Enabling Haplotype-Aware Proteomics to Better Connect Human Genomes and Proteomes

Jakub Vašíček

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2026

UNIVERSITY OF BERGEN

Enabling Haplotype-Aware Proteomics to Better Connect Human Genomes and Proteomes

Jakub Vašíček

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 29.01.2026

© Copyright Jakub Vašíček

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2026

Title: Enabling Haplotype-Aware Proteomics to Better Connect Human Genomes and Proteomes

Name: Jakub Vašíček

Print: Skipnes AS / University of Bergen

Scientific environment

The work in this thesis was carried out at the Department of Clinical Science, University of Bergen, in a research group led by Professor Marc Vaudel. The research group is affiliated with the Mohn Center for Diabetes Precision Medicine and the Computational Biology Unit at the University of Bergen. The candidate was supervised by Professor Marc Vaudel, and co-supervised by Professor Lukas Käll (SciLifeLab, KTH – Royal Institute of Technology, Stockholm, Sweden) and Professor Stefan Bruckner (Chair of Visual Analytics, University of Rostock, Germany).

The candidate conducted several short research visits to Professor Käll at SciLifeLab in Stockholm during the project, and a research stay at the Immunoproteomics laboratory headed by Professor Anthony Purcell at the Biomedicine Discovery Institute, Monash University, Melbourne, Australia from September 2024 until February 2025.

The PhD fellowship and the research stay abroad were granted by the Faculty of Medicine, University of Bergen, with additional funding from the Research Council of Norway (project 301178 to Marc Vaudel) and the Trond Mohn Foundation (Mohn Center for Diabetes Precision Medicine).

Acknowledgements

I would like to express my gratitude to everyone who was involved in this project, and made this work possible.

First, I would like to thank my supervisor Marc Vaudel for all the support in the past years. It was thanks to your trust in my ability to adapt to a new field that I was able to start this PhD. I felt supported and valued through every step of the journey, which is something not to be taken for granted, and that has allowed me to learn and grow both in science and personally. My thanks extend to the whole research group – Dafni, Ksenia, Miguel, and all that come work with us as visitors – you are a joy to be around, and I am grateful to have you as colleagues and friends. The everyday interactions and positive attitude in the whole center, and between the many research groups, not only foster science and learning, but make me look forward to coming to work, which is a great privilege.

I am also thankful to Lukas Käll for his encouraging support, and for including me in his team and even welcoming me in his group in Stockholm a few times. Having experienced one of the best research environments for molecular biology and bioinformatics, and participating in courses and meetings of the MedBioInfo research school, helped me see the broader scope of life sciences, and also sometimes envy the food they get at KI. I thank Stefan Bruckner for the continued collaboration, even though I decided to switch to a different field. Thank you for introducing me to the science in Bergen during my Master's project, and for continuing our work onward. Your ability to provide constructive support even in topics that are outside your usual scope of work is an inspiration.

I am thankful to Tony Purcell and all the many members of the lab at Monash University – Nicole, Murad, Ian, Sanjay, Hayley, Josh, Rochelle, and everyone else. From one unsolicited email to a heap of new knowledge and experiences, the stay in Melbourne was a special time for me. Thanks to you I know that there is so much more to proteomics than just data that come out of the machine. Murad, thank you for showing

me all the steps in the lab, and letting me learn something I have never tried before. Your patience and wisdom is unmatched. I sincerely hope to visit again one day!

None of this would have been possible without support from friends and family. Thanks to my parents, I have learned to approach the world with curiosity, and only with their support I was able to complete my university studies leading up to the PhD. Thank you for being here for me even though I moved to another country.

My thanks go to all friends who helped me feel at home here in Bergen, and to the lovely people at Framtiden i våre hender for having me as a volunteer, and helping me understand Norway a little better. Finally, to my wife Laura, I am beyond grateful to have you by my side. Thank you for listening to me talk about boring things, and for showing me that as long as there's comedy, good food and cute animals in the world, it's all worth it. I promise we will get a cat. Please don't heckle me during my defense.

