell line had only allele A or B at this locus and was scored as A A or B B, then this discordancy would reduce the match score to the original patient sample with the genotype A B. Also, the argument for counting single peaks twice is that it supposedly reflects a minimum genotype; however, as discussed,above cell lines are anueploid with the number of chromosomes being as few as one. For these reasons, we recommend against counting single alleles twice. The use of 13 STR markers represents a sufficient improvement by itself to recommend the counting of homozygous alleles as one. It is also less confusing to compile sets of data that consist of only what is seen and not what is assumed but not known.

The identification of the causes of the spurious match scores for related and unrelated cell line pairings illustrates the strength of the Tanabe scoring algorithm. If match scores exceed 80% match with 13 STR loci, then it is very likely that the two samples are related. If purportedly related cell lines show match scores that are less than 80%, then their provenance should be investigated to ascertain the causes for the low scores as was summarized in Table 2. It should be noted that a pair of cell lines having a high match score greater than 810%, or even a score of 100%, does not mean that the two lines are genetically identical. Kleensang *et al.* [26] compared two identical samples of MCF7 (CVCL_0031) grown under identical conditions with identical 8-loci STR profiles and found that two samples of the cell line that differed by only 3 passages behaved very differently. Ben-David *et al.* [27] delved into this problem much more extensively, finding that MCF7, an MSI-stable breast cancer cell line, evolves in culture much more than previously anticipated. Such changes can adversely affect the reproducibility between experiments within a laboratory and between laboratories.

The issue of patient confidentiality arises as more STR loci are analyzed. To address this issue, institutions (e.g., MD Anderson Cancer Center) and some countries (e.g., Japan) limit the publication of STR data to only 8 STR loci, such as those of the Promega PowerPlex 1.2 kit. The new General Data Protection Regulation of the European Union does not specifically address this issue, only requiring that the genetic data be "pseudonymized" to protect patient privacy and the extent to which this applied is at the discretion of individual EU members [28]. As a compromise, using 13 CODIS STR loci seems to be optimal for cell line identification, while minimizing exposure of patient confidentiality. However, different countries, journals, and research organizations may need to modify this in accordance with local regulations.

In summary, from this examination the effects of different STR profiling search parameters on cell line authentication based on the STR profiles of 2,284 cell lines, we drew the following conclusions for consistent authentication results and best discrimination between related and unrelated cell lines. First, the Tanabe algorithm was better than either version of Masters algorithms for this purpose. For optimal results, using 13 STR loci was better than only 8 STR loci. Including the amelogenin alleles did not improve distinguishing between related and unrelated lines. Counting single alleles twice did not improve this process, but may complicate the analyses and result in false negative and positive matches. Finally, it is important to confirm the provenance of both query and reference STR profiles to avoid spurious results and obtain consistent and reliable identification of cell lines when comparing STR profiles of cell lines. This analytical approach undoubtedly applies not only to the analysis of human samples, but also to other organisms for which there are STR data.

References

[1] Stacey, G. N. Cell contamination leads to inaccurate data: we must take action now. *Nature* **403**, 356 (2000).

[2] Hughes, P., Marshall, D., Reid, Y., Parkes, H. & Gelber, C. The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? *BioTechniques* **43**, 577–578 (2007).

[3] Lorsch, J. R., Collins, F. S. & Lippincott-Schwartz, J. Cell Biology. Fixing problems with cell lines. *Science* **346**, 1452–1453 (2014).

[4] Korch, C. & Varella-Garcia, M. Tackling the human cell line and tissue misidentification problem is needed for reproducible biomedical research. *Advances in Molecular Pathology* **1**, 209–228 (2018).

[5] Capes-Davis, A. *et al.* Cell lines as biological models: Practical steps for more reliable research. *Chemical Research in Toxicology* **32**, 1733–1736 (2019).

[6] Fusenig, N. E., Capes-Davis, A., Bianchini, F., Sundell, S. & Lichter, P. The need for a worldwide consensus for cell line authentication: Experience implementing a mandatory requirement at the International Journal of Cancer. *PLoS Biology* **15**, e2001438 (2017).

[7] Freedman, L. P. *et al.* Reproducibility: changing the policies and culture of cell line authentication. *Nature Methods* **12**, 493–497 (2015).

[8] Almeida, J. L., Cole, K. D. & Plant, A. L. Standards for Cell Line Authentication and Beyond. *PLoS Biology* **14**, e1002476 (2016).

[9] ATCC SDO. *Authentication Of Human Cell Lines: Standardization Of STR Profiling* (2011 (accessed May 2019)). URL https://webstore.ansi.org/RecordDetail.aspx?sku=ANSI%2FATCC+ASN-0002-2011.

[10] Barallon, R. *et al.* Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues. *In Vitro Cell Dev Biol Anim* **46**, 727–732 (2010).

[11] Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *Journal of Biomolecular Techniques* **29**, 25–38 (2018).

[12] Robin, T., Capes-Davis, A. & Bairoch, A. CLASTR: The Cellosaurus STR similarity search tool - A precious help for cell line authentication. *International Journal of Cancer* **146**, 1299–1306 (2020).

[13] Capes-Davis, A. *et al.* Match criteria for human cell line authentication: where do we draw the line? *International Journal of Cancer* **132**, 2510–2519 (2013).

[14] Masters, J. R. *et al.* Short tandem repeat profiling provides an international reference standard for human cell lines. *Proceedings of the National Academy of Sciences* **98**, 8012–8017 (2001).

[15] Tanabe, H. *et al.* Cell line individualization by STR multiplex system in the cell bank found cross-contamination between ECV304 and EJ-1/T24. *Tissue Culture Research Communications* **18**, 329–338 (1999).

[16] Thompson, S. L. & Compton, D. A. Chromosomes and cancer cells. *Chromosome Research* **19**, 433–444 (2011).

[17] Lee, C. C., Carette, J. E., Brummelkamp, T. R. & Ploegh, H. L. A reporter screen in a human haploid cell line identifies CYLD as a constitutive inhibitor of NF-Î°B. *PLoS One* **8**, e70339 (2013).

[18] Yu, M. *et al.* A resource for cell line authentication, annotation and quality control. *Nature* **520**, 307–311 (2015).

[19] Roschke, A. V., Stover, K., Tonon, G., Sch?ffer, A. A. & Kirsch, I. R. Stable karyotypes in epithelial cancer cell lines despite high rates of ongoing structural and numerical chromosomal instability. *Neoplasia* **4**, 19–31 (2002).

[20] Korch, C. *et al.* Authentication of M14 melanoma cell line proves misidentification of MDA-MB-435 breast cancer cell line. *Int J Cancer* **142**, 561–572 (2018).

[21] Hu, W. E. *et al.* HeLa-CCL2 cell heterogeneity studied by single-cell DNA and RNA sequencing. *PLoS One* **14**, e0225466 (2019).

[22] Francès, F. *et al.* Amelogenin test: From forensics to quality control in clinical and biochemical genomics. *Clinica Chimica Acta* **386**, 53–56 (2007).

[23] Sullivan, K. M., Mannucci, A., Kimpton, C. P. & Gill, P. A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *BioTechniques* **15**, 636–638 (1993).

[24] Liang-Chu, M. M. *et al.* Human biosample authentication using the high-throughput, cost-effective SNPtrace(TM) system. *PLoS One* **10**, e0116218 (2015).

[25] Bady, P. *et al.* DNA fingerprinting of glioma cell lines and considerations on similarity measurements. *Neuro-Oncology* **14**, 701–711 (2012).

[26] Kleensang, A. *et al.* Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. *Scientific Reports* **6**, 28994 (2016).

[27] Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).

[28] Pormeister, K. Genetic research and applicable law: the intra-EU conflict of laws as a regulatory challenge to cross-border genetic research. *Journal of Law and the Biosciences* **5**, 706–723 (2018).

3.2. Concluding Remarks

This study provides an answer to the option choice for each parameter available in STR similarity searches. Except for the count of homozygous and hemizygous loci as one allele being more based on biological rationale, the other selected search parameter options have shown to have an important influence on the results of the similarity search in terms of distinguishing related from unrelated cell lines. They thus provide regulating guidelines for the authentication of cell samples through the interpretation of STR profiles. As such, the arguments and conclusions of this analysis can be used as evidence for the planned revision of the original ATCC/ANSI ASN-0002 standard. It will also be used as a basis to inform CLASTR users of the recommended search parameters and their justification.

Additionally, this study features the use of Docker to run the whole analysis workflow. It thus enables anyone to regenerate in a single command line all the raw data, tables and figures we presented. This demonstrates how this technology can contribute to improving the state of reproducibility in scientific research. It notably enforces researchers to explicitly describe all the processing steps and avoid manual data manipulation, which is known to impede the reproduction of results. Please refer to the Discussion Section for a more in-depth presentation of the use of Docker in science.

Chapter 4

Reanalysis of HeLa Proteomics Data

4.1. Overview

Through the establishment of guidelines and corresponding data formats, proteomics data sharing and interoperability greatly improved in recent years. Nowadays, a wealth of standardized data is available to the proteomics community in specialized databases and repositories. Despite this increased accessibility in published data, few studies are based on the reanalysis of mass spectra. In this study, we selected for reanalysis 40 data sets from the PRIDE database, which all contained protein identification results from experiments carried out on the HeLa cell line. Our main goal was to identify single nucleotide variants from sequencing data at the protein level while trying to assess their impact on levels of protein expression. In the scope of the *Human Proteome Project*, we applied strict FDR levels and paid special attention to the identification of peptides from so-called *missing proteins*. Additionally, we also targeted new phosphorylation and acetylation sites, as we suspected that they were overlooked in most of the original experiments. This study was partially made possible thanks to the development of the MzVar tool, which I wrote at the beginning of this thesis. This proteogenomics software enables the compilation of customized variant protein and peptide databases from sequence variants and transcript sequences.

Unfortunately, it came to our attention that some of the mass spectra that we reanalyzed were not originating from the HeLa cell line. The error occurred early when retrieving the raw data files from the PRIDE database, as a consequence of a faulty command line. Of note, the metadata annotations on PRIDE were correct and were not the source of the problem. As only 27 files out of a total of 1,233 were concerned, we could reapply the complete workflow on the files that were containing genuine HeLa data sets and reinterpret the results. The identifications were only slightly affected, and the initial biological conclusions remained unchanged. The new anal-

ysis was released as an erratum, which is located just after the original publication in this thesis.

Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 4160−4170

pubs.acs.org/jpr

Large-Scale Reanalysis of Publicly Available HeLa Cell Proteomics Data in the Context of the Human Proteome Project

Thibault Robin,[†,‡,∥,⊥] Amos Bairoch,[†,⊥] Markus Müller,[§] Frédérique Lisacek,[‡,∥,#] and Lydie Lane[*,†,⊥]

[†]CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, CH-1211 Geneva, Switzerland

[‡]Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, CH-1211 Geneva, Switzerland

[§]Vital-IT Group, SIB Swiss Institute of Bioinformatics, Genopode Building, Quartier Sorge, CH-1015 Lausanne, Switzerland

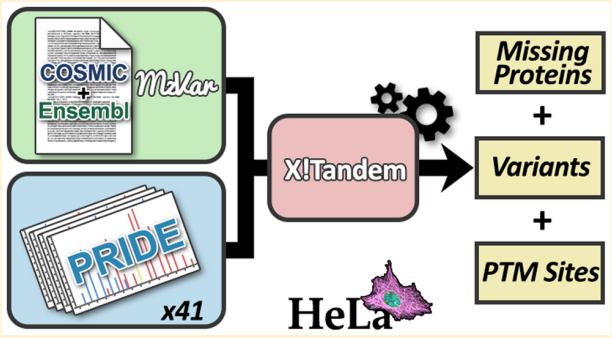
[∥]Computer Science Department, University of Geneva, CH-1211 Geneva, Switzerland

[⊥]Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, CH-1211 Geneva, Switzerland

[#]Section of Biology, University of Geneva, CH-1211 Geneva, Switzerland

Supporting Information

ABSTRACT: The practice of data sharing in the proteomics field took off and quickly spread in recent years as a result of collective effort. Nowadays, most journal editors mandate the submission of the original raw mass spectra to one of the databases of the ProteomeXchange consortium. With the exception of large institutional initiatives such as PeptideAtlas or the GPMDB, few new studies are however based on the reanalysis of mass spectrometry data. A wealth of information is thus left unexploited in public databases and repositories. Here, we present the large-scale reanalysis of 41 publicly available data sets corresponding to experiments carried out on the HeLa cancer cell line using a custom workflow. In addition to the search of new post-translational modification sites and "missing proteins", our main goal is to identify single amino acid variants and evaluate their impact on protein expression and stability through the spectral counting quantification approach. The X!Tandem software was selected to perform the search of a total of 56 363 701 tandem mass spectra against a customized variant protein database, compiled by the application of the in-house MzVar tool on HeLa-specific somatic and genomic variants retrieved from the COSMIC cell line project. After filtering the resulting identifications with a 1% FDR threshold computed at the protein level, 49 466 unique peptides were identified in 7266 protein entries, allowing the validation of 5576 protein entries in accordance with the HPP guidelines version 2.1. A new "missing protein" was observed (FRAT2, NX_O75474, chromosome 10), and 189 new phosphorylation and 392 new protein N-terminal acetylation sites could be identified. Twenty-four variant peptides were also identified, corresponding to 21 variants in 21 proteins. For three of the nine heterozygous cases where both the variant peptide and its wild-type counterpart were detected, the application of a two-tailed sign test showed a significant difference in the abundance of the two peptide versions.



KEYWORDS: *data reanalysis, proteomics, mass spectrometry, HeLa cell line, variants, identification, spectral counting, phosphorylation, N-acetylation, bioinformatics*

INTRODUCTION

Omics-based cancer research has flourished in recent years through the increased availability and refinement of next-generation sequencing methods.[1,2] With the multiplication of whole genome, exome and transcriptome sequencing studies, numerous structural and sequence variants have been identified, and their involvement in the oncogenic process was assessed.[3] While these variants are well described at the gene and transcript level, there is a lack of knowledge about their effective impact on protein function, expression and stability. The interpretation of genomic data thus tends to lag behind its massive production. It is also known that genetic information alone is not sufficient to decipher the complexity of protein biology.[4] For instance, assessing protein abundance from mRNA quantification data remains a challenging task, as many factors such as protein degradation and heterogeneous synthesis rates can alter the correlation.[5] A DREAM computational proteogenomics challenge focusing on this particular issue was launched by NCI-CPTAC last year and won by Yuanfang Guan and Hongyang Li (University of Michigan) using machine learning methods.

Special Issue: Human Proteome Project 2018

Received: May 30, 2018
Published: September 3, 2018

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

These limitations contributed to the emergence of proteogenomics, a field of research located at the intersection of genomics and proteomics. Its original purpose was to take advantage of mass spectrometry-based proteomics technologies to annotate protein-coding genes.[6] The term evolved since and is nowadays commonly used to describe any application using a proteogenomic approach to interpret mass spectrometry proteomics data.[7] This method usually consists of the database search of a customized protein database compiled from genomic/transcriptomic sequencing data.[8] As solely the sequences present in the database can be identified, the genomic information enables the detection of sample specific sequence variation in proteomics data.

Many tools were developed over the years to generate such custom databases, including PGA,[9] customProDB,[10] sapFinder,[11] Galaxy-P[12] and QUILTS.[13] They all rely on the same core principle, but implementations differ with respect to the types of variants supported and the content of the generated database. Although these tools all handle single amino acid substitutions, some specialized in more complex variants such as large insertions/deletions, frameshifts or splice variants. Recent efforts have also focused on the establishment of workflows and guidelines for the detection of variants in mass spectrometry data.[14,15] While the protocols for validating the presence of genomic variants in proteomics data are nowadays well-defined, a lot of work remains to be done regarding their influence on proteins. Furthermore, evaluating the direct impact of variants on protein abundance is not straightforward, as they may indirectly alter peptide detectability.

The production of mass spectrometry-based proteomics data has grown exponentially in recent years mainly through the improvement of instrumentation. Following the establishment of the ProteomeXchange consortium,[16] data sharing has rapidly spread in the proteomics field, despite an initial period of reluctance. A typical proteomics research project is nowadays expected to make publicly available both the original raw mass spectrometry data and the corresponding identifications, enabling the validation of the results and the reanalysis of the mass spectra by other research groups. Despite the wealth of data that is now standardized, findable and accessible in online databases and repositories, published proteomics data is seldom reused.

In order to evaluate the impact of single amino acid variants on protein abundance, we developed a new proteogenomics workflow to perform the identification of peptides harboring single amino acid variants and their relative quantification compared to wild-type counterparts. MzVar, our in-house tool, was used to compile a customized variant protein database. The workflow was applied to a total of 41 HeLa cancer cell line proteomics data sets from the PRIDE repository.[17] The HeLa choice was motivated by the fact that it is one of the most studied human cell lines in research, with around 90 000 articles being listed in PubMed with the "HeLa cells" MeSH term. In order to have enough sensitivity to perform the peptide quantification and related statistical analyses, many proteomics data sets were required. According to the Cellosaurus knowledgebase,[18] HeLa is the human cancer cell line with the highest number of proteomics data sets available, with 5-fold more data sets in PRIDE than for MCF-7. We also took advantage of this large amount of data to look for new peptides from proteins that had not been previously experimentally validated, called "missing proteins" by the HUPO Human Protein Project participants,[19] and for new phosphorylation and protein N-terminal acetylation sites, as they were not investigated or reported in most of the original experiments.

MATERIAL AND METHODS

Retrieval and Filtering of Single Amino Acid Variants

As part of the COSMIC Cell Line Project, the HeLa cancer cell line exome was sequenced (https://cancer.sanger.ac.uk/cell_lines/help/desc). In this work, the single nucleotide variants (SNVs) of HeLa were called against the GRCh37 human genome assembly using the CaVEMan algorithm and made subsequently available on the COSMIC FTP server in the form of files in the VCF format. The retrieved HeLa VCF file contained 15 608 individual SNVs.

On the basis of CaVEMan flags contained in the filter column of the VCF file, exclusion filters were applied in order to discard low quality variants, corresponding to (i) variants in which less than 1/3 of the mutant alleles were of high quality; (ii) variants in which the mean mapping quality of the mutant alleles was insufficient; (iii) variants in which most mutant alleles were on one strand with insufficient base quality; (iv) variants in which the coverage was below nine and no mutant alleles were in the first 2/3 of the read; (v) variants in which the mutant alleles were on the same strand on the second half of the read containing the GGC[AT]G motif followed by an insufficient mean base quality. As a result of the filtering steps, the number of variants in the VCF file was reduced of about 66% to 6815 SNVs.

Retrieval and Filtering of Transcript Sequences

As the genomic coordinates of the VCF file retrieved from COSMIC were based on the GRCh37 human assembly, the corresponding transcripts were downloaded from the latest Ensembl version that made use of this genome assembly (v75, February 2014) using the BioMart online tool.[20] The entries corresponding to proteins predicted but not validated in UniProtKB/Swiss-Prot were discarded. The transcripts were retrieved in the form of cDNA sequences in the FASTA format. The headers were formatted to contain the chromosome name, strand, start/stop coordinates of the CDS, introns and exons. In a further step, the transcripts corresponding to proteins whose existence was annotated as uncertain (PE5) in neXtProt[21] (release 2018.01.17) were filtered out.

Compilation of the Customized Variant Protein Database

The HeLa customized variant protein database was compiled using the MzVar Java tool developed in-house (https://bitbucket.org/sib-pig/mzvar-public). This software generates variant protein and peptide databases in the FASTA format using a VCF file as variant input and a FASTA file as transcript input. To summarize briefly, all the individual variants are first extracted from the VCF file and inserted in their corresponding transcript sequences at the correct location using the indicated chromosomes and base pair positions. Then, the whole set of transcripts is translated into protein sequences. Finally, the variants are described at the protein level by comparing the amino acid sequences before and after the variant insertion into the transcripts.

In this work, MzVar was run on both the filtered SNVs from the HeLa VCF file and the GRCh37 transcript sequences with the "Protein Database" mode enabled. Both the variant-containing and nonvariant-containing protein sequences were included in the database. To account for contaminant proteins that are commonly introduced in proteomics experiments, the

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

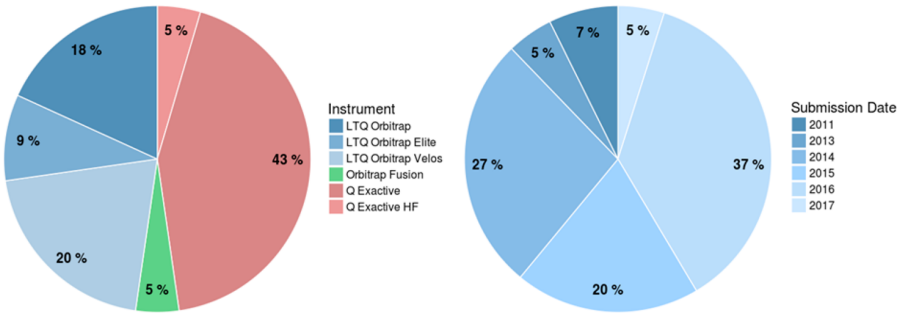


Figure 1. Distribution by instrument and submission date of the 41 selected HeLa data sets retrieved from the PRIDE database. All mass spectrometers belong to the Thermo family, evenly divided between LTQ Orbitrap and Q Exactive instruments. Most of the data sets were submitted between 2014 and 2016, while the oldest date back to 2011.

cRAP (The GPM, 2015.13.01) sequences were downloaded in the FASTA format and merged to the compiled database.

Retrieval of Tandem Mass Spectra

According to the current HPP guidelines, only data deposited in ProteomeXchange can be used. Querying "HeLa" with the PRIDE search engine retrieved 112 unique data sets. After manual verification, three data sets that were not performed on HeLa cells were removed. In an attempt to control data set heterogeneity, experiments performed with instruments not belonging to the Thermo Orbitrap family were discarded (Figure 1), reducing the number of data sets to 98. To further limit the complexity of the subsequent database search step, any data set that involved the introduction of a mass tag (e.g., SILAC, iTRAQ or TMT) for quantification analysis or the use of the SWATH technology were not taken into account, lowering the number of selected data sets to 54. Finally, experiments performed on multiple cell lines in parallel with no dedicated annotation of the HeLa cell result files or strictly focused on specific proteins (e.g., histones) were not included in the selection, resulting in a final number of 41 compatible HeLa data sets. We limited our study to data sets available in PRIDE in order to simplify the data set compilation. Of note, MassIVE presently reports 64 data sets for HeLa cells. 33 of them are also in PRIDE, and 8 were included in our selection of 41 high quality data sets. Out of the 31 data sets that are not in PRIDE, only five would match our quality criteria. IProX, that joined the ProteomeXchange consortium in December 2017, and jPOST, that opened to public in March 2018, respectively contain 2 and 14 additional HeLa data sets from which 1 and 6 would match our quality criteria.

Most of the selected data sets were recently submitted to the PRIDE database: 62% after 2015, while the oldest dated back to 2011 (Figure 1). The Thermo mass spectrometers were mainly divided into LTQ Orbitrap and Q Exactive instruments. The HCD fragmentation method was used in almost half of the data sets (20), while the rest (21) used CID. Most of the HCD data sets were produced on Q Exactive instruments.

Studies carried out on HeLa cells rarely indicate which derivative of the original cell line was used. Only five out of the 41 studies we reanalyzed did mention a HeLa subline (4 used HeLa-Kyoto and 1 used HeLa-S3). This cell line heterogeneity was not taken into account in our workflow as a filtering criterion since the different HeLa cell lines (HeLa from ATCC (CCL-2), HeLa-S3 and HeLa-Kyoto) were shown to have minimal variability in terms of SNVs.[22,23]

For each of the 41 selected experiments, the Thermo binary raw files were downloaded from the PRIDE FTP repository. In the few cases where the raw files were not accessible, the tandem mass spectra were downloaded instead in the mzXML or MGF format according to what was available. The 1233 acquired files were then converted into the mzML open format using the msconvert tool of the ProteoWizard suite[24] (version 3.0.10800, 2017.04.27). Only the MS2 scans were extracted due to file size concerns.

Database Search

The database search was performed using the X!Tandem open source search engine[25] (version Vengeance, 2015.12.15.2). Carbamidomethylation of cysteine (C) was set as fixed modification while oxidation of methionine (M) and phosphorylation of serine, threonine and tyrosine (STY) were set as variable modifications in addition to the default ones (N-terminal acetylation of proteins, water and ammonia loss from N-terminal glutamic acid (E) and glutamine (Q) respectively). Trypsin was selected as digestion enzyme and up to two missed cleavages were tolerated. The parent monoisotopic mass error was set to ±15 ppm and the fragment monoisotopic mass error to ±0.5 Da for the CID sets and ±0.05 Da for the HCD sets. All the computations were performed at the University of Geneva on the Baobab cluster (https://plone.unige.ch/distic/pub/hpc/baobab_en).

Identification and Validation of Peptides and Proteins

The X!Tandem BIOML result files were processed using the MzJava open source library[26] to extract and validate the reported identifications. All the identified peptides were filtered based on their size to remove those having a length smaller than seven amino acids. To verify their uniqueness in the human proteome, the neXtProt peptide uniqueness checker tool[27] was run on protein sequences of the neXtProt database (release 2018.01.17). The peptides that could be mapped to more than one entry (without taking SNPs into account) were discarded. In addition, any identification that could have been explained by a peptide originating from a cRAP contaminant was removed.

To control the amount of false positives in the identifications, a false discovery rate (FDR) was calculated globally for all data together following the target-decoy approach. Prior to the X!Tandem database search, reversed decoy protein sequences were concatenated to the customized protein database. The FDR was estimated at the PSM, peptide and protein levels using the X!Tandem hyperscore as the reference score. The hyperscore threshold was adjusted so that the global protein FDR would not be higher than 1%. At the peptide level, the best PSM hyperscore was assigned to each peptide.

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

At the protein level, the best peptide hyperscore was assigned to each protein. To be considered as identified in accordance with the HPP guidelines version 2.1,[28] proteins were required to have at least two non-nested unique peptides with a length equal or superior to nine amino acids. In all cases, the FDR was estimated using the following formula,[29] where N_{target} and N_{decoy} are respectively the number of target and decoy hits:

$$\mathrm{FDR} = \frac{2 \times N_{decoy}}{N_{target} + N_{decoy}}$$

Identification and Validation of PTM Sites

The X!Tandem search engine does not provide any probability score for PTMs. In the case of phosphorylations, multiple potential sites are frequently found close to each other. The exact location of a phosphorylation within a peptide can thus be ambiguous, and distinct sites may have the same probability. In order to determine with enough confidence which phosphorylation sites can be considered as identified, the PTMProphet tool of the Trans-Proteomic Pipeline[30] (TPP, version 5.1.0) was applied on the X!Tandem BIOML result files after their conversion into the pepXML format. Only the phosphorylation sites that had the maximum probability score (1.000) were subsequently taken into account.

Concerning the protein N-terminal acetylations, X!Tandem takes into account the protein N-terminal methionine excision that often occurs in eukaryotic organisms. The resulting new N-terminal residue is frequently acetylated. By default, the search engine removes this new N-terminal residue and searches acetylation once again.

Quantification of Peptides

The peptide quantification was performed following the spectral counting label-free approach. This method consists of summing the total number of PSMs for a given peptide, with or without modifications. This enables the comparison of the abundance of variant peptides against their wild-type counterparts. If a difference in the spectral counts is observed, it may provide insight into the variant impact on the protein abundance. For this analysis, only the heterozygous variants were taken into account, since both peptide versions need to be observable for their respective intensities to be compared. The cases where the wild-type peptide was not observed were also excluded, as the proof of existence of both versions was necessary for a fair comparison. To determine which peptide pairs had a significant difference in their spectral counts, a two-tailed sign test was independently applied on each variant. This statistical test establishes if there is a consistent difference between a pair of observed values throughout a set of replicates. Under the null hypothesis (H0), there is no difference between the observations, while there is a significant difference under the alternative hypothesis (H1).

In this statistical test, the spectral counts of the variant and wild-type peptides for a given variant are first separated according to the experiments from which they originated. The number of times where the variant peptide spectral count is greater than that of the wild-type in a given experiment, and conversely, are summed. The cases where the counts are equal are ignored. Then, the p-value is estimated from a binomial distribution where the number of successes is the smallest of either sum, the number of trials is the number of cases where there is a difference and the success probability is 0.5. Finally, if the p-value is less than half of the α level of 5%, the null hypothesis is rejected and the alternative hypothesis is accepted, meaning that there is a statistically significant difference between the spectral counts of the variant peptide and the wild-type peptide.

Data Availability

The following 41 HeLa PRIDE data sets were reanalyzed: PRD000525, PRD000526, PRD000527, PXD000212, PXD000243, PXD000396, PXD000589, PXD000759, PXD000895, PXD000999, PXD001061, PXD001249, PXD001258, PXD001333, PXD001374, PXD001381, PXD001426, PXD001441, PXD001660, PXD002252, PXD002277, PXD002395, PXD002815, PXD002880, PXD002987, PXD003209, PXD003370, PXD003530, PXD003560, PXD003917, PXD004182, PXD004273, PXD004613, PXD004900, PXD004940, PXD005018, PXD005181, PXD005366, PXD005509, PXD005712 and PXD006112. The Supporting Information contains a tab-delimited table listing all the validated PSMs, an Excel table in the xlsx format listing all the proteins, peptides, variant and PTM site identifications, as well as the customized variant protein database in the FASTA format.

RESULTS AND DISCUSSION

Identification of Proteins and Peptides

By applying our workflow (Figure 2) to the 1233 raw files (56 363 701 tandem mass spectra), we identified 49 466 unique peptides with 1 225 780 PSMs mapping to a total of 7266 protein entries. Notably, 355 peptides matched alternative splice isoforms that were not recorded as canonical in neXtProt. Using the stringent HPP criteria for protein validation,[28] this led to the identification of 5576 protein entries. These results were obtained after the implementation of a strict FDR threshold to prevent an unsatisfactory number of false positives among the identifications. A hyperscore threshold of 60.4 was set so that

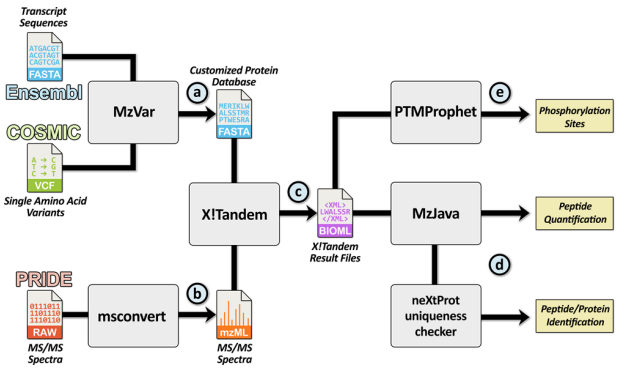


Figure 2. Global overview of the custom workflow. A customized variant protein database was first compiled using the MzVar tool on single amino acid variants belonging to the HeLa cancer cell line (a). The HeLa variants were retrieved in the form of a VCF file from the COSMIC cell line project and inserted in transcript sequences from Ensembl before their translation into proteins. Then, 41 distinct MS proteomics data sets produced on HeLa cells were retrieved from PRIDE and converted in the mzML open format using msconvert (b). X!Tandem was subsequently used to search all the MS/MS spectra against the database (c), and the resulting identifications were processed using the MzJava library along with the neXtProt uniqueness checker tool (d). Additionally, PTMProphet was used for the validation of the identified phosphorylation sites (e).

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

the global FDR would be fixed to 1.0% at the protein level. This corresponds to a global FDR of 0.0086% at the PSM level and 0.14% at the peptide level. This means we can expect 1 225 675 true and 105 false positives for the PSMs, 49 397 true and 69 false positives for the peptides and 5520 true and 56 false positives for the proteins. However, it is important to note that the FDR is only an estimate, which may fail to correctly predict the correct amount of false positives. In many cases, the actual FDR is significantly higher than the computed one. The difference is partially explained by the shortcomings of the target-decoy approach, which does not properly model all the different error types that can occur. This emphasizes the necessity to recognize that a protein passing the FDR threshold does not mean that its identification is certain.

Identification of "Missing Proteins"

"Missing proteins" are defined as confidently predicted proteins lacking evidence at the protein level in neXtProt (PE2−4). Among the identified peptides in our data set, three originated from two missing proteins (neXtProt release 2018.01.17) (Table 1): FRAT2 (NX_O75474) and ZCCHC18 (NX_P0CG32). None of the three peptides mapped with an additional entry. Both FRAT2 peptides were found in the same study[31] (PXD005712). One had not been previously identified in biological samples according to the PeptideAtlas and neXtProt databases. Interestingly, spectra obtained with the corresponding synthetic peptide are recorded in SRMAtlas. We aligned the SRMAtlas consensus spectrum (Consensus Library Spectrum ID: 10693449) produced by Agilent 6530 QTOF instruments with the spectrum produced in the PXD005712 experiment by a Thermo Q Exactive instrument (Figure 3). We also computed a spectral dot-product score (SDPscore)[32] to assess the strength of the spectral correlation. The SDPscore was calculated using the intensities of the singly charged b and y ions and a value of 0.971 was returned, indicating that the fragmentation pattern of the two spectra is alike and increasing the confidence in the identification of FRAT2.

The second peptide for NX_O75474 had been already observed in other human samples and reported in PeptideAtlas and neXtProt (PAp00759928). It was matched in our workflow by two doubly charged mass spectra. Since the corresponding consensus mass spectrum from SRMAtlas (Consensus Library Spectrum ID: 10696262) is triply charged, its comparison against the two identified mass spectra showed a weak correlation (SDPscores of 0.509 and 0.486). However, performing the comparison against a doubly charged mass spectrum obtained on kidney cancer[33] and reported in PeptideAtlas (mzspec:PXD006482::kidney_6_2:scan:5525: AVAAVAATGPASAPGPGGGR/2) showed a good correlation (Figure 4) (SDPscores of 0.789 and 0.782). These correlation scores are actually higher than the one of the mass spectrum of PeptideAtlas against the consensus spectrum from SRMAtlas (SDPscore of 0.688).

The PXD005712 data set in which the two FRAT2 peptides were identified was published in 2017[31] and will be part of the PeptideAtlas 2019 build. Given the high quality of these identifications, one can expect that they will be validated by the PeptideAtlas pipeline, and subsequently integrated in neXtProt (February 2019 release). The current rule applied by neXtProt to validate a protein as PE1 is, in accordance with HPP guidelines, the presence of 2 non-nested peptides of at least 9 aa. Since the peptide AVAAVAATGPASAPGP-GGGR is already annotated in neXtProt, the integration of

Table 1. List of Identified PSMs Mapping to "Missing Proteins"

protein	gene	peptide	PE	reported in neXtProt	score	universal spectrum identifier	PMID	date
O75474	FRAT2	AVAAVAATGPASAPGPGGGR	PE2	Yes	61.6	mzspec:PXD005712:20150123_HN_Shotaro_Parl_IP_S277A_HIS_IP_17:scan:5448	28288130	2017
O75474	FRAT2	AVAAVAATGPASAPGPGGGR	PE2	Yes	61.7	mzspec:PXD005712:20152002_RG_150218_Saita_Ctrl_3:scan:5748	28288130	2017
O75474	FRAT2	ARPPAVPLLLPPASAETVGPAPSGALR	PE2	No	66.9	mzspec:PXD005712:20152002_RG_150218_Saita_Ctrl_3:scan:11606	28288130	2017
P0CG32	ZCCHC18	GRARPLDQVLVIDSPNNSGAQSLSTSGGSGYKNDGPGNIR	PE3	No	60.4	mzspec:PXD002987:LFQ_Control_pass2_rep2:scan:33629	26560067	2015

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

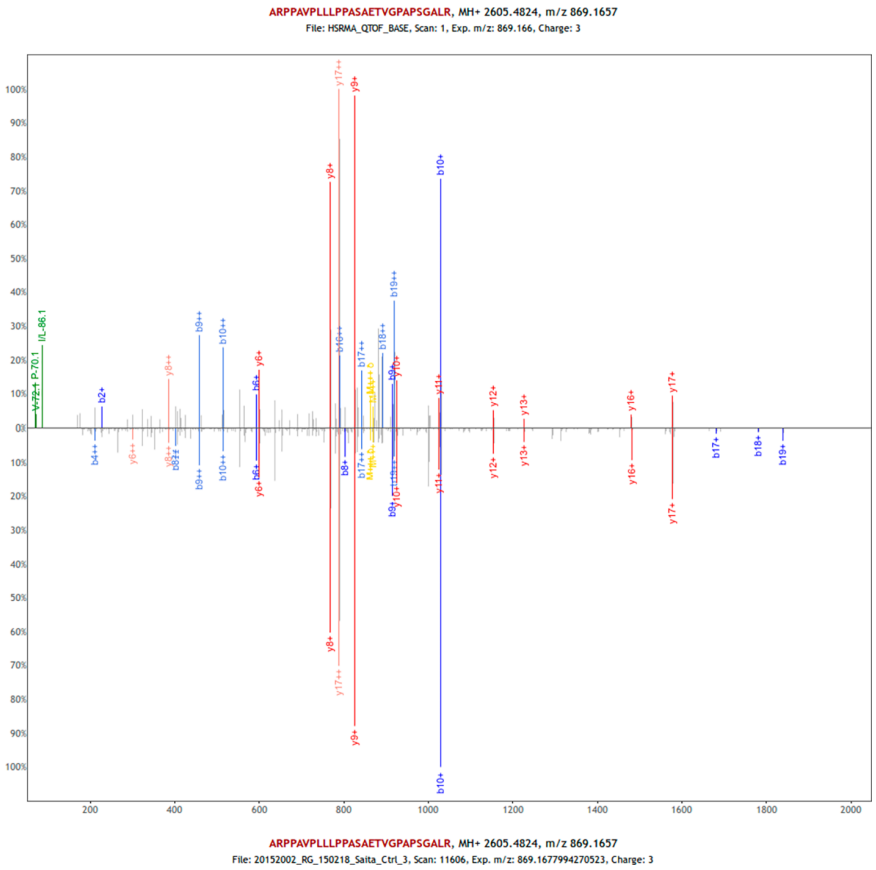


Figure 3. Spectral alignment for the ARPPAVPLLLPPASAETVGPAPSGALR peptide from the NX_O75474 PE2 "missing protein". At the top features the consensus spectrum from SRMAtlas produced by Agilent 6530 QTOF instruments. At the bottom features the spectrum identified by our workflow in the PXD005712 data set, produced by a Thermo Q Exactive instrument. The two spectra show a strong spectral correlation with a SDPscore of 0.971.

ARPPAVPLLLPPASAETVGPAPSGALR will trigger the modification of NX_O75474 status as PE1. FRAT2 is a poorly characterized protein that may act in canonical and non-canonical Wnt-signaling pathways involved in development, tissue homeostasis and cancer.[34] The peptide that was reported previously was observed in stem cells,[35] kidney cancer[33] and MCF-7 breast cancer cells,[36] suggesting that FRAT2 expression is not specific to HeLa cells or cancer, but might be related to cell proliferation. This functional hypothesis could be validated by targeted genome editing in HeLa cells.

For ZCCHC18, the only peptide observed in HeLa cells has two miscleavages and the spectrum could not be validated due to its absence in SRMAtlas. No additional peptide meeting the HPP requirements for protein validation was reported in neXtProt in any human biological sample. The reclassification of the NX_P0CG32 entry as PE1 will require further investigation.

Identification of PTM Sites

The identified N-terminal acetylation and phosphorylation sites were searched in neXtProt (release 2018.01.17) to identify those that were previously unknown. For phosphorylation, 189 previously unknown sites (120 serines (S), 57 threonines (T) and 12 tyrosines (Y)) were identified out a total of 4921 sites representing 3511 modified peptides in 1779 proteins (Table 2). For example, a new phosphorylation site (Tyr-7) was found in the first α helix of profilin-1, and a new phosphorylation site (Ser-44) was found in the RGS-like domain of ARHGEF1. Interestingly, no PTM had been reported on this domain, nor in ARHGEF11 and ARHGEF12, the two other ARHGEF protein family members that contain such a domain. Further studies would be needed to assess the biological roles of these new phosphorylation sites. For N-terminal acetylation, 392 previously unknown sites were identified out a total of 1121 sites, representing 1320 modified peptides in 1086 proteins. The percentage of novel N-terminal acetylation sites is higher than that of phosphorylation, which can be explained by the lesser effort invested in the annotation of N-terminal acetylation sites in recent years, despite the fact that up to 80% of human proteins are potentially N-acetylated.[37] The detailed list of PTM identifications can be found in the Supporting Information.

Moreover, about half (1697, 48.3%) of the 3511 peptides harboring phosphorylations were also identified without. Since phosphorylation is a reversible event, this might be due to the different biological states or sample preparation protocols used for the analyses. In contrast, only 74 of the 1320 N-terminal acetylated peptides (5.61%) were additionally observed without the modification.

Identification of Single Amino Acid Variants

We identified 24 variant peptides in 21 proteins corresponding to a total of 21 variants (Table 3), representing 2.2% of the 948 nonsynonymous single amino acid variants originally contained in the customized HeLa protein database. Among the 21

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

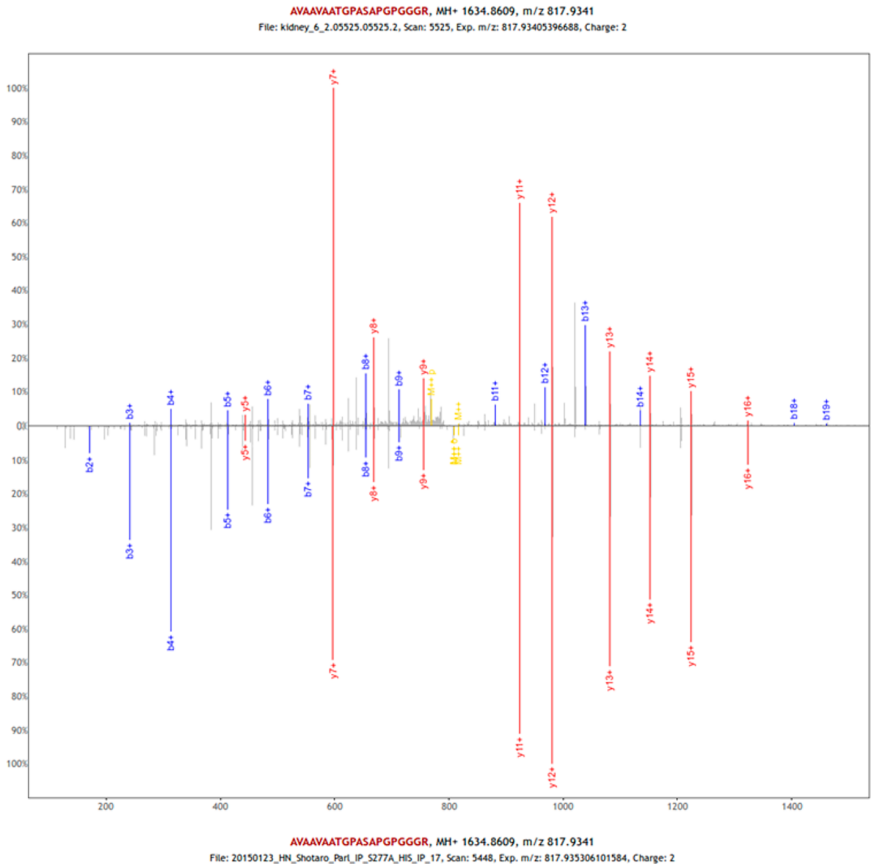


Figure 4. Spectral alignment for the AVAAVAATGPASAPGPGGGR peptide from the NX_O75474 PE2 "missing protein". At the top features the spectrum from PeptideAtlas produced by an LTQ Orbitrap Velos instrument. At the bottom features one of the two spectra identified by our workflow in the PXD005712 data set, produced by a Thermo Q Exactive instrument (mzspec:PXD005712:20150123_HN_Shotaro_Parl_IP_-S277A_HIS_IP_17:scan:5448). The two spectra show a good spectral correlation with a SDPscore of 0.789.

Table 2. Summary of the PTM Site Identifications

PTM	number of modified proteins	number of modified peptides	number of sites	number of novel sites	percent of novel sites
phosphorylation	1779	3511	4921	189	3.84%
N-terminal acetylation	1086	1320	1121	392	35.0%

variants, 16 were heterozygous and 5 homozygous. Although the number of variant identifications may seem low, it is essential to note that many factors challenge variant analysis. First of all, not all of the 20 230 human proteins are expressed at the same time in a given cell type. As seen within the results, unique peptides mapping to 7266 protein entries were detected, representing 35.9% of the total number of proteins. Among the 774 proteins harboring at least one variant that were present in the customized protein database, peptides mapping to 285 could be identified by our workflow, representing a similar rate of 36.8%. This indicates that there is no particular bias affecting the identification of variant proteins, but also that almost two-thirds of the variants in the original database had a very low probability to be observed. Moreover, not all peptides for a given protein are identified in a mass spectrometry experiment, even if it is highly expressed. Even if the variant protein is observed, only the peptides carrying the variant can bring evidence for its presence and justify further analysis.

When we compare our results to other studies using similar proteogenomic approaches on different human cell lines, we observe the same order of magnitude in term of variant identifications. For example, in 2011, Li et al. applied a bioinformatics workflow on three colorectal cancer cell lines (SW480, RKO, and HCT-116) and identified 84 variants peptides.[38] In 2016, Lobas et al. identified 112 genomics variants of the HEK-293 human kidney cell line in shotgun proteomics data[39] and Ruggles et al. reported in 2016 that around 10% of the SNVs detected by both DNA and RNA sequencing were identified at the peptide level in their analysis of breast-cancer-patient-derived xenografts (PDX) tumors.[13] While our rate of variant identification (2.2%) is slightly lower than expected based on the results of similar studies, it should be stressed that we used a high level of stringency on a very large search space.

Quantification of Single Amino Acid Variants

Nine of the identified variant peptides had their corresponding wild-type peptides also identified (Table 3). The spectral counts of both peptide versions were computed to compare their relative abundance. The application of a sign test took advantage of the large amount of data reanalyzed, as a difference in the spectral counts is expected to be more significant if it is present consistently throughout the data sets. Applying the two-tailed sign test on these variants, we obtained three cases for which the p-value was smaller than 0.025, meaning that the difference was statistically significant. In two cases

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

Table 3. List of HeLa Variant Identifications[a]

Protein	Gene	Variant	Genotype	PolyPhen prediction	Observed variant peptides	Observed wild-type peptides	Variant spectral count	Wild-type spectral count	Sign test p-value
Q8NE71	ABCF1	p.S293P	het	probably damaging	AANAAENDFSVSQAEMSPR	AANAAENDFSVSQAEMSSR	2	142	0.0002594
Q15654	TRIP6	p.S135C	het	benign	TGCLKPNPASPLPASPYGGPTPASYTTASTPAGPAFPVQVK	TGSLKPNPASPLPASPYGGPTPASYTTASTPAGPAFPVQVK,TGSLKPNPAS(p)PLPASPYGGPTPASYTTASTPAGPAFPVQVK	35	51 (incl. 2 phosphorylated)	0.003906
P26639	TARS	p.T453I	het	possibly damaging	LADFGVLHRNELSGALTGLIR, NELSGALTGLIR	LADFGVLHRNELSGALTGLTR	132	3	0.005859
P47755	CAPZA2	p.A225S	het	possibly damaging	DIQDSLTVSNEVQTSK	DIQDSLTVSNEVQTAK	2	7	0.125
P53597	SUCLG1	p.K81R	het	benign	QGTFHSQQALEYGTR	QGTFHSQQALEYGTK	35	26	0.2266
Q9C0C2	TNKS1BP1	p.G1451S	het	possibly damaging	CPARPPPSSSQGLLEEMLAASSSK	CPARPPPSGSQGLLEEMLAASSSK	3	9	0.5
Q9Y3Z3	SAMHD1	p.P26L	het	benign	TPSNTLSAEADWSPGLELHPDYK	TPSNTPSAEADWSPGLELHPDYK,TPSNTPSAEADWS(p)PGLELHPDYK	2	6 (incl. 2 phosphorylated)	0.5
P12270	TPR	p.M1293I	het	benign	ERLEQDLQQIQAK, LEQDLQQIQAK	ERLEQDLQQMQAK, LEQDLQQMQAK	11	6	0.6875
Q15633	TARBP2	p.P144A	het	benign	SPAMELQPPVSPQQSECNPVGALQELVVQK	SPPMELQPPVSPQQSECNPVGALQELVVQK	1	1	1
O43896	KIF1C	p.G421R	het	probably damaging	GALPAVSSPPAPVSPS(p)SPTTHNR,GALPAVSSPPAPVSPSST(p)THNR		2	0	*NA*
O75792	RNASEH2A	p.I253T	het	benign	EAEDVTWEDSASENQEGLR		2	0	*NA*
Q13427	PPIG	p.S257G	het	benign	SASGESEAENLEAQPQSTVRPEEIPPIPENR,S(p)AS(p)GESEAENLEAQPQSTVRPEEIPPIPENR,SASGES(p)EAENLEAQPQSTVRPEEIPPIPENR		12 (incl.6 phosphorylated)	0	*NA*
Q13505	MTX1	p.K190N	het	probably damaging	VHNISNPWQSPSGTLPALR		2	0	*NA*
Q5T9A4	ATAD3B	p.M619V	het	benign	ICSWVGTGLCPGPLSPR		1	0	*NA*
Q96RL1	UIMC1	p.R536W	het	probably damaging	HAMYCNGLMEEDTVLTWR		4	0	*NA*
Q9NP74	PALMD	p.D232N	het	probably damaging	SVYAVSSNHSAAYNGTNGLAPVEVEELLR		8	0	*NA*
O15231	ZNF185	p.E164V	hom	possibly damaging	RSSTSGDTEEEEEVEVVPFSSDEQK,RSSTSGDT(p)EEEEEVEVVPFSSDEQK		3 (incl 1 phosphorylated)	0	*NA*
P21333	FLNA	p.S1012L	hom	possibly damaging	IVGPLGAAVPCK,IVGPLGAAVPCKVEPGLGADNSVVR		44	0	*NA*
P52701	MSH6	p.G39E	hom	benign	AAAAPEASPSPGGDAAWSEAGPGPRPLAR		2	0	*NA*
P78318	IGBP1	p.R275Q	hom	benign	VFGAGYPSLPTMTVSDWYEQHQK		13	0	*NA*
Q6W2J9	BCOR	p.G1140D	hom	benign	KVS(p)DDSSHTETTAEEVPEDPLLK		7	0	*NA*

[a]The identified variant and wild-type (phospho)peptides are reported along with their corresponding spectral counts. For variants where both peptide versions were observed, a two-tailed sign-test with an α level of significance of 5% was applied. Three cases emerged as having a statistically significant abundance difference, potentially induced by the variant insertion.