Bergen, October 2025

Jakub Vašíček

List of abbreviations

1kGP – 1000 Genomes Project

DDA – data-dependent acquisition

DIA – data-independent acquisition

DNA – deoxyribonucleic acid

FAIR – findability, accessibility, interoperability, and reusability

FDR – false discovery rate

GRC – Genome Reference Consortium

GWAS – genome-wide association study

HGVS – Human Genome Variation Society

HLA – human leukocyte antigen

HPRC – Human Pangenome Reference Consortium

HRC – Haplotype Reference Consortium

LC – liquid chromatography

LD – linkage disequilibrium

NGS – next-generation sequencing

MS – mass spectrometry

MS/MS – tandem mass spectrometry

OSS – open-source software

ORF – open reading frame

PEP – posterior error probability

pQTL – protein quantitative trait locus

PRS – polygenic risk score

PSM – peptide-spectrum match

RNA – ribonucleic acid

SNP – single nucleotide polymorphism

UTR – untranslated region

USI – universal spectrum identifier

VCF – variant call (file) format

WES – whole-exome sequencing

WGS – whole-genome sequencing

Abstract in English

Mass spectrometry-based proteomics identifies peptides by matching spectra against a sequence database, which for humans is typically a collection of reference protein sequences. However, this approach overlooks protein-altering genetic variants, making peptides encoded by alternative sequences undetectable even when they are present in the sample and captured by the instrument. Proteogenomics integrates genomic and proteomic analyses to address this limitation, developing tools to create protein sequence databases from lists of individual variants or sequencing data. However, such tools have primarily been designed to study rare or pathogenic variation, and the influence of common haplotypes on the human proteome is still largely unexplored.

In **Paper 1**, we investigated how protein haplotypes—unique protein sequences encoded by combinations of alleles inherited together from a parent—affect the proteomic search space and the detectability of variant peptides. We introduced the concept of the multivariant peptide, defined as a peptide encoded by two or more alternative alleles within the same haplotype. Using a published protein haplotype database, accounting exclusively for single nucleotide polymorphisms within the 1000 Genomes Project, we estimated that 7.82% of the human proteome can be mapped to common variant peptides, and up to 12.42% of the amino acid substitutions are discoverable in multivariant peptides. Reanalysis of raw proteomic data from healthy tonsil tissue identified thousands of variant peptides, including multivariant peptides, and demonstrated that haplotype-aware databases may prevent misassignment of spectra to canonical proteins.

Paper 2 introduces ProHap, a Python-based software pipeline for generating haplotype-resolved protein sequence databases from phased genotype datasets. These databases account not only for amino acid substitutions, but include insertions, deletions, frameshifts, and loss of stop codons. Using ProHap, we generated databases from the 1000 Genomes Project, the Haplotype Reference Consortium, and the Human Pangenome Reference Consortium datasets. In the database based on the 1000 Genomes Project, we found that at least 9% of the proteome in all major popula-

tions could be attributed to variant peptides, with higher diversity in individuals with African ancestry. Application to a blood plasma proteomic dataset enabled identification of specific variant peptides in multiple individuals, and personalized analysis of stem cell data using donor genotypes confirmed the detection of peptides encoded by both the reference and alternative allele at heterozygous loci.

Paper III presents ProHap Explorer, a web-based visualization platform for integrative analysis of proteogenomic data. The tool shows variation across gene, transcript, and protein, and maps peptides identified in a public mass spectrometry dataset to their genomic origin. ProHap Explorer features interactive views for both global and gene-level exploration, with exportable data tables. The tool enables detailed exploration and validation of variant peptides, building on the findings and resources developed in Papers 1 and 2.

By integrating haplotype-aware database generation, population-based analysis, and interactive visualization into proteomic studies, these contributions enable deeper insights into the diversity of the human proteome, and open new opportunities for interrogating its functional implications in biology and medicine.