(ABCF1 and TRIP6), the variant peptide was significantly less abundant than the wild-type peptide. In contrast, the p.T452I variant peptide from TARS was more abundant.

The PolyPhen-2 predictor tool[40] was additionally applied on each identified variant. This software, based on multiple protein sequence alignment and machine learning algorithms, provides a damage prediction score (benign, possibly damaging and probably damaging) for a given protein variant. About half of the 21 variants were predicted to be damaging to proteins. While it is difficult to observe a direct correlation with so limited data, it is noteworthy that the p.S293P variant, for which the most significant difference in spectral counts was observed, was also predicted to be probably damaging.

Another possible effect of an amino-acid substitution is to remove a post-translational modification site that might be important for protein function, stability, localization, or interaction with other proteins. TRIP6 is a nucleocytoplasmic protein of the zyxin family that has different functions depending

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

Table 4. Results of the CONSeQuence Tool Applied on the Four Peptide Pairs for the Three Variants That Passed the Sign Test[a]

Protein	Gene	Variant	Variant peptide	Wild-type peptide	Variant score	Wild-type score	Score difference
Q8NE71	ABCF1	p.S293P	AANAAENDFSVSQAEMSPR	AANAAENDFSVSQAEMSSR	0.570	0.441	0.129
Q15654	TRIP6	p.S135C	TGCLKPNPASPLPASPYGGPT PASYTTASTPAGPAFPVQVK	TGSLKPNPASPLPASPYGGPT PASYTTASTPAGPAFPVQVK	0.517	0.516	0.001
P26639	TARS	p.T453I	NELSGALTGLIR	NELSGALTGLTR	0.695	0.688	0.007
P26639	TARS	p.T453I	LADFGVLHRNELSGALTGLIR	LADFGVLHRNELSGALTGLTR	0.292	0.318	-0.026

[a]The scores, representing the peptide detectability by mass spectrometry, were compared between the variant and wild-type peptides to assess whether the presence of the variant alters the peptide ability to be observed.

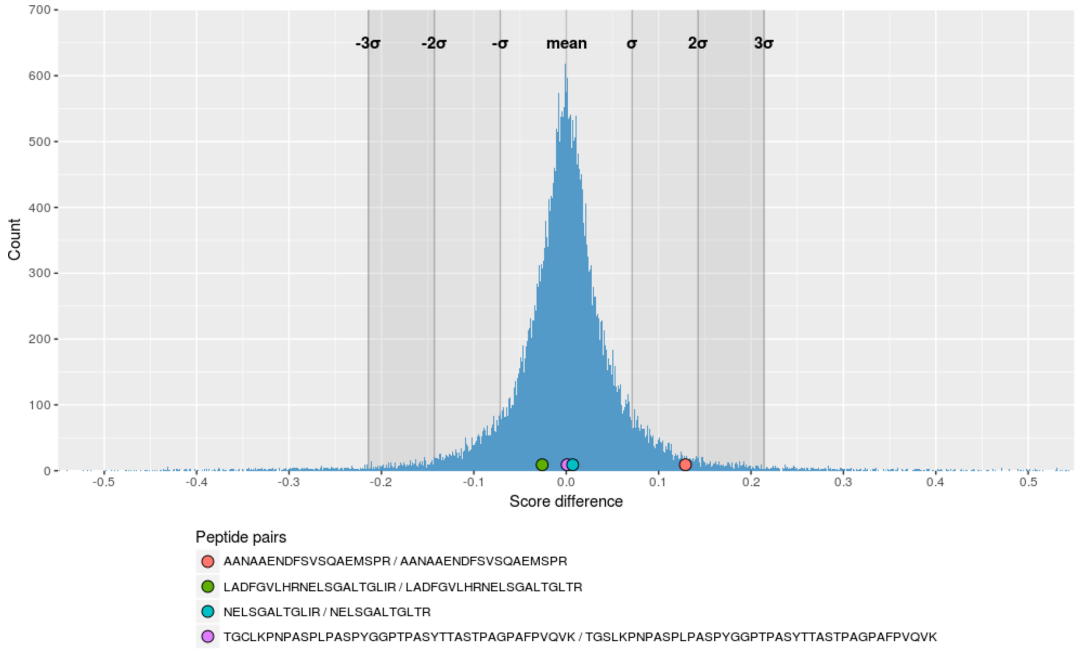


Figure 5. CONSeQuence score difference distribution. The histogram represents the distribution of the score difference for the 50 000 peptide pairs that were randomly generated. The colored dots represent the actual value of the pairs of interest.

on the cellular context and its basal expression level.[41] Ser-135 has been reported to be phosphorylated in hTERT-RPE1 cells[42] and in stem cells,[43] but the function of this phosphorylation is not known, in contrast to the well-established role of phosphotyrosine 55.[44] One can speculate that replacing Ser-135 by a cysteine might have structural and functional consequences on TRIP6 protein.

Despite a significant difference in the spectral counts of several variants, the interpretation of those results is challenging as many factors need to be taken into consideration. The main concern is that the introduction of a variant inside a peptide can modify its physicochemical properties, affecting in turn its detectability. The CONSeQuence tool[45] was used to test this hypothesis. It is based on the combination of four different machine learning algorithms predicting the peptides that are highly detectable by mass spectrometry. The tool was applied on the variant and wild-type peptides that passed the sign test with the "rank score" prediction type (Table 4).

Assuming that a significant score difference between peptide pairs indicates a bias potentially affecting the corresponding spectral counts, we plotted the CONSeQuence score difference distribution (Figure 4). To generate the plot, 1000 tryptic peptides with a randomized sequence 7 to 40 amino acid-long were produced. For each peptide, 50 random nonsynonymous single amino acid variants were independently inserted at random positions, leading to a total of 50 000 variant peptides. CONSeQuence was then applied on all the generated peptides, and the score difference between each variant peptide and its original version was plotted. The most extreme differences tend to be due to the insertion or removal of a cleavage site, as they produce a large variation in peptide sizes.

The score differences for the four peptide pairs were added on the same plot (Figure 5). All values were comprised in a range of two standard deviation (0.143) from the average of 0.000. The largest difference is found in the peptide pair for the p.S293P variant with a value of 0.129. However, this higher value of the variant peptide that was predicted as more

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

detectable by CONSeQuence had actually a lower spectral count than its wild-type counterpart. Based on these results, none of the three variants passing the sign test is predicted to have a bias in its spectral count pairs through modification of peptide detectability.

Variants may also modify the sensitivity of the protein to trypsin cleavage. For example, TARS Thr-453 has been reported to be phosphorylated in HeLa cells.[46] Since this site is adjacent to the trypsin cleavage site, cleavage could be impaired so that only the dephosphorylated form of the peptide would be detectable. That would explain the very low spectral count of the WT peptide compared to the p.T453I variant peptide which is more responsive to trypsin.

CONCLUSIONS

By reanalyzing 41 publicly available data sets obtained on the HeLa cancer cell line, we identified 49 466 unique peptides in 7266 protein entries with a 1% FDR threshold computed at the protein level and validated 5576 protein entries in accordance with the HPP guidelines version 2.1. The identified proteins included a new "missing protein", for which two non-nested peptides of length greater than nine amino acids were found. Alternative explanations, such as the presence of a variant, were taken into account to ensure that the identifications were not ambiguous. 189 new phosphorylation and 392 new N-terminal acetylation sites were also identified, highlighting the interest of reanalyzing public data for new findings, as most experiments generally focus on specific aims.

The spectral counting quantification approach was used to evaluate the impact of single amino acid variants on protein expression and stability. This required a large number of mass spectra to reach a satisfactory level of statistical power. Finally, the spectral counts of only three of the 21 identified variant peptides were significantly different from those of their wild-type counterparts after the application of a two-tailed sign test. Although we anticipated most of the common biases that may influence the spectral counts in the variant analysis, it is important to note the moderate accuracy of the spectral counting approach. A dedicated SRM experiment could further validate these three cases with higher precision. Currently, the HeLa cell line has by far the highest number of associated data sets available in PRIDE (116). In order to validate other missing proteins, find new PTM sites, or the functional effects of other SNPs, the same protocol could be applied to HEK293 (27 data sets), MCF-7 (23 data sets), Jurkat (15 data sets), U2OS (13 data sets), or HCT 116 (10 data sets), for which exome sequencing data is available.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00392.

Descriptions of file contents (PDF)
Supporting data (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: lydie.lane@sib.swiss. Tel: +41 (0) 22 379 58 41.

ORCID

Thibault Robin: 0000-0001-6548-709X
Frédérique Lisacek: 0000-0002-0948-4537
Lydie Lane: 0000-0002-9818-3030

Author Contributions

TR, AB and LL conceived and designed the analyses. MM supervised the development of the MzVar Java tool and TR wrote the code. TR and LL performed the analyses. LL and FL co-coordinated the study. TR and LL wrote the manuscript. All the authors revised the manuscript and approved its final version.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

All authors were supported by the University of Geneva. The authors would like to thank Alain Gateau for help with checking peptide uniqueness, the neXtProt team for help with neXtProt API, David Shteynberg for his assistance with the PTMProphet tool and Jean-Luc Falcone for his advice about statistics.

REFERENCES

(1) Morozova, O.; Marra, M. A. Applications of Next-Generation Sequencing Technologies in Functional Genomics. *Genomics* **2008**, *92* (5), 255−264.

(2) Metzker, M. L. Sequencing Technologies - the next Generation. *Nat. Rev. Genet.* **2010**, *11* (1), 31−46.

(3) Cirulli, E. T.; Goldstein, D. B. Uncovering the Roles of Rare Variants in Common Disease through Whole-Genome Sequencing. *Nat. Rev. Genet.* **2010**, *11* (6), 415−425.

(4) Ansong, C.; Purvine, S. O.; Adkins, J. N.; Lipton, M. S.; Smith, R. D. Proteogenomics: Needs and Roles to Be Filled by Proteomics in Genome Annotation. *Briefings Funct. Genomics Proteomics* **2008**, *7* (1), 50−62.

(5) Greenbaum, D.; Colangelo, C.; Williams, K.; Gerstein, M. Comparing Protein Abundance and MRNA Expression Levels on a Genomic Scale. *Genome Biol.* **2003**, *4* (9), 117.

(6) Jaffe, J. D.; Berg, H. C.; Church, G. M. Proteogenomic Mapping as a Complementary Method to Perform Genome Annotation. *Proteomics* **2004**, *4* (1), 59−77.

(7) Nesvizhskii, A. I. Proteogenomics: Concepts, Applications and Computational Strategies. *Nat. Methods* **2014**, *11* (11), 1114−1125.

(8) Sheynkman, G. M.; Shortreed, M. R.; Cesnik, A. J.; Smith, L. M. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annu. Rev. Anal. Chem.* **2016**, *9* (1), 521−545.

(9) Wen, B.; Xu, S.; Zhou, R.; Zhang, B.; Wang, X.; Liu, X.; Xu, X.; Liu, S. PGA: An R/Bioconductor Package for Identification of Novel Peptides Using a Customized Database Derived from RNA-Seq. *BMC Bioinf.* **2016**, *17* (1), 244.

(10) Wang, X.; Zhang, B. CustomProDB: An R Package to Generate Customized Protein Databases from RNA-Seq Data for Proteomics Search. *Bioinformatics* **2013**, *29* (24), 3235−3237.

(11) Wen, B.; Xu, S.; Sheynkman, G. M.; Feng, Q.; Lin, L.; Wang, Q.; Xu, X.; Wang, J.; Liu, S. SapFinder: An R/Bioconductor Package for Detection of Variant Peptides in Shotgun Proteomics Experiments. *Bioinformatics* **2014**, *30* (21), 3136−3138.

(12) Sheynkman, G. M.; Johnson, J. E.; Jagtap, P. D.; Shortreed, M. R.; Onsongo, G.; Frey, B. L.; Griffin, T. J.; Smith, L. M. Using Galaxy-P to Leverage RNA-Seq for the Discovery of Novel Protein Variations. *BMC Genomics* **2014**, *15*, 703.

(13) Ruggles, K. V.; Tang, Z.; Wang, X.; Grover, H.; Askenazi, M.; Teubl, J.; Cao, S.; McLellan, M. D.; Clauser, K. R.; Tabb, D. L.; et al. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol. Cell. Proteomics* **2016**, *15* (3), 1060−1071.

(14) Krasnov, G. S.; Dmitriev, A. A.; Kudryavtseva, A. V.; Shargunov, A. V.; Karpov, D. S.; Uroshlev, L. A.; Melnikova, N. V.; Blinov, V. M.; Poverennaya, E. V.; Archakov, A. I.; et al. PPLine: An

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics. *J. Proteome Res.* **2015**, *14* (9), 3729−3737.
(15) Alfaro, J. A.; Ignatchenko, A.; Ignatchenko, V.; Sinha, A.; Boutros, P. C.; Kislinger, T. Detecting Protein Variants by Mass Spectrometry: A Comprehensive Study in Cancer Cell-Lines. *Genome Med.* **2017**, *9* (1), 62.
(16) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−226.
(17) Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; et al. 2016 Update of the PRIDE Database and Its Related Tools. *Nucleic Acids Res.* **2016**, *44* (D1), D447−456.
(18) Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *Journal of Biomolecular Techniques: JBT* **2018**, jbt.18-2902-002.
(19) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013−2014 and Strategies for Finding Missing Proteins. *J. Proteome Res.* **2014**, *13* (1), 15−20.
(20) Kinsella, R. J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; et al. Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space. *Database* **2011**, *2011*, bar030.
(21) Lane, L.; Argoud-Puy, G.; Britan, A.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gaudet, P.; Gleizes, A.; Masselot, A.; et al. NeXtProt: A Knowledge Platform for Human Proteins. *Nucleic Acids Res.* **2012**, *40*, D76−83.
(22) Adey, A.; Burton, J. N.; Kitzman, J. O.; Hiatt, J. B.; Lewis, A. P.; Martin, B. K.; Qiu, R.; Lee, C.; Shendure, J. The Haplotype-Resolved Genome and Epigenome of the Aneuploid HeLa Cancer Cell Line. *Nature* **2013**, *500* (7461), 207−211.
(23) Liu, Y.; Mi, Y.; Mueller, T.; Kreibich, S.; Williams, E. G.; Van Drogen, A.; Borel, C.; Germain, P.-L.; Frank, M.; Bludau, I. Genomic, Proteomic and Phenotypic Heterogeneity in HeLa Cells across Laboratories: Implications for Reproducibility of Research Results. *bioRxiv* **2018**, DOI: 10.1101/307421.
(24) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, *24* (21), 2534−2536.
(25) Craig, R.; Beavis, R. C. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.
(26) Horlacher, O.; Nikitin, F.; Alocci, D.; Mariethoz, J.; Müller, M.; Lisacek, F. MzJava: An Open Source Library for Mass Spectrometry Data Processing. *J. Proteomics* **2015**, *129*, 63−70.
(27) Schaeffer, M.; Gateau, A.; Teixeira, D.; Michel, P.-A.; Zahn-Zabal, M.; Lane, L. The NeXtProt Peptide Uniqueness Checker: A Tool for the Proteomics Community. *Bioinformatics* **2017**, *33* (21), 3471−3472.
(28) Deutsch, E. W.; Overall, C. M.; Van Eyk, J. E.; Baker, M. S.; Paik, Y.-K.; Weintraub, S. T.; Lane, L.; Martens, L.; Vandenbrouck, Y.; Kusebauch, U.; et al. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **2016**, *15* (11), 3961−3970.
(29) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.
(30) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a Standardized Data Processing Pipeline for Large-Scale Reproducible Proteomics Informatics. *Proteomics: Clin. Appl.* **2015**, *9* (7−8), 745−754.
(31) Saita, S.; Nolte, H.; Fiedler, K. U.; Kashkar, H.; Venne, A. S.; Zahedi, R. P.; Krüger, M.; Langer, T. PARL Mediates Smac Proteolytic Maturation in Mitochondria to Promote Apoptosis. *Nat. Cell Biol.* **2017**, *19* (4), 318−328.
(32) Ye, D.; Fu, Y.; Sun, R.-X.; Wang, H.-P.; Yuan, Z.-F.; Chi, H.; He, S.-M. Open MS/MS Spectral Library Search to Identify Unanticipated Post-Translational Modifications and Increase Spectral Identification Rate. *Bioinformatics* **2010**, *26* (12), i399−406.
(33) Peng, X.; Xu, F.; Liu, S.; Li, S.; Huang, Q.; Chang, L.; Wang, L.; Ma, X.; He, F.; Xu, P. Identification of Missing Proteins in the Phosphoproteome of Kidney Cancer. *J. Proteome Res.* **2017**, *16* (12), 4364−4373.
(34) van Amerongen, R.; Nawijn, M. C.; Lambooij, J.-P.; Proost, N.; Jonkers, J.; Berns, A. Frat Oncoproteins Act at the Crossroad of Canonical and Noncanonical Wnt-Signaling Pathways. *Oncogene* **2010**, *29* (1), 93−104.
(35) Hou, J.; Tobe, B. T. D.; Lo, F.; Blethrow, J. D.; Crain, A. M.; Wolf, D. A.; Snyder, E. Y.; Singec, I.; Brill, L. M. Combined Total Proteomic and Phosphoproteomic Analysis of Human Pluripotent Stem Cells. *Methods Mol. Biol.* **2013**, *1029*, 163−189.
(36) Zhao, M.; Wei, W.; Cheng, L.; Zhang, Y.; Wu, F.; He, F.; Xu, P. Searching Missing Proteins Based on the Optimization of Membrane Protein Enrichment and Digestion Process. *J. Proteome Res.* **2016**, *15* (11), 4020−4029.
(37) Arnesen, T.; Van Damme, P.; Polevoda, B.; Helsens, K.; Evjenth, R.; Colaert, N.; Varhaug, J. E.; Vandekerckhove, J.; Lillehaug, J. R.; Sherman, F.; et al. Proteomics Analyses Reveal the Evolutionary Conservation and Divergence of N-Terminal Acetyltransferases from Yeast and Humans. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (20), 8157−8162.
(38) Li, J.; Su, Z.; Ma, Z.-Q.; Slebos, R. J. C.; Halvey, P.; Tabb, D. L.; Liebler, D. C.; Pao, W.; Zhang, B. A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics. *Mol. Cell. Proteomics* **2011**, *10* (5), M110.006536.
(39) Lobas, A. A.; Karpov, D. S.; Kopylov, A. T.; Solovyeva, E. M.; Ivanov, M. V.; Ilina, I. Y.; Lazarev, V. N.; Kuznetsova, K. G.; Ilgisonis, E. V.; Zgoda, V. G.; et al. Exome-Based Proteogenomics of HEK-293 Human Cell Line: Coding Genomic Variants Identified at the Level of Shotgun Proteome. *Proteomics* **2016**, *16* (14), 1980−1991.
(40) Adzhubei, I.; Jordan, D. M.; Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **2013**, *76*, 7.20.1.
(41) Willier, S.; Butt, E.; Richter, G. H. S.; Burdach, S.; Grunewald, T. G. P. Defining the Role of TRIP6 in Cell Physiology and Cancer. *Biol. Cell* **2011**, *103* (12), 573−591.
(42) Oppermann, F. S.; Grundner-Culemann, K.; Kumar, C.; Gruss, O. J.; Jallepalli, P. V.; Daub, H. Combination of Chemical Genetics and Phosphoproteomics for Kinase Signaling Analysis Enables Confident Identification of Cellular Downstream Targets. *Mol. Cell. Proteomics* **2012**, *11* (4), O111.012351.
(43) Phanstiel, D. H.; Brumbaugh, J.; Wenger, C. D.; Tian, S.; Probasco, M. D.; Bailey, D. J.; Swaney, D. L.; Tervo, M. A.; Bolin, J. M.; Ruotti, V.; et al. Proteomic and Phosphoproteomic Comparison of Human ES and IPS Cells. *Nat. Methods* **2011**, *8* (10), 821−827.
(44) Lai, Y.-J.; Chen, C.-S.; Lin, W.-C.; Lin, F.-T. C-Src-Mediated Phosphorylation of TRIP6 Regulates Its Function in Lysophosphatidic Acid-Induced Cell Migration. *Mol. Cell. Biol.* **2005**, *25* (14), 5859−5868.
(45) Eyers, C. E.; Lawless, C.; Wedge, D. C.; Lau, K. W.; Gaskell, S. J.; Hubbard, S. J. CONSeQuence: Prediction of Reference Peptides for Absolute Quantitative Proteomics Using Consensus Machine Learning Approaches. *Mol. Cell. Proteomics* **2011**, *10* (11), M110.003384.
(46) Zhou, H.; Di Palma, S.; Preisinger, C.; Peng, M.; Polat, A. N.; Heck, A. J. R.; Mohammed, S. Toward a Comprehensive Characterization of a Human Cancer Cell Phosphoproteome. *J. Proteome Res.* **2013**, *12* (1), 260−271.

DOI: 10.1021/acs.jproteome.8b00392
J. Proteome Res. 2018, 17, 4160−4170

Journal of proteome research

Cite This: J. Proteome Res. 2019, 18, 1926−1927

Addition/Correction

pubs.acs.org/jpr

Correction to "Large-Scale Reanalysis of Publicly Available HeLa Cell Proteomics Data in the Context of the Human Proteome Project"

Thibault Robin, Amos Bairoch, Markus Müller, Frédérique Lisacek, and Lydie Lane*

J. Proteome Res. **2018**, *17* (12), 4160−4170. DOI: 10.1021/acs.jproteome.8b00392

Supporting Information

PROBLEM DESCRIPTION

While designing a new experiment, we realized that we had a data handling problem in the results of our previous study "Large-Scale Reanalysis of Publicly Available HeLa Cell Proteomics Data in the Context of the Human Proteome Project". Out of the 1233 tandem mass spectrometry files that were originally processed in that work, 27 turned out to be associated with cell lines other than HeLa. The problematic files come from two distinct data sets: PXD001426 (3/3 files, originating from the HCT 116 cell line) and PXD002395 (24/42 files, originating from the HepG2 and HEK293 cell lines). The PXD001426 data set was not produced on HeLa cells and was improperly selected to be analyzed by our workflow. Even if HeLa was used in a subpart of this study, the actual raw files available on PRIDE all belong to HCT 116. The PXD002395 data set was produced on a large panel of 11 distinct cell lines including HeLa. Of note, the raw files were properly annotated in the PRIDE database. Our oversight was a loose regular expression "He*", which caused the retrieval of files matching cell line names starting with He and that we omitted to double-check.

REANALYSIS

The workflow previously described was reapplied on the 1206 files from the 40 data sets associated with the HeLa cell line after the removal of the 27 incriminated files (representing a loss of 1 103 531 tandem mass spectra, 2% of the total amount). The methods detailed in the original article were identically reapplied. The FDR threshold was recomputed at the different levels, and the N-terminal acetylation and phosphorylation sites were validated again. The variant and "missing protein" identifications were also revalidated, and the sign-test statistical analysis was rerun. The different tables from the Supporting Information were rebuilt.

RESULTS

The reanalysis of the 40 HeLa data sets led to the identification of 48 583 (−883) unique peptides in 7174 (−92) protein entries, allowing the validation of 5508 (−68) protein entries in accordance with the HPP guidelines version 2.1. The X! Tandem hyperscore threshold remained at 60.4 after recomputation, ensuring that the global FDR would be set to 1.0% at the protein level. All of the missing protein identifications were conserved. The validation of two N-terminal acetylation sites was lost, reducing the number of validated sites to 390, while no phosphorylation site validation was lost (Table 1). The identification of the wild-type peptide of the TARS p.T453I variant was lost (Table 2), removing as a consequence the variant from the sign-test and reducing the number of heterozygous cases where a significant difference was detected to two out of eight. Overall, the removal of the 27 files that did not belong to the HeLa cell line only slightly alters our results and does not change the conclusions that we drew in the original article.

Table 1. Summary of the PTM Site Identifications

PTM	number of modified proteins	number of modified peptides	number of sites	number of novel sites	percent of novel sites
phosphorylation	1779	3511	4921	189	3.84%
N-terminal acetylation	1048	1246	1083	390	36.0%

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00113.

Descriptions of file contents (PDF)
Supporting data (ZIP)

Published: February 25, 2019

Table 2. List of HeLa Variant Identifications[a]

Protein	Gene	Variant	Genotype	PolyPhen prediction	Observed variant peptides	Observed wild-type peptides	Variant spectral count	Wild-type spectral count	Sign test p-value
Q8NE71	ABCF1	p.S293P	het	probably damaging	AANAAENDFSVSQAEMS**P**R	AANAAENDFSVSQAEMS**S**R	2	140	0.0002594
Q15654	TRIP6	p.S135C	het	benign	TG**C**LKPNPASPLPASPYGGPTPASYTTASTPAGPAFPVQVK	TG**S**LKPNPASPLPASPYGGPTPASYTTASTPAGPAFPVQVK,TG**S**LKPNPA**S(p)**PLPASPYGGPTPASYTTASTPAGPAFPVQVK	35	51 (incl. 2 phosphorylated)	0.003906
P47755	CAPZA2	p.A225S	het	possibly damaging	DIQDSLTVSNEVQT**S**K	DIQDSLTVSNEVQT**A**K	2	7	0.125
P53597	SUCLG1	p.K81R	het	benign	QGTFHSQQALEYGT**R**	QGTFHSQQALEYGT**K**	35	26	0.2266
Q9C0C2	TNKS1BP1	p.G1451S	het	possibly damaging	CPARPPPS**S**SQGLLEEMLAASSSK	CPARPPPS**G**SQGLLEEMLAASSSK	3	9	0.5
P12270	TPR	p.M1293I	het	benign	ERLEQDLQQIQAK, LEQDLQQ**I**QAK	ERLEQDLQQ**M**QAK, LEQDLQQ**M**QAK	11	6	0.6875
Q15633	TARBP2	p.P144A	het	benign	SP**A**MELQPPVSPQQSECNPVGALQELVVQK	SP**P**MELQPPVSPQQSECNPVGALQELVVQK	1	1	1
Q9Y3Z3	SAMHD1	p.P26L	het	benign	TPSNT**L**SAEADWSPGLELHPDYK	TPSNT**P**SAEADWSPGLELHPDYK,TPSNT**P**SAEADW**S(p)**PGLELHPDYK	2	4 (incl. 2 phosphorylated)	1
O43896	KIF1C	p.G421R	het	probably damaging	GALPAVSSPPAPVSP**S(p)**SPTTHN**R**,GALPAVSSPPAPVSPSSP**T(p)**THN**R**		2	0	*NA*
O75792	RNASEH2A	p.I253T	het	benign	EAEDV**T**WEDSASENQEGLR		2	0	*NA*
P26639	TARS	p.T453I	het	possibly damaging	LADFGVLHRNELSGALTGL**I**R, NELSGALTGL**I**R		132	0	*NA*
Q13427	PPIG	p.S257G	het	benign	SAS**G**ESEAENLEAQPQSTVRPEEIPPIPENR,**S(p)**A**S(p)**G**E**S**EAENLEAQPQSTVRPEEIPPIPENR,SAS**G**E**S(p)**EAENLEAQPQSTVRPEEIPPIPENR		12 (incl.6 phosphorylated)	0	*NA*
Q13505	MTX1	p.K190N	het	probably damaging	VH**N**ISNPWQSPSGTLPALR		2	0	*NA*
Q5T9A4	ATAD3B	p.M619V	het	benign	ICSW**V**GTGLCPGPLSPR		1	0	*NA*
Q96RL1	UIMC1	p.R536W	het	probably damaging	HAMYCNGLMEEDTVLT**W**R		4	0	*NA*
Q9NP74	PALMD	p.D232N	het	probably damaging	SVYAVSSNHSAAYNGT**N**GLAPVEVEELLR		8	0	*NA*
O15231	ZNF185	p.E164V	hom	possibly damaging	RSSTSGDTEEEEE**V**EVVPFSSDEQK,RSSTSGD**T(p)**EEEEE**V**EVVPFSSDEQK		3 (incl 1 phosphorylated)	0	*NA*
P21333	FLNA	p.S1012L	hom	possibly damaging	IVGPLGAAVPCK,IVGP**L**GAAVPCKVEPGLGADNSVVR		44	0	*NA*
P52701	MSH6	p.G39E	hom	benign	AAAAP**E**ASPSPGGDAAWSEAGPGPRPLAR		2	0	*NA*
P78318	IGBP1	p.R275Q	hom	benign	VFGAGYPSLPTMTVSDWYEQH**Q**K		13	0	*NA*
Q6W2J9	BCOR	p.G1140D	hom	benign	KV**S(p)**DDSSHTETTAEEVPEDPLLK		7	0	*NA*

[a]The identified variant and wild-type (phospho)peptides are reported along with their corresponding spectral counts. For variants where both peptide versions were observed, a two-tailed sign-test with an α level of significance of 5% was applied. Two cases emerged as having a statistically significant abundance difference, potentially induced by the variant insertion.

DOI: 10.1021/acs.jproteome.9b00113
J. Proteome Res. 2019, 18, 1926−1927

4.2. Concluding Remarks

This extensive analysis of publicly available proteomics data from the HeLa cell line showcased the wealth of dormant information in online databases and repositories. As proteomics experiments commonly have restricted aims, data sharing enables the design of new studies with novel goals such as the investigation of specific proteoforms. However, the large-scale reanalysis of proteomics data creates numerous technical challenges. The metadata information provided in the PRIDE database is for instance not curated nor standardized, making it difficult when searching data sets that have specific criteria (e.g., species, tissue of origin or disease). It is thus required to validate such information using either publications, which are not always referenced, or raw binary files when they are available. Nowadays proteomics journals and editors enforce the deposition of all raw and identification data prior to publication. The case of data reanalysis, especially when a large number of data sets was processed, is however poorly supported in terms of infrastructure. Following the ProteomeXchange guidelines, one cannot submit processed identification data without submitting the corresponding raw mass spectra. In the case of our study, it would have been impractical to resubmit so many raw data sets that were already stored in the database and additionally would risk raising authorship issues. Exceptionally, our article was exempted from the obligation of depositing the data that we used since no practical solution to avoid the above-mentioned issues existed at the time this study was carried out. There is consequently still room for improvement in the reanalysis of publicly available data sets.

Chapter 5

GlyConnect Compozitor

5.1. Overview

In this chapter, we introduce the Glyconnect Compozitor web application allowing visualizing, selecting, and exporting glycan compositions. The wealth of glycosylation information stored in the Glyconnect database is structured based on distinct entities, such as taxonomy, biological source, cell line, protein, glycosylation site, and disease. Although each corresponding entry page provides the known glycans in the form of structures and compositions, it can be challenging to determine if they are specific to this context or commonly found. Moreover, a growing proportion of glycans lacks fully resolved structures when resulting from high-throughput glycoproteomics experiments. To identify intact glycopeptides, a set of glycan compositions is a required input in glycoproteomics search engines in order to define the search space. However, there is currently a lack of consensus between studies regarding the optimal choice of compositions. Glyconnect Compozitor enables researchers to explore and compare the known relative glycomes of multiple biological entities and detect the compositions best fitted to their needs. They can subsequently be exported in various formats to be used by other bioinformatics tools.

MCP RESEARCH

⌘ Author's Choice

Examining and Fine-tuning the Selection of Glycan Compositions with GlyConnect Compozitor

Thibault Robin[1,2,3,4], Julien Mariethoz[1,2], and Frédérique Lisacek[1,2,5,*]

A key point in achieving accurate intact glycopeptide identification is the definition of the glycan composition file that is used to match experimental with theoretical masses by a glycoproteomics search engine. At present, these files are mainly built from searching the literature and/or querying data sources focused on posttranslational modifications. Most glycoproteomics search engines include a default composition file that is readily used when processing MS data. We introduce here a glycan composition visualizing and comparative tool associated with the GlyConnect database and called GlyConnect Compozitor. It offers a web interface through which the database can be queried to bring out contextual information relative to a set of glycan compositions. The tool takes advantage of compositions being related to one another through shared monosaccharide counts and outputs interactive graphs summarizing information searched in the database. These results provide a guide for selecting or deselecting compositions in a file in order to reflect the context of a study as closely as possible. They also confirm the consistency of a set of compositions based on the content of the GlyConnect database. As part of the tool collection of the Glycomics@ExPASy initiative, Compozitor is hosted at https://glyconnect.expasy.org/compozitor/ where it can be run as a web application. It is also directly accessible from the GlyConnect database.

One of the most interesting current challenges in proteomics research is estimating the extent of protein diversity as recently summarized in (1). To encompass all molecular forms of a protein following genetic variation, alternative splicing and posttranslational processing, the term "proteoform" was recently proposed (2) and rapidly adopted by the community. Yet, the array of posttranslational modifications is not fully characterized and the particular case of glycosylation raises many technical issues that glycoproteomics has started to address on a large-scale basis only a few years ago. Recent examples include cancer (3, 4) or mouse brain (5) N-glycome profiling studies. These studies rely on proteomics software adapted to identifying glycopeptides such as Mascot (6) and ProteinProspector (7) or glycoproteomics dedicated software developed in recent years as reviewed in (8).

A key point in achieving accurate glycopeptide identification is the selection of a glycan composition file that will be used to match experimental with theoretical masses as is usually implemented in MS (MS) search engines. The definition of this composition file differs in the range of software commonly used. It is generally selected to cover as many monosaccharide combinations as possible and is seldom customized to reflect the expectable constraints imposed on glycan expression in a specific tissue or organism. For example, in the very popular Byonic search engine (9), the default list of N-glycan compositions is set to 309 items and usage reported in publications more often than not shows a selection of default values when human samples are processed. Nonetheless, the list can be customized and for example, the mouse brain study (5) describes the selection of compositions extracted from the literature with relevant criteria (*e.g.* relevant species and tissue). At present, searching the literature and/or querying other data sources such as Unimod (http://www.unimod.org) appear to be the most frequent approach to customizing a composition file.

In 2017, the HUPO Human Glycoproteomics Initiative launched an interlaboratory study to assess the performance of glycoproteomics software for automated intact N- and O-glycopeptide identification from high resolution MS/MS data. Benchmark datasets were provided to participants split as developers (write software) or users (use software tool selected from the range of existing ones). Data were analyzed and the results sent back to the challenge organizers for evaluation. This revealed widespread variations

From the [1]Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva, Switzerland; [2]Computer Science Department, Faculty of Science, University of Geneva, Switzerland; [3]CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva, Switzerland; [4]Microbiology and Molecular Medicine Department, Faculty of Medicine, University of Geneva, Switzerland; [5]Section of Biology, Faculty of Science, University of Geneva, Switzerland

This article contains supplemental data.

*For correspondence: Frédérique Lisacek, frederique.lisacek@sib.swiss.

DOI 10.1074/mcp.RA120.002041

TABLE I
Common glycan composition residues and corresponding abbreviations

Residue Type	Short Name	GlyConnect Notation	Byonic Notation	Condensed Notation
Monosaccharide	Hexose	Hex	Hex	H
Monosaccharide	N-Acetylhexosamine	HexNAc	HexNAc	N
Monosaccharide	Deoxyhexose	dHex	dHex	F
Monosaccharide	N-Acetylneuraminic acid	NeuAc	NeuAc	S
Monosaccharide	N-Glycolylneuraminic acid	NeuGc	NeuGc	G
Monosaccharide	Pentose	Pent	Pent	P
Monosaccharide	Hexuronic acid	HexA	HexA*	A
Monosaccharide	Ketodeoxyoctonic acid	Kdn	Kdn*	K
Monosaccharide	Ketodeoxynononic acid	Kdo	Kdn*	O
Substituent	Acetyl	Ac	Acetyl	a
Substituent	Methyl	Me	Me*	m
Substituent	Phosphate	Ph	Phospho	p
Substituent	Sulfate	Su	Sulfo	s

*Based on the official documentation, these residues lacking from Byonic were replaced by those of the GlyConnect notation.

in the definition of composition files and the subsequent variations in the quality and extent of glycopeptide identification.

We describe here a web-based tool destined to assist glycoproteomics software users in selecting appropriate N- or O-linked glycan compositions with respect to sample specifications encompassing species, tissue or cell line type and disease. This tool named *Compozitor,* relies on the data collected in the GlyConnect resource (10), which includes glycomics and glycoproteomics data. In fact, Compozitor reveals global glycomic information associated with a species, a glycoprotein, a cell or a tissue that cannot be captured when reading through the corresponding GlyConnect entries. In particular, a glycome is often provided as a list of glycan structures or a list of compositions or a mix thereof, as if the items were independent when they obviously are not. The interface of GlyConnect described in (10) was a first attempt to link and visualize glycomic and proteomic data, *e.g.* addressing the question of which glycan(s) is/are attached to which protein(s). Navigation in the database did not support the comparative investigation of structural data dependences, *e.g.* addressing the question of detecting glycan compositional trends within and similarities across protein(s) or tissues.

As part of the tool collection of the Glycomics@ExPASy initiative (11), Compozitor is hosted on the ExPASy server (12) of the SIB Swiss Institute of Bioinformatics at https://glyconnect.expasy.org/compozitor/ where it can be run as a web application. It is also referenced in GlyConnect glycoprotein, source, reference and disease pages to offer a new view on the data. It appears in the cross-reference section of these pages. The present article describes the tool and demonstrates its use through a series of use cases arising from the exploration of the GlyConnect database content. It also tackles the comparison of composition files used in several intact glycopeptide search engines and suggests options for rationalizing a selection.

MATERIALS AND METHODS

GlyConnect Glycosylation Data—The GlyConnect database stores curated data on glycosylation, glycans and glycoproteins extracted from literature. It was built upon the wealth of information contained in the GlycoSuiteDB database (13). GlyConnect was then enriched with data and annotations provided by Nicki Packer's group through collaborative work (10, 11) and by the recent integration of selected published work on high throughput MS experiments using a range of identification software in various applications. The May 2020 release of GlyConnect contains 246 species, 2662 proteins, 5675 glycosites, 1041 compositions, 3609 defined and 451 ambiguously defined structures. There are twice as many N- than O-linked recorded in general and as a reflection of the current bias in the recent literature, 53% of site-specific data corresponds to human *N*-linked glycans. Note that 6% of the reported glycosylation is relative to released glycans.

The data are provided to users through a web application. Two types of representation are available. The application is either serving HTML web pages for human readability or JSON (JavaScript Object Notation) with a RESTful API (Application Programming Interface) for software applications. The latter is used by GlyConnect Compozitor to populate the menu options and return the results of the composition search query. The RESTfull API can also be used to convert different composition notation formats or return GlyConnect or GlyTouCan identifiers as well as monoisotopic masses.

Composition Notation—There is no single way of representing glycan compositions and Compozitor currently implements three of the most common notations, especially in mammalian studies. The residue set is composed of monosaccharides and substituents (see Table I). In contrast to resolved glycan structures, the number of monosaccharides in compositions is reduced to a smaller set of residues that have identical molecular masses (e.g.180 Da for one hexose that is valid for galactose and glucose).

The GlyConnect notation is inherited from one of the oldest software tools managing compositional data known as GlycoMod (14). It uses an abbreviated code for residues and a semicolon as a separator between this code and the corresponding number of monosaccharides. Then each such pair is separated from the next by a space. For example, the N-glycan core made of two N-acetylhexosamine and three hexoses is represented as: Hex:3 HexNAc:2.

The Byonic notation is included to ease export from the interface to a file directly importable in the analysis software for customizing compositions. It uses an abbreviated code for residues like the

GlyConnect notation, but the corresponding number of residues for this code is between brackets and no space separates code-number pairs. The N-glycan core is represented as: Hex(3)HexNAc(2).

Finally, the condensed format uses a one-letter code immediately followed by the corresponding number of residues and does not use separators for code-number pairs. The N-glycan core is represented as: H3N2. In contrast to the two other formats, the condensed notation is case-sensitive with the monosaccharide using uppercase letters and substituents using lowercase letters. Therefore, some residues may share the same letter with a different case (*e.g.* upper S for N-acetylneuraminic acid and lower s for sulfate).

The order in which the residue is reported is consistent. It starts with the hexoses, followed by N-acetylhexosamine, deoxyhexoses, sialic acids, and substituents in all notations. Of note, this order is a tacit consensus with no biological relevance attached to it.

In the rest of this article we shall refer to the GlyConnect, Byonic and condensed notations to specify which format is being used in which context.

Search Criteria—Search criteria correspond to expected sources of variation in glycan expression such as species, tissue or disease, and they also reflect the information stored in the GlyConnect database. The interface of Compozitor offers six tabs labelled with search criteria, namely: protein, source, cell line, disease, custom and advanced.

In the *protein*, *source* and *disease* tabs, a "species" drop-down list can first be used to select the relevant species. This selection step is skipped in the case of cell lines, as they represent only a small number of entries. Once the species is chosen, the available options for the corresponding biological entities are proposed in a second drop-down list. In the case of the source tab, three distinct drop-down lists are provided corresponding to the three biological source categories available in GlyConnect: tissue, cell type and cell component. Of note, several sources can be selected at the same time when available. In a third step, a "glycan type" drop-down list enables to filter the retrieved glycan compositions based on the type of linkage (*i.e. N*-linked, O-linked or C-linked). In the protein tab specifically, an additional "sites" drop-down list is instantiated each time a glycan type is selected, displaying all known glycosylation sites of the corresponding protein. Compositions recorded on these sites are all included by default yet, any site can be checked or unchecked by the user. These four tabs all depend on the "glycosylations" public route of the GlyConnect RESTful API to retrieve the corresponding glycosylation information. Direct access to this route is provided via the *advanced* tab, where specific queries can be input in the form of URL parameters as detailed in the corresponding API help page: https://glyconnect.expasy.org/api/docs.

To expand the usage of the tool beyond the content of the GlyConnect database, a *custom* tab was implemented. It allows inputting a composition list in either of the three supported formats by copy-pasting in the appropriate text field or uploading a text file. These custom compositions are mapped to those contained in GlyConnect, in order to provide basic information for each match, such as GlyConnect and GlyTouCan identifiers, molecular masses and known glycan structures. These conversion and mapping services can also be used externally as described in the API help page.

In either of the selected tab, once the different search criteria have been chosen, the query is activated by clicking on the "Add to selection" button. This selection is visually summarized as a combination of search criteria (*i.e.* the species, biological entity, glycan type and glycosylation sites for proteins), followed by the corresponding number of glycan compositions recorded in the GlyConnect database. Each biological entity is cross-linked via an accession number to its matching reference database/ontology as detailed in supplemental Table S1. A given selection can be discarded at any time by clicking on its cross-icon.

In the current version of Compozitor, up to three search results can be simultaneously active. In other words, the respective known glycomes of multiple combinations can be compared, such as, three proteins or two proteins and a tissue, etc. Each search result is assigned a label in the form of a block letter. The chronological order of successive queries is matched with the alphabetical order of block letter labels (A represents the first, B the second and C the third set of results). The "Compute graph "button triggers the graphic display of glycan compositions selected by queries and enables further investigation of the results. Alternatively, users only interested in downloading composition lists can use the "Export selection" option.

Graphical Result Display—Frequently enough, a glycan may be fully contained in one or more other compositions. To account for this dependence, the search results are displayed as a directed graph in which each node is a composition and two nodes are connected if they differ by one residue (as listed in Table I). The graph generation is powered by D3.js (https://d3js.org), a flexible JavaScript library providing data visualization for web applications. The layout of the graph is based on the force-directed algorithm implemented in D3.js. When queries yield a large number of unique compositions, all computations are performed in background threads using *web workers, i.e.* protocols for web pages to execute tasks in the background. This prevents the user interface from becoming unresponsive when a graph contains hundreds of nodes.

Nodes—The graph nodes represent the unique set of glycan compositions contained in one or several query results. A single set of results is labelled A and shown with blue nodes. In the case of multiple queries, each set of results is assigned a distinct color (blue A, red B and green C). When a composition is shared between several sets of results another color scheme is used (magenta AB, yellow AC, cyan BC and black ABC). A legend located on the upper left side of the display keeps track of the search criteria, as well as the color and label codes. The occurrence count of each specific color node in the graph is shown in the legend as the number that labels each colored node.

In the graph, each node is identifiable with a glycan composition in the condensed notation. Other features reflect information stored in the GlyConnect database. Each node is labelled inside with the number of glycan structures matching the corresponding composition in the context of the search criteria. Node size varies based on the number of publications in which the composition was detected. Mousing over a node label prompts a popup window containing finer details of the composition: cross-links to GlyConnect and GlyTouCan (15), the monoisotopic mass in Da, and related glycan structure cartoons in the Symbol Nomenclature for Glycans (SNFG) format (16). A small window offering a "zoom on" opens a drop-down list of all nodes of the graph categorized as *root*, *leaf*, *unconnected* and *other*. It is located on the top right of the main display to facilitate searching for a particular composition in the graph especially when search criteria generate a large and crowded graph in which specific nodes may be difficult to find. Once a given composition is selected, the graph view centers on the corresponding node.

One of the key features of Compozitor lies in the addition of grey "virtual nodes" to increase the connectivity of a graph. These nodes are computed by performing the systematic pairwise comparison of all compositions in a graph. If two compositions differ from exactly two residues, the two corresponding nodes are tentatively connected through an intermediary node that is only one residue away from each. Then, this tentative node is added only if it meets two conditions:

- It does not already exist in the graph.
- None of its children has a non-virtual parent node (i.e., an alternative path is possible in the graph to connect two nodes differing from exactly two residues).

TABLE II
Reported glycan properties by linkage type

Glycan Type	Properties	Rules
N-linked/O-linked	*neutral*	does not contain any sialic acid (N-Acetylneuraminic acid (NeuAc) or N-Glycolylneuraminic acid (NeuGc)) or sulfate (Su)
N-linked/O-linked	*fucosylated*	contains at least one deoxyhexose (dHex)
N-linked/O-linked	*sialylated*	contains at least one sialic acid (N-Acetylneuraminic acid (NeuAc) or N-Glycolylneuraminic acid (NeuGc))
N-linked/O-linked	*fuco-sialylated*	contains at least one deoxyhexose (dHex) AND at least one sialic acid (N-Acetylneuraminic acid (NeuAc) or N-Glycolylneuraminic acid (NeuGc))
N-linked	*oligomannose*	contains at least 5 hexoses (Hex), no more than two N-Acetyl hexosamine (HexNAc) and no more than one deoxyhexose (dHex)
O-linked	*sulfated*	contains at least one sulfate (Su)

The second selective rule is implemented so as to avoid overcrowding the graph with unnecessary nodes.

Paths—The edges of the graph are represented as grey unidirectional arrows showing the addition of a residue. Similarly to nodes, each edge is labelled with the letter of the condensed notation corresponding to the added residue. To evoke the SNFG coloring scheme, the addition of one fucose is displayed in SNFG red and that of one sialic acid in SNFG purple. Although less strictly defined, SNFG blue was assigned to the addition of one N-acetylhexosamine and SNFG green to that of one hexose.

In an attempt to estimate the connectivity of the graph and the consistency of relatedness between compositions, the reachability of each node is also shown when mousing over it. Entering paths are highlighted in cyan whereas exiting paths in orange. Each node can be moved interactively by dragging it in the desired direction. It is also possible to follow paths associated with the addition of a specific residue. Mousing over a path labelled with the monosaccharide that is added will trigger the highlight of all paths labelled with that particular monosaccharide addition. Finally, zoom buttons are available at the bottom of the page for magnifying or reducing the graph.

Glycan Properties—Glycans of the GlyConnect database are associated with general properties commonly used to qualify or categorize them. In the current version of Compozitor, three of these (*neutral, fucosylated, sialylated*) qualify compositions irrespective of their type and are complemented with *oligomannose* as N-glycan specific and *sulfated* as O-glycan specific. Then, five properties are assigned to either type of glycan compositions according to rules defined upon the advice of glycobiologists and detailed in Table II.

Glycan properties are directly inferred from the presence/absence of specific monosaccharides in compositions except. The *oligomannose* property that is deduced from monosaccharide counts and as such, is not as reliably assigned as the others. Note that *fucosylated* is assigned upon the presence of a deoxyhexose, but deoxyhexoses may correspond to monosaccharides other than fucose (*e.g.* quinovose or rhamnose) in some nonmammal species. It is consequently correct for most entries in the GlyConnect database, but can be erroneous in a few specific cases.

A global view of the compositions contained in the graph is captured in an interactive bar plot of the five-property counts, displayed next to the graph on the bottom right. Mousing over any bar of the bar plot highlights in orange all nodes contributing to the corresponding frequency.

Custom Datasets—As mentioned earlier, the user is given the option of inputting custom compositions. In that case, launching the search involves a systematic comparison with the GlyConnect database content irrespective of any selection criterion other than matching the input compositions. Each item of the input list is searched in the database composition entries and when a match is found, the corresponding information is kept to be mapped on the graph. In the end, a graph is generated with nodes that are exactly those of the input list. Their respective size reflects the number of recorded publications supporting the corresponding composition and their label is the number of solved structures stored in the database for the corresponding composition. In this way, the graph provides a rough estimate of how realistic a composition set may be.

Export—The glycan compositions of each selected query can be exported as a text file in all supported formats. As mentioned earlier export can be launched before the generation of the graph to create composition lists, but when it is launched after displaying the graph, all compositions or composition subsets can be selected. More precisely, compositions that are common to several sets of results or those that correspond to virtual nodes can be singled out, therefore saved or discarded. In all cases, basic metadata information is added in the form of a header at the beginning of the exported files. It describes the tool version and GlyConnect release, date of generation, composition format, and selected queries. A shortcut copy icon also allows a quick copy-paste in the clipboard. Additionally, the graph may be exported as a vectorial image in the SVG format.

RESULTS AND DISCUSSION

Most of the results presented in this section were obtained with human N-glycomes simply because the vast majority of current glycoproteomics publications have this focus. Consequently, sampling problems tend to be minimized in this case.

Visualization and Interpretation of a Glycome Graph—The description of the tool in the Material and Method section emphasizes the array of information that is displayed in the Compozitor interface. In particular, the graph representation of the glycome of a protein, a tissue or a cell line is more amenable to capturing potential compositional biases. The bar plot of N-glycan overall properties is the first summary view. In absence of a reference for an expected distribution depending on species or tissue, it is at best informative on its own and mostly useful in the comparative mode.

Even though the output graphs tend to simplify the reality of the underlying enzymatic network at work in synthesizing glycans, this representation may reveal (dis)continuous paths and provide clues regarding the likelihood of a structure in a given context. Common sense would suggest that if a

composition occurs in the graph, it will stem from compositions that have one less monosaccharide. This is precisely the information brought by the graph. In the end, connectivity reflects the accessibility of a node via other nodes and emphasizes the consistency of synthetic reactions. From a formal point of view, node reachability provides a classical estimate of the graph properties. For example, some GlyConnect protein records contain associated compositions that produce scattered and poorly connected nodes whereas others generate a fully connected graph. The introduction of virtual nodes plays an important role in bridging the gaps. Examples illustrating these variations are shown in Fig. 1. Figs. 1*A*–1*C* show the N-glycome graphs of several well characterized human glycoproteins: alpha-fetoprotein (P02771), coagulation factor XI (P03951) and interferon gamma (P01579). These examples highlight both the variety of N-glycomes as confirmed by the variability of property bar plots and their respective consistency. Note that these examples mainly emphasize that with a similar size (about 20 nodes) the introduction of virtual nodes has distinct effects. No virtual node can enhance the tightly connected graph of alpha-fetoprotein (H4N2F1 remains isolated because this composition is too distant from all others). Only two virtual nodes are needed to join two clusters of coagulation factor XI, whereas four are needed to connect the three clusters of interferon gamma.

The glycome size does not correlate with an increase in virtual nodes. Supplemental Figs. S1*A*–S1*B* show the larger N-glycomes of human thrombospondin-1 (P07996) and nonrecombinant erythropoietin (P01588). The 83 compositions of thrombospondin-1 form an almost fully connected graph leaving out two nodes: H3N2 and H12N2. The virtual node completion integrates the former in the graph through connections to initially missing H3N3 and H4N2. However, H12N2 is three monosaccharides away from H9N2, the terminal node of the path adding hexoses to H3N2. The 58 compositions of erythropoietin form scattered clusters that require 16 virtual nodes to transform the graph into an almost complete one that only leaves out a sulfated as well as an acetylated composition.

Computed graphs are interactive and can be scrutinized to detect the potential pivotal role of some nodes. A composition can be central in one graph and a root or a terminal leaf in another. This is illustrated in supplemental Fig. S2*A*, S2*B* where H6N5F1S2 obviously plays a different role in the respective N-glycome graphs of human (supplemental Fig. S2*A*) erythropoietin (P01588) and (supplemental Fig. S2*B*) decorin (P07585). The removal of H6N5F1S2 in (supplemental Fig. S2*A*) would break six connections in the erythropoietin graph thereby disrupting a path between H7N6F1S2 and H6N5F1S1 (virtual nodes H7N5F1S2 and H6N6F1S2 would be unjustified) and would create an isolated cluster with only H6N5F1S3 and H6N5F1S3a1. In contrast, discarding H6N5F1S2 in (supplemental Fig. S2*B*) is of minimal consequence because all paths are preserved and only H6N5F1S3 is left as an isolated node.

Likewise, a node may be virtual in a graph and steadily mapped in another. We interpret the oscillation of a node between virtual and real in similar contexts as indicative of the relevance of its role. In some cases, a missing composition may point at missing data and suggest a possible check of experimental data where the composition may have been below threshold in processed data results. An example is given in supplemental Fig. S3*A*–S3*B*. H4N4F2 is virtual in (supplemental Fig. S3*A*) the extracellular matrix (ECM) protein decorin (P07585) and real in (supplemental Fig. S3*B*) thrombospondin-1 (P07996) also an ECM protein. Nonetheless, the two nodes play a comparable role with a similar connectivity in the graph. Ten cyan (outward) and 12 orange (inward) links stem from H4N4F2 in the thrombospondin-1 graph whereas in the decorin graph, H4N4F2 gives rise to nine cyan and 11 orange links. Furthermore, when virtual nodes are not included, then H4N4F2 is not included in the N-glycome graph of human decorin and H4N4F1 and H4N4F3 are remote as visible in supplemental Fig. S3*C*. No path joins them and they are therefore unreachable from one another. However, as shown in supplemental Fig. S3*D*, when H4N4F2 is introduced as one of the seven virtual nodes, H4N4F1 and H4N4F3 are logically connected through the H4N4F2 virtual node. In the end, not only is H4N4F2 as virtual in one glycome, comparable to its real counterpart in another glycome but when included as virtual, it connects nodes that should be mutually reachable in the graph. This suggests that H4N4F2 is likely to be real in the decorin glycome and may just have been below threshold in glycoproteomics results.

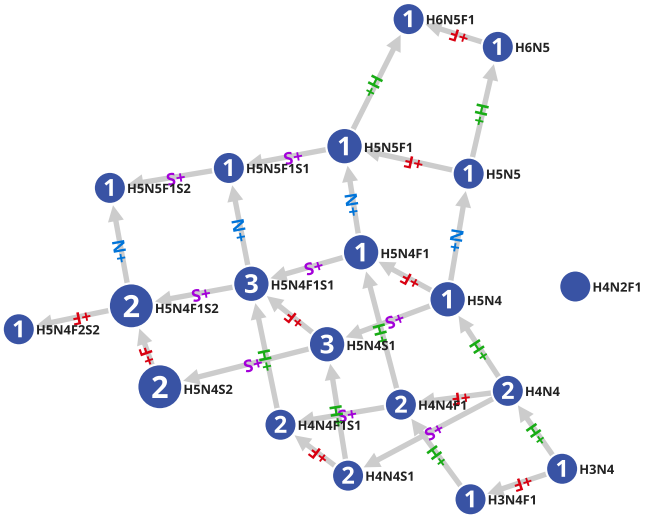
Furthermore, path highlighting reveals the contribution of a node to the graph. For example, some nodes bridge two parts of the graph as H5N4F1 occurring in the graph of the N-glycome of decorin shown again in supplemental Fig. S4*A*. Three paths end on H5N4F1 and four start from it. This highlights that when multiple routes go through one node, the impact of removing that node may substantially alter pathways. This information is often key to appraising the role of the corresponding composition. In contrast, H5N4F2S1 in the same graph shown in supplemental Fig. S4*B* is manifestly terminal. All paths end on that node and do not extend any further. In that case, the role of the node is more difficult to interpret yet it highlights the reality of node categories mentioned earlier, namely, central, root, or terminal.

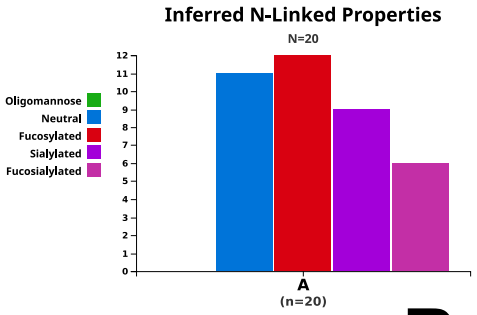
Note that in many of the figures, we have used the flexible interface to move a few nodes around to unpack dense regions and improve readability.

Comparison of Entities

The GlyConnect database contains ~2600 protein records and comparing their respective glycomes can be informative. For example, several protein entries describe the glycosylation

20 (A) Homo sapiens | Alpha-fetoprotein | N-Linked | Asn-197,Asn-251,Undefined

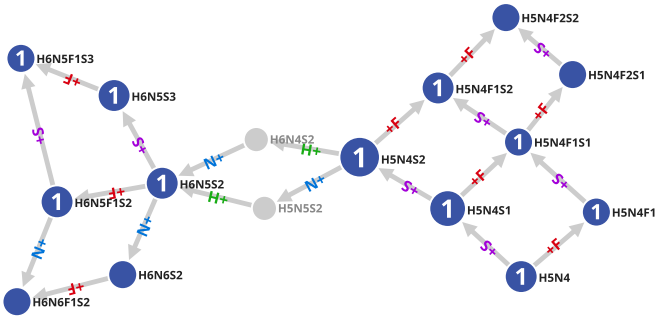


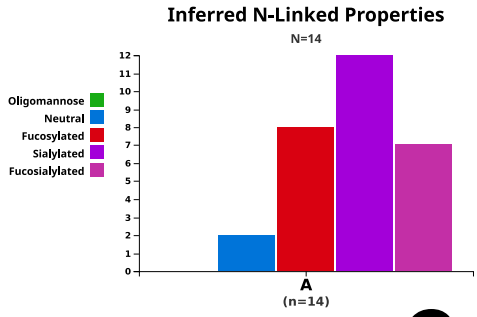


A

14 (A) Homo sapiens | Coagulation factor XI | N-Linked | Asn-90,Asn-124,Asn-126,Asn-163,Asn-450,Asn-491

2 Virtual

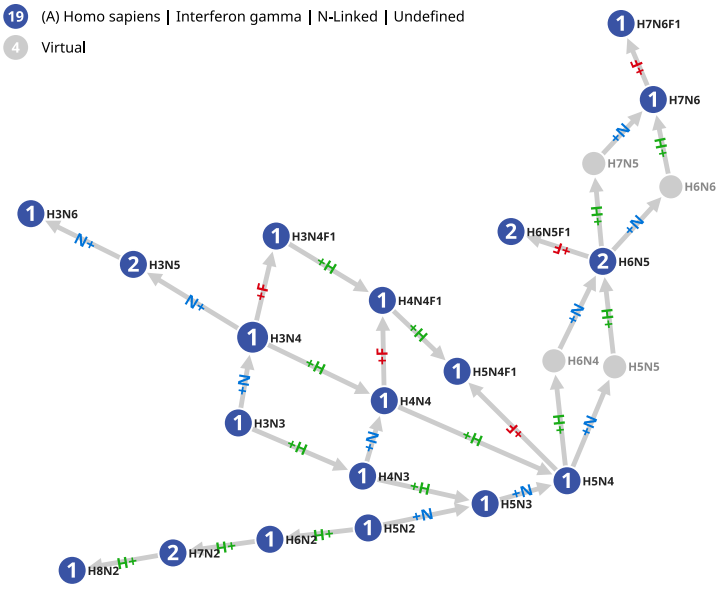


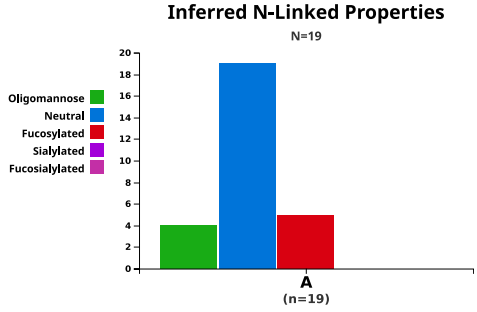


B

19 (A) Homo sapiens | Interferon gamma | N-Linked | Undefined

4 Virtual





C

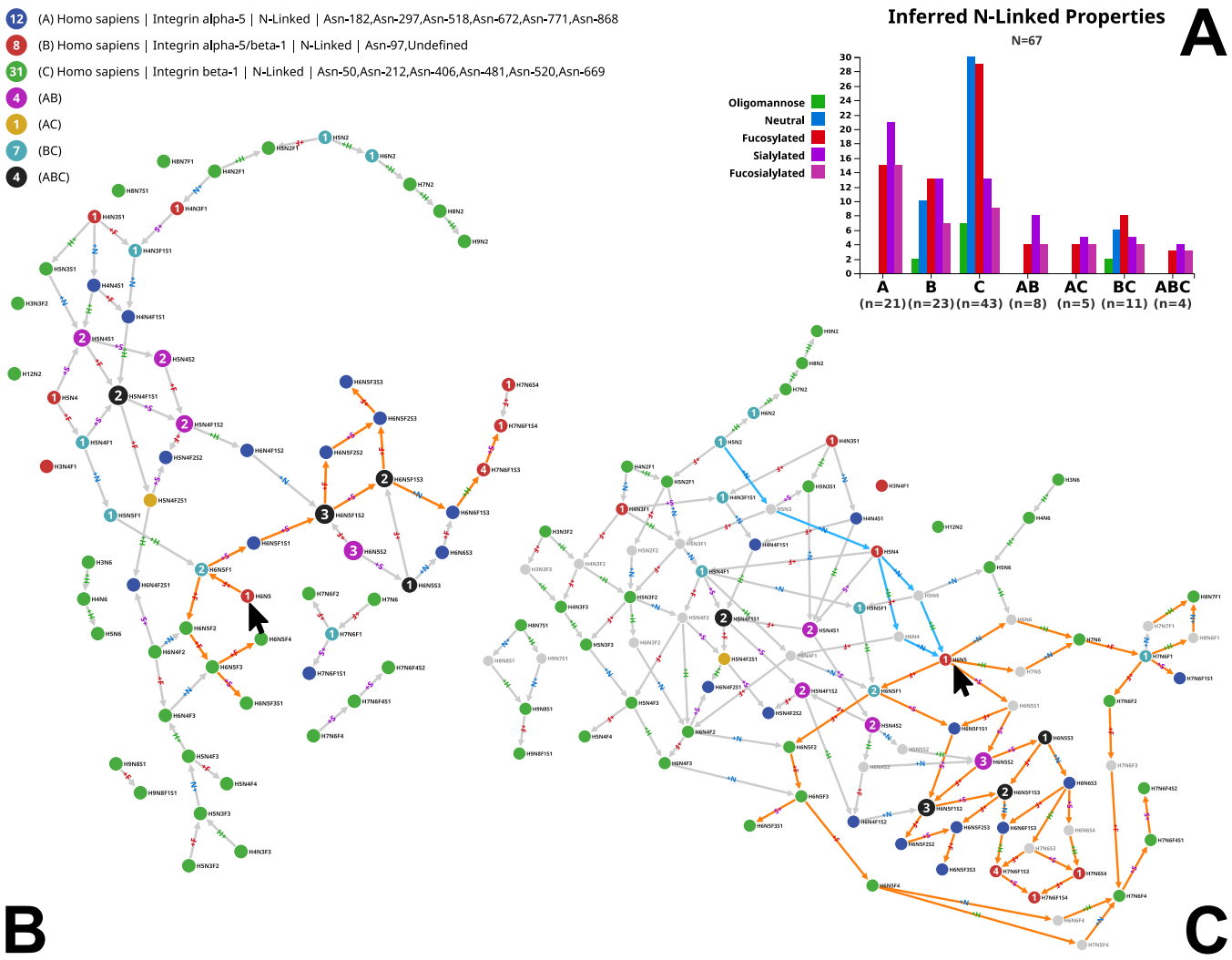


FIG. 2. **N-glycomes of two individual and complexed integrin chains.** GlyConnect protein entry ID 283 describes the glycosylation of the alpha 5/beta 1 integrin complex whereas protein entries ID 1407 and ID 1414 separately report the individual glycosylation of integrin alpha 5 and integrin beta 1 identified as intact glycopeptides. The N-glycome of each of the three entries were compared. *A*, The N-glycome properties of each protein entry are plotted and display different distributions. *B*, The resulting graph with no virtual nodes displays many clusters and limited accessibility of nodes, highlighted with the status of node H6N5 (*arrow*) only linking out to 14 nodes via orange paths. *C*, The connectivity of the resulting graph with 24 virtual nodes is enhanced as highlighted with the status of H6N5 (*arrow*) accessible from two real and two virtual nodes (cyan paths) and reaching out to 26 nodes via orange paths.

of integrins. In particular, as reported in the literature and transcribed in the corresponding UniProtKB/Swiss-Prot entries integrin alpha 5 (P08648) and integrin beta 1 (P05556) form a complex (ITGA5:ITGB1) that acts as a receptor for fibronectin, fibrinogen and fibrillin-1 among other recorded functions. The glycosylation of the complex was studied (17) and the results are reported in the corresponding GlyConnect entry (ID: 283). High throughput glycoproteomics experiments also report the identification of intact glycopeptides of integrin alpha 5 (ID:1407) and integrin beta 1 (ID: 1414) separately. We compared the N-glycomes of the three GlyConnect entries to estimate the overlap between these independent experiments. The results are shown in Fig. 2. The bar plot in Fig. 2*A* emphasizes the distinct profiles of each N-glycome in respectively integrin alpha 5 (21 compositions mostly sialylated), integrin alpha 5/beta 1 (23 compositions) and integrin beta 1 (43 compositions mostly neutral and fucosylated). The legend of the graph (top left) indicates poor overlap, which is to be expected

FIG. 1. **Examples of protein N-glycomes and connecting role of virtual nodes.** *A*, N-glycome of alpha-fetoprotein (P02771) where virtual nodes are not needed to fully connect corresponding compositions, *B*, N-glycome of coagulation factor XI (P03951) where only two virtual nodes are needed to fully connect corresponding compositions, *C*, N-glycome of interferon gamma (P01579) where four virtual nodes are needed to fully connect corresponding compositions. In each case the bar plot of glycan properties is distinctive.

between integrin alpha 5 and integrin beta 1 but less between the complex and each participant. This trend is confirmed by the high number of virtual nodes required to connect the clusters visible in Fig. 2*B* into a larger graph shown in Fig. 2*C*. The introduction of 24 grey/virtual nodes in Fig. 2*C* increases the connectivity by 40%. As an example, the reachability of H6N5 is highlighted in both Fig. 2*B* and 2*C*. In Fig. 2*B*, H6N5 is reaching out to 14 nodes (orange paths) but cannot be reached from any other node. In Fig. 2*C*, H6N5 can be reached in several steps from H5N2 (cyan paths) and links to 26 nonvirtual nodes (orange paths). In all cases, the peripheral role of green nodes of integrin beta 1 is noticeable. In contrast, compositions shared between datasets play a central role.

Finer grain information of glycosylation can be visualized through selecting one site at a time in the interface. Supplemental Fig. S5 shows the comparison of conserved Asn-72 in human alpha-1 acid glycoprotein 1 (P02763) and alpha-1 acid glycoprotein 2 (P19652) that are 89.5% identical in amino acid sequence. The graph shows that the 19 compositions reported on Asn-72 of alpha-1 acid glycoprotein 2 (ID: 14) are included in the 26 reported on Asn-72 of alpha-1 acid glycoprotein 1 (ID: 718). The seven compositions unique to alpha-1 acid glycoprotein 1 are either neutral or sialylated. The latter tend to be terminal and the former add consistency to the graph by providing a new root and multiple connecting paths via virtual nodes.

Other combinations, such as mapping the glycome of protein with that of a specific tissue or a specific disease, can be explored. The integration of Compozitor in GlyConnect reference pages provides the display of results in each published article stored in the database. Interestingly, the corresponding graphs are often consistent and do not require the addition of many virtual nodes (data not shown).

Consistency of Well-Characterized Glycomes—We have used Compozitor to estimate both the extent and the consistency of some of GlyConnect glycomes. We illustrate this approach with two examples representative of glycome mapping, *i.e.* the N-glycome of the CHO (Chinese Hamster Ovary) cell lines and that of the human immunoglobulin gamma (IgG).

N-Glycome of CHO Cell Lines—The N-glycome of the generic CHO (CVCL_0213) cell line included in GlyConnect is made of 79 N-glycan compositions summarizing the information extracted from 27 publications. These data are visualized in Compozitor by querying "CHO" in the cell line tab. In the resulting graph shown without virtual nodes in Fig. 3*A* maps the 79 compositions in three clusters and a few isolated nodes. Introducing virtual nodes generates an almost fully connected graph except for the N1F1 disaccharide that remains several monosaccharides away from other compositions/structures, as shown in Fig. 3*B*. Sixteen virtual nodes are necessary to connect 78 compositions in a single graph and among these, two achieve a missing connection between H6N5 and H7N6. Two alternative paths are suggested either via H6N6 or H7N5. When another CHO cells data set is added for comparison, for example, CHO-DG44 (CVCL_7180) a cell line in which the dihydrofolate reductase (DHFR) gene locus was removed, nine new compositions are introduced whereas 27 are common to the first data set. H6N6 is one of nine and the corresponding node in the new graph is transformed from virtual to full. This in turn cancels out the H7N5 alternative and keeps H6N6 as the only path between H6N5 and H7N6. This example illustrates how the combination of two datasets stabilizes the graph connectivity.

N-Glycome of Human Immunoglobulin Gamma (IgG)—Automated set-ups described by several groups (18–20) have generated large datasets of the unique glycosite in peptide EEQ[F/Y]**N**ST[F/Y]R of the human IgG heavy chain (the amino acid sequence is shown here as a tryptic peptide because it appears this way in the vast majority of glycoproteomics studies). We have compiled a collection of 82 potential compositions identified on this glycopeptide (see list in supplemental material) and also used reviews such as (21, 22) that cite many relevant and useful references. Finally, additional data could be found in a recently published collaborative study (23). This data set was input in the custom tab of Compozitor. The resulting graph is fully connected and does not require the introduction of virtual nodes.

Glycosite data described in (24) along with an HTP serum glycopeptide profiling method include results on IgG glycosylation and are stored in GlyConnect. However, this reference was not used in our compilation. The authors refer to 365 different N-glycan compositions that were entered in a customized glycan database for human serum. We did not consider this large set but only the identified intact glycopeptides of Igs reported in this article with a total of 81 compositions. This information is directly accessible in GlyConnect: https://glyconnect.expasy.org/browser/references/2857. Then, using the protein tab of Compozitor to select data corresponding to Ig gamma 1 and gamma 4, the graph generated with the 82 compositions of the compilation was enriched. Note that we did not select Ig gamma 2 because it does not add further information, nor gamma 3 that contains an additional glycosite to the regular EEQ[F/Y]**N**ST[F/Y]R.

The new graph resulting from combining these different sources is composed of 93 nodes. The 11 (=93-82) additional compositions are listed in Table III. Interestingly, most of the corresponding nodes in the new graph are terminal or preterminal or they simply cause adding a single path (one arrow in, one arrow out) in the graph. In other words, they extend the graph as opposed to strengthening connectivity and as such, their removal has very limited impact on the overall topology of the graph. These nodes are circled in grey in Fig. 4. Information stored in the GlyConnect corresponding records is summarized in the "comments" column of Table III. Terminal or preterminal nodes are compositions that are either seen in other species but human, generate ambiguity with O-linked structures or were only seen in a single study. The only exception is H4N6S1. A single structure (no

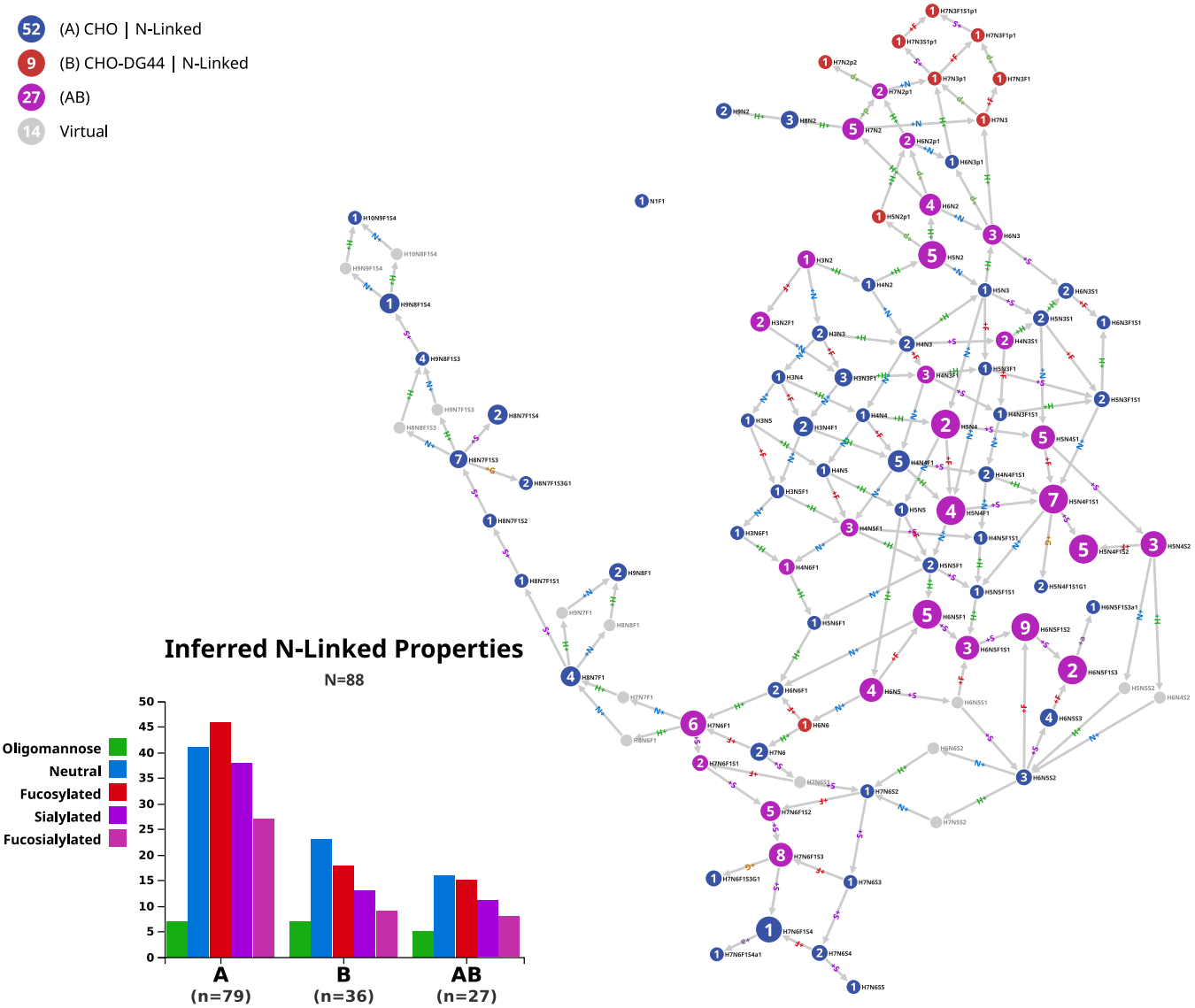


FIG. 3. **Comparison of N-glycomes of two CHO cell lines.** Two N-glycome datasets of CHO cell lines are compared. Overall glycome properties are slightly different (bar plots). Common compositions are represented as magenta nodes and appear central in the graph. In relation to supplemental Fig. S6 that shows graph differences arising from the introduction of virtual nodes in the N-glycome of the generic CHO cell line (CVCL_0213), the role of H6N6 is highlighted. H6N6 does not exist in CHO (CVCL_0213) but is present in CHO-DG44 (CVCL_7180) and improves connectivity.

GlyTouCan ID available) matching H4N6S1 was described in (25) on Asn-69 of PSA (prostate specific antigen) and 32 human glycoprotein records linked to H4N6S1 originate from large-scale glycoproteomics studies. At this stage, these observations are inconclusive regarding the relevance of including this composition in the human IgG glycome.

Similar uncertainty applies to the first single path node. H4N4S2 was observed as O-linked on human milk mucins and *N*-linked on seven human glycoproteins from large-scale glycoproteomics studies. H3N4S1 is possibly more interesting. It is reported as *N*-linked in human hormones from low throughput studies with two definite structures (GlyTouCan ID: G33876UV, G53933HU) and in 20 human glycoprotein records from large-scale glycoproteomics studies.

The last two additional nodes (H6N4F2 and H6N4F2S1 circled in red in Fig. 4) have a stronger impact on the topology of the graph and their records in GlyConnect show more frequent occurrence. These combined features favor the inclusion of these two compositions in the human IgG glycome.

Tissue Typing of Human Unspecified Mucins—The GlyConnect database is populated with references that are spread over the past five decades and some earlier work was performed with limited knowledge of amino acid sequences. In that respect, the case of mucins is challenging because of

TABLE III
Status of new nodes in the IgG graph of combined data sets

Gamma 1	GC ID	Node Status	Comments	Gamma 4	GC ID	Node Status	Comments
H3N4F2	138	pre-terminal	*N*-linked in species other than human, otherwise O-linked	H3N4F2	138	pre-terminal	*N*-linked in species other than human, otherwise O-linked
H3N4F2S1	925	terminal	only in (24)	H3N5F2	380	terminal	mainly O-linked
H3N4S1	64	single path	related to 23 glycoproteins	H4N6S1	483	pre-terminal	related to 33 glycoproteins
H4N3F3	985	terminal	only in (24)	H6N4F2	304	re-route	related to 100 glycoproteins
H4N4S2	53	single path	related to 8 glycoproteins and ambiguity with O-linked	H6N4F2S1	944	re-route	related to 41 glycoproteins
H4N6	53	pre-terminal	recorded in species other than human				
H4N7S1	931	terminal	only in (24)				

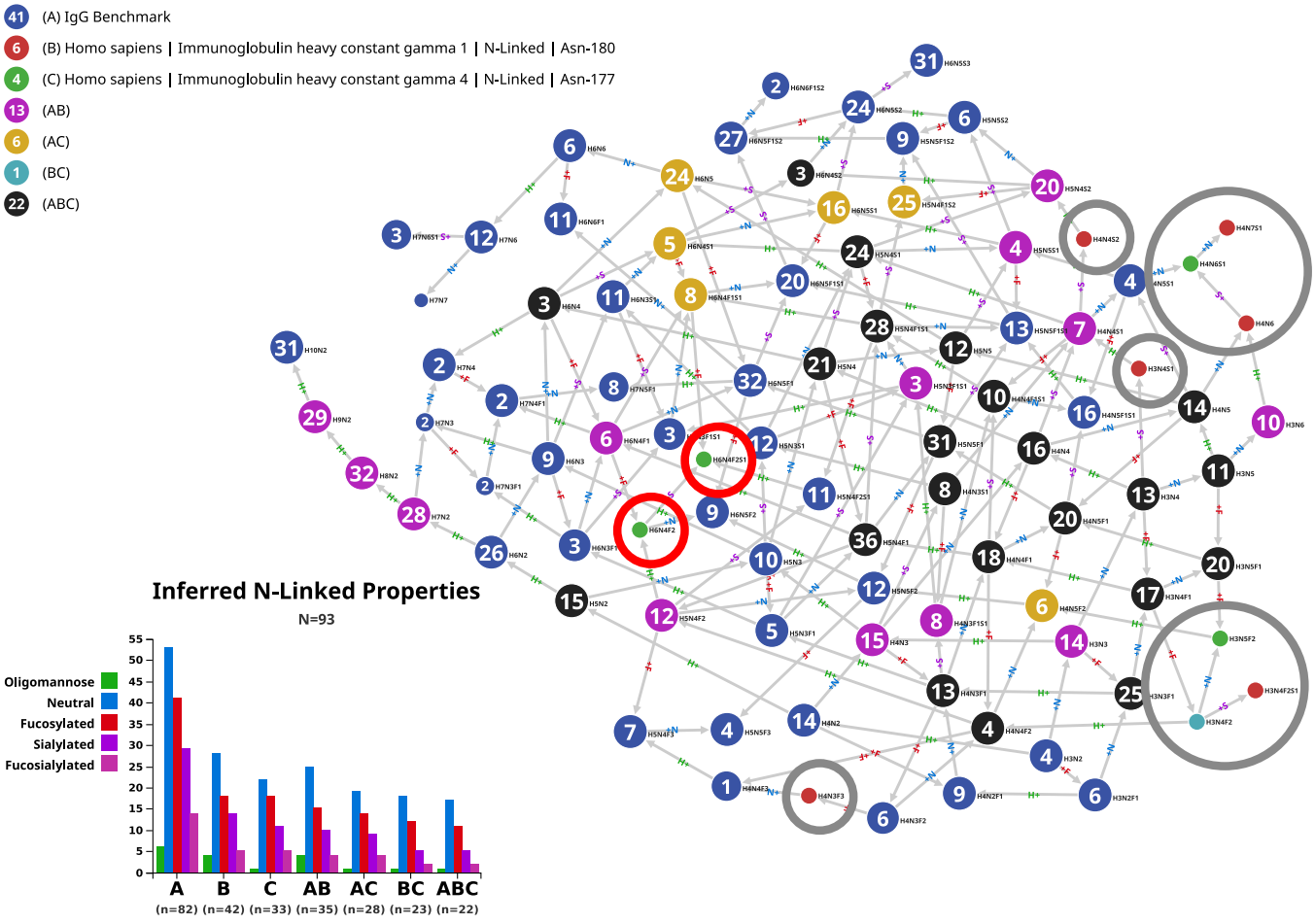


FIG. 4. **Comparison of N-glycomes of "theoretical" IgG N-glycome with two experimental datasets.** Result of the combination of a custom list of 82 compositions potentially found in immunoglobulin gamma with two individual datasets submitted to Compozitor. Only ten compositions of the individual datasets are not covered by the custom list. Most of these are peripheral in the graph (circled in grey) and as such do not contribute to reinforcing connectivity. Only H6N4F2 and H6N4F2S1 (circled in *red*) appear as relevant connectors.

the unusually large size and highly repetitive nature of these proteins particularly so, prior to the spread of high throughput DNA sequencing and the advent of genomics. As a result, a significant number of studies focused on mucosa, report glycan structures with a poor characterization of the mucin carrier. In GlyConnect, corresponding data were harvested in 32 references published between 1980 and 2003 and are collected in a "unspecified mucin" page (GlyConnect

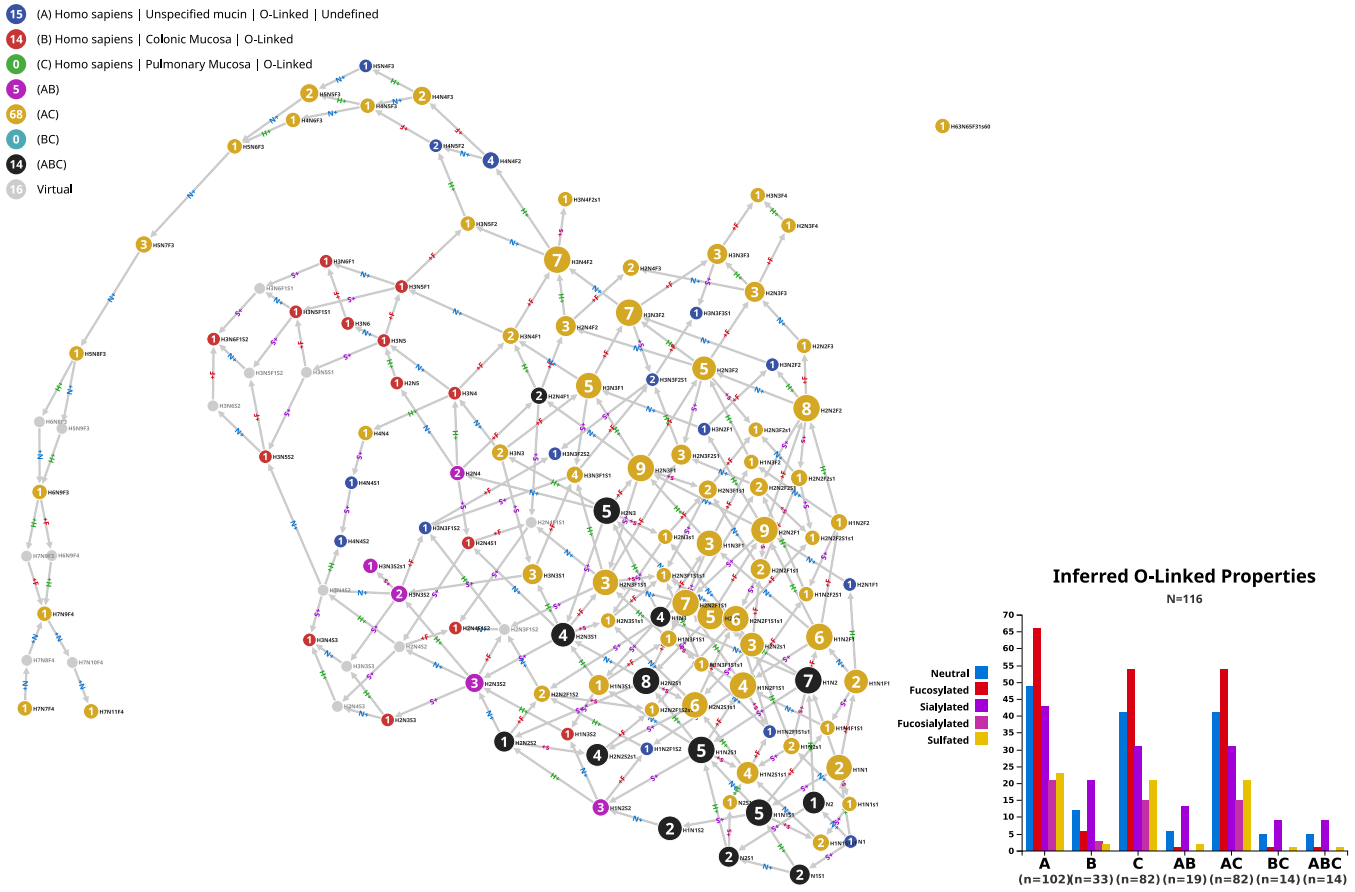


FIG. 5. **Comparison of unspecified mucin expression.** Graphic result of 102 glycan compositions reported across 32 published articles involving the study of human mucosa and mucins compared with tissue information regarding mucosa as stored in the GlyConnect database. The graph shows an overwhelming presence of glycans expressed in pulmonary mucosa (*yellow* nodes) as opposed to colonic mucosa (*red* and magenta nodes). Fourteen black nodes represent compositions common to both but not highly specific of mucosa except for H2N2S2s1.

ID: 401), which contains 251 O-linked structures matched to 102 compositions. Not a single amino acid sequence is specified, let alone a defined glycosite. Compozitor was used to visualize tissue information associated with these compositions. The "unspecified mucin" was first selected in the protein tab. Then, switching to the "source tab," "colonic mucosa" and "pulmonary mucosa" were successively selected. The graph output of this combined query is shown in Fig. 5. Both the graph and the bar plot show that most glycans of "unspecified mucins" originate from pulmonary mucosa where the domination of fucosylated glycans is obvious and in contrast with the highly sialylated colonic mucosa glycans. Admittedly, there are more compositions associated with pulmonary (82) than colonic (33) mucosa. Nonetheless, trends are observable.

Red and magenta nodes (colonic-specific) tend to cluster together independently of yellow nodes (pulmonary-specific). Fourteen black nodes represent the compositions common to both tissues. All of these nodes have a strong outgoing connectivity except for H2N2S2s1 that is terminal and exclusive to pulmonary and colonic mucosa. Only one black node is fucosylated (H2N4F1) whereas the majority is sialylated. To confirm the consistency of that subset, the fourteen compositions were extracted via the Export function and pasted in the custom tab. The resulting graph (not shown) interconnects the thirteen nonfucosylated compositions with each other and only requires a virtual node from H2N3 to include H2N4F1. In the graph of Fig. 5, H2N4 connects H2N3 and H2N4F1 but it appears in magenta as opposed to black, therefore it is not seen in pulmonary mucosa. It could be considered as a potential candidate to maintain consistency.

The annotation of the remaining fifteen blue nodes (nonpulmonary and noncolonic) reveals that they are spread between stomach and ovarian mucosa, milk/meconium and amniotic fluid. As summarized in Table IV, all stomach and ovarian mucosa compositions are fucosylated but not sialylated whereas all milk, meconium and amniotic fluid compositions are sialylated. Milk compositions are all terminal or pre-terminal. Note that N1 is the fifteenth node that is not considered in this discussion as it lacks specificity.

With this example, we show how Compozitor provides cues to refine the characterization of a tissue glycome. It also

TABLE IV
Tissue expression of mucin glycans

Tissue	Composition	Number of structures	Connectivity: outgoing/incoming links	Node Status
Stomach mucosa	H5N4**F**3	1	**30**/12	single path
	H5N4**F**2	2	**33**/14	connects 2 clusters
	H4N4**F**2	3	**28**/17	connects 2 clusters
	H3N2**F**1	1	9/**38**	Uncharacteristic
	H2N1**F**1	1	3/**65**	single path
Ovarian mucosa	H4N4**F**2	1	**28**/17	connects 2 clusters
	H3N2**F**2	1	12/**27**	single path
Amniotic fluid	H3N3**F**2**S**2	1	**50**/0	Terminal
	H3N3**F**1**S**2	1	**39**/1	Terminal
Milk	H4N4**S**2	1	**29**/0	Terminal
	H4N4**S**1	1	**11**/1	pre-terminal
	H3N3F3**S**1	1	**43**/0	Terminal
	H3N3F2**S**1	2	**39**/2	pre-terminal
Meconium	H1N2F1**S**2	1	**14**/6	Uncharacteristic
	H1N2F1**S**1s1	1	**17**/5	Uncharacteristic

demonstrates that data accumulation is key to identifying characteristic features. Needless to say, Compozitor graphic and interactive outputs need further interpretation and are intended as early steps in building a more refined picture of glycomes.

Comparison of High Throughput Datasets—One of the key features of Compozitor is to allow for the assessment of various composition datasets that are used in intact glycopeptide identification search engines. Compozitor compares the content of these datasets with information recorded in GlyConnect so as to potentially rationalize their extension to other compositions or their reduction to rationally designed subsets. We collected three datasets: a default Byonic data set of 305 N-glycan compositions (the default set of 309 was nowhere to be found so we reconstituted the file from supplemental material found in publications citing Byonic to reach 305 and missing four compositions), a GPQuest data set of 181 N-glycan compositions obtained from the authors as used in (26, 27), which results are included in GlyConnect and deciphered the composition file of N-glycan compositions used with Mascot Distiller (where fucosylation is considered as a variable modification) in (3) the results of which are also integrated in GlyConnect. In the latter case, we estimated the number of possible glycan compositions to 205 given the granted possibility of accepting fucosylation as a function of the number of HexNAcs. Each of these were separately input in the custom tab of the interface to visualize the outline of the corresponding graphs and compare the respective bar plots of overall properties. These results are shown in Fig. 6 along with data extracted from GlyConnect. Querying GlyConnect for all human N-glycan compositions outputs 472 items visualized in Compozitor by inputting "taxonomy=homo sapiens&glycan_type=*N*-linked" in the advanced tab of the interface. Note that about 60 of the additional compositions found in GlyConnect contain chemical groups such as sulfate, phosphate or acetyl that are not usually included in lists processed by search engines. Needless to say, the resulting graph is crowded and difficult to interpret yet, the summary bar plot for each case provides general information. In contrast with both Byonic and GlyConnect datasets that share a similar distribution of properties, the Mascot data set lacks sialylated compositions along with the GPQuest data set in which neutral compositions also seem slightly overrepresented. T

Then, to get another type of snapshot view, we compared the graph topologies of each search engine data set as well as the effect of adding virtual nodes. To that end, we submitted each composition file to Compozitor via the custom tab. Supplemental Figs. S7–S9 show the outline of the respective graphs that were generated without and with virtual nodes. The 205 compositions in the Mascot file were manifestly produced by the systematic addition of monosaccharides to the N-glycan core as seen in the regular shape and mesh-line graph (supplemental Fig. S7*A*). As confirmed by the size and labels of the nodes, the majority of the selected structures are well documented in the GlyConnect database. Only three virtual nodes are needed to merge the two clusters created originally (supplemental Fig. S7*B*). In contrast, many clusters characterize the GPQuest and Byonic outlines in the absence of virtual nodes (supplemental Fig. S8*A* and S9*A*) and both display a strongly connected core with many side extensions when virtual nodes are included (supplemental Fig. S8*B* and S9*B*). Some very large high mannose (*e.g.* H12N2 in GPQuest) or hybrid (*e.g.* H11N11S1 in Byonic) compositions remain unconnected to the main graph.

We scrutinized virtual nodes in each search engine data set and observed that the connectivity of a specific node in different contexts is variable. Summary figures are provided in Table V where numbers quoted above along with the amount of compositions that are identified by the engines are

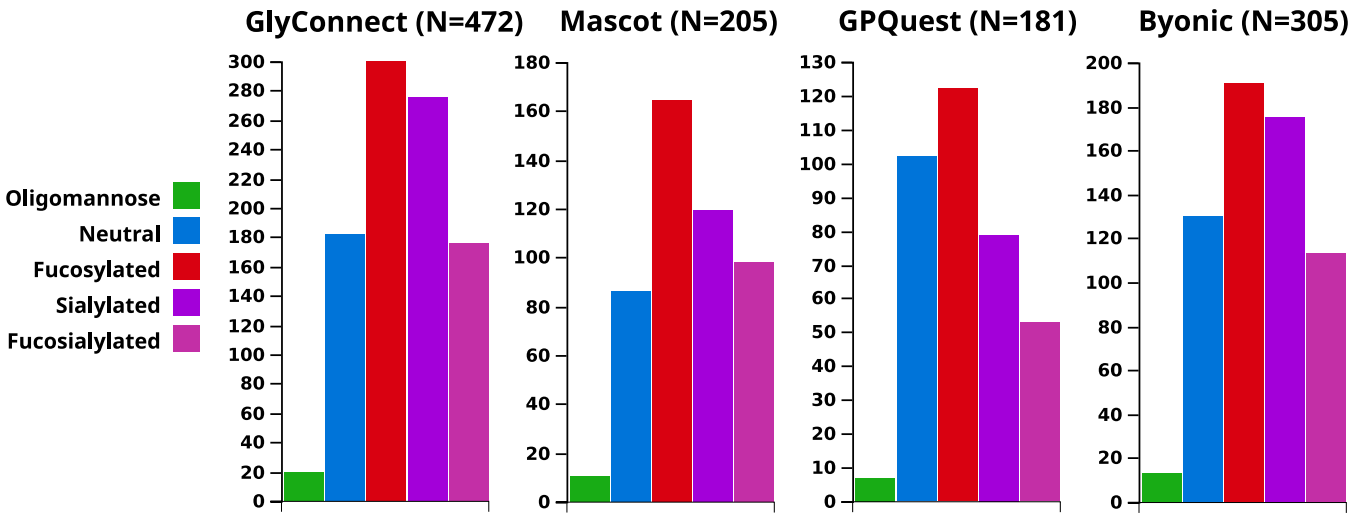


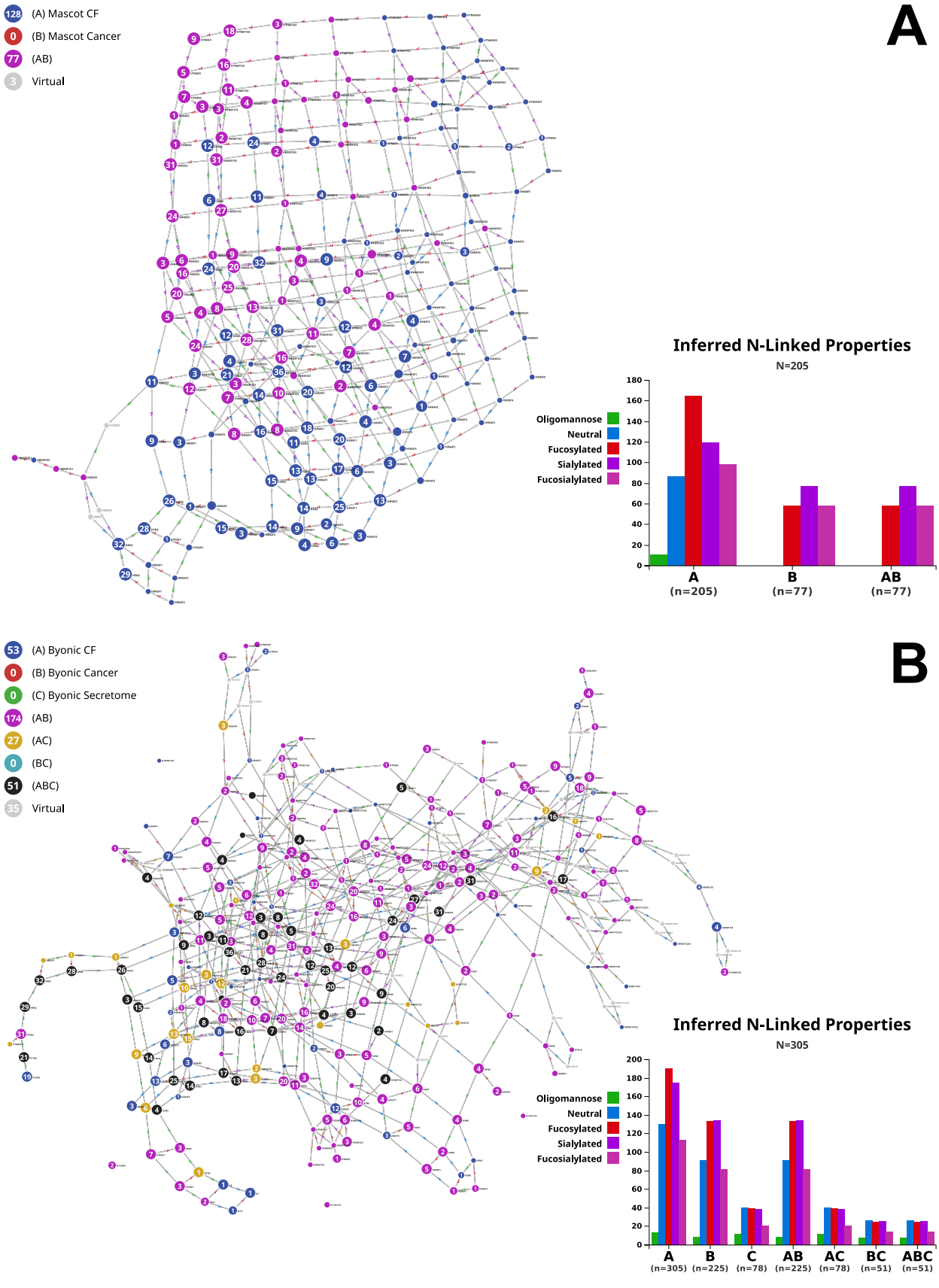
FIG. 6. **Bar plots showing properties of four selected composition files.** Plotted glycan composition properties of four datasets: (1) the current collection of human N-glycans in the GlyConnect database (472 in total) (2) the estimated data set of all possible N-glycan compositions processed by Mascot Distiller in (3) (205 in total) (3) the composition file used by GPQuest in (26, 27) as communicated by authors (181 in total) and iv) the reconstituted default composition file for N-glycans in Byonic, missing 4 compositions (305 in total).

TABLE V
Summary of composition file content in a selection of search engines

	Initial Data set	Identified Compositions	Identified Proteins	Virtual Nodes
Mascot	205	77 in (3)	257 in (3)	3
GPQuest	181	166 in (26), 64 in (27)	934 in (26), 244 in (27)	25
Byonic	305	225 in (4), 78 in (28)	97 in (4), 58 in (28)	35

shown. We extracted the information from GlyConnect in the original publications cited as well as two additional where GPQuest (27) and Byonic (28) were used. The ratio between theoretical (*i.e.* composition file provided to the search engine) and identified compositions is obviously hyper variable (from 64 to 166 in two GPQuest related articles and from 78 to 225 in two Byonic related articles) and apparently not correlated on the number of identified proteins. Further data integration is needed to appraise these variations and we are heading in this direction by planning an extensive inclusion of new data in 2020. The percentage of virtual nodes is limited and their stability may reveal relevant information on their meaning. Stability is defined here as follows: if a virtual node matches a composition already known in other contexts then this node may not be virtual in another graph and is called unstable. Conversely, if a virtual node matches a composition that never occurs in other contexts then it will remain virtual. In this latter case, the missing composition may be suggested as one to supplement a composition data set upon the user's judgment of its realism.