Sammendrag

Massespektrometri-basert proteomikk identifiserer peptider ved å sammenligne spektra mot en sekvensdatabase, som for mennesker vanligvis er en samling av referanse-proteinsekvenser. Denne tilnærmingen overser imidlertid proteinendrende genetiske varianter, noe som gjør at peptider kodet av alternative sekvenser ikke blir oppdaget selv om de er til stede i prøven og kan fanges opp av instrumentet. Proteogenomikk integrerer genomiske og proteomiske analyser for å løse denne begrensningen, og utvikler verktøy for å lage proteinsekvensdatabaser fra lister over individuelle varianter eller sekvenseringsdata. Slike verktøy har hovedsakelig vært utviklet for å studere sjeldne eller sykdomsrelaterte varianter, og innflytelsen av vanlige haplotyper på det menneskelige proteomet er fortsatt understudert.

Artikkel 1 undersøkte hvordan protein-haplotyper—unike proteinsekvenser kodet av kombinasjoner av alleler som arves sammen fra en forelder—påvirker proteomet og deteksjonen av variantpeptider. Vi introduserte begrepet multivariant peptid, definert som et peptid kodet av to eller flere alternative alleler innenfor samme haplotype. Ved å bruke en protein-haplotype-database, som utelukkende tar hensyn til enkelt-nukleotid-polymorfismer fra 1000 Genomes Project, estimerte vi at 7,82% av det menneskelige proteomet kan knyttes til vanlige variantpeptider, og opptil 12,42% av aminosyresubstitusjonene kan oppdages i multivariant-peptider. Reanalyse av rå proteomikkdata fra friskt mandelvev identifiserte tusenvis av variantpeptider, inkludert multivariant-peptider, og viste at haplotype-bevisste databaser kan forhindre feiltilordning av spektra til kanoniske proteiner.

Artikkel 2 introduserer ProHap, en Python-basert programvarepipeline for å generere haplotype-beviste proteinsekvensdatabaser fra fasete genotypedata. Disse databasene tar ikke bare hensyn til aminosyresubstitusjoner, men inkluderer også insersjoner, delesjoner, rammeskift og tap av stoppkodon. Ved bruk av ProHap genererte vi databaser fra 1000 Genomes Project, Haplotype Reference Consortium og Human Pangenome Reference Consortium. I databasen basert på 1000 Genomes Project fant vi at minst 9% av proteomet i alle større populasjoner kunne tilskrives variantpeptider, med høyere

diversitet hos individer med afrikansk opphav. Ved anvendelse av ProHap på et proteomikkdatasett på blodplasma klarte vi å identifisere spesifikke variantpeptider hos flere individer. Videre bekreftet vi deteksjon av peptider kodet av både referanse- og alternativt allel på heterozygote loci ved bruk av personspesifikke stamcelldata med donorens genotyper.

Artikkel 3 presenterer ProHap Explorer, en nettbasert visualiseringsplattform for integrert analyse av proteogenomiske data. Verktøyet viser variasjon på tvers av gen, transkript og protein, og kobler peptider identifisert i et fritt tilgjengelig massespektrometri-datasett til deres genomiske opphav. ProHap Explorer kan visualisere data både globalt og på gennivå, med eksporterbare datatabeller. Verktøyet muliggjør detaljert utforskning og validering av variantpeptider, og bygger på funnene og ressursene utviklet i de to første artikkelene.

Ved å integrere haplotype-bevisst databasegenerering, populasjonsbasert analyse og interaktiv visualisering i proteomikkstudier, gir disse bidragene dypere innsikt i mangfoldet i det menneskelige proteomet og åpner nye muligheter for å undersøke dets funksjonelle betydning for biologi og medisinsk forskning.

List of publications

The following publications are included in this thesis:

Paper 1

Vašíček, J., Skiadopoulou, D., Kuznetsova, K. G., Wen, B., Johansson, S., Njølstad, P. R., Bruckner, S., Käll, L., Vaudel, M. **Finding Haplotypic Signatures in Proteins**. *GigaScience* 2023, 12, giad093.

Paper 2

Vašíček, J., Kuznetsova, K. G., Skiadopoulou, D., Unger, L., Chera, S.;, Ghila, L. M., Bandeira, N., Njølstad, P. R., Johansson, S., Bruckner, S., Käll, L., Vaudel, M. **ProHap Enables Human Proteomic Database Generation Accounting for Population Diversity**. *Nature Methods* 2025, 12, 273–277.

Paper 3

Vašíček, J., Skiadopoulou, D., Kuznetsova, K. G., Käll, L., Vaudel, M., Bruckner, S. **ProHap Explorer: Visualizing Haplotypes in Proteogenomic Datasets**. *IEEE Computer Graphics and Applications* 2025, 45(5), 64-77.

The reprint of Paper 1 is permitted under the CC-BY 4.0 license. The reprint of Paper 2 and Paper 3 is permitted by Springer Nature and IEEE. All rights reserved.

The following publications are related to this thesis:

Additional Paper 1

Skiadopoulou, D., Vašíček, J., Kuznetsova, K., Bouyssié, D., Käll, L., Vaudel, M. **Retention Time and Fragmentation Predictors Increase Confidence in Identification of Common Variant Peptides**. *J. Proteome Res.* 2023, 22 (10), 3190–3199.

Additional Paper 2

Kuznetsova, K. G., Vašíček, J., Skiadopoulou, D., Molnes, J., Udler, M., Johansson, S., Njølstad, P. R., Manning, A., Vaudel, M. **Bioinformatics Pipeline for the Systematic Mining Genomic and Proteomic Variation Linked to Rare Diseases: The Example of Monogenic Diabetes**. *Plos one* 2024, 19 (4), e0300350.

Contents

1 Introduction

1.1 Precision medicine and omics technologies

Modern medicine aims to understand disease and the human body at a deep level, integrating insights from molecular biology and clinical research to develop more targeted therapies and treatments[1]. This approach has led to significant advances in the diagnosis and management of complex conditions.

Just over 20 years ago, a 15-month-old boy was admitted to a hospital in Wisconsin, USA, with a severe Crohn's disease-like autoimmune condition[2]. After initial treatment and maintenance, he returned at the age of four with new, severe complications. When all standard forms of treatment proved unsuccessful, the medical team identified genetic variants responsible for the disease. Based on this finding, a cell transplant was recommended, and the patient subsequently regained the ability to eat and drink, with no further recurrence of the disease recorded[2]. This case was one of the first instances where sequencing technology was applied directly in the clinic, enabling timely diagnosis and treatment of a potentially life-threatening condition.

Around the same time, it was discovered that for a substantial proportion of diabetes cases diagnosed before three months of age and requiring permanent treatment, the release of insulin from the insulin-producing cells in the pancreas is impaired due to rare variants in a single gene[3]. Following the discovery of the exact molecular mechanism, an alternative oral medication was proposed instead of permanent insulin injections[4]. Similar mechanisms were observed in several types of maturity-onset diabetes of the young (MODY)[5,6], and patients with both subclasses of diabetes can now be routinely identified by genetic testing[7,8], and benefit from the improved treatment with lower impact on their quality of life.

These two examples illustrate the principles of *precision medicine* – a term popularized by President Obama in his 2015 State of the Union address, which outlined the vision for a national Precision Medicine Initiative in the United States[1]. While therapeutics are rarely developed for single individuals (an approach known as *personalized medicine*), precision medicine defines subgroups of individuals and develops targeted treatments for these groups[1,9].

Discoveries like these have been greatly enabled by the development of novel methods in molecular biology and advances in downstream data interpretation. The mechanistic understanding of the subclasses of diabetes, followed by the implementation of genetic testing in the diagnostic process, is allowed by the ability to systematically characterize genetic variants at a decreasing cost. Similarly, the precise diagnosis of the rare inflammatory bowel disease was achieved thanks to the ability to rapidly sequence all protein-coding genes in an individual and compare them to reference sequences.