Fig. 7 shows the comparison of identified *versus* possible/theoretical compositions used as input in a search engine. The version of the graph with virtual nodes was selected for that purpose. Fig. 7*A* illustrates the case with Mascot in (3) and Fig. 7*B* with Byonic in (4, 28). Interestingly, in Fig. 7*A* all of the 77 compositions identified with Mascot, as confirmed by the associated bar plot, are either fucosylated or sialylated and grouped in the upper left part of the graph. Mousing over "neutral" listed in the bar plot highlights most of the blue nodes that correspond to compositions of the input file not identified in the study. This partition is likely to reflect the unfavorable effect of titanium dioxide enrichment - described in (3) - on neutral glycans and brings out the contrast between the identified and unidentified glycans. In Fig. 7*B*, 51 compositions are common to both studies using Byonic whereas 27 yellow nodes are specific to the endothelial cell secretome and 124 magenta nodes to prostate cancer. The bar plots indicate a steady presence of neutral compositions and a slight decrease in fucosylation. Yet, the interpretation of these results is made difficult by the gap in identification between the two studies. The remaining 53 blue nodes representing compositions not seen in any of the two cited studies are of interest if it is assumed that the reduction of the input composition file may improve the accuracy of identification. A new graph of 252 nodes was generated leaving out these



53 compositions and in which a total of 25 virtual nodes were necessary to maximize connectivity (see list in supplemental material). Five compositions were suggested as virtual nodes and matched one of the 53 "unseen" Byonic compositions: H3N6F1, H6N6, H7N7, H7N7F1S2 and H3N5F1G1. This somehow confirms their relevance in the original data set. In contrast, twenty-three out of 53 compositions contain NeuGc and could possibly be discarded when analyzing human samples (note that 22/23 are fucosylated). Then, the 30 nonNeuGc containing compositions are mostly fucosylated (20/30), have either a small (<430 Da) or large (>2400 Da) molecular weight, are seldom sialylated (5/30) or contain unusual counts of the same monosaccharide (*e.g.* nine HexNAc or five dHex). These very specific characteristics may explain why these compositions are unmatched and deserve further attention in their selection in the first place.

As mentioned earlier in this Results section, the connectivity and the graph location of a virtual node are criteria to be considered for determining the relevance of selecting the corresponding composition in further analyses. We selected examples in the graph of Fig. 7*B*. H7N7 is an illustration of a potentially interesting case that questions the relevance of its inclusion when processing human data. It is not in GlyConnect yet included in the Byonic initial data set. This means that there is no evidence of a human glycopeptide carrying H7N7 according to publications stored in GyConnect irrespective of which engine was used for identification. Nonetheless, a structure is reported in GlyTouCan (ID: G85304RG) and the composition is recorded as well (ID: G57748MK). The GlyTouCan details of G85304RG reveal that information is inherited from CarbBank (23) and two references support the existence of the structure (1) in rat kidney and (2) in rat brain and bovine plasma. Whether to keep H7N7 in the input file remains an open question that may be context-dependent. From another angle, H5N7 is a virtual node that substantially modifies the connectivity of the graph. To illustrate this point, differential connectivity depending on the introduction of virtual nodes is highlighted in the comparative display of excerpts of the graph in Fig. 7*B* and its counterpart without virtual nodes (full graph not shown). In supplemental Fig. S10 specific edge count (outgoing in cyan and incoming in orange) labels each displayed node. Supplemental Fig. S10*A* (resp. supplemental Fig. S10*B*) is a close-up on the graph without virtual nodes (resp. with virtual nodes). The increase in edge count associated with yellow nodes (secretome data) is particularly striking with the introduction of H5N7. H5N7S1 and H5N7S2 are hardly reachable in the absence of H5N7 whereas the whole path from H3N2 is established if H5N7 is accounted for. Interestingly, H5N7 is recorded in the GlyConnect database (composition ID:79) but only found in chicken (29). These examples emphasize the benefit of inspecting closely a virtual node neighborhood to determine the relevance of its inclusion.

CONCLUSION

We have introduced Compozitor, a new software tool to visualize and compare glycome data based on compositional information. In the absence of a single representation of glycan compositions, the tool processes various notations and can easily accommodate new ones if necessary. Compozitor provides a range of usage through an interactive graphic interface. To begin with, it offers a different view on the content of the GlyConnect database and the option of comparing glycomes whether defined at the level of a glycosite, a glycoprotein, a cell line or a tissue. Second, it enables the customization of a composition file prior to using a search engine in a glycoproteomic experiment. The content of an input file can be better rationalized from comparing glycomes. This comparison will improve with time and the planned growth of the GlyConnect database. The next major challenge is the inclusion of site-specific quantitative data. The Compozitor interface was designed to step up in that direction. Currently, the node size of a glycome graph reflects the number of associated published articles supporting the existence of a glycan composition. Once a critical mass of quantitative studies will be published, it will be easy to change this parameter and correlate the node size with the observed expression of the corresponding glycan. As other tools released as part of the Glycomics@ExPASy initiative, we also plan to improve the tool from collected feedback of users.

Acknowledgments—The authors would like to thank Kathirvel Alagesan, Daniel Kolarich and Nicolle H. Packer for their feedback on using Compozitor. The ExPASy portal is maintained by SIB Swiss Institute of Bioinformatics and hosted at the Vital-IT Competency Centre.

Funding and additional information—This work is supported by Swiss National Science Foundation [SNSF 31003A_179249].

Author contributions—T.R. and F.L. designed research; T.R., J.M., and F.L. performed research; T.R., J.M., and F.L.

FIG. 7. **Comparison of input *versus* identified N-glycan compositions in selected studies.** *A*, Output of the identified (magenta nodes) and estimated data set (*blue* nodes) of N-glycan compositions processed by Mascot Distiller in (3) submitted to Compozitor. Magenta nodes tend to group together and cover the part of the graph where non-neutral and sialylated compositions lie. *B*, Output of the N-glycan compositions identified in a prostate cancer study (4) (magenta nodes) and in the secretome of endothelial cells (28) (*yellow* nodes) in which the Byonic search engine was used, mapped along with our version of the Byonic default N-glycan composition file (CF) missing four items (*blue* nodes).

contributed new reagents/analytic tools; T.R., J.M., and F.L. analyzed data; T.R., J.M., and F.L. wrote the paper.

Conflict of interest—Authors declare no competing interests.

Abbreviations—The abbreviations used are: API, application programming interface; CHO, Chinese Hamster ovary; DHFR, dihydrofolate reductase; ECM, extracellular matrix; HUPO, human proteome organization; IgG, immunoglobulin Gamma; JSON, JavaScript object notation; MS, mass spectrometry; MS/MS, tandem mass spectrometry; REST, representational state transfer; SNFG, symbol nomenclature for glycans.

Received March 16, 2020, and in revised form, July 1, 2020 Published, MCP Papers in Press, July 7, 2020, DOI 10.1074/mcp.RA120.002041

REFERENCES

1. Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., Costello, C. E., Cravatt, B. F., Fenselau, C., Garcia, B. A., Ge, Y., Gunawardena, J., Hendrickson, R. C., Hergenrother, P. J., Huber, C. G., Ivanov, A. R., Jensen, O. N., Jewett, M. C., Kelleher, N. L., Kiessling, L. L., Krogan, N. J., Larsen, M. R., Loo, J. A., Ogorzalek Loo, R. R., Lundberg, E., MacCoss, M. J., Mallick, P., Mootha, V. K., Mrksich, M., Muir, T. W., Patrie, S. M., Pesavento, J. J., Pitteri, S. J., Rodriguez, H., Saghatelian, A., Sandoval, W., Schlüter, H., Sechi, S., Slavoff, S. A., Smith, L. M., Snyder, M. P., Thomas, P. M., Uhlén, M., Van Eyk, J. E., Vidal, M., Walt, D. R., White, F. M., Williams, E. R., Wohlschlager, T., Wysocki, V. H., Yates, N. A., Young, N. L., and Zhang, B. (2018) How many human proteoforms are there? *Nat. Chem. Biol*. **14,** 206–214
2. Smith, L. M., and Kelleher, N. L. and Consortium for Top Down Proteomics, (2013) Proteoform: a single term describing protein complexity. *Nat. Methods* **10,** 186–187
3. Bollineni, R. C., Koehler, C. J., Gislefoss, R. E., Anonsen, J. H., and Thiede, B. (2018) Large-scale intact glycopeptide identification by Mascot database search. *Sci. Rep* **8,** 2117
4. Kawahara, R., Ortega, F., Rosa-Fernandes, L., Guimarães, V., Quina, D., Nahas, W., Schwämmle, V., Srougi, M., Leite, K. R. M., Thaysen-Andersen, M., Larsen, M. R., and Palmisano, G. (2018) Distinct urinary glycoprotein signatures in prostate cancer patients. *Oncotarget* **9,** 33077–33097
5. Riley, N. M., Hebert, A. S., Westphall, M. S., and Coon, J. J. (2019) Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nat. Commun*. **10,** 1311
6. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
7. Chalkley, R. J., and Baker, P. R. (2017) Use of a glycosylation site database to improve glycopeptide identification from complex mixtures. *Anal. Bioanal. Chem*. **409,** 571–577
8. Hu, H., Khatri, K., and Zaia, J. (2017) Algorithms and design strategies towards automated glycoproteomics analysis: algorithms and design strategies. *Mass Spectrom. Rev*. **36,** 475–498
9. Bern, M., Kil, Y. J., Becker, C., (2012) in *Current Protocols in Bioinformatics*, eds Baxevanis, A. D., Petsko, G. A., Stein, L. D., and Stormo, G. D. (John Wiley & Sons, Inc., Hoboken, NJ, U. S. A.), p 13.20.1–13.20.14
10. Alocci, D., Mariethoz, J., Gastaldello, A., Gasteiger, E., Karlsson, N. G., Kolarich, D., Packer, N. H., and Lisacek, F. (2019) GlyConnect: glycoproteomics goes visual, interactive, and analytical. *J. Proteome Res*. **18,** 664–677
11. Mariethoz, J., Alocci, D., Gastaldello, A., Horlacher, O., Gasteiger, E., Rojas-Macias, M., Karlsson, N. G., Packer, N. H., and Lisacek, F. (2018) Glycomics@ExPASy: bridging the gap. *Mol. Cell. Proteomics* **17,** 2164–2176
12. Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., and Stockinger, H. (2012) ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res*. **40,** W597–W603
13. Cooper, C. A., Joshi, H. J., Harrison, M. J., Wilkins, M. R., and Packer, N. H. (2003) GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res*. **31,** 511–513
14. Cooper, C. A., Gasteiger, E., and Packer, N. H. (2001) GlycoMod–a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics*. **1,** 340–349
15. Tiemeyer, M., Aoki, K., Paulson, J., Cummings, R. D., York, W. S., Karlsson, N. G., Lisacek, F., Packer, N. H., Campbell, M. P., Aoki, N. P., Fujita, A., Matsubara, M., Shinmachi, D., Tsuchiya, S., Yamada, I., Pierce, M., Ranzinger, R., Narimatsu, H., and Aoki-Kinoshita, K. F. (2017) GlyTouCan: an accessible glycan structure repository. *Glycobiology* **27,** 915–919
16. Varki, A., Cummings, R. D., Aebi, M., Packer, N. H., Seeberger, P. H., Esko, J. D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J. J., Schnaar, R. L., Freeze, H. H., Marth, J. D., Bertozzi, C. R., Etzler, M. E., Frank, M., Vliegenthart, J. F., Lütteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K. F., Dell, A., Narimatsu, H., York, W., Taniguchi, N., and Kornfeld, S. (2015) Symbol nomenclature for graphical representations of glycans. *Glycobiology* **25,** 1323–1324
17. Nakagawa, H., Zheng, M., Hakomori, S., Tsukamoto, Y., Kawamura, Y., and Takahashi, N. (1996) Detailed oligosaccharide structures of human integrin alpha 5 beta 1 analyzed by a three-dimensional mapping technique. *Eur. J. Biochem*. **237,** 76–85
18. Stöckmann, H., Adamczyk, B., Hayes, J., and Rudd, P. M. (2013) Automated, high-throughput IgG-antibody glycoprofiling platform. *Anal. Chem*. **85,** 8841–8849
19. Baković, M. P., Selman, M. H. J., Hoffmann, M., Rudan, I., Campbell, H., Deelder, A. M., Lauc, G., and Wuhrer, M. (2013) High-throughput IgG Fc N-glycosylation profiling by mass spectrometry of glycopeptides. *J. Proteome Res*. **12,** 821–831
20. Falck, D., Jansen, B. C., de Haan, N., and Wuhrer, M. (2017) in *High-Throughput Glycomics and Glycoproteomics*, eds Lauc G, Wuhrer M (Springer New York, New York, NY), 31–47
21. Jennewein, M. F., and Alter, G. (2017) The immunoregulatory roles of antibody glycosylation. *Trends Immunol*. **38,** 358–372
22. Trbojević-Akmačić, I., Vilaj, M., and Lauc, G. (2016) High-throughput analysis of immunoglobulin G glycosylation. *Expert Rev. Proteomics*. **13,** 523–534
23. De Leoz, M. L. A., Duewer, D. L., Fung, A., Liu, L., Yau, H. K., Potter, O., Staples, G. O., Furuki, K., Frenkel, R., Hu, Y., Sosic, Z., Zhang, P., Altmann, F., Gru Nwald-Grube, C., Shao, C., Zaia, J., Evers, W., Pengelley, S., Suckau, D., Wiechmann, A., Resemann, A., Jabs, W., Beck, A., Froehlich, J. W., Huang, C., Li, Y., Liu, Y., Sun, S., Wang, Y., Seo, Y., An, H. J., Reichardt, N.-C., Ruiz, J. E., Archer-Hartmann, S., Azadi, P., Bell, L., Lakos, Z., An, Y., Cipollo, J. F., Pucic-Bakovic, M., Štambuk, J., Lauc, G., Li, X., Wang, P. G., Bock, A., Hennig, R., Rapp, E., Creskey, M., Cyr, T. D., Nakano, M., Sugiyama, T., Leung, P.-K. A., Link-Lenczowski, P., Jaworek, J., Yang, S., Zhang, H., Kelly, T., Klapoetke, S., Cao, R., Kim, J. Y., Lee, H. K., Lee, J. Y., Yoo, J. S., Kim, S.-R., Suh, S.-K., de Haan, N., Falck, D., Lageveen-Kammeijer, G. S. M., Wuhrer, M., Emery, R. J., Kozak, R. P., Liew, L. P., Royle, L., Urbanowicz, P. A., Packer, N. H., Song, X., Everest-Dass, A., Lattová, E., Cajic, S., Alagesan, K., Kolarich, D., Kasali, T., Lindo, V., Chen, Y., Goswami, K., Gau, B., Amunugama, R., Jones, R., Stroop, C. J. M., Kato, K., Yagi, H., Kondo, S., Yuen, C. T., Harazono, A., Shi, X., Magnelli, P. E., Kasper, B. T., Mahal, L., Harvey, D. J., O'Flaherty, R., Rudd, P. M., Saldova, R., Hecht, E. S., Muddiman, D. C., Kang, J., Bhoskar, P., Menard, D., Saati, A., Merle, C., Mast, S., Tep, S., Truong, J., Nishikaze, T., Sekiya, S., Shafer, A., Funaoka, S., Toyoda, M., de Vreugd, P., Caron, C., Pradhan, P., Tan, N. C., Mechref, Y., Patil, S., Rohrer, J. S., Chakrabarti, R., Dadke, D., Lahori, M., Zou, C., Cairo, C., Reiz, B., Whittal, R. M., Lebrilla, C. B., Wu, L., Guttman, A., Szigeti, M., Kremkow, B. G., Lee, K. H., Sihlbom, C., Adamczyk, B., Jin, C., Karlsson, N. G., Örnros, J., Larson, G., Nilsson, J., Meyer, B., Wiegandt, A., Komatsu, E., Perreault, H., Bodnar, E. D., Said, N., Francois, Y.-N., Leize-Wagner, E., Maier, S., Zeck, A., Heck, A. J. R., Yang, Y., Haselberg, R., Yu, Y. Q., Alley, W., Leone, J. W.,

Yuan, H., and Stein, S. E. (2020) and Stein, S. E. (2020) NIST interlaboratory study on glycosylation analysis of monoclonal antibodies: comparison of results from diverse analytical methods. *Mol. Cell. Proteomics* **19,** 11–30
24. Liu, M., Zhang, Y., Chen, Y., Yan, G., Shen, C., Cao, J., Zhou, X., Liu, X., Zhang, L., Shen, H., Lu, H., He, F., and Yang, P. (2014) Efficient and accurate glycopeptide identification pipeline for high-throughput site-specific N-glycosylation analysis. *J. Proteome Res*. **13,** 3121–3129
25. Song, E., Mayampurath, A., Yu, C.-Y., Tang, H., and Mechref, Y. (2014) Glycoproteomics: identifying the glycosylation of prostate specific antigen at normal and high isoelectric points by LC–MS/MS. *J. Proteome Res*. **13,** 5570–5580
26. Hu, Y., Shah, P., Clark, D. J., Ao, M., and Zhang, H. (2018) Reanalysis of global proteomic and phosphoproteomic data identified a large number of glycopeptides. *Anal. Chem*. **90,** 8065–8071
27. Sun, S., Hu, Y., Jia, L., Eshghi, S. T., Liu, Y., Shah, P., and Zhang, H. (2018) Site-specific profiling of serum glycoproteins using N-linked glycan and glycosite analysis revealing atypical *N* -glycosylation sites on albumin and α-1B-glycoprotein. *Anal. Chem*. **90,** 6292–6299
28. Yin, X., Bern, M., Xing, Q., Ho, J., Viner, R., and Mayr, M. (2013) Glycoproteomic analysis of the secretome of human endothelial cells. *Mol. Cell. Proteomics* **12,** 956–978
29. Harvey, D. J., Wing, D. R., Küster, B., and Wilson, I. B. (2000) Composition of N-linked carbohydrates from ovalbumin and co-purified glycoproteins. *J. Am. Soc. Mass Spectrom*. **11,** 564–571

Supplementary Material

Category	Reference	URL
Protein	UniProtKB/Swiss-Prot DB	uniprot.org
Tissue	Uberon ontology	uberon.org
Plant Tissue	BRENDA ontology	brenda-enzymes.org/ontology.php?ontology_id=3
Cell Type	Cell ontology	cellontology.org
Cell Component	Gene ontology	geneontology.org
Cell line	Cellosaurus DB	web.expasy.org/cellosaurus/
Disease	Disease ontology	disease-ontology.org

Supp Table 1: Controlled vocabularies and ontologies used in the GlyConnect database

List of 82 IgG potential compositions:

Hex:3 HexNAc:2, Hex:3 HexNAc:2 dHex:1, Hex:3 HexNAc:3, Hex:3 HexNAc:3 dHex:1, Hex:3 HexNAc:4, Hex:3 HexNAc:4 dHex:1, Hex:3 HexNAc:5, Hex:3 HexNAc:5 dHex:1, Hex:3 HexNAc:6, Hex:4 HexNAc:2, Hex:4 HexNAc:2 dHex:1, Hex:4 HexNAc:3, Hex:4 HexNAc:3 dHex:1, Hex:4 HexNAc:3 dHex:1 NeuAc:1, Hex:4 HexNAc:3 dHex:2, Hex:4 HexNAc:3 NeuAc:1, Hex:4 HexNAc:4, Hex:4 HexNAc:4 dHex:1, Hex:4 HexNAc:4 dHex:2, Hex:4 HexNAc:4 dHex:1 NeuAc:1, Hex:4 HexNAc:4 dHex:3, Hex:4 HexNAc:4 NeuAc:1, Hex:4 HexNAc:5, Hex:4 HexNAc:5 dHex:1, Hex:4 HexNAc:5 dHex:1 NeuAc:1, Hex:4 HexNAc:5 dHex:2, Hex:4 HexNAc:5 NeuAc:1, Hex:5 HexNAc:2, Hex:5 HexNAc:3, Hex:5 HexNAc:3 dHex:1, Hex:5 HexNAc:3 dHex:1 NeuAc:1, Hex:5 HexNAc:3 NeuAc:1, Hex:5 HexNAc:4, Hex:5 HexNAc:4 dHex:1, Hex:5 HexNAc:4 dHex:1 NeuAc:1, Hex:5 HexNAc:4 dHex:1 NeuAc:2, Hex:5 HexNAc:4 dHex:2, Hex:5 HexNAc:4 dHex:2 NeuAc:1, Hex:5 HexNAc:4 dHex:3, Hex:5 HexNAc:4 NeuAc:1, Hex:5 HexNAc:4 NeuAc:2, Hex:5 HexNAc:5, Hex:5 HexNAc:5 dHex:1, Hex:5 HexNAc:5 dHex:1 NeuAc:1, Hex:5 HexNAc:5 dHex:1 NeuAc:2, Hex:5 HexNAc:5 dHex:2, Hex:5 HexNAc:5 dHex:3, Hex:5 HexNAc:5 NeuAc:1, Hex:5 HexNAc:5 NeuAc:2, Hex:6 HexNAc:2, Hex:6 HexNAc:3, Hex:6 HexNAc:3 dHex:1, Hex:6 HexNAc:3 dHex:1 NeuAc:1, Hex:6 HexNAc:3 NeuAc:1, Hex:6 HexNAc:4, Hex:6 HexNAc:4 dHex:1, Hex:6 HexNAc:4 dHex:1 NeuAc:1, Hex:6 HexNAc:4 NeuAc:1, Hex:6 HexNAc:4 NeuAc:2, Hex:6 HexNAc:5, Hex:6 HexNAc:5 dHex:1, Hex:6 HexNAc:5 dHex:1 NeuAc:1, Hex:6 HexNAc:5 dHex:1 NeuAc:2, Hex:6 HexNAc:5 dHex:2, Hex:6 HexNAc:5 NeuAc:1, Hex:6 HexNAc:5 NeuAc:2, Hex:6 HexNAc:5 NeuAc:3, Hex:6 HexNAc:6, Hex:6 HexNAc:6 dHex:1, Hex:6 HexNAc:6 dHex:1 NeuAc:2, Hex:7 HexNAc:2, Hex:7 HexNAc:3, Hex:7 HexNAc:3 dHex:1, Hex:7 HexNAc:4, Hex:7 HexNAc:4 dHex:1, Hex:7 HexNAc:5 dHex:1, Hex:7 HexNAc:6, Hex:7 HexNAc:6 NeuAc:1, Hex:7 HexNAc:7, Hex:8 HexNAc:2, Hex:9 HexNAc:2, Hex:10 HexNAc:2.

Supplemental figure legends

Supp Figure 1: Examples of protein large N-glycomes and connecting role of virtual nodes

The glycome size does not correlate with an increase in virtual nodes. (A) The N-glycome of human thrombospondin-1 is composed of 83 compositions but only two virtual nodes are needed to fully connect corresponding compositions (P07996). (B) In contrast, the N-glycome of non-recombinant erythropoietin (P01588) is composed of 58 compositions but sixteen virtual nodes are needed to fully connect corresponding compositions.

Supp Figure 2: Differential connectivity of the same node in two N-glycomes

A composition node can be central in one graph and a terminal leaf in another. Its structuring role in the graph is then different. (A) In the N-glycome graph of human erythropoietin (P01588) H6N5F1S2 connects two regions of the graph. Its removal would cause the collapse of parts and create two separate clusters. (B) In the N-glycome graph of human decorin (P07585) H6N5F1S2 is pre-terminal. Its removal would only result in isolating the leaf node H6N5F1S3.

Supp Figure 3: Differential roles of same node in two extracellular matrix (ECM) proteins

A composition node can be virtual in one graph and real in another. (A) H4N4F2 is a virtual node (9 cyan links and 11 orange links) in the N-glycome graph of human decorin (P07585) (B) H4N4F2 is a regular node (10 cyan links and 12 orange links) in the N-glycome graph of human thrombospondin-1 (P07996) (C) when virtual nodes are omitted in the N-glycome graph of human decorin (P07585) H4N4F1 and H4N4F3 are remote (D) when virtual nodes are included in the N-glycome graph of human decorin (P07585) then H4N4F2 occurs and it connects logically H4N4F1 and H4N4F3.

Supp Figure 4: Connectivity in the graph of human decorin N-glycome

Path highlighting reveals the contribution of a node to a graph as seen in the N-glycome of human decorin (P07585). (A)The H5N4F1 node is a connector between two areas of the graph (17 cyan links and 12 orange links) (B) The H5N4F2S1 node is terminal (22 cyan links)

Supp Figure 5: Comparison of conserved glycosylated asparagine in two highly similar human proteins

Asn-72 is a conserved glycosite in 89.5% similar alpha-1 acid glycoprotein 1 (P02763) and alpha-1 acid glycoprotein 2 (P19652). However, the glycome comparison of the two respective glycosites shows the total inclusion of the latter in the former (no red nodes). The corresponding bar plots show differences mainly in proportions of neutral and sialylated compositions.

Supp Figure 6: CHO cell N-glycome with and without virtual nodes

The N-glycome of the generic CHO cell line (CVCL_0213) includes 79 compositions. They are mapped in Compozitor with 16 virtual nodes in a single graph that only leaves N1F1 isolated since this composition is particularly small, as shown in (A) and without virtual nodes creating three clusters and leaving three nodes isolated, as shown in (B).

Supp Figure 7: Mascot composition file with and without virtual nodes as submitted in Ref.3

A set of 205 potential compositions was estimated from details of (3) where Mascot was used for intact glycopeptide identification. The mesh-like regular structure of the graph reflects the systematic approach for generating compositions (A). This regularity is confirmed by the very low number of virtual nodes needed to close the graph (B).

Supp Figure 8: GPQuest composition file with and without virtual nodes as submitted in Ref 26 and 27

A file of 181 compositions was communicated by the authors of (26)(27) and input in Compozitor. The outline of the graph almost closed with 27 virtual nodes (A) is similar to many biological networks where several extensions stem from central nodes. A few clusters of large compositions (#H >10) do not fit in. Node scattering is greater when virtual nodes are not included (B).

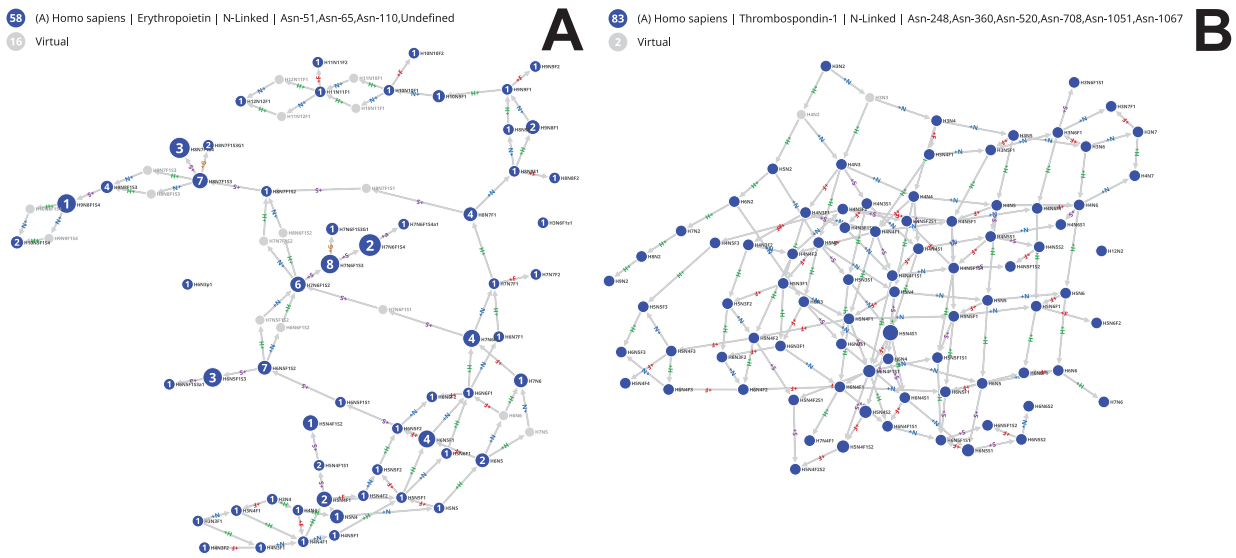
Supp Figure 9: Byonic composition file with and without virtual nodes as submitted in Ref 4 and 28

305 of default 309 compositions provided with the Byonic search engine were input in Compozitor. The outline of the graph almost closed with 35 virtual nodes (A) is similar to many biological networks where several extensions stem from central nodes. A few clusters of large compositions (#H >10) do not fit in. Node scattering is greater when virtual nodes are not included (B).

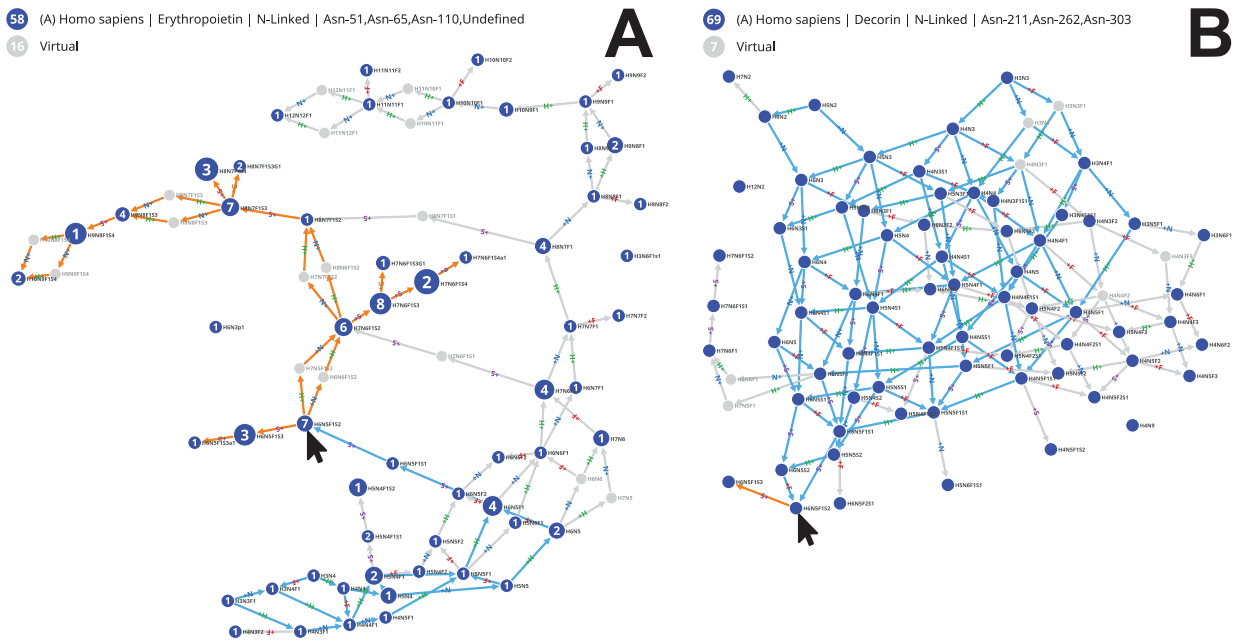
Supp Figure 10: Comparison of node connectivity in Figure 6B graph with and without virtual nodes

Neighbourhood of H6N7 in (A) an excerpt of the graph shown in Figure 6B and (B) its counterpart with no virtual nodes. Each node is labelled with the number of outgoing (cyan) and incoming (orange) paths connecting it to other nodes in the graph. Yellow nodes corresponding to identified compositions in the secretome of endothelial cells are particularly impacted by the introduction of virtual node H5N7.

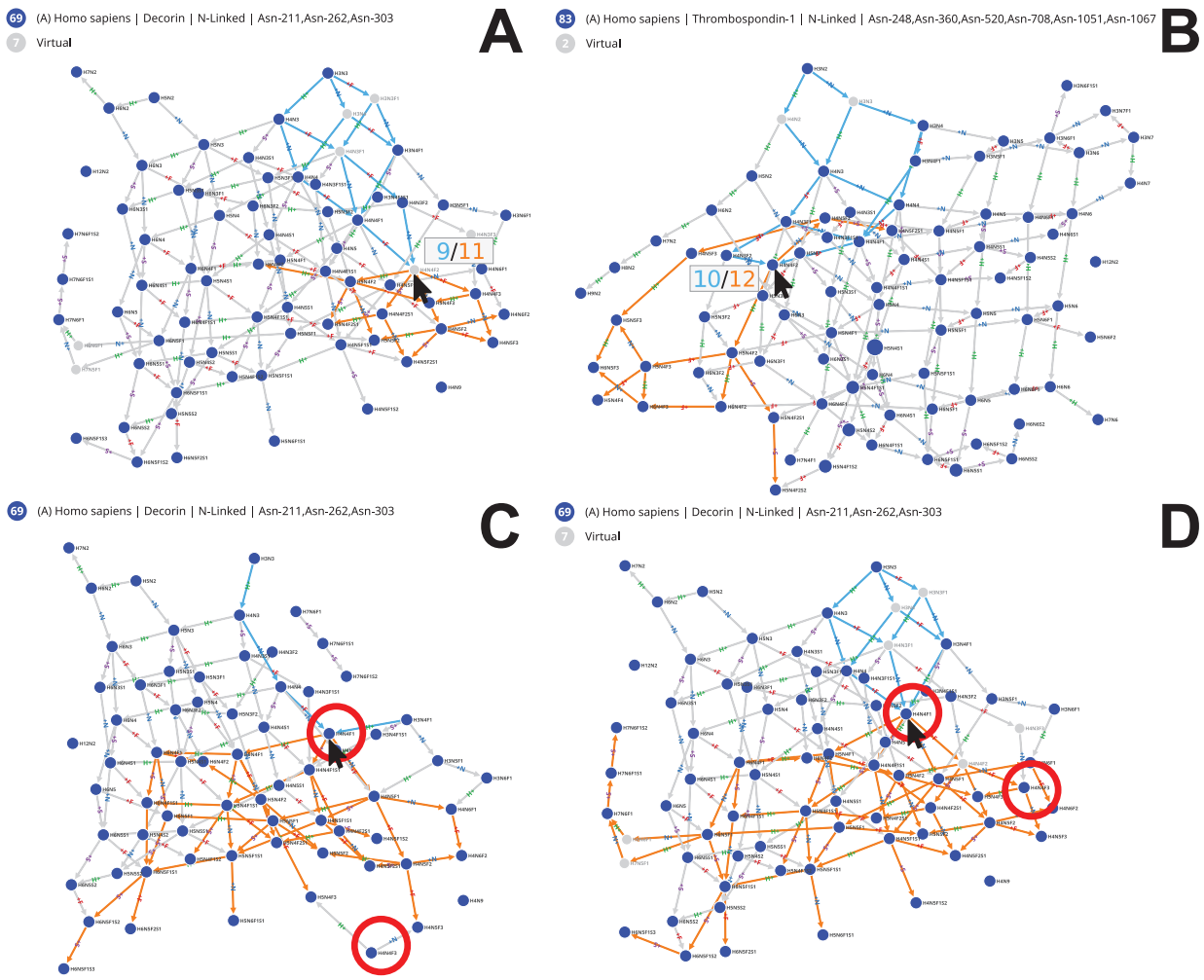
Supp Figure S1



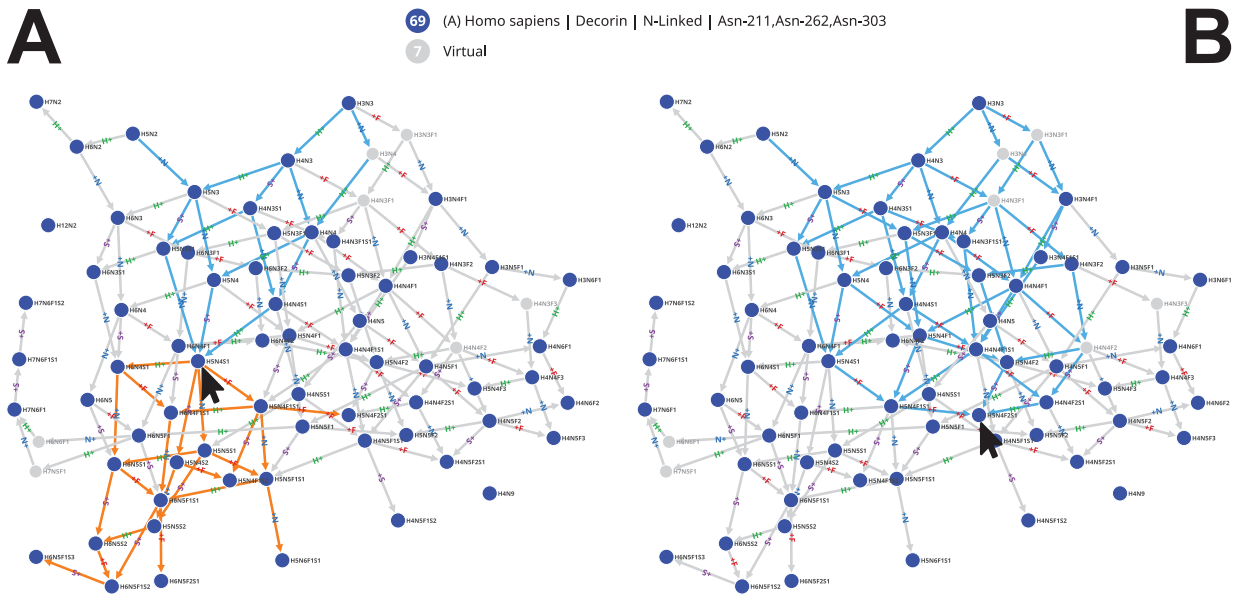
Supp Figure S2



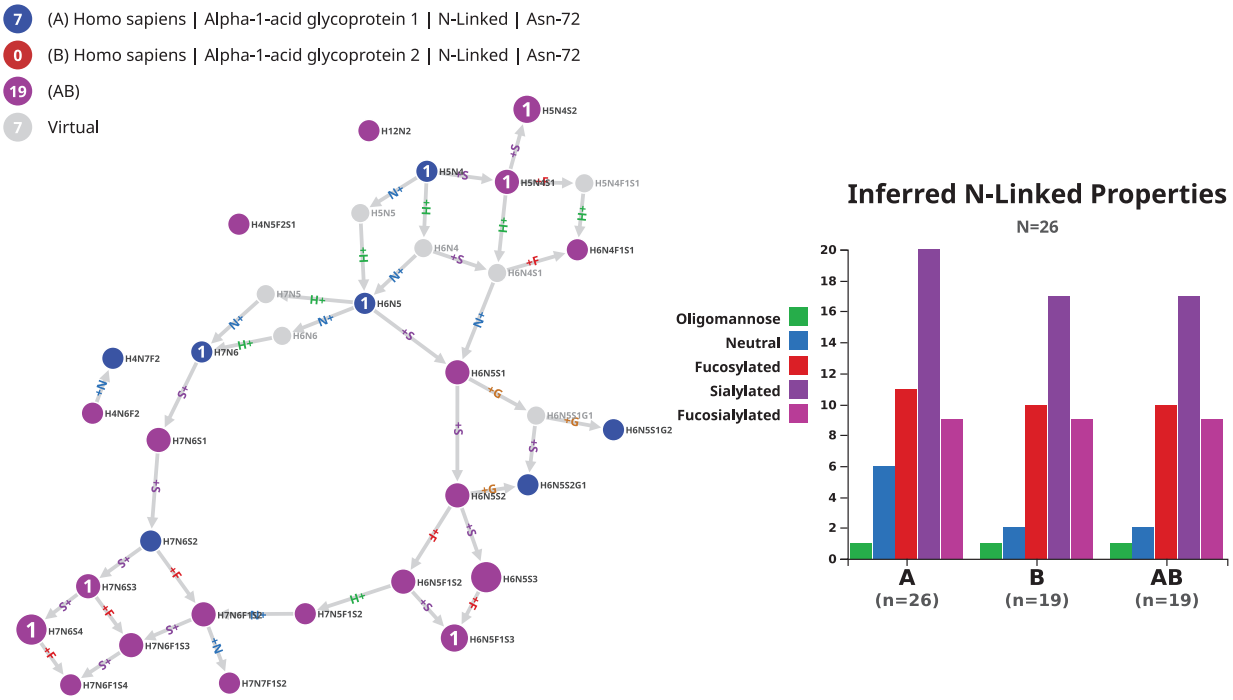
Supp Figure S3



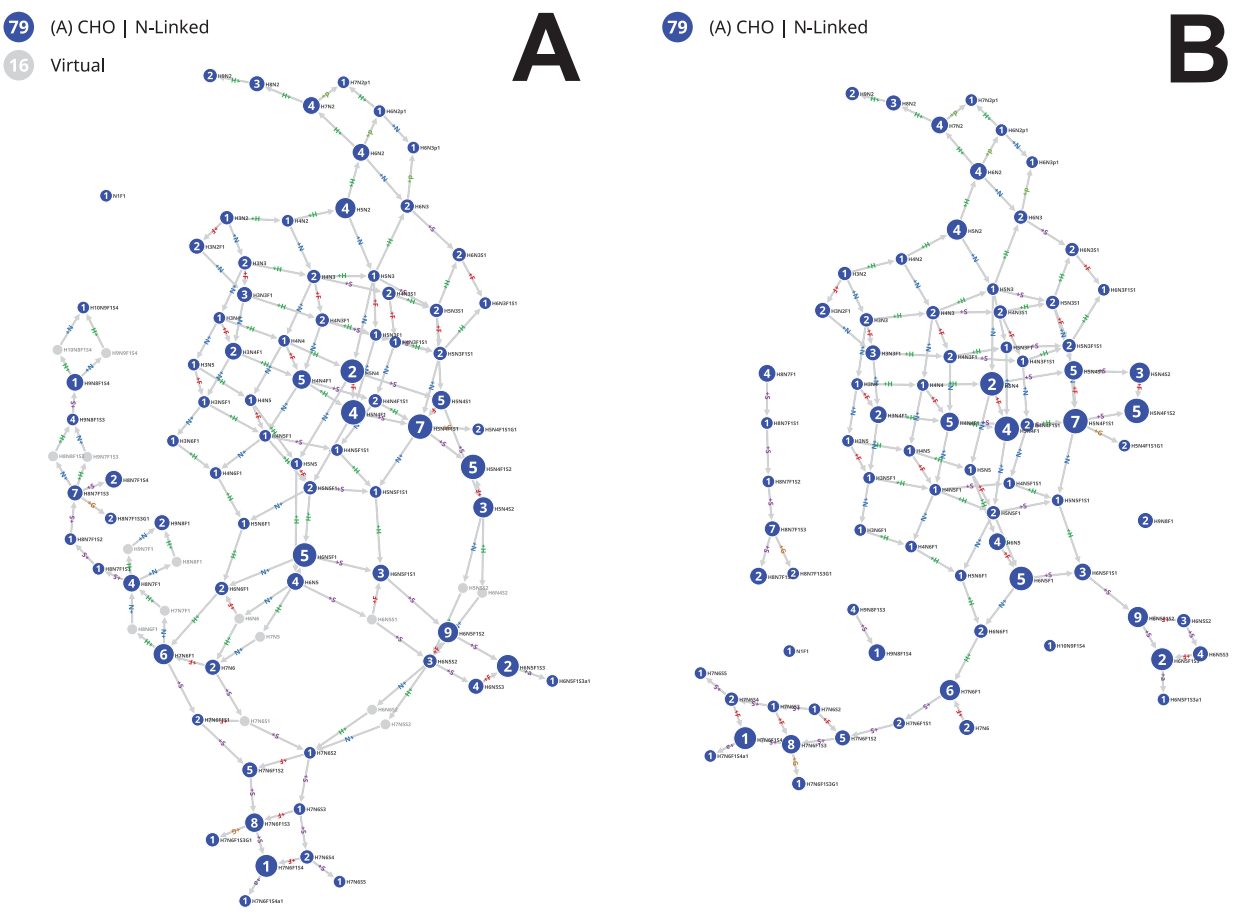
Supp Figure S4



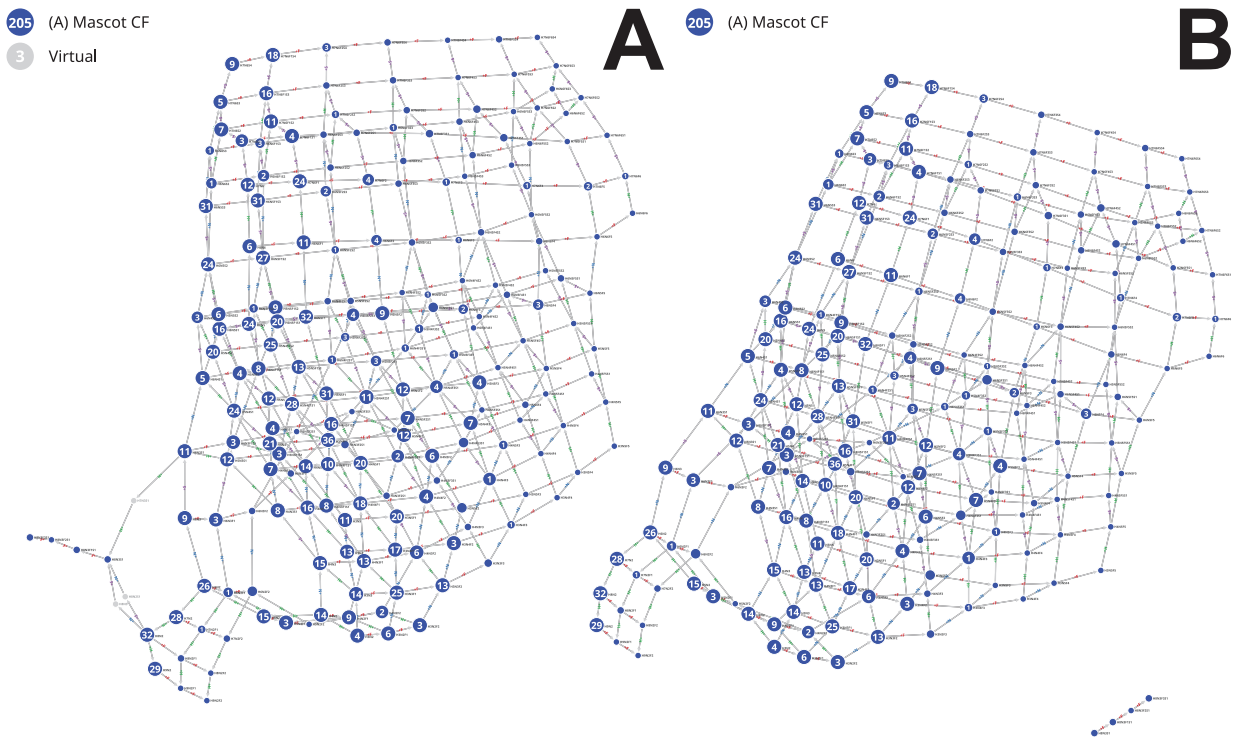
Supp Figure S5



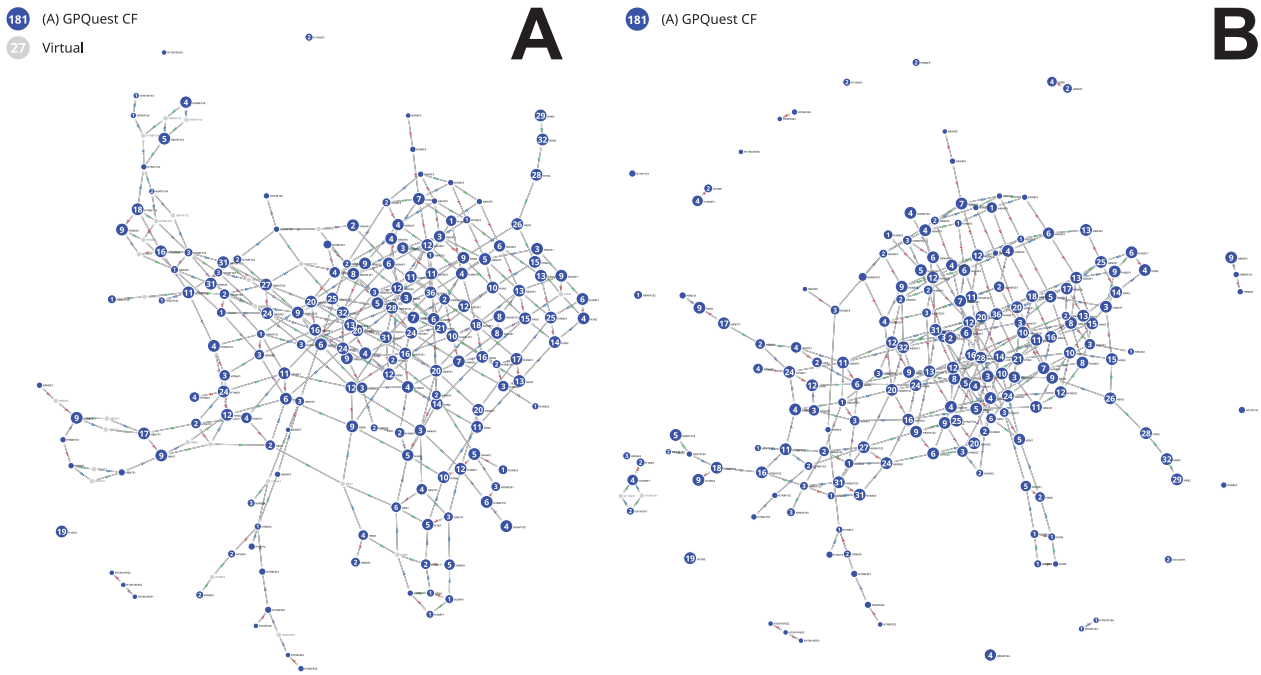
Supp Figure S6



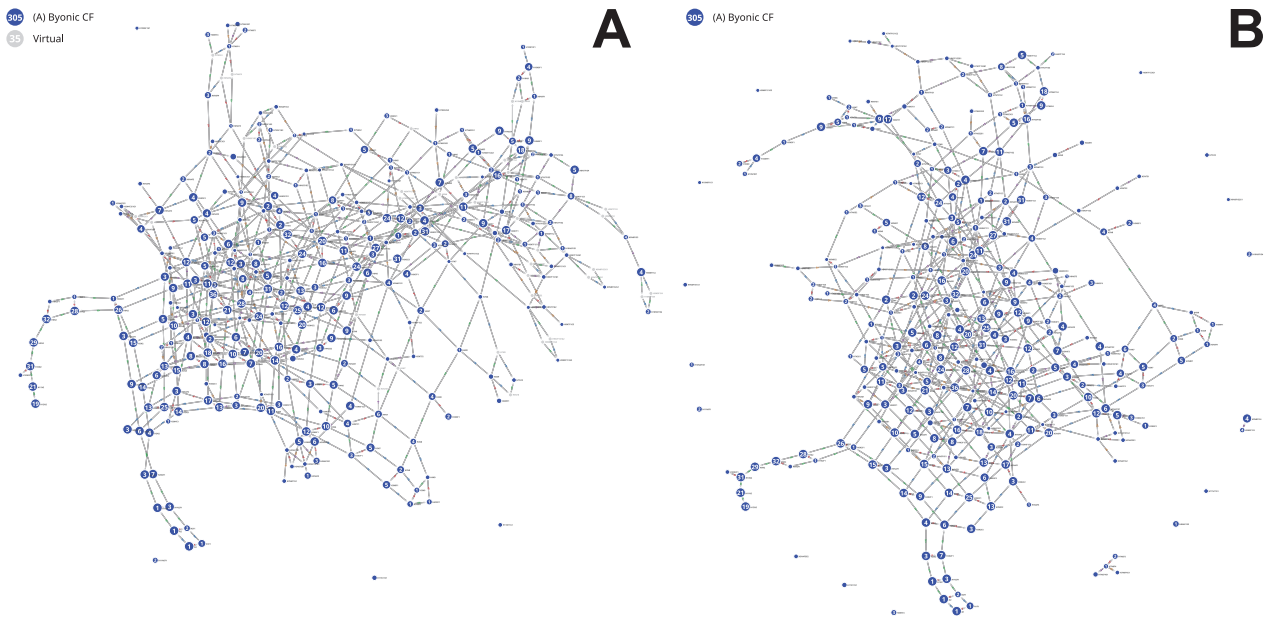
Supp Figure S7



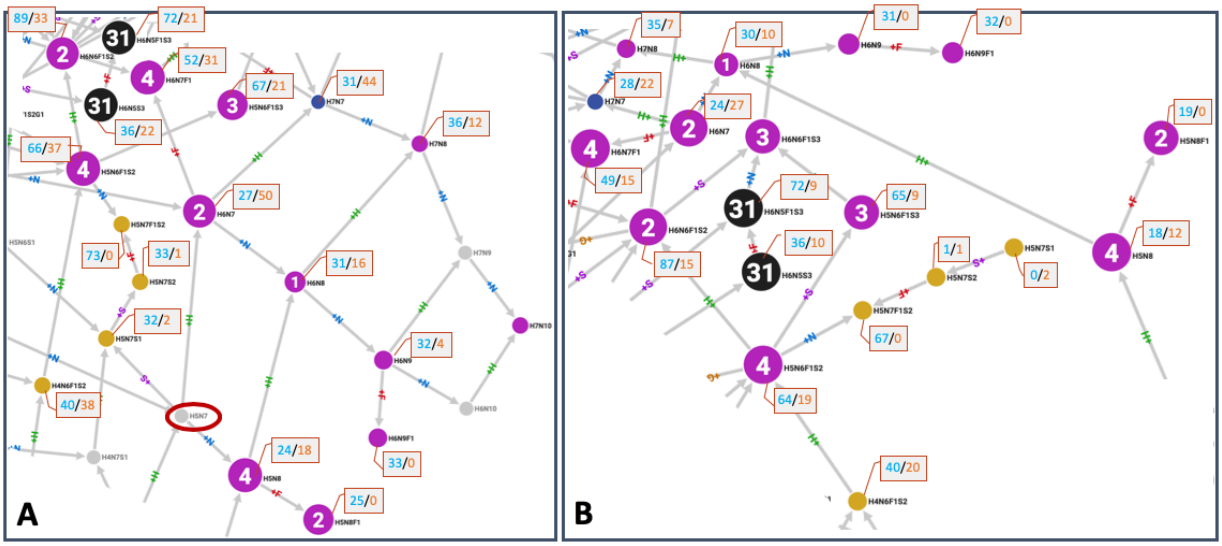
Supp Figure S8



Supp Figure S9



Supp Figure S10



5.2. Concluding Remarks

The analyses showcased in this study demonstrate how data visualization software such as GlyConnect Compozitor can add value to the content of a database. Without having to include additional data, such tool can highlight existing but hidden or implicit information. In the case of Compozitor, the comparison of the respective glycomes of distinct biological entities allows assessing their consistency and detecting specific glycan composition differences that may potentially have biological relevance. Nonetheless, visualization tools in bioinformatics are too often considered as an end. In our case, we have emphasized the ability to export any subset of compositions to various formats. They can subsequently be used in glycoproteomics identification software, whose default sets show inconsistencies. GlyConnect Compozitor is a significant contribution to untangling the complexity of glycosylation.