In another example, the principle of tailoring medical advice to specific subgroups becomes crucial. Genetic testing is now routinely used in risk stratification for a heart condition called hypertrophic cardiomyopathy (HC)[10]. Between 2005 and 2007, several individuals of either African or unspecified ancestry received reports that they carried pathogenic variants associated with HC. In such cases, patients and their relatives are advised to modify their lifestyles and undergo prolonged risk screenings, which may cause stress and economic burden. However, the genetic variants originally identified as pathogenic were actually common among the Black American population, while being rare among individuals of European ancestry, and were later reclassified as benign[10].

Instead of targeting advice and treatment to a specific group, an assumed "one-size-fits-all" reference panel was used, violating the principle of precision medicine. This highlighted the problem of lacking diversity in genetic studies[11]; had individuals of African ancestry been included in the control cohorts, the misclassification of variants could have been prevented[10].

1.1.1 High-throughput omics technologies

An essential asset of modern biomedical research is methodologies referred to as omics. These are large-scale approaches aiming for comprehensive measurements of biological entities within cells, tissues, or organisms. Such approaches aim to capture the entirety of a specific type of biological information, such as genes (genomics), RNA transcripts (transcriptomics), proteins (proteomics), or metabolites (metabolomics), among others[12–14] (Figure 1.1).

The initial completion of the Human Genome Project catalyzed a new research paradigm, referred to as *discovery science*[15], and currently known as *data-driven* approach. Complementary to *hypothesis-driven* science, data-driven approaches provide an unbiased view by considering all the elements of a biological system (such as genes) without prior assumptions about their function[15]. Omics technologies allow us not only to determine the elements of a system, but also to generate hypotheses about their biologically important properties, which can be tested in hypothesis-driven research[15,16].

The frequent use of these approaches in medical research and clinical practice is enabled by high-throughput experimental technologies. These technologies, such as next-generation sequencing (NGS), are able to analyze thousands to millions of molecules in parallel at a rapidly

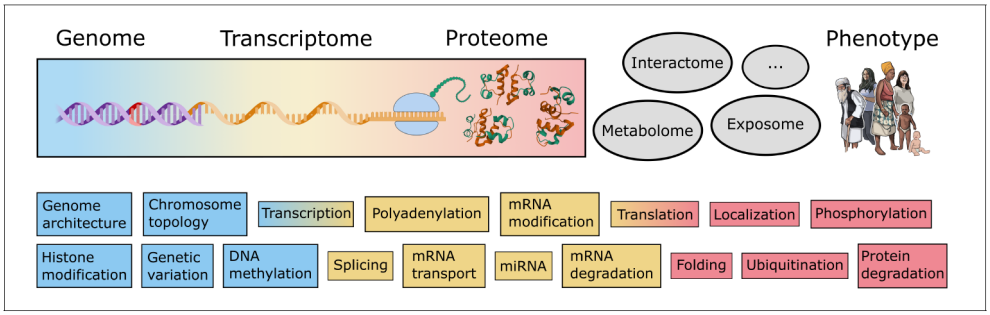


Figure 1.1: The path from the genome to the observable and measurable characteristics or traits of an organism, known as the phenotype. Figure adapted from[20], illustrations downloaded from NIH BioArt Source (bioart.niaid.nih.gov), protein model downloaded from Protein Data Bank (PDB)[21] entry 5HQI.

decreasing cost[13–17]. These massively-parallel technologies allow for the discovery of biomarkers (measurable indicators of biological processes or responses to interventions[18]), disease mechanisms, and therapeutic targets that would be missed by traditional approaches[13,17,19].

1.1.2 The role of computer science

Handling the data produced by high-throughput instruments required establishing downstream software tools that enable researchers to analyze and interpret the vast volumes of data generated by experiments[12,22]. Before the advent of high-throughput technologies, software and algorithms were primarily employed to align sequenced DNA fragments and assemble genomic maps[23]. To manage the computational complexity of these early genomic efforts, techniques such as Yeast Artificial Chromosomes and Bacterial Artificial Chromosomes allowed scientists to efficiently clone, map, and organize large fragments of human DNA, which could then be systematically sequenced[24]. This proved instrumental in the success of the Human Genome Project.

As omics data grew in volume and heterogeneity, the need for more advanced computational approaches and infrastructure became clear[12,25,26]. Key algorithmic innovations in sequence alignment, such as the use of suffix trees, followed by Burrows-Wheeler Transform-based methods, have greatly improved computational speed and memory efficiency, enabling the growing scope and scale of genomics[27,28].

In state-of-the-art biomedical research, the overall success of a project depends as much on data interpretation as on sample processing[22]. Furthermore, data generated by experiments is commonly deposited into shared repositories, with data volumes growing exponentially[12,29,30]. Increasingly, re-analysis of existing datasets using new tools and approaches can yield novel biological insights[29,30], reducing both the cost and labor required for new research projects. As a result, fluency in computational methods is often essential for biologists to perform their work effectively. While multidisciplinary teamwork is common, a disconnect between data analysis

and interpretation can lead to errors, such as mistaking artifacts for genuine findings[31]. It is therefore crucial to develop scientific software that is user-friendly and well-documented[31–33].

Taken together, these developments highlight the increasingly central role of computational methods in modern biomedical research. As the field continues to evolve, the integration of robust software tools and interdisciplinary expertise will be essential for translating complex data into meaningful discoveries.

1.2 Human genetics and genomics

The study of the laws of heredity, which evolved into the field of genetics, has occupied biologists for over a century[23]. Genetics focuses on genes as the fundamental units responsible for passing on traits, studying the modes of transmission of traits between generations, mechanisms by which genes affect the physical properties of organisms, and the distribution of heritable traits within populations[34]. Major milestones include the discovery of chromosomes, the structure of the DNA double helix[35], and the development of DNA sequencing technologies[36,37].

Early sequencing methods could only analyze short stretches of DNA and RNA[37]. However, advances in computational sequence alignment and increased sequencing speed eventually made it possible to sequence the complete DNA of entire organisms (i.e., their genomes)[38] (Figure 1.2). This enabled the emergence of the field of genomics, which studies the genome in its entirety, an approach that is better suited to address complex traits resulting from the combined effect of many genes and environmental factors.

The first genomes to be sequenced were those of bacteria, viruses, and yeast[37]. In 2001, the first draft of the human genome was published, covering 94% of the genome and requiring 15 months to complete. However, its completion was preceded by more than a decade of preparation and legwork[23]. Two years later, a sequence covering 99% of the human genome was published[39], and the Human Genome Project was declared essentially complete.

1.2.1 The reference human genome

The first human genome was assembled from the sequencing of samples from 20 donors; however, 70% of the sequence was derived from a single individual[41]. As a result, the published human genome is not a true “standard”, but rather an assembly of long stretches of individual genomes. For example, it has been shown that the donor most represented in the initially published human genome sequence had a high risk for developing type 1 diabetes[42].

Nevertheless, this published sequence is referred to as the *reference genome* and serves as the