Chapter 6

Discussion

The four last chapters have covered all the bioinformatics tools and analyses that were published or submitted to peer-reviewed scientific journals. In this chapter, we start by discussing what has been achieved globally during this thesis, while highlighting some of the difficulties specifically arising from the different fields of application. Subsequently, we introduce a technical discussion showcasing the core technical features of the software written to address the needs of the different projects.

6.1. Achievements

This thesis encompasses some of the many facets of biological data analysis and interpretation performed with the help of bioinformatics software. Through our projects, we had the opportunity to both process data ourselves and develop data processing tools for the scientific community. Thus, we could evaluate the different challenges that arise throughout the life-cycle of biological data. In the following, we discuss the different aspects of data handling in bioinformatics we had to put into practice.

6.1.1 Data Annotation

As presented in Subsection 1.1.1, metadata represents a critical element for the interpretation of any biological data. Without this information, it is impossible to put data back in the context of its generation. Metadata provides essential attributes describing both the biological sample and the methodology that was used to produce the data. Metadata information can however be missing, incomplete, or erroneous. As such, the use of standards, ontologies and controlled vocabularies

is recommended to provide a regulatory framework for the disseminated metadata information and its encoding. While it is relatively easy to control metadata formatting when using tools and workflows compliant to standard data formats, the knowledge extracted from literature can be sub-optimal in terms of standardization. For instance, the names and identifiers for biological entities can often contain issues (e.g., spelling errors, duplications, or empty values) when they are manually inputted in result tables. As a result, one or several steps of curation are often required to validate and standardize the annotation of biological data. It was notably the case for the GlyConnect database, which required significant refactoring to improve some of its data annotations.

As a direct heritage of the GlycoSuiteDB database [115], the GlyConnect annotations for biological sources were originally underwhelming. They consisted in up to four terms describing the origin of glycans and glycoproteins. All of these terms were labeled as *tissues*, even when they aimed to characterize cell types or cell components. Additionally, they were not relying on any ontology or controlled vocabulary, leading to numerous duplications and incoherent names. Many entries originating from the same tissue also had inconsistencies in the level of detail of their annotations. As GlyConnect Compozitor required precise biological source annotations for its query system, substantial refactoring was needed before developing the tool. To this end, I manually reviewed and refactored the previous annotations based on reference biological ontologies. In the new system, the sources consist in the combination of three distinct categories when known: tissues, which are based on the Uberon [117] ontology (i); cell types, which are based on the Cell Ontology [118] (ii); or cell components, which are based on the Gene Ontology [119] (iii). Of note, the tissue annotations additionally use the BRENDA Tissue Ontology (BTO) [120] in the case of plant organisms, as plant tissues are mostly lacking from Uberon. A few sources, such as the royal jelly from honey bees, could however not be associated with any ontology. They were still retained as they contain a wealth of information about glycosylation. Furthermore, the disease annotations were refactored based on the Disease Ontology [121].

Like in any other omics field, numerous glycomics and glycoproteomics experiments have been performed on cell lines. While the cell lines used can be found in the data extracted from publications, the information was never exposed in the GlyConnect database. Such information can however be of strong interest as cell lines represent valid model systems, while being responsible for the generation of unreliable experimental results when they are misidentified or contaminated. After reviewing the original data from the GlycoSuiteDB database, a total of 112 cell lines could be added as a supplementary fourth category of biological sources (GlyConnect release of October 2019). For all cell lines, a cross-reference to their corresponding

Cellosaurus entry is now provided. The naming of cell lines is known to often be problematic and can prevent correct identification, as several cell lines of distinct origin can end up sharing the same name [53]. Thus, the cell line annotations in GlyConnect were adapted to use the recommended names from the Cellosaurus. Of note, this review and integration process enabled the creation of 14 new Cellosaurus entries. Among the total of 112 cell lines referenced in GlyConnect, 8 are reported as problematic (2 misidentified and 6 contaminated, including 4 by the HeLa cell line alone) by the Cellosaurus. These problematic cases are clearly labeled both in GlyConnect and its Compozitor tool in order to inform the user about the potential issues of using the related data.

6.1.2 Data Standardization

Bioinformatics being a fundamentally data-based scientific field, the support of a wide range of data formats is a crucial aspect to take into account when developing computer software. When they exist, standard formats should always be preferred so as to reliably propagate data and metadata information. Unfortunately, such standards are not always available in some of the more niche research fields. Another issue is that some standard formats may be limited in their data compatibility, either by design or through a lack of update to adapt to newer methodology, instrumentation, or software. As a result, researchers encountering such limitations can resort to develop new formats or extend existing ones to bypass the problem. While this can temporarily and locally improve the situation, it results in the multiplication of available formats. These formats can differ both in terms of technology and content, which can make the development of converter tools a challenge.

In comparison to other omics disciplines such as genomics, transcriptomics or proteomics, glycomics is a newer field of investigation that only started to gain a widespread interest in recent years. Thus, researchers in glycomics had a chance to follow in the footsteps of the standardization efforts that have been previously achieved in the other fields. This notably led to the inception of the MIRAGE (minimum information required for a glycomics experiment) initiative [122] in 2014, based on the success of the other minimum information standards launched in the previous years. This initiative provided a regulatory framework for glycomics experimental data generated in a wide range of experimental approaches, such as mass spectrometry (MS), chromatography and glycan arrays. Regrettably, the situation is more complicated regarding the formats describing the glycans themselves, for which standardization has been a long-lasting issue. A multitude of data formats encoding glycan structures emerged over the years. In most cases, these glycan structure formats were designed to answer the needs of a tool or database and did not origi-

nate from a consensus. As a result, they mostly vary in terms of architecture and technology, while containing the same core information. Converter tools play consequently a major function in glycomics, despite rarely being comprehensive and only supporting a limited subset of data formats.

The information complexity of glycan compositions is significantly reduced compared to glycan structures. Consequently, much less effort has been invested in the development of corresponding formats. The composition of a given glycan consists in the enumeration of all its constituting residues (monosaccharides and substituents) along with their corresponding frequency counts and ignores linkage information. As such, glycan composition formats represent essentially one-line string notations that vary in their denomination of residues and in their separator characters. Similarly to glycan structures, the existing notations for glycan compositions mostly emerged in parallel to the development of bioinformatics tools and databases. Three distinct notations are currently implemented in the GlyConnect Compozitor tool: the GlyConnect notation derived from the GlycoMod [123] tool (i); the Byonic notation used in the Byonic [124] search engine (ii); and a condensed notation that can be found in the scientific literature (iii).

Nevertheless, the three formats are not fully inter-compatible in terms of the composition residues they support. Some residues can thus be specific to a notation and not be described in the others. These limitations represented a challenge for GlyConnect Compozitor, since the tool required to handled the different formats in a cohesive manner. As an export function is supposed to provide the same core information in the different data formats it proposes, excluding these problematic cases was not a satisfactory solution. Instead, we opted for using the naming convention in use in the GlyConnect notation for the missing Byonic residues. In addition, a comment is provided next to the affected compositions to indicate that they may not be supported by the Byonic search engine. Moreover, the condensed notation has never been properly defined as a standard since it is mainly used to report concisely compositions in publications. As such, we had to extend this notation for it to be able to support all of the residues contained in the compositions stored in the GlyConnect database. This highlights the issues encountered in glycomics concerning data standardization. While defining a standard for glycan compositions would be the preferred approach, it would require to mobilize the glycomics community for the achievement of this endeavor.

6.1.3 Data Reanalysis

Through collective efforts, the public sharing of omics data sets is nowadays considered a good scientific practice. The availability of biological data in the public do-

main contributed to the improvement of the reproducibility of scientific research by enabling the regeneration of experimental results. This consequently led to a drastic increase in the amount of data stored in databases and repositories, requiring in turn the establishment of standards and guidelines to ensure that the accumulated information could remain manageable and easily reusable at all times. In addition to the minimum information standards that were successively established in the different omics fields, FAIR principles [125] were collaboratively elaborated in 2016 to regulate data and data resources. These guiding principles decree that any research object should be Findable, Accessible, Interoperable and Reusable (FAIR) by both individuals and computers. Fifteen principles were thus defined, recommending for instance the use of unique and persistent identifiers for data and metadata. To further improve the findability and accessibility of biological data sets, several aggregating resources such as OmicsDI [126] and DataMed [127] were put online in the last years. They enable data discovery across multiple repositories by searching for specific data sets meeting requested criteria (e.g., data type, instrument, organism, tissue, or disease).

Data reanalysis can be performed either punctually by independent research groups or as a routine process by online resources. For instance, the PeptideAtlas [128] database systematically reprocesses the data sets newly submitted to the resources part of the ProteomeXchange [74] consortium to extract novel protein and peptide identifications. A recent study [129] estimated that a given omics data set is reanalyzed 2.3 times on average. Interestingly, proteomics data sets were shown to have a higher reanalysis rate (5.9) than genomics (1.26) or transcriptomics (1.31) data sets. This can potentially be explained by the great complexity of proteomes. A great variety of orthogonal studies can be designed upon existing proteomics data sets, characterizing the many aspects of proteoforms.

Although the establishment of regulatory guidelines and standard data formats widely democratized the reuse of published omics data sets, data reanalysis remains a challenging task for individual researchers. As we experienced first hand in our reanalysis of proteomics data (see Chapter 4), numerous issues can arise throughout the reanalysis process. This is particularly true in the case of large-scale studies, as the available computational resources can become a supplementary limitation. The reanalysis of a large amount of data often requires the use of computer clusters to process it in a reasonable time span. We were privileged to run our pipeline on the Baobab high-performance computing (HPC) cluster of the University of Geneva for our study in contrast with many researchers worldwide who cannot even access such computational infrastructure. While online computational services exist, they often represent a significant financial investment depending on the number of CPU hours required. This also represents an issue for the repro-

ducibility of scientific research. Even if a large-scale experiment is theoretically reproducible through sharing all the raw data, methods, and computer scripts, in practice reproduction may be difficult for many researchers that do not have access to the computational resources. Setting up a workflow can also require advanced technical skills and take a significant amount of time. The use of containerization technologies (see the Docker Containerization Subsection) can greatly facilitate this workflow setup process, but it does not resolve the computational issues.

Another problem that can be encountered when aiming to reanalyze omics data is that the application of the newly established guidelines is not always retroactive. Although converter tools are available, previously deposited data sets are not always converted to the more recent standard data formats. As we found out in our reanalysis of HeLa proteomics data, numerous older data sets in the PRIDE [72] repository are for instance only available as peak list files. As such, the files do not contain any metadata information, forcing users to rely on the annotations provided in the free text description of the experiment which is one of the components of a PRIDE entry. Furthermore, these annotations are not manually curated and their quality consequently depends on the goodwill of the submitter. For these cases, it is therefore essential to go back to the original publication, if possible, to validate the provided annotation information.

6.2. Technical Discussion

Prior to starting this thesis, my programming experience was mostly restricted to the Python language. As many students new to bioinformatics, I mainly used Python as a scripting language to write simple data analysis scripts. These scripts could however be combined to design larger workflows automating the analysis process. Python became, alongside R, one of the most used programming languages in bioinformatics. This popularity is mainly related to the availability of numerous bioinformatics frameworks, such as Biopython [130] for Python and Bioconductor [131] for R. As a high-level dynamically-typed language, Python can also be more easily accessible for beginners.

Despite Python supporting, among others, the object-oriented programming (OOP) paradigm, it is rarely taught in detail in courses and lectures targeting biologists. Nonetheless, this paradigm procures substantial advantages over scripting when developing complex applications. The modular nature of OOP enhances code reusability between projects while facilitating developers to spot and fix bugs in their software. Under the condition that the code is properly documented, OOP promotes code maintainability which is critical for large-scale projects. Additionally, Python support for the OOP enables programmers used to other object-oriented languages, such as C++ or Java, to adapt to Python more easily.

I switched early in the thesis from Python to Java, as most of the code base of the SIB PIG group was written in this programming language. Despite Java being less popular than Python or R for bioinformatics software development, numerous libraries for omics data analysis and processing are available. Most of the tools developed by the PIG group are based on the in-house MzJava library [132]. This open-source Java library provides functionalities and methods for the analysis of mass spectrometry (MS) data generated in high-throughput proteomics and glycomics experiments. It was extensively used to process the MS identification data from the analysis presented in Chapter 4. Although none of the tools presented here directly use MzJava, they are all, at least in part, written in Java. In the following, we present the core technical features of the tools developed during this thesis.

6.2.1 From Desktop to Web Applications

Three main applications have been successively developed in the context of this thesis: MzVar, CLASTR and GlyConnect Compozitor (Table 6.1). In addition to answering distinct biological questions from their respective fields, the three tools dif-

fer in terms of design and architecture. While MzVar was written as a Java desktop application implementing a JavaFX graphical user interface (GUI), CLASTR and GlyConnect Compozitor are web applications using a Java backend and a JavaScript frontend. This transition from desktop to web applications has several advantages, at the cost of greater technical investment. For instance, the configuration and installation of the application are not delegated to users. Under the condition that it was properly developed, a web application can be accessed by any device, operating system and web browser combination with no time needed to set it up. This is particularly relevant to bioinformatics, as the target audience is composed of very distinct profiles: biologists, computer scientists, and bioinformaticians whose skills, expectations and work habits differ significantly.

Tool	Year	Category	Function	Language
MzVar	2017	Desktop Application	Database compiler	Java
CLASTR	2019	Web Application	Search engine	Java, JavaScript
Compozitor	2020	Web Application	Visualization tool	Java, JavaScript

Table 6.1: Characteristics of the tools developed in this thesis.

Another important aspect of software user experience consists in graphical user interfaces (GUIs). Web applications provide one by default but many bioinformatics desktop applications are only operated using command lines in a terminal. It was for instance the case of the first versions of MzVar, limited to few input files and parameters. As detailed in Chapter 4, the purpose of this tool is to compile customized variant protein and peptide databases from transcript sequences and variant call format (VCF) files. Despite being initially developed to answer the needs of one of our own projects, it was also intended for usage by the scientific community. Many biologists are however not comfortable with tools operating in a terminal window. The addition of a GUI coded in JavaFX was a simple way to make the tool accessible (Figure 6.1) and not requiring the redesign of the application structure. Furthermore, the same executable JAR file that launches the JavaFX application interface can be executed as command lines. Thus, the application does not constrain usage in any way. Keeping the application executable in command lines can still be advantageous, as it can be integrated into workflows or run on servers.

The CLASTR and GlyConnect Compozitor applications were successively developed in a very short time interval. As a result, their global architecture and supporting technologies are very similar. Furthermore, both tools directly rely on the data stored in a database (Cellosaurus and GlyConnect respectively). In this respect, a web application was an obvious choice. This enabled in both cases the addition of

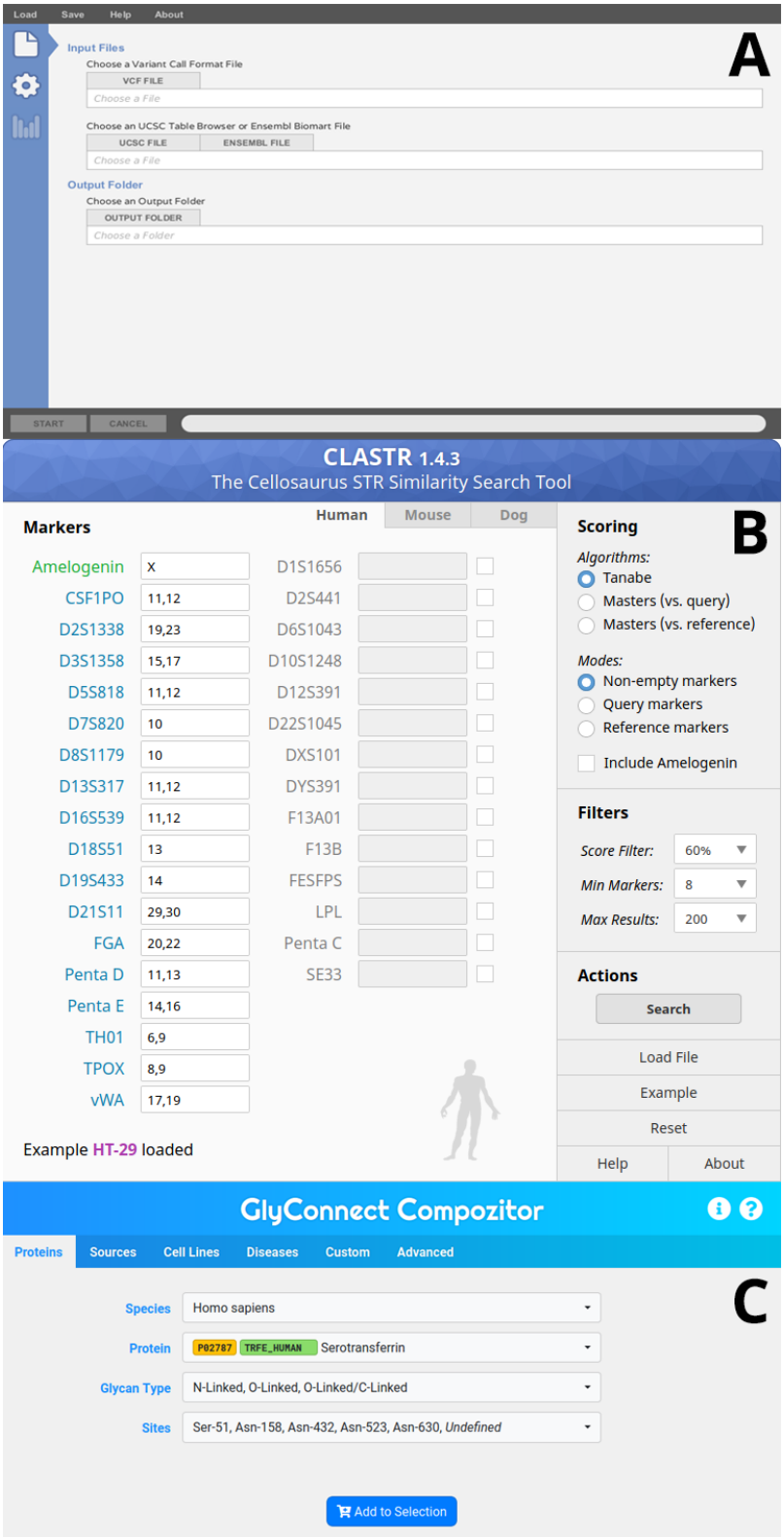


Figure 6.1: User interface of the tools developed in this thesis: MzVar (a); CLASTR (b); and GlyConnect Compozitor (c).

cross-reference URLs from the database entries to the tools, automatically loading all the relevant information into the application input forms. A major difference in the design lies in the way the database information is obtained. In the case of Compozitor, the glycosylation data is retrieved through calls to the GlyConnect RESTful API based on the queries the user is making. In CLASTR, the STR profile information is preloaded and the application starts with parsing the XML version of the Cellosaurus available on the ExPASy FTP server. In both applications, most of the processing is performed by a backend written in Java. The backend is exposed through the use of an Apache Tomcat application server and a RESTful API. While the Compozitor API was developed to support the fronted and not initially meant to be openly used, the CLASTR API was designed as a service. All core features provided by the CLASTR web interface can be accessed through the API and retrieved in various formats.

6.2.2 Modern Web Development

With the advent and popularization of web technologies such as JavaScript and its numerous libraries, web development significantly evolved in recent years. Using JavaScript to interact with the document object model (DOM), the content of an HTML page can be altered with no need to refresh it or make server calls. This enables the design of single-page applications (SPAs), in which the content of a unique web page is dynamically updated based on user actions. The main advantage of developing such applications is the overall enhanced user experience resulting from improved web page usability and performance. However, the most significant change in modern web development is arguably the wider use of JavaScript frameworks to design web sites and applications. These frameworks (e.g. Angular or React) facilitate development by providing pre-written web components while ensuring that the user interface is synchronized with the inner state of the application. Frameworks reveal their full potential when developing large-scale applications or SPAs with complex user interfaces. Writing code using a given framework is however not straightforward, often requiring an initial time investment to understand modalities and nomenclatures. Additionally, they facilitate the design of responsive applications, whose user interfaces adapt themselves in function of the device viewport size. This aspect is however somewhat less relevant to scientific software, as a large majority of researchers uses desktop computers in their workplace and not smartphones or tablets.

To keep the CLASTR application simple, it was developed as a SPA without the use of a JavaScript framework. The choice was further justified by also wanting a simple user interface: a form to input the STR profile data and an HTML table

to present the corresponding search results. Several libraries were however used, including SheetJS and Papa Parse to parse the different input file formats, and JQuery to perform AJAX requests to the backend. With this experience, the subsequent development of GlyConnect Compozitor was handled differently. The main architectural distinction between the two applications is that GlyConnect Compozitor heavily relies on the Bootstrap frontend framework to display its interface. Bootstrap is notoriously lightweight and offers numerous functionalities that decrease development time. It enabled the design of a clean and functional interface composed of multiple views in a very short span of time.

6.2.3 Docker Containerization

Docker is a free and open-source platform providing tools to automate and facilitate the development, deployment, and execution of computer software. Using Docker, applications are packaged into containers bundling all the required code, dependencies and configuration files. This isolates the running processes from their environment, ensuring they will behave consistently regardless of where they are deployed. Multiple containers can thus be launched in parallel on the same server without the risk of having dependency or configuration conflicts. The deployment is performed using a single command line without having to consider pre-installation requirements other than the Docker daemon itself. This simplified procedure allows easily moving any application from a given computing environment to another, which can prove to be particularly useful when deploying from test to production. The containers are created from Docker images, which are files representing snapshots of the file system once all the necessary system libraries and tools have been installed. The content of a Docker image is specified by a *Dockerfile*, which is a small script file containing all the command instructions required to build the image. Containers are built from images (as objects are instances of classes in the OOP), and multiple containers can be launched using the same image. This enables to easily scale an application by running more instances of the same container under the control of a load balancer.

Docker is made possible by two features from the Linux kernel called *namespacing* and *cgroups*, enabling respectively the isolation of resources per process and the limitation of the amount of resources allocated per process. On Linux, Docker can run without the use of virtual machines, making it lightweight in terms of hardware resources required to function. On the Windows and Mac operating systems, the Docker engine runs on a Linux virtual machine in order to access to the features provided by the Linux kernel. Of note, Docker does not work on the *home* version of Windows 10, as the Microsoft Hyper-V hypervisor handling virtual ma-

chines is not availableon this platform. This however is not much of an issue as most servers use Linux, and the *Enterprise* version of the Docker engine supports Windows servers.

The use of Docker in the context of this thesis was prompted by the instability of some applications that are part of the Glycomics@ExPASy initiative [133]. The GlyConnect [114], SugarBind [134] and UniCarb-DB [135] databases all inherited the same common architecture: they were developed using the Play framework, which requires the installation of Scala and its SBT build tool in addition to a Java runtime environment. Each of the three databases needs a specific combination of version dependencies in order to be able to be compiled and launched. As all the database applications were deployed on a single server, the different versions of the same dependencies were seemingly causing conflicts that were making them periodically crash. In order to resolve the issue, we ported the applications to Docker by writing a specific Dockerfile for each of them. Using the Docker compose tool, we made sure that the Docker daemon would automatically restart any container that would shutdown unexpectedly in case they would still be unstable. With this newly acquired experience, I additionally ported the CLASTR and GlyConnect Compozitor tools to Docker. As they are fairly similar in design, the same Docker architecture (Figure 6.2) was used to deploy them on the SIB ExPASy server [116].

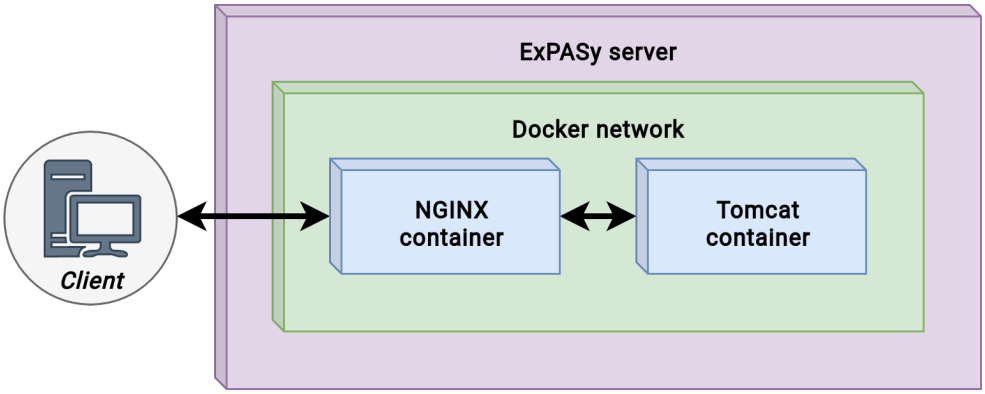


Figure 6.2: Scheme of the Docker implementation shared by CLASTR and Glyconnect Compozitor, as deployed on the ExPASy server. The two blue boxes represent the Docker containers: one for the NGINX web server handling the front end HTML/CSS and JavaScript files; and another for the Tomcat application server handling the back end written in Java. The two containers communicate by share the same network in bridge mode as represented by the green box. Only the NGINX container is directly accessible from the exterior.

The interest of Docker for bioinformatics, and more generally for research in science, is twofold. In addition to the containerization of applications detailed above, Docker can be used to run workflows and pipelines in containers. Thus, this technology can

contribute to the improvement of the reproducibility of scientific research [136]. When an analysis workflow featured in a scientific publication uses Docker (or any other containerization system, such as Singularity or OpenVZ), it enables other researchers to easily reproduce the presented results. For instance, all the data and plots presented in the study featured in Chapter 3 can be regenerated by anyone using a single Docker command in a console. All the tools and scripts that were used to generate the results of an article are not always available in the corresponding supplementary material. Additionally, setting up a whole analysis workflow can be time-consuming and require computer skills that wet lab researchers may be lacking. Dockerfiles being significantly smaller than Docker images, they represent good candidates to be included in the supplementary information. One downside is that Dockerfiles rely on the availability of dependencies to be able to build images. There is no strict guarantee that a Dockerfile would work after a prolonged period of time, even if it is unlikely when using trusted dependency providers (e.g., Google APIs, Cloudflare or Apache Maven). As such, there is an interest to also provide the corresponding image to ensure sustainability, despite them potentially weighting several gigabytes as a result of being file system snapshots. In addition to the reproduction of experimental results, workflow containerization also enables to apply the same implemented methodology on new data sets with little effort.

Chapter 7

Conclusion

In the previous chapters, we focused on the work achieved in the course of this thesis. We discussed the problems that were encountered in the different biomolecular disciplines we covered while detailing the computer software solutions that were designed to overcome them. In this final chapter, we conclude by listing some of the tasks that can still be undertaken to further improve what we accomplished, as both omics and computer technologies are constantly evolving.

7.1. CLASTR

Since its public release on the SIB ExPASy server [116] in March 2019, the CLASTR tool has been regularly updated to fix minor issues and implement new functionalities. For example, the Microsoft Excel XLSX format was integrated as an export choice in both the web interface and RESTful API. While not being a standard file format, it remains commonly used by researchers and allows keeping the same color code and comments than the ones proposed in the application web interface. As originally planned, the species support was additionally extended to mouse and dog cell lines, which are the only species other than human for which STR markers have been selected. Overall, CLASTR has been particularly stable and provides the scientific community with a reliable service to perform STR similarity searches. As a result, there is little room for improvement in the current tool version (1.4.3). The main possible enhancements are in the Cellosaurus data itself, with the integration of new STR profiles and the annotation of cell line entries when new cases of misidentification or cross-contamination are reported.

Nonetheless, there is currently no standard data format for the exchange of STR profiles between the different stakeholders, that is, the laboratories carrying out STR profiling, the researchers, the journals and online resources such as the Cel-

losaurus. The existence of such a standard would strongly benefit and promote the sharing of STR profiles between data producers and consumers. The implementation of standardized file formats to report experimental results has been successfully achieved in numerous omics disciplines [137] and invitroomics [138] would benefit from such a development. This new standard format should be designed to enable storing and regulating metadata information that is relevant to the generation of an STR profile. It could notably contain annotation information about the data producer identity, the cell line characteristics and culture conditions, the profiling methodology, and the dates of submission and analysis. The data core would consist of one or more STR profiles, indicating the STR loci and corresponding alleles. In terms of architecture, the file format could be either text or XML based as commonly the case in bioinformatics. One advantage of the XML format is the ability to validate that a file is compliant with the standard through the use of XSD files describing the schema. Once the standard specifications have been defined, the file parser of CLASTR would have to be extended to support the extraction of STR profile information from the file and its loading into the input form.

7.2. MzVar

In comparison to other software solutions proposed in this thesis, MzVar has a simpler design and architecture due to its more limited scope. Additionally, numerous tools proposing similar functionalities emerged around the time frame of its development, such as PGA [139] or QUILTS [140]. MzVar in its present state is complete in terms of core features. Nonetheless, an interesting addition would be the integration of the recently developed PEFF format [87]. MzVar already uses FASTA headers to annotate sequence variants introduced in the protein or peptide sequences. Thus there would be a strong incentive to use a standardized format to regulate such information. This could improve the compatibility of MzVar with existing tools and workflows while ensuring the variant annotation is reliably encoded.

7.3. GlyConnect Compozitor

Through the implementation of multiple search criteria, GlyConnect Compozitor enables researchers to visually explore most of the content stored in the GlyConnect database. A notable exception lies in the glycopeptides that were observed in MS-based experiments, which were the last data type to be integrated into the database. As a consensus regarding validation criteria has yet to be reached, the quality of an-

notation is difficult to define. In comparison to traditional proteomics approaches, the validation of identified glycopeptides presents additional challenges [141], such as the usually low amount of glycopeptides identified per run and the complex fragmentation patterns they produce with the different fragmentation methods. Nonetheless, some GlyConnect peptide entries contain essential information about genetic variation (with cross-references to the dbSNP database [142] and UniProtKB/Swiss-Prot [85] variant pages when available) in relation to glycosylation sites. In some cases, a variant can create a new glycosylation site, by either changing the amino acid directly anchoring the glycan or other amino acids involved in the glycosylation pattern. For instance, the Beta-2-glycoprotein 1 (P02749) is annotated in GlyConnect as having a substitution from a serine (Ser) to an asparagine (Asn) residue at position 107 (dbSNP:rs1801692). As a result, a new glycosylation site was created at this position, for which three distinct compositions were detected in a recent glycoproteomics experiment [143]. However, such glycosylation sites are not yet indicated as variants in the site category of GlyConnect. They could have a strong biological relevance by their potential involvement in the development of genetic-based diseases such as cancer. Exposing this information would enable Compozitor to highlight the related glycan compositions, allowing the investigation of some specific glycans associated with these novel sites.

Much work remains to be done regarding the annotation of cell lines in online bioinformatics resources, particularly when they were deemed as misidentified or contaminated. For instance, the HEK (CVCL_M624) cell line was originally described as being derived from a human embryonic kidney. As such, the glycan structures and compositions that were identified on this cell line are annotated as originating from kidney in GlyConnect. As reported in the Cellosaurus, HEK was however contaminated by the HeLa human cervical cancer cell line. As a result, the tissue annotation of any results generated through the experimental use of this cell line should be corrected to cervix. In addition, a disease annotation should be included to specify they were generated from an adenocarcinoma sample. While it is an important step in the right direction, settling for cross-referencing the Cellosaurus is thus insufficient for databases and resources handling cell line data. Moreover, numerous databases still only use the cell line name to refer to it, which was shown to be highly unreliable [53]. As achieved for GlyConnect and the Compozitor tool, it is additionally important to inform users when the cell line is problematic. As future work, the systematic correction of tissue annotations of contaminated cell lines will have to be performed in the GlyConnect database. The Cellosaurus has planned to refine the tissue sources of cell lines using the Uberon ontology [117]. Once this is completed, the tissue annotations in GlyConnect could thus be reliably corrected. Additional efforts will have to be carried out to revise the disease and cell type annotations.

7.4. Final Thoughts

To conclude, the development of the GlyConnect Compozitor and CLASTR web applications brought significant value to their respective databases. With Compozitor, the glycan compositions stored in GlyConnect can now be accessed in a new manner that allows their context dependent comparison and subsequent export. With CLASTR, the STR profiles listed in the Cellosaurus can be easily searched by researchers in need of authenticating the cell lines they are working with. It constitutes an important contribution in preventing the publication of more unreliable experimental results in the literature. This also directly affects the reproducibility of experimental results, which requires above all the provision of all material, methods, and software that were used. Although the deposition of experimental data in public repositories is nowadays a common and well-accepted practice in all omics fields, the quality of its annotation can still hinder accessibility and reusability. Initiatives such as OmicsDI [126] and ProteomeXchange [74] allow searching for specific data sets across multiple online resources, but they rely on metadata annotations that are not always standardized or can even be missing. As we experienced first-hand during our reprocessing of proteomics data from the HeLa cell line, the lack of data set metadata standardization in repositories such as PRIDE can be a significant obstacle to data reanalysis and reproducibility. We thus hope that significant regulation and curation efforts will be undertaken to tackle this issue, while inviting researchers to be cautious and more rigorous about annotations when submitting their data sets to public resources.

Appendix A

Other Publications

A.1. Proteomics Data Representation and Databases

This chapter of the *Elsevier Encyclopedia of Bioinformatics and Computational Biology* reviews the main standard data formats and databases commonly used in MS-based proteomics research.

Proteomics Data Representation and Databases

Thibault Robin, University of Geneva, CUI, Carouge, Switzerland

Introduction

Proteomics distinguishes itself from other life sciences by the variety and the complexity of the data generated. Compared to genomics, many biological events such as post-translational modifications, protein isoforms, and protein degradation add challenges to data analysis and interpretation (Altelaar *et al.*, 2012). Furthermore, proteins rarely act independently but rather interact with each other in large intricate, dynamic and plastic networks (Baker, 2012; Altelaar *et al.*, 2012).

In recent years, significant breakthroughs were achieved in proteomics technologies (Altelaar *et al.*, 2012). Mass spectrometry (MS) in particular has emerged as the reference approach for the high-throughput identification of peptides and proteins. This progress was made possible through improvements in sample preparation, instrumentation and computational tools (Altelaar *et al.*, 2012). The widespread adoption of next-generation mass spectrometers by the proteomics community yielded to a massive increase in the amount of data generated, raising concerns about their storage and sharing policies. Despite some initial reluctance, the need for open access data in proteomics is nowadays fully embraced by the community. Although data sharing was introduced a long time ago in other omics fields, it was not effortless in proteomics and required a collective input over the years to convince both researchers and instrument manufacturers (Prince *et al.*, 2004; Verheggen and Martens, 2015; Deutsch *et al.*, 2017).

To meet the demand, a large panel of databases with their respective underlying purpose and content specificities was developed over time. Despite the undeniable biological value provided, the resulting diversity can prove to be troublesome for users who have to grasp the extent of the various data formats for both input and output files. Another consequence of such fragmented information is that long-term database maintenance is not guaranteed. Some databases may be financially impacted when the only source of income lies in research grants and donations. A recent setback was the successive discontinuation of the Peptidome (2011) (Slotta *et al.*, 2009; Csordas *et al.*, 2013) and Tranche (2013) (Smith *et al.*, 2011; Science Signaling, 2013) databases due to the lack of funding, resulting in the partial loss of each dataset (Deutsch *et al.*, 2017). These incidents led the ProteomeXchange consortium to make data sustainability one of its priorities (Deutsch *et al.*, 2017).

This article will describe the proteomics field in regard to the available databases and the data they contain. First, the principal data types are presented along with the corresponding standard formats that were established through the years. Next, the main categories of databases are detailed and illustrated, with representative examples that have the greatest impact on the field today. Then, the ProteomeXchange consortium is detailed to present the significant contribution it brought to the proteomics field. Finally, the prospects of proteomics data and databases are examined.

Data Types and Standard Formats

The data generated in most MS-based protemics experiments can be divided into two main categories depending on their origin: raw data and processed data. Metadata is a separate though significant category, present throughout the analysis process. Its purpose is to store and track a wide range of technical and biological information, essential to proper data handling and interpretation. Metadata information is especially essential in the proteomics field, since a broad range of approaches and technologies are being used (**Table 1**).

Raw Data

In its primal form, proteomics raw data is mass spectra. A mass spectrum consists at its core in a collection of mass-to-charge ratios (m/z) plotted against their corresponding intensities (Aebersold and Mann, 2003). Supplementary information is usually provided, such as the precursor ion m/z value and charge along with the retention time when a chromatographic separation step was performed.

Proprietary formats

Mass spectrometers usually produce raw data in the form of proprietary binary files presenting various levels of compression (Martens *et al.*, 2005b). Most of the time, the exact specifications of the format encoding-schemes are not publicly available. Proprietary binary files cannot consequently be accessed by users without software supplied by the instrument vendor. Likewise, third-party software, such as MSConvert from the ProteoWizard suite (Kessner *et al.*, 2008), will require the implementation of specific proprietary libraries to decode the information contained in the data files. The data formats also change drastically from one vendor to another, or even sometimes between instruments, raising important standardization concerns. Despite all the restrictions, proprietary formats contain the greatest amount of information and present overall good performances when being processed by compatible software.

 doi:10.1016/B978-0-12-809633-8.20273-2

Table 1 Summary of the main open data formats in the proteomics field

Name	*Supported Data*	*Organization*	*URL*
mzData	Raw mass spectra	Proteomics Standards Initiative, Human Proteome Organization	http://www.psidev.info/mzdata
mzXML	Raw mass spectra	Seattle Proteome Center, Institute for Systems Biology	http://tools.proteomecenter.org/wiki/index.php?title=Formats:mzXML
mzML	Raw mass spectra	Proteomics Standards Initiative, Human Proteome Organization	http://www.psidev.info/mzml
mz5	Raw mass spectra	Proteomics Center, Boston Children's Hospital	http://software.steenlab.org/mz5
TraML	SRM transition lists	Proteomics Standards Initiative, Human Proteome Organization	http://www.psidev.info/traml
mzIdentML	Identification data	Proteomics Standards Initiative, Human Proteome Organization	http://www.psidev.info/mzidentml
mzQuantML	Quantification data	Proteomics Standards Initiative, Human Proteome Organization	http://www.psidev.info/mzquantml
pepXML	Identification and quantification data	Seattle Proteome Center, Institute for Systems Biology	http://tools.proteomecenter.org/wiki/index.php?title=Formats:pepXML
protXML	Identification and quantification data	Seattle Proteome Center, Institute for Systems Biology	http://tools.proteomecenter.org/wiki/index.php?title=Formats:protXML
mzTab	Identification and quantification data	Proteomics Standards Initiative, Human Proteome Organization	http://www.psidev.info/mztab

Peak list formats

The limitations of the proprietary formats favored the common practice of extracting the mass spectra from the binary files as peak lists stored in text files at the cost of an overall reduced amount of information. In addition to the almost complete loss of metadata information, the mass spectra usually go through denoising, deisotoping, centroiding and charge deconvolution steps during the format conversion (Martens *et al.*, 2005b). The resulting peak list files are therefore significantly more compact while becoming easily manageable for users. The lack of metadata represents the main shortcoming of this format, making it poorly adapted to data storing and sharing.

Many peak list formats were developed through the years, such as the Mascot generic format (MGF), the Micromass peak lists (PKL) and the SEQUEST files (DTA). Despite slight variations in the amount of information they contain, these formats are easily inter-convertible without significant data loss using publicly available software (Martens *et al.*, 2005b).

mzData and mzXML formats

In order to determine a middle solution between proprietary binary files and basic peak lists, new open data formats were proposed at the beginning of the 2000s owing to the expansion of XML (Extensible Markup Language) in bioinformatics. For several years, the mzData (HUPO-PSI, 2006) and mzXML (Pedrioli *et al.*, 2004) formats used to be the most common open data formats to store and share raw data information (Martens *et al.*, 2011). Although sharing the same XML architecture, the two formats mainly differed in regard to their ontology policies (Martens *et al.*, 2011).

On the one hand, the mzData format was developed by the HUPO Proteomics Standards Initiative (PSI) as an open format standard to share and archive data. Its metadata annotation was based on a controlled vocabulary that could be regularly updated (Martens *et al.*, 2011). On the other hand, the mzXML format was developed at the Institute for Systems Biology (ISB) as an open data format accompanied by a suite of tools (Pedrioli *et al.*, 2004). Its metadata annotation was based on a fixed schema that required modification along with the corresponding software to incorporate new annotations (Martens *et al.*, 2011). Although nowadays deprecated, the mzData and mzXML formats are still compatible with most current software and are being used at a smaller scale by the proteomics community.

mzML format

With the purpose of format unification, the mzML format was developed as part of a collective effort under the guidance of the HUPO-PSI to replace both the mzData and mzXML formats (Martens *et al.*, 2011). The coexistence of two open formats having similar functionality was perceived at the time as confusing for users. The goal was then to design a unique XML-based format supporting all past features while providing a consensus way to encode the information (Martens *et al.*, 2011). In addition to instrument support improvement, the mzML format features notably metadata annotation following a flexible controlled vocabulary that allows the inclusion of custom new terms by users without requiring the modification of the XML schema (Martens *et al.*, 2011). The mzML format quickly became compatible with most tools available in proteomics, strengthening its format unification goal.

mz5 format

More recently, the mz5 open data format was developed at the Proteomics Center of the Boston Children's Hospital with the aim of providing a more speedy and storage efficient alternative to mzML while preserving the same ontology (Wilhelm *et al.*, 2012).

Distinguishing itself from its predecessors, the mz5 format is designed based on the Hierarchical Data Format (HDF5) instead of the traditional XML architecture. Despite the partial base-64 encoding being used in mzML, XML was designed to remain easily manipulable by users and consequently struggles to handle massive data files. With data benchmarks, the mz5 format was shown to increase 3–4 times read and write performances while halving data file size in comparison to its XML-based counterparts (Wilhelm *et al.*, 2012). Despite its obvious benefits, the low amount of compatible software prevented the widespread adoption of the mz5 format by the proteomics community.

Processed Data

In most high-throughput proteomics experiments, peptide and protein identification is achieved through the database search approach (Nesvizhskii, 2006). In this method, the experimental spectra are aligned against theoretical spectra inferred from the fragmentation patterns of the peptides contained in a reference protein sequence database. The tools performing the search, referred to as search engines, additionally provide numerous statistical scores assessing the quality of the resulting peptide-spectrum matches (PSMs), allowing to rank the peptides matching a given experimental spectrum (Nesvizhskii, 2006). In a further step, the identified peptides are used to determine the proteins from which they originated. Protein inference can however prove to be challenging, especially in eukaryote species, because peptides can belong to several distinct proteins and multiple protein isoforms may coexist (Nesvizhskii, 2005). The notion of proteotypic peptide was then introduced to single out those peptides that are unique to a protein.

Protein identification is the main task in most MS data processing workflows. Nonetheless, protein quantification is also frequently performed. Quantification experiments usually have the aim of comparing protein expression across multiple samples prepared in distinct conditions, although absolute quantification of proteins can also be achieved. Two mutually exclusive strategies have been established through time for quantification in proteomics: stable isotope labeling and label-free quantification (Bantscheff *et al.*, 2007). In stable isotope labeling approaches such as SILAC (Ong *et al.*, 2002), a mass tag is inserted in proteins through a wide range of experimental protocols. The induced mass shift between the light and heavy forms of peptides is detectable by mass spectrometers, leading to the relative quantification of the proteins of interest through the comparison of the signal intensities (Bantscheff *et al.*, 2007). The label-free approaches expanded more recently with the emergence of more accurate instruments. Being usually based either on spectral counting or on precursor ion intensity, label-free approaches do not include any isotope-tagging step and consequently reduce the sample preparation complexity (Bantscheff *et al.*, 2007).

PepXML and protXML formats

The pepXML and protXML formats (Keller *et al.*, 2005) were originally developed with the purpose of improving and facilitating data transfer between the various tools being part of the Trans-Proteomic Pipeline (TPP) (Keller *et al.*, 2005). The pepXML format stores and shares peptide identification and quantification data while the protXML format focuses solely on the corresponding protein data (Deutsch *et al.*, 2015). Some search engines can directly output their results into files in the pepXML format to be further processed by the TPP. Converter tools are also available in the TPP for the other formats. Although the pepXML and protXML formats were not defined with the underlying aim of becoming official standard formats, many tools outside the TPP adopted them over the years (Deutsch *et al.*, 2015).

mzIdentML and mzQuantML formats

To provide a reliable open standard format for the distribution of peptide and protein identification data, mzIdentML was developed by the HUPO-PSI alongside with a suite of converters for most open and proprietary formats available (Jones *et al.*, 2012). As in the case of pepXML and protXML, search engines that evolve over the years offer the option to export directly their results in the mzIdentML format (Jones *et al.*, 2012).

Based on the success of mzIdentML, the mzQuantML format was launched shortly thereafter under the direction of the HUPO-PSI as a quantitative-focused counterpart (Walzer *et al.*, 2013). It was developed to answer the pressing need for a standard format in quantitative proteomics, especially since the techniques and approaches may diverge significantly between individual experiments (Walzer *et al.*, 2013). The mzQuantML format notably features the ability to cross-reference other XML-based open formats such as mzML and mzIdentML. These cross-links help to keep track of the numerous data and metadata produced throughout the analysis workflow (Walzer *et al.*, 2013). Both formats are based on the same controlled vocabulary that was introduced by the HUPO-PSI for mzML (Jones *et al.*, 2012; Walzer *et al.*, 2013), thus providing a robust yet flexible way to annotate metadata information.

mzTab format

The mzTab format was recently launched by the HUPO-PSI as a complement to both the mzIdentML and mzQuantML formats (Griss *et al.*, 2014). The underlying purpose was to design a simple tabular format summarizing identification and quantification data from proteomics and metabolomics experiments. As the strength of the format resides in the easy report of experimental results, it usually contains an overall lower amount of information compared to most XML-based formats (Griss *et al.*, 2014). Files in the mzTab format do not require specific parsing software since no special encoding is involved, which significantly simplifies their handling by users.

TraML format

Through the years, more specialized open formats also made their apparition in the proteomics field. The TraML (Deutsch *et al.*, 2012) standard format developed by the HUPO-PSI is such an example, capturing solely the transition lists resulting from selected reaction monitoring (SRM) experiments. In SRM approaches, a transition represents the pair of *m/z* values of the targeted peptide precursor ion and of one of its corresponding fragment ions. TraML is designed around the same principles that were established for the mzML and mzIdentML formats, using an XML-based architecture along with a controlled vocabulary encoding the metadata information (Deutsch *et al.*, 2012).

Databases and Repositories

Databases in the proteomics field differ strongly in terms of both their purpose and the data they contain. While some solely serve as repositories for the dissemination of experimental datasets, others have been developed with more advanced functions such as sequence annotation or targeted research goals.

Proteomics databases and repositories are still under substantial development in comparison to other life science fields. This delay can partially be explained by the initial skepticism of a part of the proteomics community in regard to sharing publicly experimental data due to data submission imposed by publishers. One of the common concerns was the possible third-party reanalysis of datasets prior to result release in a publication of the original submitter. To address this issue, most databases feature nowadays the ability to keep dataset submission private for the duration of the reviewing process (Vizcaíno *et al.*, 2014).

Needless to say, UniProt (The UniProt Consortium, 2017) remains arguably the most popular and internationally renowned resource for proteins, playing a central role in the proteomics database landscape as the ever-present reference for protein annotation. As UniProt is however a knowledge database that contains almost to no MS data, it will not be presented in details here (**Table 2**). Likewise, the species-specific resources will also not be discussed in this article.

GPMdb

The Global Proteome Machine database (Craig *et al.*, 2004) (GPMdb) was initially designed by Beavis Informatics for the purpose of storing the reprocessed data produced by the GPM data analysis servers. In this automated pipeline, the MS raw data submitted by users or contained in other databases are reanalyzed using the online version of the X!Tandem (Craig and Beavis, 2004) search engine. The input files submitted to the GPMdb have to be formatted either as peak list (DTA, PKL or MGF) or in a compatible open standard format (mzXML or mzData). In addition to a large panel of search parameters, the users can choose prior to the database search if they authorize the inclusion of the dataset to the GPMdb. With the accumulation of a large amount of data over the years, the GPMdb has diversified its functions and is today commonly used by researchers to compile spectral libraries and develop related search tools. This notably led to the inception of the X!Hunter (Craig *et al.*, 2006) online search engine, which can perform the search of experimental mass spectra against consensus spectral libraries built from the GPMdb.

PeptideAtlas

The PeptideAtlas database (Desiere *et al.*, 2006) was originally developed in 2004 at the Institute for Systems Biology (ISB) as a resource providing genome annotations from the automatic reprocessing of MS raw data by the TPP pipeline. PeptideAtlas presents the peculiarity of being composed of distinct builds according to the organism or tissue of interest (Deutsch *et al.*, 2008). Another noteworthy feature of PeptideAtlas consists in the computation of an observability score for each proteotypic peptide represented in the database (Deutsch *et al.*, 2008). This offers the option of determining the peptides that are the most likely to be detected in a MS experiment for a given protein. As in the case of the GPMdb, the PeptideAtlas database surpassed its initial annotation purpose and is nowadays frequently used as a research database for the compilation of spectral libraries.

Table 2 Summary of the main public databases and repositories in the proteomics field

Name	*Supported Data*	*Organization*	*URL*
PRIDE Archive	Raw and processed data	European Molecular Biology Laboratory, European Bioinformatics Institute	https://www.ebi.ac.uk/pride/archive
MassIVE	Raw and processed data	Center for Computational Mass Spectrometry, University of California San Diego	https://massive.ucsd.edu
jPOSTrepo	Raw and processed data	National Bioscience Data Center, Japan Science and Technology Agency	http://jpostdb.org
PASSEL	SRM raw data	Seattle Proteome Center, Institute for Systems Biology	http://www.peptideatlas.org/passel
GPMdb	Processed data	Beavis Informatics	http://gpmdb.thegpm.org
PeptideAtlas	Processed data	Seattle Proteome Center, Institute for Systems Biology	http://www.peptideatlas.org

PASSEL

The PeptideAtlas SRM Experiment Library (PASSEL) was developed in the framework of the PeptideAtlas project (Farrah *et al.*, 2012). It was designed as a receiving repository for data produced in SRM experiments. In addition, the raw data submitted to PASSEL are automatically reprocessed by specific tools such as mQuest from the mProphet suite (Reiter *et al.*, 2011). The results of this workflow are stored in a separate database that can be accessed through the PASSEL web interface. Among all the obtained transitions, those considered to be the best based on quality metrics are included in the SRMAtlas database (Picotti *et al.*, 2008).

PRIDE/PRIDE Archive

The Proteomics Identifications (PRIDE) database (Martens *et al.*, 2005a) was established in 2004 at the European Bioinformatics Institute (EBI) as a structured repository designed to propagate experimental proteomics data. The underlying intent was to provide a resource that would keep the submitted data intact, preserving it from any kind of control or alteration. PRIDE contains both the raw MS data and the resulting peptide and protein identifications, along with the corresponding biological and experimental metadata (Martens *et al.*, 2005a). As of 2014, the PRIDE database was entirely redesigned and replaced by PRIDE Archive (Vizcaíno *et al.*, 2016). The most significant structural changes brought by PRIDE Archive lie in a new data storage architecture and submission system along with a reworked web interface (Vizcaíno *et al.*, 2016).

The Peptidome database (Slotta *et al.*, 2009) was a public data repository developed in 2009 at the National Center for Biotechnology Information (NCBI) that aimed to allow the exchange of experimental proteomics data. Peptidome was thus closely related to PRIDE in terms of both its aim and features. After the discontinuation of Peptidome in 2011 due to lack of funds, a coordinated effort of both teams led to transfer most of its content to the PRIDE repository (Csordas *et al.*, 2013).

MassIVE

The Mass spectrometry Interactive Virtual Environment (MassIVE) repository was developed at the Center for Computational Mass Spectrometry (CCMS) as a public data repository for datasets produced in MS-based proteomics experiments. Compared to other repositories, MassIVE aims to enhance the interactivity of its content. A noteworthy feature is the ability to add comments and reanalyzes for a given submitted dataset. MassIVE also provides the possibility to directly reprocess the raw data that were submitted using several online workflows.

Tranche (Smith *et al.*, 2011) was a public data repository developed in 2005 as part of the Proteome Commons network that aimed to transfer proteomics raw datasets at a large scale. Since the repository had to handle a high data traffic in relation with the large file size, a special effort was made in the design of an efficient peer-to-peer infrastructure coupled with a reliable client-server architecture (Smith *et al.*, 2011). After the funding shortfall that led to the discontinuation of Proteome Commons and Tranche in 2013 (Science Signaling, 2013), as many data sets as possible were recovered and transferred to MassIVE (Deutsch *et al.*, 2017).

jPOST

The Japan Proteome Standard (jPOST) is an emerging resource that ultimately aims to provide both a repository and a database for the worldwide proteomics community. On the one hand, the jPOST repository (jPOSTrepo) (Okuda *et al.*, 2017) was launched at the National Bioscience Data Center (NBDC) in 2015 as a platform for the global exchange of datasets produced in MS-based proteomics experiments. The underlying aim was to offer an alternative for the Asia and Oceania regions to the existing proteomics repositories. Indeed, most public MS data resources are either hosted in Europe (PRIDE Archive) or in the United States (PASSEL, MassIVE). This can prove to be problematic in terms of data transfer speed for researchers who agree to share their experimental datasets but are located far away from these regions. A special effort was thus additionally undertaken to deploy an efficient file upload system, further facilitating data exchange (Okuda *et al.*, 2017).

On the other hand, the jPOST database (jPOSTdb) is still under development and has not been made publicly available yet. Eventually, jPOSTdb will include the automatic reprocessing of the MS raw data submitted to jPOSTrepo through a customized workflow using a combination of several search engines.

ProteomeXchange Consortium

The ProteomeXchange (PX) consortium (Vizcaíno *et al.*, 2014) was originally formed in 2006 and officially launched in 2011 as a collective effort to coordinate data exchange and submission between MS-based proteomics databases and repositories. Since the partner resources had been developed independently at distinct research institutions, it resulted in limited coordination in regard to their respective data submission guidelines and policies (Vizcaíno *et al.*, 2014). This proved to be troublesome for researchers willing to share their experimental data, struggling to determine where and in which form datasets should be submitted. Likewise, data comparison through the use of several sources could turn out to be challenging. The purpose of the PX consortium was to remove these obstacles by providing a regulated infrastructure and framework (Vizcaíno *et al.*, 2014). Such an initiative had already been successfully implemented in other omics fields, as for instance with the development in genomics of the International Nucleotide Sequence Database Collaboration (INSDC) (Cochrane *et al.*, 2016) to regulate the dissemination of DNA sequencing data.

In its first inception, the PX consortium had only two core members: PRIDE and PeptideAtlas (along with its PASSEL resource) (Vizcaíno *et al.*, 2014). PRIDE was used at the time as the unique entry point for data produced in shotgun proteomics experiments, whereas SRM data were redirected to PASSEL. In the pursuit of its unification role, more receiving repositories progressively joined the PX consortium in recent years. This led notably to the inclusion of the MassIVE repository in 2014, followed shortly by the addition of jPOST in 2016 (Deutsch *et al.*, 2017). More emerging resources are also expected to become members of the PX consortium in the future, further improving the state of data sharing in the proteomics field.

Data submission in the PX consortium can be subdivided in two distinct workflows: complete and partial (Deutsch *et al.*, 2017). In a complete submission, the format of all the submitted files can be recognized and parsed by the receiving repository. It is the recommended submission process in most cases, as it allows searching directly through the results with precise queries. A partial submission contains files that are in a non-compatible format, thus preventing the receiving repository to read them. As with the complete submissions, the files can however still be searched using the associated submission metadata. Despite its limitations, the partial submission process remains essential to support the inclusion of datasets coming from a broad range of proteomics experiments, including marginal or emerging techniques (Deutsch *et al.*, 2017).

Both complete and partial workflows require the submission of the raw data, the processed data and the corresponding biological and technical metadata information. The latter is encoded in an XML format that was specifically developed for the PX consortium (PX XML) (Vizcaíno *et al.*, 2014). After submission, each dataset is assigned a unique PX identifier. A receiving repository specific identifier is additionally assigned, except for PRIDE (Deutsch *et al.*, 2017). If a dataset was submitted as complete, a digital object identifier (DOI) is also created to allow proper referencing in case of reanalysis in another project (Vizcaíno *et al.*, 2014). Since data reuse is also one of the main focuses of the PX consortium, distinct identifiers are specifically generated to track reprocessed datasets and store corresponding results (Deutsch *et al.*, 2017).

The publicly available datasets stored in the resources of the PX consortium are accessed and queried through the ProteomeCentral online portal, which was deployed from the first version. ProteomeCentral notably features the ability to filter the datasets using metadata keywords, which can be for instance used to retrieve solely the set of results matching a given organism, tissue or MS instrument (Deutsch *et al.*, 2017). A RSS feed was also set up to push updates about new publicly available datasets that are released in the PX consortium, providing links to both the ProteomeCentral entry and the corresponding PX XML file storing the submission metadata (Vizcaíno *et al.*, 2014).

Despite the essential contribution brought by PX consortium to the proteomics field, there is still room for improvement. Some work remains for example to be done regarding the compatibly of complete submission with data produced by increasingly popular experimental approaches, such as the data independent acquisition, top-down proteomics or MS-based imaging techniques (Deutsch *et al.*, 2017). The MassIVE repository already started this process by becoming compatible with some DIA-based workflows (Deutsch *et al.*, 2017). The PX consortium is arguably the most impactful initiative that occurred in recent years for databases and repositories in proteomics, bringing together the different resources in the field and standardizing data submission, processing and dissemination.

Conclusion and Prospects

The successful outcome of collective efforts such as the ProteomeXchange initiative has significantly boosted data sharing in the proteomics field. An obvious consequence is the increasing number of journal editors who nowadays mandate the deposition of the original raw datasets to enhance the publication of scientific articles and allow for evidence checks. Further into the future, more integration of the different omics fields is to be expected. Progress has already been made in this direction with the recent development of the Omics Discovery Index (OmicsDI) portal (Perez-Riverol *et al.*, 2016). OmicsDI aims to provide an online platform for the dissemination and access of datasets from genomics, transcriptomics, metabolomics and proteomics resources. Tighter collaboration between proteomics and metabolomics is also likely, based in particular on the fact that both fields rely predominantly on MS-based approaches and techniques to produce their data. Some open standard formats such as mzTab have already started to become compatible with metabolomics workflows.

Acknowledgement

The author would like to thank the Dr. Frédérique Lisacek and Emma Ricart Altimiras for their support.

See also: Bioinformatics Data Models, Representation and Storage. Biological Database Searching. Clinical Proteomics. Data Storage and Representation. Experimental Platforms for Extracting Biological Data: Mass Spectrometry, Microarray, Next Generation Sequencing. Identification of Proteins from Proteomic Analysis. Information Retrieval in Life Sciences. Natural Language Processing Approaches in Bioinformatics. Protein Structure Databases. Proteomics Mass Spectrometry Data Analysis Tools. Quantification of Proteins from Proteomic Analysis. Standards and Models for Biological Data: FGED and HUPO. Supervised Learning: Classification. The Evolution of Protein Family Databases

References

Aebersold, R., Mann, M., 2003. Mass spectrometry-based proteomics. Nature 422 (6928), 198–207.

Altelaar, A.F.M., Munoz, J., Heck, A.J.R., 2012. Next-generation proteomics: Towards an integrative view of proteome dynamics. Nature Reviews Genetics 14 (1), 35–48.

Baker, M., 2012. Proteomics: The interaction map. Nature 484 (7393), 271–275.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., 2007. Quantitative mass spectrometry in proteomics: A critical review. Analytical and Bioanalytical Chemistry 389 (4), 1017–1031.

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., International Nucleotide Sequence Database Collaboration, 2016. The international nucleotide sequence database collaboration. Nucleic Acids Research 44 (D1), D48–D50.

Craig, R., Beavis, R.C., 2004. TANDEM: Matching proteins with tandem mass spectra. Bioinformatics. 20 (9), 1466–1467.

Craig, R., Cortens, J.P., Beavis, R.C., 2004. Open source system for analyzing, validating, and storing protein identification data. Journal of Proteome Research 3 (6), 1234–1242.

Craig, R., Cortens, J.C., Fenyo, D., Beavis, R.C., 2006. Using annotated peptide mass spectrum libraries for protein identification. Journal of Proteome Research 5 (8), 1843–1849.

Csordas, A., Wang, R., Ríos, D., *et al.*, 2013. From peptidome to PRIDE: Public proteomics data migration at a large scale. Proteomics 13 (10–11), 1692–1695.

Desiere, F., Deutsch, E.W., King, N.L., *et al.*, 2006. The PeptideAtlas project. Nucleic Acids Research 34 (Database issue), D655–D658.

Deutsch, E.W., Chambers, M., Neumann, S., *et al.*, 2012. TraML– A standard format for exchange of selected reaction monitoring transition lists. Molecular & Cellular Proteomics 11 (4), (R111.015040-R111.015040).

Deutsch, E.W., Csordas, A., Sun, Z., *et al.*, 2017. The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. Nucleic Acids Research 45 (D1), D1100–D1106.

Deutsch, E.W., Lam, H., Aebersold, R., 2008. PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. EMBO Reports 9 (5), 429–434.

Deutsch, E.W., Mendoza, L., Shteynberg, D., *et al.*, 2015. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. PROTEOMICS – Clinical Applications 9 (7–8), 745–754.

Farrah, T., Deutsch, E.W., Kreisberg, R., *et al.*, 2012. PASSEL: The PeptideAtlas SRM experiment library. Proteomics 12 (8), 1170–1175.

Griss, J., Jones, A.R., Sachsenberg, T., *et al.*, 2014. The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Molecular & Cellular Proteomics 13 (10), 2765–2775.

HUPO-PSI, 2006. The mzData standard. Available at: http://www.psidev.info/mass-spectrometry#mzdata (accessed 27.03.17).

Jones, A.R., Eisenacher, M., Mayer, G., *et al.*, 2012. The mzIdentML data standard for mass spectrometry-based proteomics results. Molecular & Cellular Proteomics 11 (7), (M111.014381-M111.014381).

Keller, A., Eng, J., Zhang, N., Li, X.K., Aebersold, R., 2005. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Molecular Systems Biology 1 (1), E1–E8.

Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., 2008. ProteoWizard: Open source software for rapid proteomics tools development. Bioinformatics 24 (21), 2534–2536.

Martens, L., Chambers, M., Sturm, M., *et al.*, 2011. mzML – A community standard for mass spectrometry data. Molecular & Cellular Proteomics 10 (1), doi:10.1074/mcp.R110.000133.

Martens, L., Hermjakob, H., Jones, P., *et al.*, 2005a. PRIDE: The proteomics identifications database. Proteomics 5 (13), 3537–3545.

Martens, L., Nesvizhskii, A.I., Hermjakob, H., *et al.*, 2005b. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. Proteomics 5 (13), 3501–3505.

Nesvizhskii, A.I., 2005. Interpretation of shotgun proteomic data: The protein inference problem. Molecular & Cellular Proteomics 4 (10), 1419–1440.

Nesvizhskii, A.I., 2006. Protein identification by tandem mass spectrometry and sequence database searching. Mass Spectrometry Data Analysis in Proteomics. New Jersey, NJ: Humana Press, pp. 87–120.

Okuda, S., Watanabe, Y., Moriya, Y., *et al.*, 2017. jPOSTrepo: An international standard data repository for proteomes. Nucleic Acids Research 45 (D1), D1107–D1111.

Ong, S.-E., Blagoev, B., Kratchmarova, I., *et al.*, 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & Cellular Proteomics 1 (5), 376–386.

Pedrioli, P.G.A., Eng, J.K., Hubley, R., *et al.*, 2004. A common open representation of mass spectrometry data and its application to proteomics research. Nature Biotechnology 22 (11), 1459–1466.

Perez-Riverol, Y., Bai, M., Leprevost, F., Squizzato, S., *et al.*, 2016. Omics Discovery Index–Discovering and Linking Public Omics Datasets.

Picotti, P., Lam, H., Campbell, D., *et al.*, 2008. A database of mass spectrometric assays for the yeast proteome. Nature Methods 5 (11), 913–914.

Prince, J.T., Carlson, M.W., Wang, R., Lu, P., Marcotte, E.M., 2004. The need for a public proteomics repository. Nature Biotechnology 22 (4), 471–472.

Reiter, L., Rinner, O., Picotti, P., *et al.*, 2011. mProphet: Automated data processing and statistical validation for large-scale SRM experiments. Nature Methods 8 (5), 430–435.

Science Signaling, 2013. ST NetWatch: Protein Databases. Available at: http://stke.sciencemag.org/resources/st-netwatch-archive/st-netwatch-protein-databases (accessed 23.03.17).

Slotta, D.J., Barrett, T., Edgar, R., 2009. NCBI peptidome: A new public repository for mass spectrometry peptide identifications. Nature Biotechnology 27 (7), 600–601.

Smith, B.E., Hill, J.A., Gjukich, M.A., Andrews, P.C., 2011. Tranche distributed repository and ProteomeCommons.org. In: Hamacher, M, Eisenacher, M, Stephan, C (Eds.), Data Mining in Proteomics. Totowa, NJ: Humana Press, pp. 123–145.

The UniProt Consortium, 2017. UniProt: The universal protein knowledgebase. Nucleic Acids Research 45 (D1), D158–D169.

Verheggen, K., Martens, L., 2015. Ten years of public proteomics data: How things have evolved, and where the next ten years should lead us. EuPA Open Proteomics 8, 28–35.

Vizcaíno, J.A., Csordas, A., del-Toro, N., *et al.*, 2016. 2016 update of the PRIDE database and its related tools. Nucleic Acids Research 44 (D1), D447–D456.

Vizcaíno, J.A., Deutsch, E.W., Wang, R., *et al.*, 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nature Biotechnology 32 (3), 223–226.

Walzer, M., Qi, D., Mayer, G., *et al.*, 2013. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. Molecular & Cellular Proteomics 12 (8), 2332–2340.

Wilhelm, M., Kirchner, M., Steen, J.A.J., Steen, H., 2012. mz5: Space- and time-efficient storage of mass spectrometry data sets. Molecular & Cellular Proteomics 11 (1), [O111.011379-O111.011379].

Further Reading

Deutsch, E.W., 2012. File formats commonly used in mass spectrometry proteomics. Molecular & Cellular Proteomics 11 (12), 1612–1621.

Martens, L., 2011. Proteomics databases and repositories. In: Wu, C.H., Chen, C (Eds.), Bioinformatics for Comparative Proteomics. Totowa, NJ: Humana Press, pp. 213–227.

Martens, L., Vizcaíno, J.A., 2017. A golden age for working with public proteomics data. Trends in Biochemical Sciences 42 (5), 333–341.
Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D., *et al.*, 2013. The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. Database 2013 (0), bat009.
Perez-Riverol, Y., Alpi, E., Wang, R., *et al.*, 2015. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. Proteomics 15 (5–6), 930–950.
Riffle, M., Eng, J.K., 2009. Proteomics data repositories. Proteomics 9 (20), 4653–4663.
Taylor, C.F., Paton, N.W., Lilley, K.S., *et al.*, 2007. The minimum information about a proteomics experiment (MIAPE). Nature Biotechnology 25 (8), 887–893.
Vaudel, M., Verheggen, K., Csordas, A., *et al.*, 2016. Exploring the potential of public proteomics data. Proteomics 16 (2), 214–225.
Verheggen, K., Martens, L., 2015. Ten years of public proteomics data: How things have evolved, and where the next ten years should lead us. EuPA Open Proteomics 8, 28–35.
Vizcaíno, J.A., Foster, J.M., Martens, L., 2010. Proteomics data repositories: Providing a safe haven for your data and acting as a springboard for further research. Journal of Proteomics 73 (11), 2136–2146.

A.2. Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update

This article published in the *Journal of Proteome Research* presents an extensive analysis aiming to detect missing proteins in the human sperm proteome, as part of the Chromosome-Centric Human Proteome Project (C-HPP). In the scope of this work, I performed some supplementary analyses using the MzMod open modification search engine [144] to identify PTMs in the human sperm proteome that were overlooked in the original search. As in most studies based on the database search approach, only the oxidation of methionines and acetylation of protein N-termini were selected as variable modifications. Thus, the rationale was that the peptides of some missing proteins could potentially be missed if they were not searched in all their modified forms. The new analysis found out that many peptides had deaminated asparagine (Asn) residues (and less frequently glutamine (Gln) residues), but no additional missing protein could be identified with this strategy.

Journal of proteome research

Article

pubs.acs.org/jpr

Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update

Yves Vandenbrouck,[*,†,‡,§,▲] Lydie Lane,[∥,⊥,▲] Christine Carapito,[#] Paula Duek,[⊥] Karine Rondel,[▽] Christophe Bruley,[†,‡,§] Charlotte Macron,[#] Anne Gonzalez de Peredo,[○] Yohann Couté,[†,‡,§] Karima Chaoui,[○] Emmanuelle Com,[▽] Alain Gateau,[⊥] Anne-Marie Hesse,[†,‡,§] Marlene Marcellin,[○] Loren Méar,[▽] Emmanuelle Mouton-Barbosa,[○] Thibault Robin,[◆] Odile Burlet-Schiltz,[○] Sarah Cianferani,[#] Myriam Ferro,[†,‡,§] Thomas Fréour,[¶,+] Cecilia Lindskog,[†,‡] Jérôme Garin,[†,‡,§] and Charles Pineau[*,▽]

[†]CEA, DRF, BIG, Laboratoire de Biologie à Grande Echelle, 17 rue des martyrs, Grenoble F-38054, France

[‡]Inserm U1038, 17, rue des Martyrs, Grenoble F-38054, France

[§]Université de Grenoble, Grenoble F-38054, France

[∥]Department of Human Protein Sciences, Faculty of Medicine, University of Geneva, 1, rue Michel-Servet, 1211 Geneva 4, Switzerland

[⊥]CALIPHO Group, SIB-Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland

[#]Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), IPHC, Université de Strasbourg, CNRS UMR7178, 25 Rue Becquerel, 67087 Strasbourg, France

[▽]Protim, Inserm U1085, Irset, Campus de Beaulieu, Rennes 35042, France

[○]Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse, CNRS, UPS, 31062 Toulouse, France

[◆]Proteome Informatics Group, Centre Universitaire d'Informatique, Route de Drize 7, 1227 Carouge, CH, Switzerland

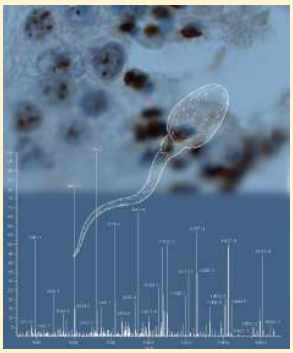
[¶]Service de Médecine de la Reproduction, CHU de Nantes, 38 boulevard Jean Monnet, 44093 Nantes cedex, France

[+]INSERM UMR1064, Nantes 44093, France

[□]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala 751 85, Sweden

Supporting Information

ABSTRACT: The Chromosome-Centric Human Proteome Project (C-HPP) aims to identify "missing" proteins in the neXtProt knowledgebase. We present an in-depth proteomics analysis of the human sperm proteome to identify testis-enriched missing proteins. Using protein extraction procedures and LC−MS/MS analysis, we detected 235 proteins (PE2−PE4) for which no previous evidence of protein expression was annotated. Through LC−MS/MS and LC−PRM analysis, data mining, and immunohistochemistry, we confirmed the expression of 206 missing proteins (PE2−PE4) in line with current HPP guidelines (version 2.0). Parallel reaction monitoring acquisition and sythetic heavy labeled peptides targeted 36 ≪one-hit wonder≫ candidates selected based on prior peptide spectrum match assessment. 24 were validated with additional predicted and specifically targeted peptides. Evidence was found for 16 more missing proteins using immunohistochemistry on human testis sections. The expression pattern for some of these proteins was specific to the testis, and they could possibly be valuable markers with fertility assessment applications. Strong evidence was also found of four "uncertain" proteins (PE5); their status should be re-examined. We show how using a range of sample preparation techniques combined with MS-based analysis, expert knowledge, and complementary antibody-based techniques can produce data of interest to the community. All MS/MS data are available via ProteomeXchange under identifier PXD003947. In addition to contributing to the C-HPP, we hope these data will stimulate continued exploration of the sperm proteome.



KEYWORDS: *human proteome project, spermatozoon, missing proteins, mass spectrometry proteomics, immunohistochemistry, bioinformatics, data mining, cilia*

INTRODUCTION

The Chromosome-Centric Human Proteome Project (C-HPP) aims to catalogue the protein gene products encoded by the human genome, in a gene-centric manner.[1] As part of this

Special Issue: Chromosome-Centric Human Proteome Project 2016

Received: May 3, 2016
Published: July 22, 2016

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

project, neXProt[2] has been confirmed as the reference knowledgebase for human protein annotation.[3] Numerous initiatives were launched worldwide to search for so-called missing proteins - proteins predicted by genomic or transcriptomic analysis but not yet validated experimentally by mass-spectrometry or antibody-based techniques. These proteins are annotated with a "Protein Existence" (PE) score of 2 when they are predicted by transcriptomics analysis, 3 when they are predicted by genomic analysis and have homologues in distant species, and 4 when they are only predicted by genomic analysis in human or other mammals. The most recent neXtProt release (2016-01-11) contains 2949 such missing proteins. It was suggested by Lane and collaborators[4] that proteins that have been systematically missed might be expressed only in a few organs or cell types. The very high number of testis-specific genes that have been described[5] supports the hypothesis that the testis is a promising organ in which to search for elements of the missing proteome.[6,7] The testis' main function is well known to produce male gametes, known as spermatozoa (commonly called sperm). Human spermatozoa are produced at a rate of ~1000 cells/s[8] by a complex, intricate, tightly controlled and specialized process known as spermatogenesis.[9,10] Spermiogenesis is the final stage of spermatogenesis, which sees the maturation of spermatids into mature, motile spermatozoa. The fact that the number of couples consulting for difficulties related to conceiving has increased in recent years and that sperm quality has been shown to be altered in one in seven men, for example, with abnormal motility or morphology,[11] makes further study of these cells even more topically relevant.

Large numbers of spermatozoa can be recovered in highly pure preparation through noninvasive procedures, making it possible to access the final proteome of the germ cell lineage and providing access to a large number of germ cell-specific proteins. Thus, MS-based proteomics studies of spermatozoa have generated highly relevant data.[12] Knowledge of the mature sperm proteome will significantly contribute to sperm biology and help us to better understand fertility issues.

In a recent study,[13] the Proteomics French Infrastructure (ProFI; www.profiproteomics.fr) described a step-by-step strategy combining bioinformatics and MS-based experiments to identify and validate missing proteins based on database search results from a compendium of MS/MS data sets. The data sets used were generated using 40 human cell line/tissue type/body fluid samples. In addition to the peptide- and protein-level false discovery rate (FDR), supplementary MS-based criteria were used for validation, such as peptide spectrum match (PSM) quality as assessed by an expert eye, spectral dot-product - calculated based on the fragment intensities of the native spectrum (endogenous peptide) and a reference spectrum (synthetic peptide) - and LC-SRM assays that were specifically developed to target proteotypic peptides.

Some of these criteria were also used in a concomitant study[14] involving trans-chromosome-based data analysis on a high-quality mass spectrometry data set to catalogue missing proteins in total protein extracts from isolated human spermatozoa. This analysis validated 89 missing proteins based on version 1.0 of the HPP guidelines (http://www.thehpp.org/guidelines/). The distribution of two interesting candidates (C2orf57 and TEX37) was further studied by immunohistochemistry in the adult testis, and their expression was confirmed in postmeiotic germ cells. Finally, on the basis of analyses of transcript abundance during human spermatogenesis, we concluded that it would be possible to characterize additional missing proteins in ejaculated spermatozoa.

The study presented in this paper originated with the Franco–Swiss contribution to the C-HPP initiative to map chromosomes 14 (France) and 2 (Switzerland) by identifying additional missing proteins. Here we combine the search for proteins that are currently classed as "missing" with an extensive examination of the sperm proteome. A single pool of human spermatozoa was treated by a range of approaches, and the most recent version of the guidelines for the identification of missing proteins was followed (Deutsch et al., submitted; http://www.thehpp.org/guidelines/). We thus performed an in-depth analysis of human sperm using different fractionation/separation protocols along with different protein extraction procedures. Through MS/MS analysis, 4727 distinct protein groups were identified that passed the 1% PSM-, peptide-, and protein-level FDR thresholds. Mapping of unique peptides against the most recent neXtProt release (2016-01-11) revealed 235 proteins (201 PE2, 22 PE3, 12 PE4) that are still considered missing by the C-HPP and 9 proteins annotated with a PE5 (uncertain) status in neXtProt. Additional MS-based strategies (spectral comparison and parallel reaction monitoring (PRM) assays) were applied to validate some of these missing proteins. Data mining was also applied to determine which proteins would be selected for validation by immunohistochemistry on human testes sections.

MATERIALS AND METHODS

Ethics and Donor Consent

The study protocol "*Study of Normal and Pathological Human Spermatogenesis*" was approved by the local ethics committee. The protocol was then registered as No. PFS09-015 at the French Biomedicine Agency. Informed consent was obtained from donors where appropriate.

Sample Collection and Preparation

Human semen samples were collected from five healthy donors of unproven fertility at Nantes University Hospital (France). The donors gave informed consent for the use of their semen for research purposes, and samples were anonymized. Semen samples were all obtained on-site by masturbation following 2 to 7 days of sexual abstinence. After 30 min of liquefaction at room temperature under gentle agitation, 1 mL of each sample was taken. Aliquots were pooled and a protease inhibitor mix (protease inhibitor cocktail tablets, complete mini EDTA-free, Roche, Meylan, France) was added according to the manufacturer's instructions. To separate sperm cells from seminal plasma and round cells, we loaded the pooled sperm sample onto 1 mL of a 50% suspension of silica particles (SupraSperm, Origio, Malov, Denmark) diluted in Sperm Washing medium (Origio, Malov, Denmark). The sample was centrifuged at 400*g* for 15 min at room temperature. The sperm pellet was then washed once by resuspension in 3 mL of phosphate-buffered saline (PBS) and centrifuged again at 400*g* for 5 min at room temperature. The supernatant was removed, and the cell pellet was flash-frozen in liquid nitrogen.

Protein Extraction, Digestion, and Liquid Chromatography–Tandem Mass Spectrometry (LC–MS/MS) Analyses

MS/MS analysis of pooled sperm was performed using four different protocols based on a range of protein extraction procedures: (i) total cell lysate followed by a 1D SDS-PAGE

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998–4019

separation (23 gel slices); (ii) separation of Triton X-100 soluble and insoluble fractions followed by a 1D SDS-PAGE separation (20 gel slices per fraction); (iii) total cell lysate, in-gel digestion, and peptides analyzed by nano-LC with long gradient runs; and (iv) total cell lysate, in-gel digestion, and peptides fractionated by high-pH reversed-phase (Hp-RP) chromatography. For all protocols, tryptic peptides were analyzed by high-resolution MS instruments (Q-Exactive). These experiments were performed by the three proteomics platforms making up ProFI (Grenoble, Strasbourg, and Toulouse). A detailed description of the protein fractionation using Triton X-100, protein extraction and digestion, and liquid chromatography-tandem mass spectrometry (LC−MS/MS) analyses performed in this study can be found in the Supporting Information.

MS/MS Data Analysis

Peak lists were generated from the original LC−MS/MS raw data using the Mascot Distiller tool (version 2.5.1, Matrix Science). The Mascot search engine (version 2.5.1, Matrix Science) was used to search all MS/MS spectra against a database composed of *Homo sapiens* protein entries from UniProtKB/SwissProt (release 2015-10-30, 84 362 protein coding genes sequences (canonical and isoforms)) and a list of contaminants frequently observed in proteomics analyses (the protein fasta file for these contaminants is available at ftp://ftp.thegpm.org/fasta/cRAP; it consists of 118 sequences). The following search parameters were applied: carbamidomethylation of cysteines was set as a fixed modification and oxidation of methionines and protein N-terminal acetylation were set as variable modifications. Specificity of trypsin digestion was set for cleavage after K or R, and one missed trypsin cleavage site was allowed. The mass tolerances for protein identification on MS and MS/MS peaks were 5 ppm and 25 mmu, respectively. The FDR was calculated by performing the search in concatenated target and decoy databases in Mascot. Peptides identified were validated by applying the target-decoy approach, using Proline software (http://proline.profiproteomics.fr/), by adjusting the FDR to 1%, at PSM and protein levels. At peptide level, only the PSM with the best Mascot score was retained for each peptide sequence. Spectra identifying peptides in both target and decoy database searches were first assembled to allow competition between target and decoy peptides for each MS/MS query. Finally, the total number of validated hits was computed as Ntarget+Ndecoy, the number of false-positive hits was estimated as 2×Ndecoy, and the FDR was then computed as 2×Ndecoy/(Ntarget + Ndecoy). Proline software automatically determined a threshold Mascot *e*-value to filter peptides and computed the FDR as described so as to automatically adjust it to 1%. At protein level, a composite score was computed for each protein group based on the MudPIT scoring method implemented in Mascot: For each nonduplicate peptide identifying a protein group, the difference between its Mascot score and its homology threshold was computed, and these "score offsets" were then summed before adding them to the average homology (or identity) thresholds for the peptide. Therefore, less significant peptide matches contributed less to the total protein score. Protein groups were filtered by applying a threshold to this MudPIT protein score to obtain a final protein-level FDR of 1%. To optimize discrimination between true-positive and true-negative protein hits, the software applies a selection scheme approach by adjusting the FDR separately for the subset of proteins identified by more than one validated peptide and then for the single-peptide hits. In accordance with version 2.0.1 of the HPP data interpretation guidelines (Deutsch et al., submitted; http://www.thehpp.org/guidelines/), individual result files from each of the five MS/MS data sets were combined, and a procedure to produce a protein-level FDR threshold of 1% was reapplied. This combination of result files created a single identification data set from a set of identification results and was performed as follows: All PSMs identified and validated at 1% were merged to create a unique combination of amino acid sequences and a list of PTMs located on that sequence that were aggregated in a single "representative" PSM. The newly created PSMs were then grouped into proteins and protein families.[41] The resulting data set therefore provides a nonredundant view of the identified proteins present in the original sample.

Detection of Missing Proteins

The sequence of each peptide identified was searched in all splicing isoform sequences present in neXtProt release 2016-01-11 using the pepx program developed in-house (https://github.com/calipho-sib/pepx). The method is based on a 6-mer amino acid index that is regenerated at each release; the 6 aa length was chosen because it significantly speeds up the mapping process. Leucine and isoleucine were considered equivalent. A peptide is considered to match an isoform sequence when all the 6-mers covering the peptide return the same sequence. Peptides were subsequently checked against the retrieved isoform sequence(s) to ensure an exact string match. All matches to splicing isoforms derived from a single entry were considered relevant for the identification of the entry.

To further validate the identification of missing proteins, we performed a second round of peptide-to-protein mapping, taking into account the 2.5 million variants described in neXtProt (SNPs and disease mutations). Currently, pepx only considers a single amino acid substitution or deletion in the 6-mer; substitutions and deletions more than 1 aa in length, as well as insertions, are not taken into account. Consequently, pepx returns a match if single amino acid variations in the isoform sequence are spaced at least five amino acids apart. Peptides matching more than one entry when variants were taken into account were excluded as they are potentially not proteotypic.

Data Availability

All MS proteomics data, including reference files (readme, search database, .dat files), form a complete submission with the ProteomeXchange Consortium.[15] Data were submitted via the PRIDE partner repository under data set identifiers PXD003947 and 10.6019/PXD003947.

Additional MS-Based Validation (MS/MS Analysis of Synthetic Peptides, Comparison of Reference/Endogenous Fragmentation Spectra and LC−PRM Analysis)

Synthetic heavy labeled peptides were purchased (crude PEPotec, Thermo Fisher Scientific) for 36 "one-hit wonder" candidates selected based on visual inspection of PSMs. The 36 peptides initially identified were synthesized along with two additional predicted proteotypic peptides per protein when possible. Thus, a total of 100 peptides were synthesized (Supplementary Table 4). The labeled peptides corresponding to the 36 peptides initially identified were mixed together and analyzed by LC−MS/MS (Q Exactive Plus, Thermo Fisher Scientific) to acquire higher energy collisional dissociation

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

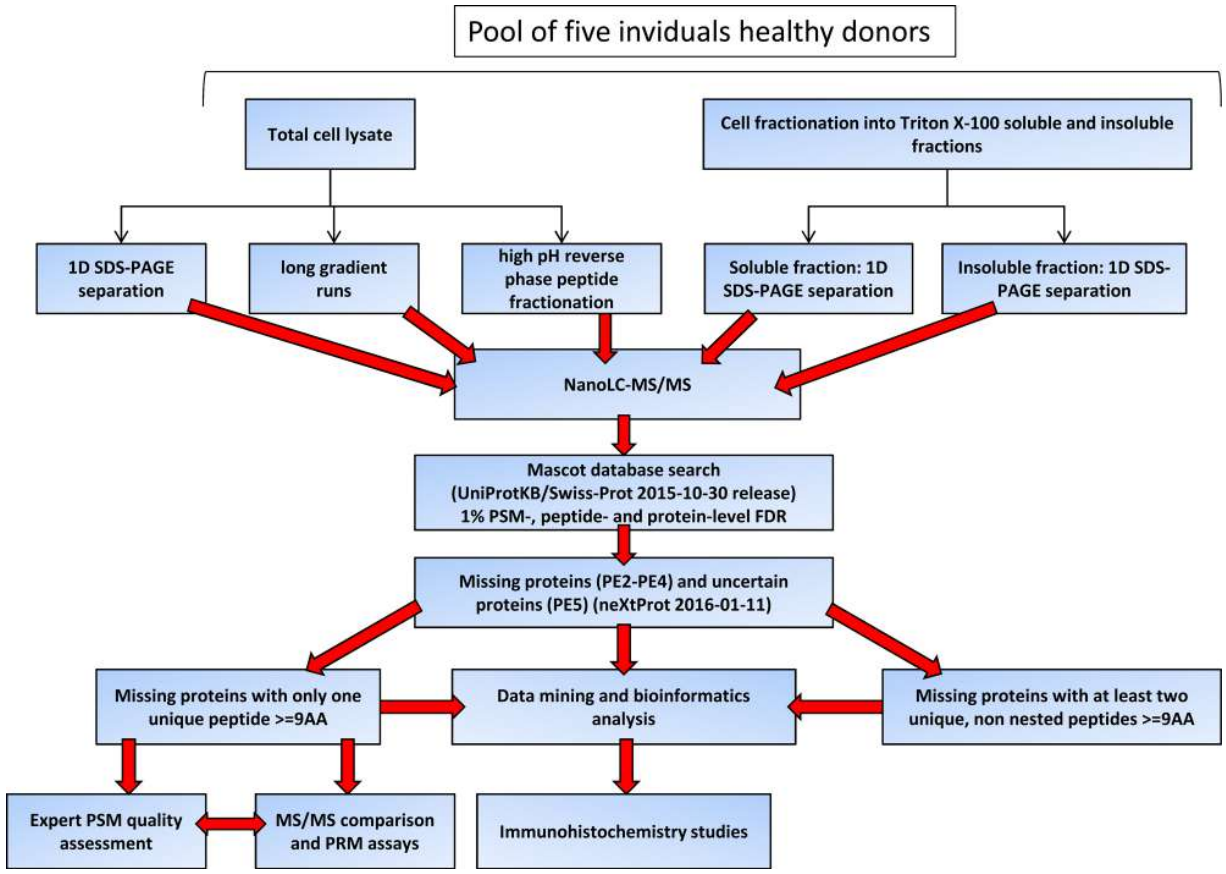


Figure 1. Flowchart illustrating the strategies used to identify and validate missing proteins detected in the human sperm proteome.

(HCD) fragmentation spectra for comparison with the initial spectra in the closest possible conditions. All MS/MS spectrum pairs are shown in Supplementary Figure 1. Following this step, targeted assays using a PRM acquisition approach were developed on the same LC–MS/MS platform to target all 100 peptides, first in a total protein fraction prepared in stacking gel bands and subsequently in gel bands obtained from 1D SDS-PAGE separation of the Triton X-100 insoluble proteins fraction. See the Supporting Information for details of MS experiments.

Data Mining to Select Missing Proteins for Further Characterization

For each protein identified by MS, the tissue expression profile based on RNA sequencing analysis was retrieved from the Human Protein Atlas (HPA) portal (version 14) (www.proteinatlas.org/). The evolutionary conservation profile was determined by a BLAST analysis using UniProtKB "Reference Proteomes" as target. In addition, homologues were systematically searched for in a number of ciliated organisms from distant groups including *Choanoflagellida* (*Salpingoeca, Monosiga*), Chlorophyta (*Micromonas, Volvox, Chlamydomonas*), Ciliophora (*Paramecium, Oxytricha, Stylonychia, Tetrahymena, Ichthyophthirius*), Trypanosomatidae (*Trypanosoma, Phytomonas, Leishmania, Angomonas, Leptomonas*), Cryptophyta (*Guillardia*), *Naegleria gruberi*, and Flagellated protozoan (*Bodo saltans*). For each protein and all its orthologs, all existing names, synonyms, and identifiers were collected from appropriate model organism databases. These names were used to query PubMed and Google. Proteins to be further validated by immunohistochemistry were selected based on a combination of criteria including antibody quality, available immunohistochemistry data in Protein Atlas (version 14), phenotype of mutant organisms, predicted or experimental biological function, tissue localization, interacting partners, and phylogenetic profile. Uncharacterized proteins that are selectively expressed in testis or ciliated tissues and well-conserved in ciliated organisms interact with testis or cilia-related proteins, for which knockout model organisms show a reproduction phenotype and for which high-quality antibodies from the HPA were available were considered the best candidates for further validation.

Immunohistochemistry

To confirm the germline expression of proteins of interest, we performed immunohistochemistry experiments on human testes fixed in 4% paraformaldehyde and embedded in paraffin, as described.[16] Normal human testes were collected at autopsy at Rennes University Hospital from HIV-1-negative cadavers.

Paraffin-embedded tissues were cut into 4 μm thick slices, mounted on slides, and dried at 58 °C for 60 min. Immunohistochemical staining, using the Ventana DABMap and OMNIMap detection kit (Ventana Medical Systems, Tucson, USA), was performed on a Discovery Automated IHC stainer. Antigen retrieval was performed using proprietary Ventana Tris-based buffer solution, CC1, at 95 to 100 °C for 48 min. Tissue sections were then saturated for 1 h with 5% BSA in TBS, and endogenous peroxidase was blocked with Inhibitor-D, 3% H_2O_2, (Ventana) for 8 min at 37 °C. After rinsing in TBS, slides were incubated at 37 °C for 60 min with

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998–4019

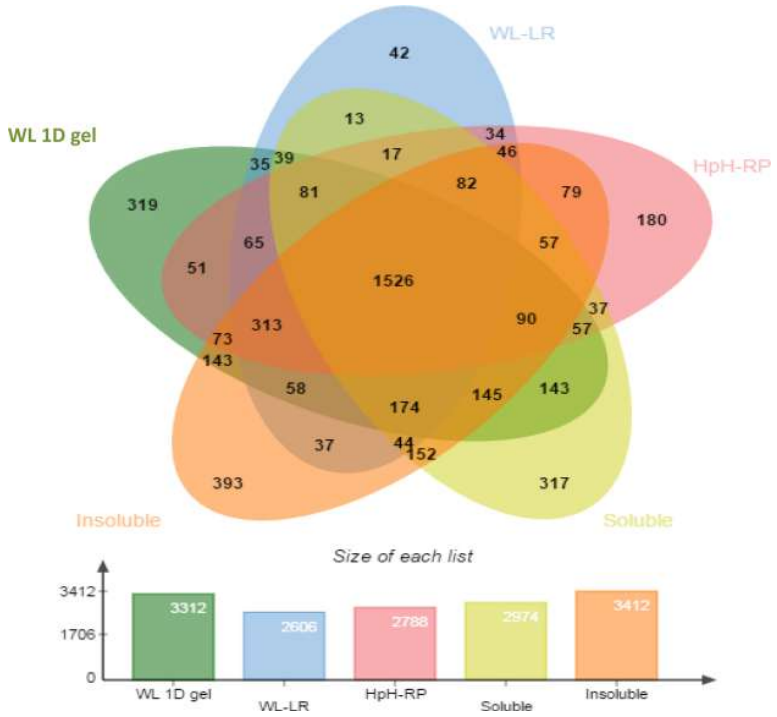


Figure 2. Contribution of the different fractionation protocols to identification of spermatozoa proteins. Upper part: Venn diagram created with the jvenn web application[39] illustrating overlap between the five fractionation protocols. WL 1D gel: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), WL LR: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs, HpH-RP: total cell lysate, in-gel digestion of proteins, and peptide fractionation by high-pH reversed-phase (HpH-RP) chromatography, Soluble and Insoluble: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by a 1D SDS-PAGE separation of proteins (20 gel slices per fraction). Lower part: bar chart representing the total number of proteins identified in each MS/MS data set.

polyclonal rabbit antibodies specific for the selected missing proteins (Atlas Antibodies) diluted in TBS containing 0.2% Tween 20 (v/v) and 3% BSA (TBST-BSA). The antibody dilutions used are listed in Supplementary Table 6. Non-immune rabbit serum (1:1000) was used as a negative control. After several washes in TBS, sections were incubated for 16 min with a biotinylated goat antirabbit antibody (Roche) at a final dilution of 1:500 in TBST-BSA. Signal was enhanced using the Ventana DABMap Kit or Ventana OMNIMap kit. Sections were then counterstained for 16 min with hematoxylin (commercial solution, Microm) and for 4 min with bluing reagent (commercial solution, Microm) before rinsing with Milli-Q water. After removal from the instrument, slides were manually dehydrated and mounted in Eukitt (Labnord, Villeneuve d'Ascq, France). Finally, immunohistology images were obtained using NDP.Scan acquisition software (v2.5, Hamamatsu) and visualized with NDP.View2 software (Hamamatsu). Representative images are shown.

RESULTS AND DISCUSSION

Overall Workflow

The overall workflow for the detection and validation of missing proteins is illustrated in Figure 1 and described in the Material and Methods, with full details of sample preparation in the Supporting Information. By applying this workflow, we produced a list of 235 "candidate missing protein" entries (PE2−4) and 9 PE5 entries. This list was divided into two distinct subsets in line with version 2.0.1 of the HPP data interpretation guidelines (Deutsch et al., submitted; http://www.thehpp.org/guidelines/): those validated by two or more distinct uniquely mapping peptide sequences of length ≥9 amino acids and those detected based on only one unique peptide of length ≥9 amino acids. Each PSM from the latter subset was then examined to seek additional MS-based evidence (PSM quality as assessed by an expert, comparison between endogenous and reference (synthetic peptide) fragmentation spectra and LC−PRM assays). In parallel, the full list of missing or uncertain protein entries (PE2−5) was mined by gathering additional information from public resources, bioinformatics analysis, and the literature. This information was used to select a subset of high-priority proteins for further immunohistochemistry analysis on human testes sections.

Analysis of the Human Sperm Proteome

Because the workflow involved a range of enrichment strategies and separation protocols, including peptide prefractionation protocols based on high pH reverse phase (HpH-RP) chromatography that have been shown to be orthogonal to subsequent online reverse-phase nano-LC separation of peptides,[17] sensitivity was high and coverage extensive. This type of "cover all bases" approach has been shown to be particularly efficient for improving the detection of missing proteins.[18] Validation was subsequently performed for each results file (.dat) through the target-decoy approach,[19] using the in-house developed Proline software (http://proline.profiproteomics.fr/), by adjusting the FDR to 1%, at PSM- and protein-level. In a second step, individual results files were combined for each data set, and a 1% protein-level FDR was applied to comply with the HPP data interpretation guidelines, version 2.0.1 (http://www.thehpp.org/guidelines/, guideline

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 1. Description of Missing Proteins (PE2−PE4) Detected in This Study

total number of missing proteins (PE2−PE4)	missing proteins with at least two unique, non-nested peptides ≥9AA	missing proteins with only one unique peptide ≥9AA	missing proteins with no annotated function in Uniprot	missing proteins with at least one transmembrane domain
235	188	47	180	56

#9) (see Materials and Methods). Supplementary Table 1 lists PSM-, peptide-, and protein-level FDR values along with the total number of true-positives and false-positives at each level for the five sperm sample preparation methods (individual results files for each fraction and after combination). After MS/MS data processing and filtering, a total of 4727 distinct protein groups passed the 1% PSM-, peptide-, and protein-level criteria. Detailed information on the proteins identified from the five MS data sets is reported in Supplementary Table 2. Protein identification and their distribution across the five MS data sets were then compared to assess their contribution to total human sperm proteome data sets (Figure 2). The Venn diagram shows that 1526 proteins detected were present in all five data sets. In addition, each fractionation/separation method used in this study provided a significant added-value in terms of proteome coverage. Thus, Triton X-100 insoluble and soluble fractions, whole cell lysate analyzed by 1D SDS-PAGE, high-pH reverse phase peptide fractionation and long gradient runs allowed gains of 8.3, 6.7, 6.7, 3.8 and 0.8%, respectively. These results clearly emphasize the complementarity of the different enrichment techniques when seeking to obtain exhaustive (or as exhaustive as possible) proteome coverage.

In 2014, Amaral et al.[20] published a sperm proteome comprising 6198 proteins identified by combined MS-based analysis based on 30 LC−MS/MS proteomics studies. Crossing identifier lists between their data and the sperm proteome produced here revealed that our analysis yielded 1140 additional proteins. However, further investigation will be necessary to ensure a fair assessment as Amaral et al. applied different validation criteria to ours (e.g., identification of at least two peptides with a protein-level FDR < 5%).

Focusing on Missing Proteins Identified from the Sperm Proteome

Missing proteins were detected as described in Materials and Methods using the most recent neXtProt release (2016-01-11). The first step in this detection took all possible splice isoforms and I/L ambiguities into account but no single amino-acid variants. This produced a list of 235 missing proteins (PE2−4) and 9 uncertain proteins (PE5). Among the PE2−4 protein entries, 188 were identified and validated (<1% FDR) by at least two or more distinct uniquely mapping peptide sequences of length ≥9 amino acids, while the remaining 47 were associated with only one unique peptide ≥9 amino acids (Table 1). We named this subset of missing proteins "one-hit wonders" and considered it separately for further MS-based analysis (see the next section). In fact, five of these protein entries were identified with two (A2RUU4, Q5GH77, Q8NG35, Q8WTQ4) or three (Q6PI97) unique peptides, but because only one of these had a length equal to or greater than nine amino acids (Table 2), these entries were nevertheless considered to be "one-hit wonders". Among the full set of 235 proteins considered to be missing by neXtProt (PE2−4), 180 have no annotated function, while 56 are predicted to have at least one transmembrane helix (TMH). Full details (description, number of unique peptides, chromosome location, etc.) on the missing proteins are reported in Table 2 and Supplementary Table 3. The Venn diagram illustrated in Figure 3A shows how each fractionation protocol contributed to the identification of the whole set of missing proteins (PE5 included). Thirty-three proteins were detected in all five data sets, whereas 63 were specifically detected in a given data set (30 in "insoluble fraction-1D gel", 9 in "soluble fraction-1D gel" gel, 17 in "whole lysate-1D gel", 6 in "peptide HpH-RP", and 1 in "whole lysate-long runs").

We noticed that among the 30 proteins only detected in the insoluble fraction, around one-third (9 proteins) were annotated with at least one TMH (see Supplementary Table 3), illustrating the benefit of preliminary subcellular fractionation for the identification of hydrophobic proteins. Unsurprisingly, only a small number of proteins (8) with at least one TMH were detected in all five sperm data sets, and an even smaller number of them were specifically detected when protocols starting with the soluble fraction or a whole cell lysate were applied (5 in "soluble fraction-1D gel", 3 in "whole lysate-1D gel", 2 in "peptide HpH-RP", and none in "whole lysate-long runs"). The missing proteins identified in sperm were found to be distributed across all chromosomes, except chromosome Y and chromosome 21, with the highest number (21 proteins) coded by genes present on chromosome 1 (Figure 3B). In terms of coverage of missing proteins (PE5 included), around 80% were supported by two or more distinct uniquely mapping peptide sequences of length ≥9 amino acids, with some proteins very well covered (up to 78 peptides). The other 20% of missing proteins (52 proteins) were identified by only one unique peptide sequence of length ≥9 amino acids (Figure 3C).

To comply with recent C-HPP guidelines (version 2.0.1; http://www.thehpp.org/guidelines/), we also considered alternative mappings of all peptides of length ≥9 amino acids mapping to PE2−5 proteins by taking the 2.5 million single amino acid variants available in neXtProt into account. This analysis indicated that 13 peptides mapping to 12 missing (PE2−4) proteins could correspond to an alternative peptide sequence. Peptide "TKMGLYYSYFK" maps uniquely to DPY19L2P1 (Q6NXN4), but if reported SNPs are considered, it could also map to the PE1 proteins DPY19L2 (Q6NUT2) and DPY19L1 (Q2PZI1). Peptide "TPPYQGDVPLGIR" maps uniquely to the PE2 protein SPAG11A (Q6PDA7), but if reported SNPs are considered, it could also map to the paralog SPAG11B (Q08648), which was also identified in this study with two other unique peptides. Likewise, the PE2 protein LRRC37A (A6NMS7) was identified by two peptides "NAFEENDFMENTNMPEGTISENTNYNHPPEADSAGTAFNLGPTVK" and "SKDLTHAISILESAK". If reported SNPs are considered, these peptides could also map to the PE2 protein LRRC37A2 (A6NM11) and the PE1 protein LRRC37A3 (O60309), respectively. Therefore, DPY19L2P1, SPAG11A, and LRRC37A will need further investigation to validate their existence at protein level. Peptides "QNVQQNEDASQYEESILTK" and "QNVQQNEDATQYEESILTK" were both validated by PRM (see below) but only differ by one residue (S or T) and map to two close paralogs, RSPH10B2 (B2RC85) and RSPH10B (P0C881), respectively. An S71T variant has been

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 2. List of Missing Proteins (PE2−PE4) Identified in the Five MS/MS Datasets[a]

entry	gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
A0AVI2	FER1L5	1	2q11.2	evidence at transcript level	1
A1A4V9	CCDC189 C16orf93	10	16p11.2	evidence at transcript level	0
A1L453	PRSS38 MPN2	2	1q42.13	evidence at transcript level	0
A2RUU4	CLPSL1 C6orf127	2	6p21.31	evidence at transcript level	0
A4D1F6	LRRD1	13	7q21.2	evidence at transcript level	0
A4D256	CDC14C CDC14B2	1	7p12.3	evidence at transcript level	1
A4QMS7	C5orf49	5	5p15.31	evidence at transcript level	0
A5D8W1	CFAP69 C7orf63	25	7q21.13	evidence at transcript level	0
A5PLK6	RGSL1 RGSL RGSL2	5	1q25.3	evidence at transcript level	1
A6H8Z2	FAM221B C9orf128	7	9p13.3	evidence at transcript level	0
A6NCJ1	C19orf71	10	19p13.3	predicted	0
A6NCL2	LRCOL1	2	12q24.33	evidence at transcript level	0
A6NCM1	IQCA1L IQCA1P1	2	7q36.1	inferred from homology	0
A6NCN8		3	12p13.33	predicted	0
A6NE01	FAM186A	6	12q13.12	evidence at transcript level	0
A6NE52	WDR97 KIAA1875	11	8q24.3	evidence at transcript level	0
A6NEN9	CXorf65	4	Xq13.1	predicted	0
A6NF34	ANTXRL	8	10q11.22	inferred from homology	1
A6NFU0	FAM187A	4	17q21.31	inferred from homology	1
A6NFZ4	FAM24A	4	10q26.13	inferred from homology	0
A6NGB0	TMEM191C	1	22q11.21	inferred from homology	1
A6NGY3	C5orf52	2	5q33.3	evidence at transcript level	0
A6NI87	CBY3	3	5q35.3	inferred from homology	0
A6NIV6	LRRIQ4 LRRC64	14	3q26.2	inferred from homology	0
A6NJI9	LRRC72	5	7p21.1	evidence at transcript level	0
A6NLX4	TMEM210	1	9q34.3	inferred from homology	1
A6NM11	LRRC37A2	3	17q21.31	evidence at transcript level	1
A6NMS7	LRRC37A LRRC37A1	2	17q21.31	evidence at transcript level	1
A6NN90	C2orf81	9	2p13.1	inferred from homology	0
A6NNE9	MARCH11	1	5p15.1	evidence at transcript level	2
A6NNW6	ENO4 C10orf134	8	10q25.3	evidence at transcript level	0
A6NNX1	RIIAD1 C1orf230	3	1q21.3	predicted	0
A7E2U8	C4orf47 Chr4_1746	12	4q35.1	evidence at transcript level	0
A8MTL0	IQCF5	3	3p21.2	evidence at transcript level	0
A8MTZ7	C12orf71	2	12p11.23	predicted	0
A8MV24	C17orf98	5	17q12	predicted	0
A8MYZ5	IQCF6	5	3p21.2	inferred from homology	0
A8MZ26	EFCAB9	5	5q35.1	inferred from homology	0
B1AJZ9	FHAD1 KIAA1937	2	1p36.21	evidence at transcript level	0
B1ANS9	WDR64	35	1q43	evidence at transcript level	0
B2RC85	RSPH10B2	1	7p22.1	evidence at transcript level	0
B2RV13	C17orf105	4	17q21.31	evidence at transcript level	0

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 2. continued

entry	gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
B4DYI2	SPATA31C2 FAM75C2	7	9q22.1	evidence at transcript level	1
C9J6K1	C19orf81	1	19q13.33	predicted	0
D6REC4	CFAP99	1	4p16.3	inferred from homology	0
H3BNL8	C6orf229	5	6p22.3	predicted	0
H3BTG2-2	C1orf234	4	1p36.12	evidence at transcript level	0
O95214	LEPROTL1My047 UNQ577/PRO1139	1	8p12	evidence at transcript level	4
O95473	SYNGR4	9	19q13.33	evidence at transcript level	4
P0C221	CCDC175 C14orf38	13	14q23.1	predicted	0
P0C5Z0	H2AFB2; H2AFB3 H2ABBD H2AFB	1	Xq28	evidence at transcript level	0
P0C7A2	FAM153B	1	5q35.2	evidence at transcript level	0
P0C7I6	CCDC159	9	19p13.2	evidence at transcript level	0
P0C7M6	IQCF3	5	3p21.2	evidence at transcript level	0
P0C7X4	FTH1P19 FTHL19	1	Xp21.1	uncertain	0
P0C874	SPATA31D3 FAM75D3	2	9q21.32	uncertain	1
P0C875	FAM228B	6	2p23.3	evidence at transcript level	0
P0C881	RSPH10B	1	7p22.1	evidence at transcript level	0
P0C8F1	PATE4	8	11q24.2	evidence at transcript level	0
P0CW27	CCDC166	6	8q24.3	predicted	0
P0DJG4	THEGL	10	4q12	evidence at transcript level	0
P0DKV0	SPATA31C1 FAM75C1	7	9q22.1	evidence at transcript level	1
P49223	SPINT3	3	20q13.12	inferred from homology	0
Q08648-4		3	8p23.1	evidence at transcript level	
Q0P670	C17orf74	13	17p13.1	evidence at transcript level	1
Q0VAA2	LRRC74A C14orf166B LRRC74	18	14q24.3	evidence at transcript level	0
Q14409	GK3P GKP3 GKTB	4	4q32.3	uncertain	0
Q14507	EDDM3A FAM12A HE3A	3	14q11.2	evidence at transcript level	0
Q17R55	FAM187B TMEM162	6	19q13.12	evidence at transcript level	1
Q2TAA8	TSNAXIP1 TXI1	20	16q22.1	evidence at transcript level	0
Q2WGJ8	TMEM249 C8orfK29	3	8q24.3	evidence at transcript level	2
Q32M84	BTBD16 C10orf87	20	10q26.13	evidence at transcript level	0
Q3KNT9	TMEM95 UNQ9390/PRO34281	1	17p13.1	evidence at transcript level	1
Q3SY17	SLC25A52 MCART2	1	18q12.1	evidence at transcript level	6
Q3ZCV2	LEXM LEM C1orf177	9	1p32.3	evidence at transcript level	0
Q494V2	CCDC37	7	3q21.3	evidence at transcript level	0
Q495T6	MMEL1MELL1MMEL2 NEP2	24	1p36.32	evidence at transcript level	1
Q499Z3	SLFNL1	18	1p34.2	evidence at transcript level	0
Q4G0N8	SLC9C1 SLC9A10	4	3q13.2	evidence at transcript level	16
Q4G1C9	GLIPR1L2	10	12q21.2	evidence at transcript level	1

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 2. continued

entry	gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
Q4ZJI4	SLC9B1 NHEDC1	3	4q24	evidence at transcript level	1
Q502W6	VWA3B	7	2q11.2	evidence at transcript level	0
Q502W7	CCDC38	15	12q23.1	evidence at transcript level	0
Q537H7	SPATA45 C1orf227 HSD-44 HSD44	3	1q32.3	inferred from homology	0
Q53FE4	C4orf17	2	4q23	evidence at transcript level	0
Q53SZ7	PRR30 C2orf53	12	2p23.3	evidence at transcript level	0
Q58FF6	HSP90AB4P	2	15q21.3	uncertain	0
Q5BJE1	CCDC178 C18orf34	3	18q12.1	evidence at transcript level	0
Q5GAN3	RNASE13	6	14q11.2	evidence at transcript level	0
Q5GH77	XKR3 XRG3	2	22q11.1	evidence at transcript level	10
Q5H913	ARL13A	3	Xq22.1	evidence at transcript level	0
Q5H9T9	FSCB C14orf155	8	14q21.2	evidence at transcript level	0
Q5I0G3	MDH1B	14	2q33.3	evidence at transcript level	0
Q5JRC9	FAM47A	3	Xp21.1	evidence at transcript level	0
Q5JU00	TCTE1	10	6p21.1	evidence at transcript level	0
Q5JU67	C9orf117	17	9q34.11	evidence at transcript level	0
Q5JWF8	ACTL10 C20orf134	7	20q11.22	evidence at transcript level	0
Q5SQS8	C10orf120	7	10q26.13	evidence at transcript level	0
Q5SY80	C1orf101	11	1q44	evidence at transcript level	1
Q5T0J7	Tex35 C1orf49	6	1q25.2	evidence at transcript level	0
Q5T1A1	DCST2	2	1q21.3	evidence at transcript level	6
Q5T1B0	AXDND1 C1orf125	19	1q25.2	evidence at transcript level	0
Q5T7R7	C1orf185	4	1p32.3	evidence at transcript level	1
Q5TBE3	C9orf153	2	9q21.33	predicted	0
Q5TEZ5	C6orf163	22	6q15	predicted	0
Q5TFG8	ZC2HC1B C6orf94 FAM164B	2	6q24.2	evidence at transcript level	0
Q5TGP6-2	MROH9 C1orf129	3	1q24.3	evidence at transcript level	0
Q5VTH9	WDR78	21	1p31.3	evidence at transcript level	0
Q5VZ72	IZUMO3 C9orf134	8	9p21.3	inferred from homology	1
Q5VZQ5	TEX36 C10orf122	6	10q26.13	evidence at transcript level	0
Q5XX13	FBXW10	2	17p11.2	evidence at transcript level	0
Q63HN1	FAM205BP C9orf144 C9orf144A FAM205B	1	9p13.3	uncertain	0
Q68DN1	C2orf16	62	2p23.3	evidence at transcript level	0
Q68G75	LEMD1	2	1q32.1	evidence at transcript level	1
Q6ICG8	WBP2NL PAWP	11	22q13.2	evidence at transcript level	0
Q6IPT2-2	FAM71E1	2	19q13.33	evidence at transcript level	0
Q6NXN4	DPY19L2P1	1	7p14.2	evidence at transcript level	3

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 2. continued

entry	gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
Q6NXP6	NOXRED1 C14orf148	2	14q24.3	evidence at transcript level	0
Q6P2C0	WDR93	6	15q26.1	evidence at transcript level	0
Q6P2D8	XRRA1	6	11q13.4	evidence at transcript level	0
Q6PDA7-2	SPAG11A EP2 HE2	1	8p23.1	evidence at transcript level	0
Q6PI97	C11orf88	3	11q23.1	evidence at transcript level	0
Q6PIY5	C1orf228 NCRNA00082	14	1p34.1	evidence at transcript level	0
Q6UW60	PCSK4 PC4 UNQ2757/PRO6496	1	19p13.3	evidence at transcript level	1
Q6UWQ5	LYZL1 LYC2 UNQ648/PRO1278	2	10p11.23	evidence at transcript level	0
Q6UXN7	TOMM20L UNQ9438/PRO34772	1	14q23.1	evidence at transcript level	1
Q6V702	C4orf22	10	4q21.21	evidence at transcript level	0
Q6ZMY6	WDR88 PQWD	4	19q13.11	evidence at transcript level	0
Q6ZNQ3	LRRC69	1	8q21.3	evidence at transcript level	0
Q6ZRH7	CATSPERG C19orf15	15	19q13.2	evidence at transcript level	1
Q6ZUB0	SPATA31D4 FAM75D4	1	9q21.32	uncertain	1
Q6ZUB1	SPATA31E1 C9orf79 FAM75E1	44	9q22.1	evidence at transcript level	1
Q6ZUG5		10	3q21.3	evidence at transcript level	0
Q6ZVS7	FAM183B	4	7p14.1	evidence at transcript level	0
Q7RTY9	PRSS41 TESSP1	1	16p13.3	uncertain	0
Q7Z2V1	C16orf82	2	16p12.1	evidence at transcript level	0
Q7Z4T8	GALNTL5 GALNT15	1	7q36.1	evidence at transcript level	1
Q7Z4W2	LYZL2	2	10p11.23	evidence at transcript level	0
Q7Z5J8	ANKAR	15	2q32.2	evidence at transcript level	1
Q7Z7B7	DEFB132 DEFB32 UNQ827/PRO1754	1	20p13	inferred from homology	0
Q86TZ1-2	TTC6	3	14q21.1	evidence at transcript level	0
Q86UG4	SLCO6A1 OATP6A1 SLC21A19	15	5q21.1	evidence at transcript level	12
Q86VE3	SATL1	2	Xq21.1	evidence at transcript level	0
Q86VS3	IQCH	9	15q23	evidence at transcript level	0
Q86WZ0	HEATR4	3	14q24.3	evidence at transcript level	0
Q86X67	NUDT13	1	10q22.2	evidence at transcript level	0
Q8IUB5	WFDC13 C20orf138 WAP13	2	20q13.12	inferred from homology	0
Q8IVL8	CPO	3	2q33.3	evidence at transcript level	0
Q8IVU9	C10orf107	7	10q21.2	evidence at transcript level	0
Q8IWF9	CCDC83 HSD9	2	11q14.1	evidence at transcript level	0
Q8IXM7	ODF3L1	12	15q24.2	evidence at transcript level	0
Q8IXW0	LMNTD2 C11orf35	10	11p15.5	evidence at transcript level	0
Q8IYJ2	C10orf67	4	10p12.2	evidence at transcript level	0

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 2. continued

entry	gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
Q8IYM0	FAM186B C12orf25	15	12q13.12	evidence at transcript level	0
Q8IYU4	UBQLNL	1	11p15.4	evidence at transcript level	0
Q8IYW2	CFAP46 C10orf123 C10orf124 C10orf92 C10orf93 TTC40	34	10q26.3	evidence at transcript level	0
Q8N0W5	IQCK	9	16p12.3	evidence at transcript level	0
Q8N309	LRRC43	10	12q24.31	evidence at transcript level	0
Q8N456	LRRC18 UNQ9338/PRO34010	7	10q11.23	evidence at transcript level	0
Q8N4B4	FBXO39 FBX39	1	17p13.1	evidence at transcript level	0
Q8N4L4	SPEM1 C17orf83	10	17p13.1	evidence at transcript level	1
Q8N4P6	LRRC71 C1orf92	15	1q23.1	evidence at transcript level	0
Q8N5S1	SLC25A41	6	19p13.3	evidence at transcript level	6
Q8N5S3	C2orf73	4	2p16.2	evidence at transcript level	0
Q8N5U0	C11orf42	3	11p15.4	evidence at transcript level	0
Q8N5W8	FAM24B	1	10q26.13	inferred from homology	0
Q8N688	DEFB123 DEFB23 UNQ1963/PRO4485	1	20q11.21	evidence at transcript level	0
Q8N6G2	TEX26 C13orf26	9	13q12.3	evidence at transcript level	0
Q8N6K0	TEX29 C13orf16	3	13q34	evidence at transcript level	1
Q8N6M8	IQCF1	6	3p21.2	evidence at transcript level	0
Q8N6V4	C10orf53	1	10q11.23	inferred from homology	0
Q8N7B9	EFCAB3	20	17q23.2	evidence at transcript level	0
Q8N7C7	RNF148	1	7q31.32	evidence at transcript level	1
Q8N7X0	ADGB C6orf103 CAPN7L	32	6q24.3	evidence at transcript level	0
Q8N7X2-4	C9orf173	9	9q34.3	evidence at transcript level	0
Q8N801	C2orf61	8	2p21	evidence at transcript level	0
Q8N9W8	FAM71D	1	14q23.3	evidence at transcript level	0
Q8N9Z9	LMNTD1 IFLTD1	1	12p12.1	evidence at transcript level	0
Q8NA56	TTC29	13	4q31.22	evidence at transcript level	0
Q8NA66	CNBD1	1	8q21.3	evidence at transcript level	0
Q8NA69	C19orf45	10	19p13.2	evidence at transcript level	0
Q8NCQ7	PROCA1	5	17q11.2	evidence at transcript level	0
Q8NCU1	LINC00521	1	14q32.12	uncertain	0
Q8ND07	BBOF1 C14orf45 CCDC176	7	14q24.3	evidence at transcript level	0
Q8ND61	C3orf20	8	3p25.1	evidence at transcript level	1
Q8NDH2	CCDC168 C13orf40	9	13q33.1	evidence at transcript level	0
Q8NE28	STKLD1 C9orf96 SGK071	14	9q34.2	evidence at transcript level	0
Q8NEA5	C19orf18	4	19q13.43	evidence at transcript level	1
Q8NEE8	TTC16	7	9q34.11	evidence at transcript level	0

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 2. continued

entry	gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
Q8NEX6	WFDC11 WAP11	1	20q13.12	inferred from homology	0
Q8NG35	DEFB105A BD5 DEFB105 DEFB5; DEFB105B	2	8p23.1	evidence at transcript level	0
Q8NHS2	GOT1L1	6	8p11.23	evidence at transcript level	0
Q8NHU2	CFAP61 C20orf26	33	20p11.23	evidence at transcript level	0
Q8NHX4	SPATA3 TSARG1	5	2q37.1	evidence at transcript level	0
Q8TBY8	PMFBP1	78	16q22.2	evidence at transcript level	0
Q8TBZ9	C7orf62	12	7q21.13	evidence at transcript level	0
Q8TD35	LKAAEAR1 C20orf201	2	20q13.33	evidence at transcript level	0
Q8WTQ4	C16orf78	2	16q12.1	evidence at transcript level	0
Q8WVZ1	ZDHHC19	2	3q29	evidence at transcript level	4
Q8WVZ7	RNF133	3	7q31.32	evidence at transcript level	1
Q8WW18	C17orf50	1	17q12	evidence at transcript level	0
Q8WWF3	SSMEM1 C7orf45	6	7q32.2	evidence at transcript level	1
Q8WXQ8	CPA5	10	7q32.2	evidence at transcript level	0
Q96E66	LRTOMT LRRC51	12	11q13.4	evidence at transcript level	0
Q96KW9	SPACA7 C13orf28	3	13q34	evidence at transcript level	0
Q96L03	SPATA17	8	1q41	evidence at transcript level	0
Q96L15	ART5 UNQ575/PRO1137	4	11p15.4	evidence at transcript level	0
Q96LI9	CXorf58	1	Xp22.11	evidence at transcript level	0
Q96LM5	C4orf45	5	4q32.1	evidence at transcript level	0
Q96LU5	IMMP1L	3	11p13	evidence at transcript level	0
Q96M20	CNBD2 C20orf152	6	20q11.23	evidence at transcript level	0
Q96M60	FAM227B C15orf33	1	15q21.2	evidence at transcript level	0
Q96M69	LRGUK	11	7q33	evidence at transcript level	0
Q96M83	CCDC7 BIOT2	7	10p11.22	evidence at transcript level	0
Q96M86	DNHD1 C11orf47 CCDC35 DHCD1 DNHD1L UNQ5781/ PRO12970	13	11p15.4	evidence at transcript level	0
Q96N23	CFAP54 C12orf55 C12orf63	38	12q23.1	evidence at transcript level	0
Q96PP4	TSGA13	4	7q32.2	evidence at transcript level	0
Q96SF2	CCT8L2 CESK1	5	22q11.1	evidence at transcript level	0
Q9BYW3	DEFB126 C20orf8 DEFB26	5	20p13	evidence at transcript level	0
Q9BZ19	ANKRD60 C20orf86	1	20q13.32	inferred from homology	0
Q9BZJ4	SLC25A39 CGI-69 PRO2163	3	17q21.31	evidence at transcript level	6
Q9GZN6	SLC6A16 NTT5	6	19q13.33	evidence at transcript level	12
Q9H1M3	DEFB129 C20orf87 DEFB29 UNQ5794/PRO19599	3	20p13	evidence at transcript level	0
Q9H1P6	C20orf85	5	20q13.32	evidence at transcript level	0

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Table 2. continued

entry	gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
Q9H1U9	SLC25A51 MCART1	1	9p13.1	evidence at transcript level	6
Q9H3M9	ATXN3L ATX3L MJDL	7	Xp22.2	uncertain	0
Q9H3V2	MS4A5 CD20L2 TETM4	1	11q12.2	evidence at transcript level	4
Q9H3Z7	ABHD16B C20orf135	2	20q13.33	inferred from homology	0
Q9H579-2	MROH8 C20orf131 C20orf132	18	20q11.23	evidence at transcript level	0
Q9H5F2	C11orf1	5	11q23.1	evidence at transcript level	0
Q9H693	C16orf95	1	16q24.2	evidence at transcript level	0
Q9H7T0	CATSPERB C14orf161	22	14q32.12	evidence at transcript level	4
Q9H8X9	ZDHHC11 ZNF399	2	5p15.33	evidence at transcript level	4
Q9H943	C10orf68	1	10p11.22	evidence at transcript level	0
Q9NUD7	C20orf96	1	20p13	evidence at transcript level	0
Q9NZM6	PKD2L2	2	5q31.2	evidence at transcript level	6
Q9P1V8	SAMD15 C14orf174 FAM15A	19	14q24.3	evidence at transcript level	0
Q9P1Z9-2	CCDC180 C9orf174 KIAA1529	22	9q22.33	evidence at transcript level	1
Q9P2S6	ANKMY1 TSAL1 ZMYND13	19	2q37.3	evidence at transcript level	0
Q9UKJ8	ADAM21	3	14q24.2	evidence at transcript level	1
Q9ULG3	KIAA1257	1	3q21.3	evidence at transcript level	0
Q9Y238	DLEC1 DLC1	22	3p22.2	evidence at transcript level	0
Q9Y581	INSL6 RIF1	2	9p24.1	evidence at transcript level	0
W5XKT8	SPACA6 SPACA6P UNQ2487/PRO5774	2	19q13.41	evidence at transcript level	1

[a]Accession numbers and number of transmembrane helices (no. TMH) were retrieved from UniprotKB; gene names and chromosome location are as referenced in neXtProt.

reported, and thus the two paralogs cannot be distinguished based on these peptides. Because RSPH10B2 and RSPH10B differ by only three amino acids in total, it is very difficult to find other suitable unique peptides for validation. However, the identification of RSPH10B2 (B2RC85) was confirmed by PRM using the additional peptide "EEEFNTWVNNTYVFFVNTLFHAYK". The PE2 protein ZDHHC11 (Q9H8X9) was identified by two unique peptides, but one of them ("GVLQQGAGALGSSAQGVK") could also map to its paralog ZDHHC11B if SNP was considered. The six other peptides that lost their unicity when SNP was taken into account map to proteins for which there were more than three peptides of length ≥9 amino acids.

In the small group of nine proteins with a PE5 status (uncertain), three were identified and validated (<1% FDR) by at least two or more distinct uniquely mapping peptide sequences of length ≥9 amino acids. ATXN3L (Q9H3M9) was identified by six peptides, of which one would lose its unicity if SNPs were considered. This protein has now been characterized as a deubiquitinylase,[22−25] and its entry in UniProtKB/Swiss-Prot is currently under revision by the curators. HSP90AB4P (Q58FF6) and GK3P (Q14409), identified, respectively, by two and three peptides ≥9 amino acids in length, are annotated as pseudogenes in most protein databases. Their status should be revised based on the results presented here.

The six others (PRSS41, LINC00521, FTH1P19, FAM205BP, SPATA31D3, and SPATA31D4) were detected with only one distinct uniquely mapping peptide of length ≥9 amino acids. PRSS41 (Q7RTY9) is the ortholog of the recently characterized testis-specific serine protease Prss41/Tessp-1[21] and should no longer be considered as a pseudogene; the PE5 status of the entry is under revision by UniProtKB/Swiss-Prot curators. The putative uncharacterized protein encoded by LINC00521 (Q8NCU1) was detected by a 26-amino acid peptide for which no other match in the human proteome was found, even when possible variants were considered. The identification of the putative pseudogene FTH1P19 (P0C7X4) is also plausible because the peptide identified could only match the validated FTH1 protein (P02794) if both the rare D172N variant and an unknown S164A variant were considered. Nevertheless, according to the current HPP guidelines, the identifications of PRSS41, LINC00521, and FTH1P19 still need to be confirmed using other peptides.

The three remaining identifications are more dubious. Indeed, the peptide identifying FAM205BP (Q63HN1) could also match FAM205A (Q6ZU69), a PE1 protein, if its very common (30%) M499 V variant form was considered.

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

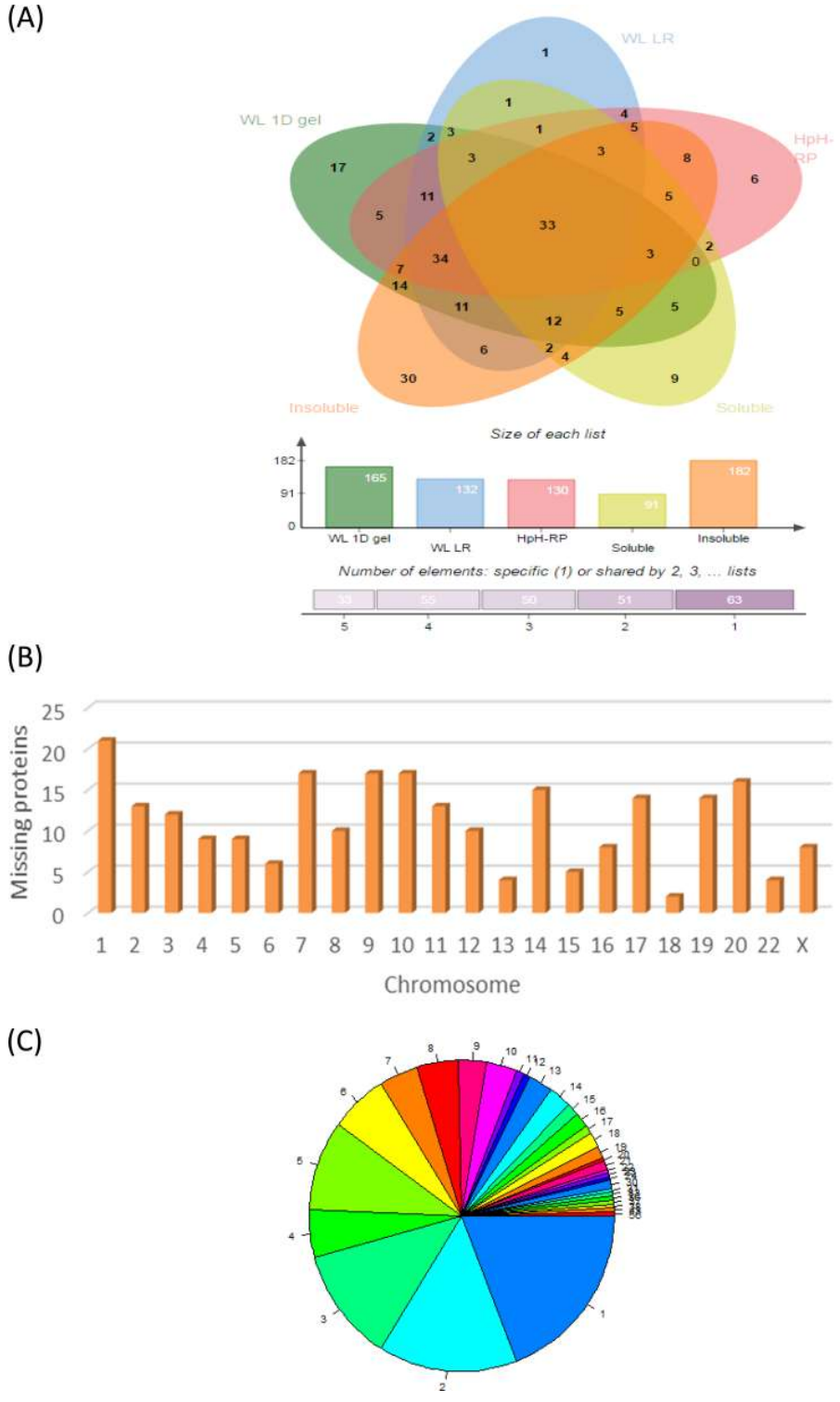


Figure 3. Missing proteins detected in the sperm proteome. A. Venn diagram illustrating the overlap between the five different fractionation protocols and the contribution of each fraction to the detection of missing proteins. Venn diagram was created with the jvenn web application.[39] WL-1D-gel: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), WL-LGR: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs, WL-HP-RP: total cell lysate, in-gel digestion of proteins, and peptide fractionation by high-pH reversed-phase (HpH-RP) chromatography, Soluble and Insoluble: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by a 1D SDS-PAGE separation of proteins (20 gel slices per fraction). Lower part: bar charts representing the number of missing proteins identified in each MS/MS data set. Under the bar chart: information related to the number of missing proteins identified specifically in each (1) or shared by 2, 3, 4 or all 5 data sets. B. Distribution of missing proteins according to the chromosomal location of their genes (retrieved from neXtProt). C. Distribution of missing proteins according to the number of proteotypic peptides: numbers beside each portion of the pie indicate the

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Figure 3. continued

number of unique peptides that mapped onto missing proteins (max. number of unique peptides: 56; see Table 2 and Supplementary Table 3 for details).

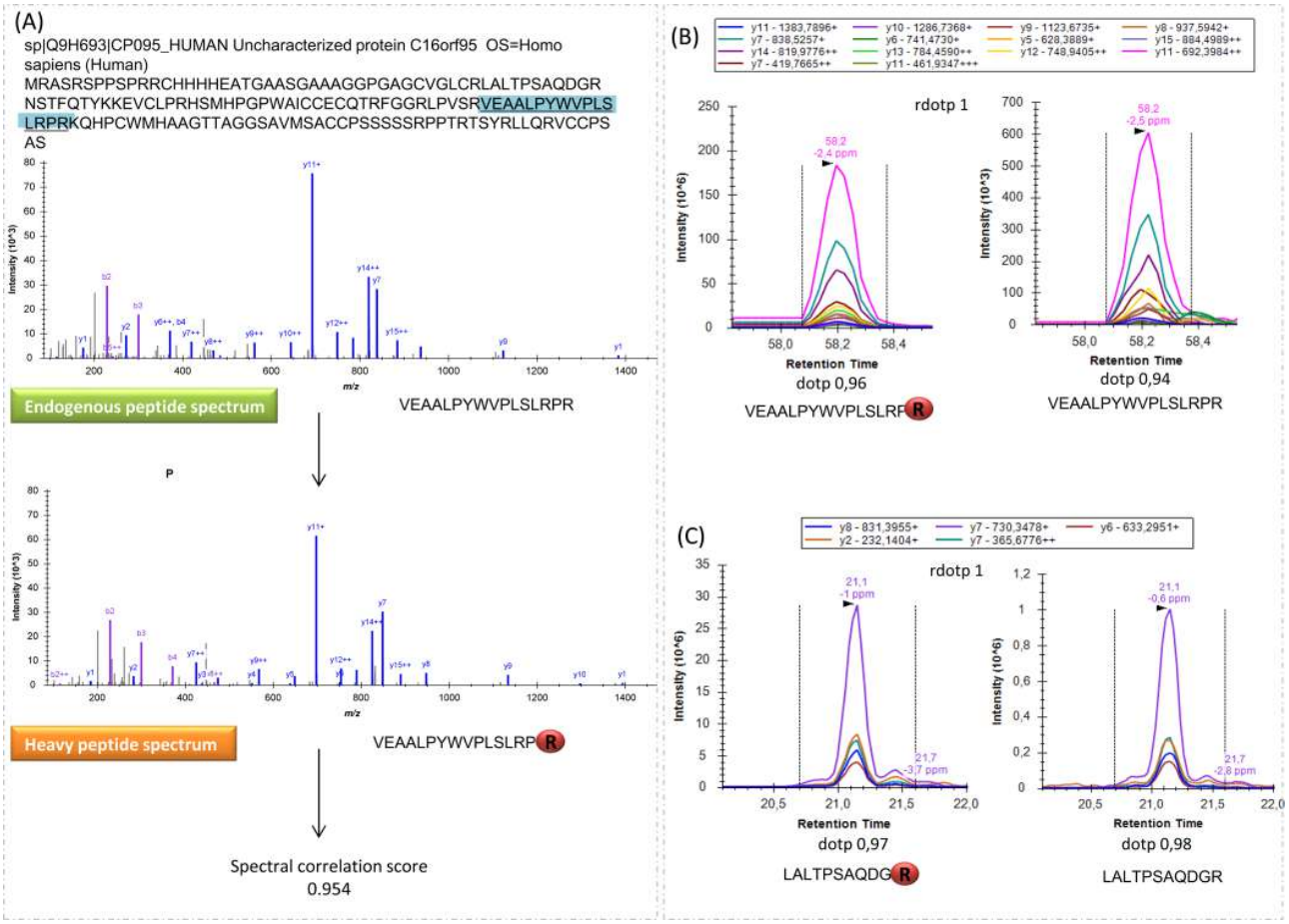


Figure 4. MS validation of one-hit wonder missing protein, example of protein Q9H693 (C16orf95). (A) Missing protein sequence with the unique peptide identified highlighted in blue; MS/MS spectrum of the endogenous peptide correlated with the MS/MS spectrum of its labeled synthetic counterpart and spectral correlation score calculated (SDPscore). (B) MS2 traces extracted from LC−PRM data from an unfractionated total protein extract with dotp calculated for the endogenous and labeled peptides and rdotp calculated for light/heavy correlation. (C) MS2 traces extracted from LC−PRM data from an unfractionated total protein extract. Analysis was targeted to detect an additional predicted peptide for protein Q9H693. The dotp for the endogenous and labeled peptide and the rdotp for light/heavy correlation were calculated.

Likewise, as discussed in Jumeau et al.,[14] the peptides mapping to SPATA31D4 (Q6ZUB0) and SPATA31D3 (P0C874) do not confidently identify one protein or the other. This issue cannot be resolved without access to data relative to the genomic sequences of the donors, as both variants SPATA31D3 R882G (dbSNP:rs815819) and SPATA31D4 G882R (dbSNP:rs138456481) may be present in the pooled sample studied.

Investigating One-Hit Wonder Missing Proteins Using MS-Based Criteria

Because one-hit wonder proteins could potentially correspond to incorrect PSM assignment or false-positives that passed the 1% FDR threshold, the recent HPP guidelines recommend that additional MS-based analyses be performed to provide further proteomics evidence (Deutsch et al., submitted; http://www.thehpp.org/guidelines/). We therefore investigated our subset of one-hit wonder missing or uncertain proteins using additional MS-based criteria, as described in ref 13. The first criterion relied on a blinded inspection of each MS/MS spectrum for the 52 missing or uncertain proteins (47 PE2-PE4; 5 PE5), all of which were identified by a unique peptide ≥9 amino acids in length.

The quality control of the PSMs corresponding to these unique peptides was carried out by visual validation by at least two mass spectrometry experts from each of the three sites of the ProFI infrastructure (Grenoble, Strasbourg, Toulouse) and from the Protim core facility (Rennes). PSM quality was classed in two categories: low and high. A high classification was based on the following spectral features: (i) the presence of y-ion and b-ion series; (ii) peak intensities; and (iii) quality of the match between the experimental and theoretical spectra. A subset of 18 peptides was assigned a "high"-quality tag by three out of the four sites and was therefore preferentially selected for further MS validation (data not shown). In addition to these 18

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

candidates, other one-hit wonders were selected that were awarded a majority rather than a consensual "high" quality attribute. A supplementary filter was applied by retaining only peptides that could be synthesized (i.e., peptides shorter than 25 amino acids). This filter was necessary as further validation steps required the availability of a synthetic peptide for each candidate to be validated. Thus, of the initial 52 "one-hit wonders" we ended up with a final list of 36 peptides mapping to 34 missing proteins (PE2−4) and 2 uncertain proteins (PE5) for further assessment (among the 18 that had a unanimous high-quality vote, 17 peptides were selected; see Supplementary Table 4). Synthetic labeled peptide versions of all 36 peptides were ordered to allow systematic comparison of identifications with a synthetic version of the peptide (i.e., same charge, same instrument fragmentation conditions). The goal of this step was to increase confidence in the identification of each peptide. To allow readers to assess spectrum quality and peak intensity patterns between the endogenous peptide and its synthetic counterpart peptide for themselves, an example spectrum for synthetic peptides is shown alongside the naturally derived peptides for protein entry Q9H693 in Figure 4A; likewise all other comparative spectra are presented in Supplementary Figure 1. To objectively assess peptide "VEAALPYWVPLSLRPR" (protein entry Q9H693; shown in Figure 4A), we calculated the spectral dot-product score (SDPscore)[26] that corresponds to the spectral correlation score calculated for the intensities of all common singly charged b- and y- ions of the reference spectrum and the native spectrum, as in our previous studies.[13] An SDPscore of 0.954 was obtained for this peptide, indicating that its MS/MS fragmentation pattern is very similar to the pattern obtained for the reference synthetic peptide.

Additional Experimental Validation Using Targeted LC−PRM Assays

In a final MS-based validation attempt, targeted MS assays were developed for the 36 candidate "one-hit wonder" proteins to try to redetect their proteotypic peptide coeluting with its synthetic labeled counterpart. Samples for these assays were prepared independently. In addition to the original unique peptide identified, two additional predicted proteotypic peptides were selected when possible and synthesized for all 36 proteins (Supplementary Table 4). A total of 100 labeled peptides were synthesized. These peptides were mixed with protein digests, and the heavy and light forms were targeted for analysis using PRM scanning on a high-resolution Q-Orbitrap mass spectrometer. In a first attempt, all proteotypic peptides were targeted in an unfractionated total protein extract prepared in stacking gel bands. This allowed us to unambiguously detect perfectly coeluting specific light/heavy transition groups for 24 proteotypic peptides corresponding to 21 of the 36 proteins. Examples of light/heavy transition group coelution are presented for protein Q9H693 for its initial peptide (Figure 4B) and for one additional predicted peptide that was detected thanks to the increased sensitivity of targeted assays (Figure 4C). Subsequently, to attempt to validate more candidates, the remaining undetected peptides were targeted in the insoluble proteins fraction after preliminary separation on 1D SDS-PAGE. Samples were prepared as previously described for the total proteome analysis (protocol ii). The insoluble fraction was chosen for these targeted assays as a majority of the one-hit wonders (18 out of the 36, Supplementary Table 4) were identified in samples prepared by this protocol. Thus, our chances of validating peptides were greater with samples prepared by this protocol. These final MS experiments unambiguously validated 27 additional peptides belonging to 20 out of the 36 proteins, not all the same as the previously validated 21 proteins. Thus, in total, 24 out of the 36 proteins were validated with additional predicted and specifically targeted peptides. This successful validation with additional peptides confirms the utility of highly sensitive targeted assays compared with nontargeted data-dependent LC−MS/MS acquisitions and shows that this approach is suitable for unambiguous validation of missing proteins. All of the results obtained with targeted LC−PRM assays can be found in Supplementary Table 4 and Supplementary Figure 1.

Bioanalysis of the Missing Proteins

For all missing proteins identified by MS-based analysis, the chromosomal location, PE status, predicted number of TMH, and functional annotation were retrieved from neXtProt (see Table 2 and Supplementary Table 3 for details). Using a previously described methodology,[14] we also extensively mined publicly available transcriptome data, that is, the "*Human testis gene expression program*" described by Chalmel and collaborators[5] to check whether the missing and uncertain proteins identified in the present study corresponded to genes carrying the testis signature, related to the onset of human spermatogenesis, or whether they corresponded to genes expressed only at the very end of spermiogenesis.

Only 132 of the 235 missing proteins and 3 of the 9 uncertain proteins identified in this study corresponded to genes referenced in the "*Human testis gene expression program*",[5] with an increasing expression in seminiferous tubules containing postmeiotic germ cells (Johnsen score ≥7). Up to 76 (+2 uncertain) of these proteins corresponded to genes specifically expressed in the testis (SET), 31 (+ 1 uncertain) corresponded to genes preferentially expressed in the testis (PET), and 25 corresponded to genes with intermediate (IE) or ubiquitous (UE) expression in the testis (Supplemental Table 3). This information was not used to select candidates for validation, but it is important to help understand the results of the immunohistochemistry experiments. Of note is that 26 and 2 proteins corresponding, respectively, to SET and PET genes are present in the current list of 1057 testis-enriched proteins in HPA, an update of the initial list from Djureinovic et al.[42] That suggests the spermatozoon has great potential for the identification of additional missing proteins. It also shows that proteins that are not considered as highly enriched in the testis might concentrate in germ cells during late spermiogenesis and become accessible in the spermatozoa. However, it is also important to note that a significant subset of missing proteins identified in the present study corresponded to genes that are testis-specific but do not belong to the "*Human testis gene expression program*" as their expression is not enriched in seminiferous tubules containing postmeiotic germ cells (data not shown).

To date, up to 111 of the 244 PE2−5 proteins identified have been subjected to extensive data and literature mining (Supplementary Table 5). The information gleaned from this data mining was used to establish priorities for further antibody-based studies. The first selection criterion was based on transcriptomics analysis. RNA sequencing results from the HPA database indicated that 88 of these 111 proteins were either specifically expressed in testis or were enriched in a small group of tissues in which testis has one of the two highest

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

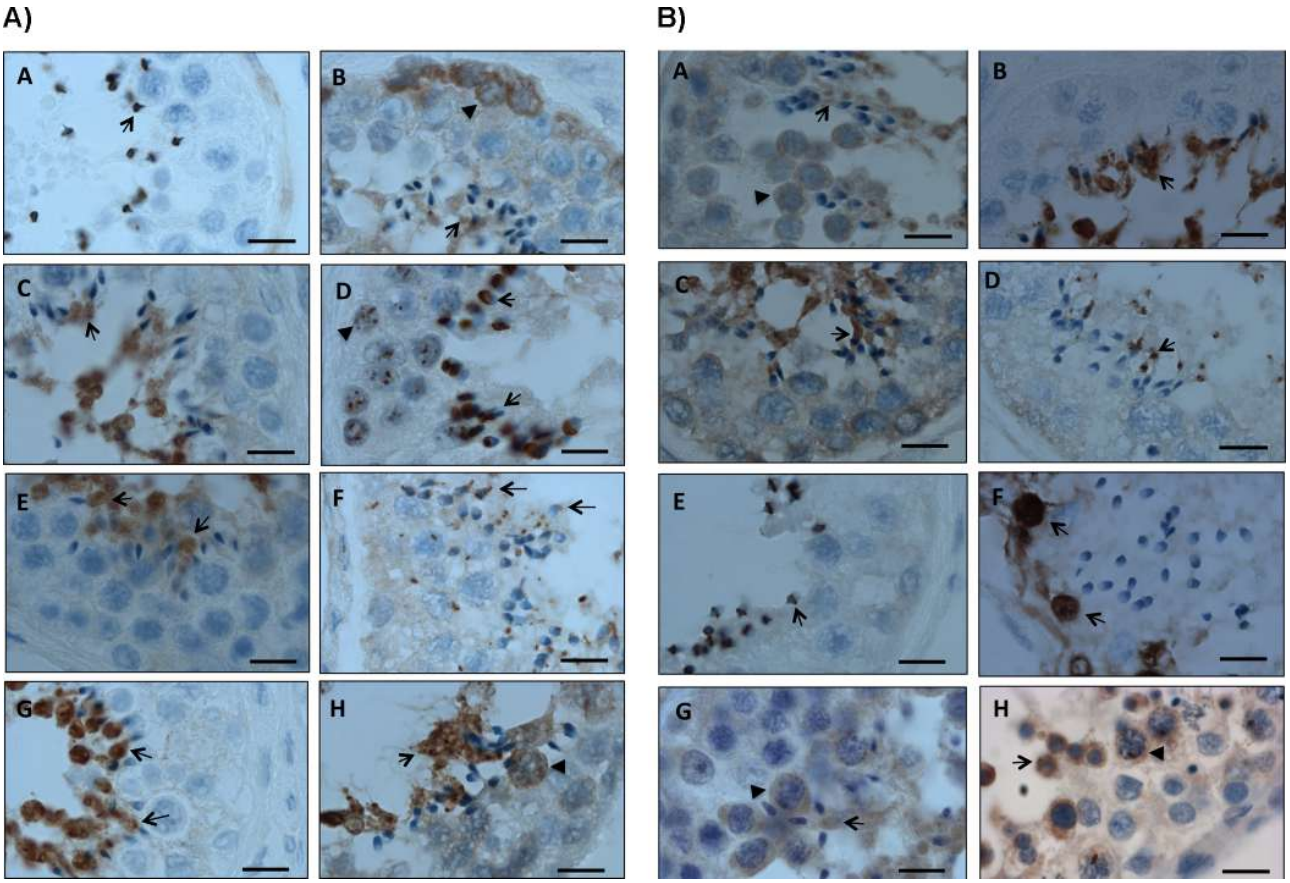


Figure 5. (A) Antibody staining for orphan proteins CXorf58 (Q96LI9), C19orf81 (C9J6K1), C17orf105 (B2RV13), C20orf85 (Q9H1P6), C10orf53 (Q8N6V4), C10orf67 (Q8IYJ2), FAM187B (Q17R55), and SSMEM1 (Q8WWF3) in adult human testis. Proteins were detected in transverse testis sections at stages IV to VI of the seminiferous epithelium[40] using polyclonal antibodies from the HPA specific for CXorf58 (HPA031543) (A), C19orf81 (HPA060238) (B), C17orf105 (HPA053028) (C), C20orf85 (HPA058271) (D), C10orf53 (HPA037951) (E), C10orf67 (HPA038131) (F), FAM187B (HPA014687) (G), and SSMEM1 (HPA026877) (H). Nonimmune serum was used as a negative control (data not shown). In all testis sections, a more or less intense antibody staining signal was visible in germ cells in all stages and for all proteins (A–H). CXorf58 immunoreactivity was very strong in the headpiece of late spermatids (A; arrow). C19orf81 presented strong staining in the cytoplasm of spermatogonia (arrowhead) and of late spermatids (B; arrow). Intense C17orf105 staining was visible in the cytoplasm of late spermatids (C; arrow). C20orf85 immunoreactivity displayed as an intense granular staining in pachytene spermatocytes (arrowhead) and concentrated with a very strong signal in the acrosome of elongating spermatids (arrows) (D). C10orf53 immunoreactivity appeared concentrated in the cytoplasm of late spermatids (E; arrows). A punctiform signal was visible for C10orf67 in the cytoplasm of elongating spermatids (F; arrows). FAM187B immunoreactivity concentrated in the cytoplasm of elongating spermatids (G; arrows). SSMEM1 displayed intense staining in pachytene spermatocytes (arrowhead) and in elongating spermatids (arrows) (H). Scale bars = 20 μm. (B) Antibody staining for CCT8L2 (Q96SF2), AXDND1 (Q5T1B0), WDR88 (Q6ZMY6), GOT1L1 (Q8NHS2), CFAP46 (Q8IYW2), SAMD15 (Q9P1V8), WDR93 (Q6P2C0), and NOXRED1 (Q6NXP6) in adult human testis. Proteins were detected in transverse testis sections at stages IV to VI of the seminiferous epithelium[40] using polyclonal antibodies from the HPA specific for CCT8L2 (HPA039268) (A), AXDND1 (HPA071114) (B), WDR88 (HPA041916) (C), GOT1L1 (HPA028778) (D), CFAP46 (HPA038034) (E), SAMD15 (HPA030673) (F), WDR93 (HPA048112) (G) and NOXRED1 (HPA055658) (H). Nonimmune serum was used as a negative control (data not shown). In all testis sections, a more or less intense signal for all proteins was visible in germ cells in all stages (A–H). CCT8L2 staining was clearly cytoplasmic in pachytene spermatocytes (arrowhead) and late spermatids (arrow) (A). AXDND1 immunoreactivity was intense in the cytoplasm of late spermatids (arrow; B). WDR88 immunoreactivity concentrated with a very strong signal in the cytoplasm of elongating spermatids (arrow; C). A very strong GOT1L1 immunostaining was observed in the cytoplasm of late spermatids (D; arrow). CFAP46 immunoreactivity was very strong and ring-shaped in the headpiece of late spermatids (arrow; E). SAMD15 immunoreactivity was intense in pachytene spermatocytes (arrow; F). A strong immunoreactivity was observed for WDR93 in the cytoplasm of pachytene spermatocytes (arrowhead) and in the cytoplasm of elongating spermatids (arrow) (G). NOXRED1 immunoreactivity was exclusively cytoplasmic and the signal increased from premeiotic germ cells to pachytene spermatocytes (arrowhead) and round spermatids (arrows) (H). Scale bars = 20 μm.

expression levels (column C). These proteins were assigned a very high score (dark green). The eight proteins that were enriched in a small group of tissues including testis, but for which expression levels in testis were not among the two highest ones, were assigned with a lower score (light green). In contrast, a low (red) priority score was assigned to six proteins that were shown either to be specifically expressed in organs other than testis or to be undetectable by RNA sequencing.[5,43] The remaining proteins are either ubiquitously expressed or expressed at low levels in testis and were assigned a neutral score (white).

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998–4019

The second selection criterion was based on phylogenetic profiling. We sought and report (column D) on the presence of homologues in *S. cerevisiae* and a number of ciliated organisms from distant groups. The 18 proteins for which homologues were present in at least three of the ciliated groups were assigned with a very high priority score (dark green) because we hypothesized that they could be involved in ciliogenesis. The 17 having homologues in one or two of the ciliated groups but not in yeast were assigned a high priority score (green). In contrast, a low priority score (red) was assigned to the three proteins that were found to be conserved in yeast because they are not expected to play a specific role in ciliogenesis or human reproduction.

The third selection criterion was based on the phenotypes observed in knockout mice (column E). The four proteins for which a deletion led to a reproduction phenotype were assigned a very high priority score (dark green). The ten proteins whose deletion had no effect on fertility were assigned a low priority score (red).

The fourth selection criterion was based on information retrieved from literature mining (gene expression regulation, protein interactions, function). This information is summarized in column F. Proteins for which published information suggested a putative role in spermatogenesis, ciliogenesis, or cilia function (85 proteins) were assigned a very high (dark green, 28 proteins) or high (green, 56 proteins) priority score. In the course of literature mining, we noticed that up to 76 of these 111 proteins had already been detected by mass spectrometry in human testis[27] or sperm.[28−30] However, these identifications have not yet been curated by PeptideAtlas or neXtProt; thus the existence of these proteins was not considered validated at the time of writing.

The fifth selection criterion was based on immunohistochemistry data retrieved from the HPA (column G). The nine proteins that were specifically observed in germ cells or associated with ciliary or cytoskeletal structures were assigned a very high priority score (dark green). Twelve other proteins observed in testis or ciliated cells were assigned a high priority score (green). Conversely, four proteins that were not seen in testis but were clearly seen in other tissues were assigned a low priority score (red).

On the basis of the combination of all these criteria, each of the 111 PE2−5 proteins was assigned a score grading the potential relevance of the protein in sperm and testis biology as high/medium/low. Seven entries that were initially classed as of high/medium interest were down-graded to low interest proteins because their localization in human sperm cells had already been published (column H). This left a total of 33 "high interest" proteins. Among them, we selected the 26 for which a HPA antibody was available that passed the Protein Arrays (PA) test with a single peak, corresponding to interaction only with its own antigen. We selected 12 additional proteins from among the 42 scored as "medium interest" and one from among the 36 scored as "low interest" (columns I and J).

Orthogonal Immunohistochemistry Evidence

Immunohistochemical studies were undertaken to provide non-MS-based evidence of the expression of the 39 missing proteins selected based on this data mining process. Specific immunohistochemistry staining in human testes was obtained for 16 missing proteins using antibodies from the HPA without need for further technical improvement (see Supplemental Table 6). Results from these experiments show that all 16 selected proteins displayed immunoreactive signals at various intensities in germ cells in all stages of their development (Figures 5A,B). No staining above background levels was visible in interstitial cells or somatic cells in the seminiferous tubules for any of the antibodies (Figures 5A,B). The staining intensity for missing proteins increased significantly from premeiotic and meiotic germ cells onward (for C19orf81, C20orf85, SSMEM1, CCT8L2, WDR88, SAMD15, WDR93, and NOXRED1) or postmeiotic germ cells onward (for CXorf58, C17orf105, C10orf53, C10orf67, FAM187B, AXDND1, GOT1L1, and CFAP46). This profile was to be expected, as all proteins, except C10orf67, C19orf81, and WDR93, corresponded to genes in the TGEP, and their expression was expected to gradually increase in later stages of sexual maturation[5] (Supplemental Tables 3 and 6). The expression levels for genes coding for C10orf67, C19orf81, and WDR93 may be below the threshold required to be part of the TGEP, even though immunoreactivity was observed in the germ cell lineage.

All 16 missing proteins whose expression was demonstrated in situ deserve further study to determine their role in sperm biology. However, because of their very specific expression patterns, five of them call for an immediate focus. Indeed, based on our immunohistochemistry data, staining for CXorf58, C20orf85, and CFAP46 was concentrated in late spermatids at the level of the acrosome under formation, a sperm-specific organelle essential for fertilization, with CFAP46 and CXorf58 displaying spectacular annular staining. Expression of FAM187B and AXDND1 displayed a slightly different profile in the adult testis, with immunoreactivity concentrated in the cytoplasmic region of elongating and elongated spermatids undergoing intense remodeling. This staining profile has previously been shown to be associated with the expression of proteins playing a role in sperm maturation.

CXorf58 is an orphan protein with no curated functional comments in UniProtKB/Swiss-Prot; TargetP and MitoProt prediction programs predict a mitochondrial localization. Its expression in sperm cells was confirmed by PRM based on endogenous peptide "SFFDEAPAFSGGR", detected only in the insoluble fraction, and the additional peptide "DISAQIIQR". The staining pattern observed in our immunohistochemistry experiment suggests that this protein is mainly located at the lower part of the head and midpiece of spermatids. This position is in favor of a link to mitochondria. Interestingly, mitochondria are grouped in the midpiece of mature spermatozoa, and numerous studies support the proposal that these organelles are important for sperm function and fertilization (for a review, see ref 31).

C20orf85 is also an orphan protein. It is expressed in the epithelium of the airways,[32] with levels increasing sharply during mucociliary differentiation. Its expression clusters with that of genes involved in regulation of cytoskeletal organization and intracellular transport.[33] In the adult testis, C20orf85 immunoreactivity was very strongly concentrated in the acrosome as it formed in elongating spermatids. The protein might migrate further down to the midpiece or flagellum in mature sperm. In yeast two-hybrid experiments, the murine C20orf85 ortholog (1700021F07Rik) was shown to interact with CCNB1IP1, a putative ubiquitin E3 ligase that is essential for chiasmata formation and hence fertility.[34] Together with these observations, the altered expression of C20orf85 in asthenozoospermic patients[20] suggests a possible role in sperm movement.

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Finally, CFAP46 is the ortholog of Chlamydomonas FAP46, which is known to be part of the central apparatus of the cilium axoneme and to play a role in cilium movement.[35] In the adult testis, immunostaining for CFAP46 was annular in late spermatids, a pattern that is typical of migration that will continue further down the midpiece or flagellum. Interestingly, CFAP46 expression has been shown to be downregulated in patients with primary ciliary dyskinesia.[36] That, together with our observations, is in favor of a central role in sperm movement.

FAM187B is an orphan protein with no curated comments in UniProtKB/Swiss-Prot, except an indication that it is a transmembrane protein.[30] The very peculiar localization of the immunohistochemistry staining for this protein, in the cytoplasmic region of elongating spermatids, suggests that FAM187B may play a role in cytoplasm displacement and elimination that take place during spermiogenesis. Interestingly, FMA187B mRNA expression in sperm has been proposed as a valuable diagnostic indicator of sperm survival, fertility, and capacity to promote early embryogenesis.[37]

AXDND1 is an intracellular protein that is selectively expressed in the nasopharynx, bronchus, testis and fallopian tubes, according to HPA immunochemistry data. It is highly conserved in vertebrates and in the choanoflagellate *Salpingoeca* and contains an axonemal dynein light chain domain (IPR019347). Interestingly, the outer arm dynein complex is the main propulsive force generator for ciliary/flagellar beating. The staining pattern for the protein, positive in the cytoplasmic region of elongating spermatids undergoing extensive remodelling, matches with a possible role for the protein in mobility of the sperm flagellum.

CONCLUSIONS

In this study, MS-based analysis of sperm samples detected 235 missing (PE2−4) and 9 uncertain (PE5) proteins. Among these, 206 missing and 4 uncertain proteins were validated with at least two or more distinct peptide sequences with ≥9 amino acids that mapped only to a single protein entry, even when possible variants were considered. In line with version 2.0 of the HPP Data Interpretation Guidelines, these 210 proteins can therefore be considered as validated. Twenty-four of these proteins were confirmed by LC−PRM assays and 16 by IHC on human testis sections. IHC studies allowed us not only to confirm the existence of the proteins in sperm but also to hypothesize a biological role for some of them (i.e., CXorf58, C20orf85, CFAP46, FAM187B, and AXDND1). The combination of LC−PRM and IHC was clearly instrumental in validating two "one-hit wonders": CXorf58 and C19orf81.

The importance of considering possible variants was illustrated by the cases of eight proteins, including three PE5, identified with peptides that lost their proteotypicity when possible variants were considered. These eight proteins therefore cannot be considered validated with our data.

The remaining 26 proteins were detected with only one unique peptide ≥9 amino acids. Six of these peptides were confirmed by LC−PRM, and for three others, manual inspection unanimously indicated high-quality LC−MS spectra. Thus, these nine identifications are reported here with confidence. However, the current HPP guidelines require MS-based validation of additional peptides for these proteins or antibody detection to definitively validate their existence. The 17 other peptides passed the FDR criterion, but visual examination of their spectra indicated insufficient quality to warrant further study. Hence, the identification of these 17 proteins can only be considered dubious.

The Swiss−French collaborative project investigating the human sperm proteome in the context of the C-HPP started 3 years ago. In our previous article,[14] we reported the detection of 94 PE2−5 proteins in sperm using an LTQ-Orbitrap XL mass spectrometer, with at least one peptide of 9 aa. This data set was submitted to proteomeXchange, reanalyzed by PeptideAtlas, combined with other data sets, and used by neXtProt to validate protein existence based on the stringent guidelines established in 2016. Finally, 54 of these 94 proteins, including 3 PE5, were validated and are now annotated PE1. It is remarkable that all 94 of these proteins were identified in the present study. Except for TMEM239 (Q8WW34), detected with a single 24 amino acid peptide, all proteins were detected with at least two peptides and often many more (Supplementary Table 7). The coverage of each protein was considerably improved by the use of cutting edge instruments (i.e., Q-Exactive; Thermo Scientific) and sample fractionation.

We are confident that the present data can be used to validate the existence of 210 missing or uncertain proteins and are looking forward to integration of these validations in neXtProt once they have been reanalyzed by PeptideAtlas and combined with data from the other C-HPP teams interested in the testis or sperm proteome. In the meantime, the investigation of the human sperm proteome continues in our laboratories together with extensive data mining on the remaining set of missing proteins presented in this study. The information gleaned will help to extend our knowledge on the potential roles of these proteins in sperm function or maturation. Indeed, some of the proteins identified here may present a high clinical potential, and could also benefit the Biology and Disease driven HPP (B/D-HPP) that aims to explore the impact of proteomic technologies applied to a focused area of life science and health.[38]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00400.

Shotgun LC−MS/MS analyses. 2. LC−MS/MS analysis of labeled synthetic peptides and comparison of fragmentation spectra. 3. How targeted LC−PRM assays were developed (details). (PDF)

Supplementary Figure 1: MS/MS spectra for the 36 endogenous peptides and their synthetic reference counterparts combined with LC−PRM results for the 36 peptides, additional predicted proteotypic peptides, and their labeled synthetic counterparts. (PDF)

Supplementary Table 1: PSM-, peptide-, and protein-level FDR values along with the total number of expected true- and false-positives at each level for each sperm proteome data set and the combined data set (tab 1: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), tab 2: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs, tab 3: total cell lysate, in-gel digestion of proteins, and peptide fractionation by high-pH reversed-phase (HpH-RP) chromatography; tabs 4 and 5: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by 1D

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

SDS-PAGE separation of proteins (20 gel slices per fraction); tab 6: combined data set corresponding to the combination of results for the five proteome data sets). (XLSX)

Supplementary Table 2: List of proteins identified and validated with a protein-level 1% FDR for each fraction (detailed information): tab 1: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), tab 2: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs, tab 3: total cell lysate, in-gel digestion of proteins, and peptide fractionation by high-pH reversed-phase (HpH-RP) chromatography; tabs 4 and 5: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by 1D SDS-PAGE separation of proteins (20 gel slices per fraction). (XLSX)

Supplementary Table 3: Missing (PE2−4) and uncertain (PE5) proteins detected in the sperm proteome: detailed information. Accession numbers, entry description, molecular weight (MW), protein length (length) number of transmembrane domains (No. TMH), subcellular location, and function (CC field) were retrieved from UniprotKB; gene names and chromosome location are as referenced in neXtProt. Coverage (protein coverage in %) and number of unique peptides mapping to missing proteins, proteins seen (yes)/not seen (not) in each of the five MS/MS data sets acquired in this study are reported. The expression annotations of the Human testis gene expression program (TGEP[5]) were also reported when available (SET: proteins produced by genes specifically expressed in the testis; PET: proteins produced by genes preferentially expressed in the testis; IE: proteins produced by genes with intermediate expression in the testis; UE: proteins produced by genes with ubiquitous expression in the testis). The ≪testis-enriched gene≫ status in the Human Protein Atlas version 15 is also provided. (XLSX)

Supplementary Table 4: tab 1: List of one-hit wonder missing proteins identified and selected for further spectral comparison (MS/MS) and PRM validation. tab 2: Extended list of 100 peptides selected for validation of the 36 one hit wonders and synthesized as crude labeled peptides. (XLSX)

Supplementary Table 5: List of 111 PE2−5 proteins for which complete data mining was performed, showing the rationale for their prioritization for subsequent antibody-based studies. Entry accession numbers (column A) and gene names (column B) were retrieved from UniProtKB, transcript abundance was retrieved from HPA (column C), phylogenetic profiles in ciliated organisms were determined by Blast analysis on UniProtKB "Reference proteomes" (column D), knockout mice phenotypes were retrieved from MGI (column E), associated publications that were not annotated in neXtProt were searched in PubMed (column F), and immunohistochemistry data was retrieved from HPA (column G). For all these criteria, a four-color grading system was adopted. Dark-green and green cells were retained as positive criteria, red cells as negative ones. Based on these criteria, a score of relevance (high/medium/low) for the implication of the proteins in spermatogenesis has been assigned (column H). The existence of a suitable antibody in HPA is reported in column I, and the list of proteins that were finally selected for IHC is provided in column J. (XLSX)

Supplementary Table 6: List of missing proteins for which IHC was successful, HPA antibody names, and dilutions used. (XLSX)

Supplementary Table 7: List of the 94 missing proteins detected in Jumeau et al. (2015)[14] and confirmed in the present study. The number of peptides identified in each study is reported in columns C and E. The protein existence (PE) status of the entries in the 2015 and 2016 neXtProt reference releases is reported in columns B and F. (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*Y.V.: E-mail: yves.vandenbrouck@cea.fr. Tel: +33 (0)4 38 78 26 74. Fax: (33) (0)4 38 78 50 32.

*C.P.: E-mail: charles.pineau@inserm.fr. Tel: +33 (0)2 23 23 52 79.

Author Contributions

Y.V., L.L., and C.P. cocoordinated the study. Y.V., L.L., C.P., A.G.P., C.C., C.B., S.C., O.B.S., M.F., and J.G. conceived and designed the experiments and analyses. T.F. and K.R. performed spermatozoa preparation. C.C., C.M., A.G.P., Y.C., M.M., and K.C. performed the sample preparation and MS/MS analysis. Y.V., C.B., A.M.H., E.M.B., L.L., T.R., and A.G. processed and analyzed MS/MS data sets. C.C. and C.M. performed MS/MS spectra comparison and PRM assays. P.D., Y.V., C.P., A.G., L.M., E.C., and L.L. performed bioinformatics analysis and data/literature mining on identified proteins and selected candidates for IHC studies. Immunohistochemical studies were done by K.R., C.P., and C.L. Y.V., L.L., P.D., C.C., C.M., and C.P. prepared the figures, tables, and Supporting Information. Y.V., L.L., C.C., and C.P. drafted the manuscript. All the authors approved the final version of the manuscript.

Author Contributions

▲Y.V. and L.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

All MS/MS data are available via ProteomeXchange under identifier PXD003947.

ACKNOWLEDGMENTS

This work was partially funded through the French National Agency for Research (ANR) (grant ANR-10-INBS-08; ProFI project, "Infrastructures Nationales en Biologie et Santé"; "Investissements d'Avenir" call). This work was also supported by grants from Biogenouest and *Conseil Régional de Bretagne* awarded to C.P. We are grateful to Monique Zahn and Maighread Gallagher-Gambarelli for suggestions on language usage. We thank Marine Seffals for technical support with immunohistochemistry experiments on the H2P2 core facility (Université de Rennes 1, US18, UMS3480 Biosit, Biogenouest, Rennes, France). We particularly thank Blandine Guével, Véronique Dupierris, Mélanie Lagarrigue, Jean-Philippe Menetrey, Mathieu Schaeffer, Régis Lavigne, and Christine Kervarrec for their technical assistance.

REFERENCES

(1) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang,

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221−3.

(2) Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Teixeira, D.; Zhang, Y.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.* **2015**, *43* (Database issue), D764−70.

(3) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.* **2015**, *14* (9), 3452−60.

(4) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013−2014 and strategies for finding missing proteins. *J. Proteome Res.* **2014**, *13* (1), 15−20.

(5) Chalmel, F.; Lardenois, A.; Evrard, B.; Mathieu, R.; Feig, C.; Demougin, P.; Gattiker, A.; Schulze, W.; Jégou, B.; Kirchhoff, C.; Primig, M. Global human tissue profiling and protein network analysis reveals distinct levels of transcriptional germline-specificity and identifies target genes for male infertility. *Hum. Reprod.* **2012**, *27* (11), 3233−48.

(6) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P. H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Ponten, F. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347* (6220), 1260419.

(7) Uhlen, M.; Hallstrom, B. M.; Lindskog, C.; Mardinoglu, A.; Ponten, F.; Nielsen, J. Transcriptomics resources of human tissues and organs. *Mol. Syst. Biol.* **2016**, *12* (4), 862.

(8) Baker, M. A.; Nixon, B.; Naumovski, N.; Aitken, R. J. Proteomic insights into the maturation and capacitation of mammalian spermatozoa. *Syst. Biol. Reprod. Med.* **2012**, *58* (4), 211−7.

(9) Eddy, E. M. Male germ cell gene expression. *Recent Prog. Horm. Res.* **2002**, *57*, 103−28.

(10) Jégou, B.; Pineau, C.; Dupaix, A. Paracrine Control of Testis Function. In *Male Reproductive Function*; Wang, C., Ed.; Kluwer Academic: Berlin, 1999; pp 41−64.

(11) Krausz, C. Male infertility: pathogenesis and clinical diagnosis. *Best practice & research. Clinical endocrinology & metabolism* **2011**, *25* (2), 271−85.

(12) Rolland, A. D.; Jégou, B.; Pineau, C. Testicular development and spermatogenesis: harvesting the postgenomics bounty. *Advances in experimental medicine and biology* **2009**, *636*, 16−41.

(13) Carapito, C.; Lane, L.; Benama, M.; Opsomer, A.; Mouton-Barbosa, E.; Garrigues, L.; Gonzalez de Peredo, A.; Burel, A.; Bruley, C.; Gateau, A.; Bouyssie, D.; Jaquinod, M.; Cianferani, S.; Burlet-Schiltz, O.; Van Dorsselaer, A.; Garin, J.; Vandenbrouck, Y. Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *J. Proteome Res.* **2015**, *14* (9), 3621−34.

(14) Jumeau, F.; Com, E.; Lane, L.; Duek, P.; Lagarrigue, M.; Lavigne, R.; Guillot, L.; Rondel, K.; Gateau, A.; Melaine, N.; Guével, B.; Sergeant, N.; Mitchell, V.; Pineau, C. Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, *14* (9), 3606−20.

(15) Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P. A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H. J.; Albar, J. P.; Martinez-Bartolome, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−6.

(16) Com, E.; Rolland, A. D.; Guerrois, M.; Aubry, F.; Jégou, B.; Vallet-Erdtmann, V.; Pineau, C. Identification, molecular cloning, and cellular distribution of the rat homolog of minichromosome maintenance protein 7 (MCM7) in the rat testis. *Mol. Reprod. Dev.* **2006**, *73* (7), 866−77.

(17) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. Orthogonality of separation in two-dimensional liquid chromatography. *Anal. Chem.* **2005**, *77* (19), 6426−34.

(18) Kitata, R. B.; Dimayacyac-Esleta, B. R.; Choong, W. K.; Tsai, C. F.; Lin, T. D.; Tsou, C. C.; Weng, S. H.; Chen, Y. J.; Yang, P. C.; Arco, S. D.; Nesvizhskii, A. I.; Sung, T. Y.; Chen, Y. J. Mining Missing Membrane Proteins by High-pH Reverse-Phase StageTip Fractionation and Multiple Reaction Monitoring Mass Spectrometry. *J. Proteome Res.* **2015**, *14* (9), 3658−69.

(19) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **2010**, *604*, 55−71.

(20) Amaral, A.; Castillo, J.; Ramalho-Santos, J.; Oliva, R. The combined human sperm proteome: cellular pathways and implications for basic and clinical science. *Hum. Reprod. Update* **2014**, *20* (1), 40−62.

(21) Yoneda, R.; Kimura, A. P. A testis-specific serine protease, Prss41/Tessp-1, is necessary for the progression of meiosis during murine in vitro spermatogenesis. *Biochem. Biophys. Res. Commun.* **2013**, *441* (1), 120−5.

(22) Weeks, S. D.; Grasty, K. C.; Hernandez-Cuebas, L.; Loll, P. J. Crystal structure of a Josephin-ubiquitin complex: evolutionary restraints on ataxin-3 deubiquitinating activity. *J. Biol. Chem.* **2011**, *286* (6), 4555−65.

(23) Buus, R.; Faronato, M.; Hammond, D. E.; Urbe, S.; Clague, M. J. Deubiquitinase activities required for hepatocyte growth factor-induced scattering of epithelial cells. *Curr. Biol.* **2009**, *19* (17), 1463−6.

(24) Sacco, J. J.; Yau, T. Y.; Darling, S.; Patel, V.; Liu, H.; Urbe, S.; Clague, M. J.; Coulson, J. M. The deubiquitylase Ataxin-3 restricts PTEN transcription in lung cancer cells. *Oncogene* **2014**, *33* (33), 4265−72.

(25) Ge, F.; Chen, W.; Qin, J.; Zhou, Z.; Liu, R.; Liu, L.; Tan, J.; Zou, T.; Li, H.; Ren, G.; Chen, C. Ataxin-3 like (ATXN3L), a member of the Josephin family of deubiquitinating enzymes, promotes breast cancer proliferation by deubiquitinating Kruppel-like factor 5 (KLF5). *Oncotarget* **2015**, *6* (25), 21369−78.

(26) Ye, D.; Fu, Y.; Sun, R. X.; Wang, H. P.; Yuan, Z. F.; Chi, H.; He, S. M. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **2010**, *26* (12), i399−406.

(27) Liu, M.; Hu, Z.; Qi, L.; Wang, J.; Zhou, T.; Guo, Y.; Zeng, Y.; Zheng, B.; Wu, Y.; Zhang, P.; Chen, X.; Tu, W.; Zhang, T.; Zhou, Q.; Jiang, M.; Guo, X.; Zhou, Z.; Sha, J. Scanning of novel cancer/testis proteins by human testis proteomic analysis. *Proteomics* **2013**, *13* (7), 1200−10.

(28) Wang, G.; Guo, Y.; Zhou, T.; Shi, X.; Yu, J.; Yang, Y.; Wu, Y.; Wang, J.; Liu, M.; Chen, X.; Tu, W.; Zeng, Y.; Jiang, M.; Li, S.; Zhang, P.; Zhou, Q.; Zheng, B.; Yu, C.; Zhou, Z.; Guo, X.; Sha, J. In-depth proteomic analysis of the human sperm reveals complex protein compositions. *J. Proteomics* **2013**, *79*, 114−22.

(29) Pilarski, L. M.; Gillitzer, R.; Zola, H.; Shortman, K.; Scollay, R. Definition of the thymic generative lineage by selective expression of high molecular weight isoforms of CD45 (T200). *Eur. J. Immunol.* **1989**, *19* (4), 589−97.

(30) Wang, G.; Wu, Y.; Zhou, T.; Guo, Y.; Zheng, B.; Wang, J.; Bi, Y.; Liu, F.; Zhou, Z.; Guo, X.; Sha, J. Mapping of the N-linked glycoproteome of human spermatozoa. *J. Proteome Res.* **2013**, *12* (12), 5750−9.

(31) Rajender, S.; Rahul, P.; Mahdi, A. A. Mitochondria, spermatogenesis and male infertility. *Mitochondrion* **2010**, *10* (5), 419−28.

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

(32) Hackett, N. R.; Butler, M. W.; Shaykhiev, R.; Salit, J.; Omberg, L.; Rodriguez-Flores, J. L.; Mezey, J. G.; Strulovici-Barel, Y.; Wang, G.; Didon, L.; Crystal, R. G. RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC Genomics* **2012**, *13*, 82.
(33) Ross, A. J.; Dailey, L. A.; Brighton, L. E.; Devlin, R. B. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **2007**, *37* (2), 169−85.
(34) Strong, E. R.; Schimenti, J. C. Evidence Implicating CCNB1IP1, a RING Domain-Containing Protein Required for Meiotic Crossing Over in Mice, as an E3 SUMO Ligase. *Genes* **2010**, *1* (3), 440−51.
(35) Brown, J. M.; Dipetrillo, C. G.; Smith, E. F.; Witman, G. B. A FAP46 mutant provides new insights into the function and assembly of the C1d complex of the ciliary central apparatus. *J. Cell Sci.* **2012**, *125* (Pt 16), 3904−13.
(36) Geremek, M.; Zietkiewicz, E.; Bruinenberg, M.; Franke, L.; Pogorzelski, A.; Wijmenga, C.; Witt, M. Ciliary genes are down-regulated in bronchial tissue of primary ciliary dyskinesia patients. *PLoS One* **2014**, *9* (2), e88216.
(37) Georgiadis, A. P.; Kishore, A.; Zorrilla, M.; Jaffe, T. M.; Sanfilippo, J. S.; Volk, E.; Rajkovic, A.; Yatsenko, A. N. High quality RNA in semen and sperm: isolation, analysis and potential application in clinical testing. *J. Urol.* **2015**, *193* (1), 352−9.
(38) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12* (1), 23−7.
(39) Bardou, P.; Mariette, J.; Escudie, F.; Djemiel, C.; Klopp, C. jvenn: an interactive Venn diagram viewer. *BMC Bioinf.* **2014**, *15*, 293.
(40) Clermont, Y. The cycle of the seminiferous epithelium in man. *Am. J. Anat.* **1963**, *112*, 35−51.
(41) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419−1440.
(42) Djureinovic, D.; Fagerberg, L.; Hallstrom, B.; Danielsson, A.; Lindskog, C.; Uhlen, M.; Ponten, F. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Mol. Hum. Reprod.* **2014**, *20* (6), 476−488.
(43) Petit, F. G.; Kervarrec, C.; Jamin, S. P.; Smagulova, F.; Hao, C.; Becker, E.; Jegou, B.; Chalmel, F.; Primig, M. Combining RNA and protein profiling data with network interactions identifies genes associated with spermatogenesis in mouse and human. *Biol. Reprod.* **2015**, *92* (3), 1−18.

DOI: 10.1021/acs.jproteome.6b00400
J. Proteome Res. 2016, 15, 3998−4019

Bibliography

[1] International Human Genome Sequencing Consortium and others. Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).

[2] Luscombe, N. M., Greenbaum, D. & Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine* **40**, 346–358 (2001).

[3] Cook, C. E. *et al.* The European Bioinformatics Institute in 2018: tools, infrastructure and training. *Nucleic Acids Research* **47**, D15–D22 (2018).

[4] Li, Y. & Chen, L. Big biological data: challenges and opportunities. *Genomics, Proteomics & Bioinformatics* **12**, 187 (2014).

[5] Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics* **29**, 365 (2001).

[6] Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology* **26**, 889 (2008).

[7] Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology* **25**, 887 (2007).

[8] Kumari, A., Kanchan, S., Sinha, R. P. & Kesheri, M. Applications of biomolecular databases in bioinformatics. In *Medical Imaging in Clinical Applications*, 329–351 (Springer, 2016).

[9] Rigden, D. J. & Fernández, X. M. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. *Nucleic Acids Research* **48**, D1–D8 (2019).

[10] Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature News* **533**, 452 (2016).

[11] Peng, R. The reproducibility crisis in science: A statistical counterattack. *Significance* **12**, 30–32 (2015).

[12] Freedman, L. P. *et al.* Reproducibility: changing the policies and culture of cell line authentication. *Nature Methods* **12**, 493 (2015).

[13] Masters, J. R. Human cancer cell lines: fact and fantasy. *Nature Reviews Molecular Cell Biology* **1**, 233 (2000).

[14] Kaur, G. & Dufour, J. M. Cell lines: Valuable tools or useless artifacts. *Spermatogenesis* **2**, 1–5 (2012).

[15] Stacey, G. Primary cell cultures and immortal cell lines. *eLS* (2001).

[16] Geraghty, R. *et al.* Guidelines for the use of cell lines in biomedical research. *British Journal of Cancer* **111**, 1021 (2014).

[17] Capes-Davis, A. *et al.* Check your cultures! A list of cross-contaminated or misidentified cell lines. *International Journal of Cancer* **127**, 1–8 (2010).

[18] Shay, J. W. & Wright, W. E. Hayflick, his limit, and cellular ageing. *Nature Reviews Molecular Cell Biology* **1**, 72 (2000).

[19] Gillet, J.-P., Varma, S. & Gottesman, M. M. The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute* **105**, 452–458 (2013).

[20] Lucey, B. P., Nelson-Rees, W. A. & Hutchins, G. M. Henrietta Lacks, HeLa cells, and cell culture contamination. *Archives of Pathology & Laboratory Medicine* **133**, 1463–1467 (2009).

[21] Hnasko, R. M. & Stanker, L. H. Hybridoma technology. In *ELISA*, 15–28 (Springer, 2015).

[22] Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).

[23] Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

[24] Mirjalili, A., Parmoor, E., Bidhendi, S. M. & Sarkari, B. Microbial contamination of cell cultures: a 2 years study. *Biologicals* **33**, 81–85 (2005).

[25] Nikfarjam, L. & Farzaneh, P. Prevention and detection of Mycoplasma contamination in cell culture. *Cell Journal (Yakhteh)* **13**, 203 (2012).

[26] Young, L., Sung, J., Stacey, G. & Masters, J. R. Detection of Mycoplasma in cell cultures. *Nature Protocols* **5**, 929 (2010).

[27] Horbach, S. P. & Halffman, W. The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. *PLoS One* **12**, e0186281 (2017).

[28] Capes-Davis, A. *et al.* Cell Lines as Biological Models: Practical Steps for More Reliable Research (2019).

[29] Hughes, P., Marshall, D., Reid, Y., Parkes, H. & Gelber, C. The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? *Biotechniques* **43**, 575–586 (2007).

[30] Nardone, R. M. Eradication of cross-contaminated cell lines: a call for action. *Cell Biology and Toxicology* **23**, 367–372 (2007).

[31] Lichter, P. *et al.* Obligation for cell line authentication: appeal for concerted action. *International Journal of Cancer* **126**, 1–1 (2010).

[32] Fusenig, N. E., Capes-Davis, A., Bianchini, F., Sundell, S. & Lichter, P. The need for a worldwide consensus for cell line authentication: experience implementing a mandatory requirement at the International Journal of Cancer. *PLoS Biology* **15**, e2001438 (2017).

[33] Defendi, V., Billingham, R., Silvers, W. K. & Moorhead, P. Immunological and karyological criteria for identification of cell lines. *Journal of the National Cancer Institute* **25**, 359–385 (1960).

[34] American Type Culture Collection Standards Development Organization Workgroup ASN-0002 and others. Cell line misidentification: the beginning of the end. *Nature Reviews Cancer* **10**, 441 (2010).

[35] Gartler, S. M. Apparent HeLa cell contamination of human heteroploid cell lines. *Nature* **217**, 750 (1968).

[36] Almeida, J. L., Cole, K. D. & Plant, A. L. Standards for cell line authentication and beyond. *PLoS Biology* **14**, e1002476 (2016).

[37] Huijsmans, R., Damen, J., van der Linden, H. & Hermans, M. Single nucleotide polymorphism profiling assay to confirm the identity of human tissues. *The Journal of Molecular Diagnostics* **9**, 205–213 (2007).

[38] Wyman, A. R. & White, R. A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences* **77**, 6754–6758 (1980).

[39] Jeffreys, A. J., Wilson, V. & Thein, S. L. Hypervariable 'minisatellite'regions in human DNA. *Nature* **314**, 67 (1985).

[40] Jeffreys, A. J., Wilson, V. & Thein, S. L. Individual-specific 'fingerprints' of human DNA. *Nature* **316**, 76 (1985).

[41] Weller, P., Jeffreys, A., Wilson, V. & Blanchetot, A. Organization of the human myoglobin gene. *The EMBO Journal* **3**, 439–446 (1984).

[42] Richard, G.-F., Kerrest, A. & Dujon, B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews* **72**, 686–727 (2008).

[43] Reid, Y., Storts, D., Riss, T. & Minor, L. Authentication of human cell lines by STR DNA profiling analysis. In *Assay Guidance Manual* (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2013).

[44] Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology*, vol. 51, 263–273 (Cold Spring Harbor Laboratory Press, 1986).

[45] Edwards, A., Civitello, A., Hammond, H. A. & Caskey, C. T. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics* **49**, 746 (1991).

[46] Godbey, W. T. *An Introduction to Biotechnology: The Science, Technology and Medical Applications* (Elsevier, 2014).

[47] Clayton, T., Whitaker, J. & Maguire, C. Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Science International* **76**, 7–15 (1995).

[48] Clayton, T. *et al.* Further validation of a quadruplex STR DNA typing system: a collaborative effort to identify victims of a mass disaster. *Forensic Science International* **76**, 17–25 (1995).

[49] Yeung, S. H. *et al.* Rapid and high-throughput forensic short tandem repeat typing using a 96-lane microfabricated capillary array electrophoresis microdevice. *Journal of Forensic Sciences* **51**, 740–747 (2006).

[50] Zahra, A., Hussain, B., Jamil, A., Ahmed, Z. & Mahboob, S. Forensic STR profiling based smart barcode, a highly efficient and cost effective human identification system. *Saudi Journal of Biological Sciences* **25**, 1720–1723 (2018).

[51] Masters, J. R. *et al.* Short tandem repeat profiling provides an international reference standard for human cell lines. *Proceedings of the National Academy of Sciences* **98**, 8012–8017 (2001).

[52] Romano, P. *et al.* Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Research* **37**, D925–D932 (2008).

[53] Bairoch, A. The Cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques* **29**, 25 (2018).

[54] Altelaar, A. M., Munoz, J. & Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics* **14**, 35 (2013).

[55] Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics* **17**, 100–107 (2001).

[56] Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nature Biotechnology* **21**, 255 (2003).

[57] Csizmok, V. & Forman-Kay, J. D. Complex regulatory mechanisms mediated by the interplay of multiple post-translational modifications. *Current Opinion in Structural Biology* **48**, 58–67 (2018).

[58] Bonetta, L. Protein–protein interactions: interactome under construction. *Nature* **468**, 851 (2010).

[59] Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198 (2003).

[60] Sidoli, S. & Garcia, B. A. Middle-down proteomics: a still unexploited resource for chromatin biology. *Expert Review of Proteomics* **14**, 617–626 (2017).

[61] Switzar, L., Giera, M. & Niessen, W. M. Protein digestion: an overview of the available techniques and recent developments. *Journal of Proteome Research* **12**, 1067–1077 (2013).

[62] Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. *Nature Biotechnology* **28**, 695 (2010).

[63] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).

[64] Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry* **60**, 2299–2301 (1988).

[65] Nadler, W. M. *et al.* MALDI versus ESI: the impact of the ion source on peptide identification. *Journal of Proteome Research* **16**, 1207–1215 (2017).

[66] El-Aneed, A., Cohen, A. & Banoub, J. Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers. *Applied Spectroscopy Reviews* **44**, 210–230 (2009).

[67] Hao, Z., Hong, Q., Zhang, F., Wu, S.-L. & Bennett, P. Current Methods for the Characterization of Posttranslational Modifications in Therapeutic Proteins

Using Orbitrap Mass Spectrometry. *Protein Analysis using Mass Spectrometry: Accelerating Protein Biotherapeutics from Lab to Patient* 21 (2017).

[68] Han, X., Aslanian, A. & Yates III, J. R. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology* **12**, 483–490 (2008).

[69] Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society* **120**, 3265–3266 (1998).

[70] Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences* **101**, 9528–9533 (2004).

[71] Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).

[72] Martens, L. *et al.* PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537–3545 (2005).

[73] Moriya, Y. *et al.* The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Research* **47**, D1218–D1224 (2018).

[74] Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **32**, 223 (2014).

[75] Nesvizhskii, A. I. Protein identification by tandem mass spectrometry and sequence database searching. In *Mass Spectrometry Data Analysis in Proteomics*, 87–119 (Springer, 2007).

[76] Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).

[77] Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **10**, 1794–1805 (2011).

[78] Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).

[79] Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).

[80] Craig, R., Cortens, J., Fenyo, D. & Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research* **5**, 1843–1849 (2006).

[81] Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: challenges and opportunities in basic and applied research. *Nature Protocols* **12**, 1289 (2017).

[82] Zhu, W., Smith, J. W. & Huang, C.-M. Mass spectrometry-based label-free quantitative proteomics. *BioMed Research International* **2010** (2009).

[83] Kirkpatrick, D. S., Gerber, S. A. & Gygi, S. P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* **35**, 265–273 (2005).

[84] Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nature Methods* **10**, 186 (2013).

[85] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515 (2018).

[86] Lane, L. *et al.* neXtProt: a knowledge platform for human proteins. *Nucleic Acids Research* **40**, D76–D83 (2011).

[87] Binz, P.-A. *et al.* Proteomics Standards Initiative Extended FASTA Format (PEFF). *Journal of Proteome Research* (2019).

[88] Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).

[89] Cummings, R. D. & Pierce, J. M. The challenge and promise of glycomics. *Chemistry & Biology* **21**, 1–15 (2014).

[90] Varki, A. & Lowe, J. B. Biological roles of glycans. In *Essentials of Glycobiology. 2nd edition* (Cold Spring Harbor Laboratory Press, 2009).

[91] Baum, L. G. & Cobb, B. A. The direct and indirect effects of glycans on immune function. *Glycobiology* **27**, 619–624 (2017).

[92] Varki, A., Kannagi, R. & Toole, B. P. Glycosylation changes in cancer. In *Essentials of Glycobiology. 2nd edition* (Cold Spring Harbor Laboratory Press, 2009).

[93] Varki, A. & Sharon, N. Historical background and overview. In *Essentials of Glycobiology. 2nd edition* (Cold Spring Harbor Laboratory Press, 2009).

[94] Prestegard, J. H., Liu, J. & Widmalm, G. Oligosaccharides and polysaccharides. In *Essentials of Glycobiology. 3rd edition* (Cold Spring Harbor Laboratory Press, 2017).

[95] Han, L. & Costello, C. E. Mass spectrometry of glycans. *Biochemistry (Moscow)* **78**, 710–720 (2013).

[96] Seeberger, P. Monosaccharide diversity. In *Essentials of Glycobiology. 3rd edition* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

[97] Smith, G. & Leary, J. A. Differentiation of stereochemistry of glycosidic bond configuration: Tandem mass spectrometry of diastereomeric cobalt-glucosyl-glucose disaccharide complexes. *Journal of the American Society for Mass Spectrometry* **7**, 953–957 (1996).

[98] Hirayama, H. *et al.* Free glycans derived from O-mannosylated glycoproteins suggest the presence of an O-glycoprotein degradation pathway in yeast. *Journal of Biological Chemistry* jbc–RA119 (2019).

[99] Stanley, P., Taniguchi, N. & Aebi, M. N-glycans. In *Essentials of Glycobiology. 3rd edition* (Cold Spring Harbor Laboratory Press, 2017).

[100] Brockhausen, I., Schachter, H. & Stanley, P. O-GalNAc glycans. In *Essentials of Glycobiology. 2nd edition* (Cold Spring Harbor Laboratory Press, 2009).

[101] Steen, P. V. d., Rudd, P. M., Dwek, R. A. & Opdenakker, G. Concepts and principles of O-linked glycosylation. *Critical Reviews in Biochemistry and Molecular Biology* **33**, 151–208 (1998).

[102] Rojas-Macias, M. A. *et al.* Towards a standardized bioinformatics infrastructure for N-and O-glycomics. *Nature Communications* **10**, 1–10 (2019).

[103] Rudd, P., Karlsson, N. G., Khoo, K.-H. & Packer, N. H. Glycomics and glycoproteomics. In *Essentials of Glycobiology. 3rd edition* (Cold Spring Harbor Laboratory Press, 2017).

[104] Sahoo, S. S., Thomas, C., Sheth, A., Henson, C. & York, W. S. GLYDE—an expressive XML standard for the representation of glycan structure. *Carbohydrate Research* **340**, 2802–2807 (2005).

[105] Aoki, K. F. *et al.* KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Research* **32**, W267–W272 (2004).

[106] Bohne-Lang, A., Lang, E., Förster, T. & von der Lieth, C.-W. LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydrate Research* **336**, 1–11 (2001).

[107] Banin, E. *et al.* A novel linear code® nomenclature for complex carbohydrates. *Trends in Glycoscience and Glycotechnology* **14**, 127–137 (2002).

[108] Lütteke, T. Translation and Validation of Carbohydrate Residue Names with MonosaccharideDB Routines. In *A Practical Guide to Using Glycomics Databases*, 29–40 (Springer, 2017).

[109] Herget, S., Ranzinger, R., Maass, K. & Lieth, C.-W. GlycoCT—a unifying sequence format for carbohydrates. *Carbohydrate Research* **343**, 2162–2171 (2008).

[110] Tanaka, K. *et al.* WURCS: the Web3 unique representation of carbohydrate structures. *Journal of Chemical Information and Modeling* **54**, 1558–1566 (2014).

[111] Harvey, D. J. *et al.* Proposal for a standard system for drawing structural diagrams of N-and O-linked carbohydrates and related compounds. *Proteomics* **9**, 3796–3801 (2009).

[112] Varki, A. *et al.* Symbol nomenclature for graphical representations of glycans. *Glycobiology* **25**, 1323–1324 (2015).

[113] Varki, A. *et al. Nematoda–Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2009).

[114] Alocci, D. *et al.* GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. *Journal of Proteome Research* **18**, 664–677 (2018).

[115] Cooper, C. A., Harrison, M. J., Wilkins, M. R. & Packer, N. H. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Research* **29**, 332–335 (2001).

[116] Artimo, P. *et al.* ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research* **40**, W597–W603 (2012).

[117] Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biology* **13**, R5 (2012).

[118] Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biology* **6**, R21 (2005).

[119] Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

[120] Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Research* **30**, 47–49 (2002).

[121] Schriml, L. M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research* **40**, D940–D946 (2012).

[122] York, W. S. *et al.* MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology* **24**, 402–406 (2014).

[123] Cooper, C. A., Gasteiger, E. & Packer, N. H. GlycoMod–a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* **1**, 340–349 (2001).

[124] Bern, M., Kil, Y. J. & Becker, C. Byonic: advanced peptide and protein identification software. *Current Protocols in Bioinformatics* **40**, 13–20 (2012).

[125] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3** (2016).

[126] Perez-Riverol, Y. *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology* **35**, 406–409 (2017).

[127] Ohno-Machado, L. *et al.* Finding useful data across multiple biomedical data repositories using DataMed. *Nature Genetics* **49**, 816–819 (2017).

[128] Desiere, F. *et al.* The peptideatlas project. *Nucleic Acids Research* **34**, D655–D658 (2006).

[129] Perez-Riverol, Y. *et al.* Quantifying the impact of public omics data. *Nature Communications* **10**, 1–10 (2019).

[130] Cock, P. J. *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

[131] Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80 (2004).

[132] Horlacher, O. *et al.* MzJava: An open source library for mass spectrometry data processing. *Journal of Proteomics* **129**, 63–70 (2015).

[133] Mariethoz, J. *et al.* Glycomics@ ExPASy: bridging the gap. *Molecular & Cellular Proteomics* **17**, 2164–2176 (2018).

[134] Mariethoz, J. *et al.* SugarBindDB, a resource of glycan-mediated host–pathogen interactions. *Nucleic Acids Research* **44**, D1243–D1250 (2015).

[135] Campbell, M. P. *et al.* Validation of the curation pipeline of UniCarb-DB: building a global glycan reference MS/MS repository. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1844**, 108–116 (2014).

[136] Boettiger, C. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review* **49**, 71–79 (2015).

[137] Chervitz, S. A. *et al.* Data standards for Omics data: the basis of data sharing and reuse. In *Bioinformatics for Omics Data*, 31–69 (Springer, 2011).

[138] Bols, N. C., Pham, P. H., Dayeh, V. R. & Lee, L. E. Invitromatics, invitrome, and invitroomics: introduction of three new terms for in vitro biology and illustration of their use with the cell lines from rainbow trout. *In Vitro Cellular & Developmental Biology-Animal* **53**, 383–405 (2017).

[139] Wen, B. *et al.* PGA: an R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. *BMC Bioinformatics* **17**, 244 (2016).

[140] Ruggles, K. V. *et al.* An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Molecular & Cellular Proteomics* **15**, 1060–1071 (2016).

[141] Hu, H., Khatri, K. & Zaia, J. Algorithms and design strategies towards automated glycoproteomics analysis. *Mass Spectrometry Reviews* **36**, 475–498 (2017).

[142] Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308–311 (2001).

[143] Hu, Y., Shah, P., Clark, D. J., Ao, M. & Zhang, H. Reanalysis of global proteomic and phosphoproteomic data identified a large number of glycopeptides. *Analytical Chemistry* **90**, 8065–8071 (2018).

[144] Horlacher, O., Lisacek, F. & Müller, M. Mining large scale tandem mass spectrometry data for protein modifications using spectral libraries. *Journal of Proteome Research* **15**, 721–731 (2015).