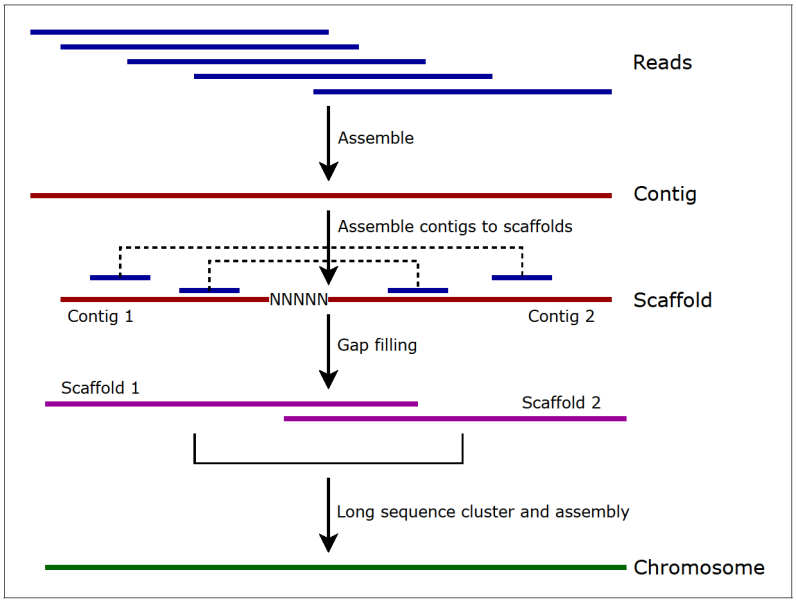


Figure 1.2: *De novo* assembly of a genome. Figure adapted from[40].

foundation for most scientific and clinical work in human genomics. The Genome Reference Consortium (GRC) is responsible for curating and improving the reference human genome[43]. Each new build incorporates additional data, corrects errors, and may include alternative representations of complex regions. Consistency is ensured through the standardization of genome assembly steps, software tools, and procedures[44].

Of particular importance was the 19th release of the human genome by the GRC, known as GRCh37 or HG19, published in 2009[44,45]. While it offered higher overall quality and fewer gaps compared to previous versions, this reference genome is derived from about 13 individuals[46]. GRCh37 has been widely used for the analysis of high-throughput sequencing data, especially with the advent of Illumina's technology[40]. As a result, GRCh37 is still commonly used today, even though more recent assemblies are available. The 20th release, GRCh38, was published in 2013 and represents a further improvement in quality, enabling more accurate analysis of human genomic sequencing data[40].

Despite these advances, some regions of the genome, such as centromeric regions and regions with highly repetitive sequences, remained unresolved. In 2022, the Telomere-to-Telomere (T2T) Consortium announced the first truly complete human genome sequence, known as T2T-CHM13[47]. This assembly is derived from the CHM13hTERT cell line, which is uniformly homozygous, meaning both copies of the genome are identical. The T2T-CHM13 genome filled in all previously missing regions, including centromeres and the short arms of chromosomes, providing the most comprehensive and accurate human reference genome to date[47].

1.2.2 Genetic variants and their consequences

With NGS technology becoming exponentially cheaper and faster, by 2015, it was estimated that almost a million human genomes had been sequenced[48]. When a genome is sequenced using NGS, the resulting reads are aligned to the reference human genome to identify differences between the reference and the sequenced genome. This process is known as *variant calling*[49], and the identified differences are referred to as variants. A variant *locus* is a specific, fixed position on a chromosome where genetic variation is observed. This variation can range from a single base pair (bp) to several thousand base pairs in length. An *allele* is one of the possible genetic sequences that can exist at a given locus. In diploid chromosomes, such as the non-sex chromosomes (called *autosomes*) in humans, an individual is expected to have two copies of each locus. If both copies carry the same allele, the individual is *homozygous* at that locus; if the two copies carry different alleles, the individual is *heterozygous* at that locus. The genetic makeup of an individual – the set of all alleles at their variant loci – is called the *genotype*. The process of determining an individual's genotype is known as *genotyping*[50].

To assess the significance of different loci in the genome and the potential consequences of genetic variation, segments and features of the genome are annotated with their presumed roles in biological processes[51]. The most notable annotated segments are genes – regions of the genome that code for proteins or functional RNA molecules (Figure 3)[51–53]. While early studies estimated between 25,000 and 40,000 human genes, current evidence suggests there are between 19,000 and 20,000 protein-coding genes[52,54,55], which account for only one to two percent of the entire human genome[56]. The vast majority of genetic variant loci are therefore located within intergenic regions, where they can still exert regulatory influence on gene expression. These regulatory effects are often mediated through mechanisms such as altered binding and post-translational modification of histone proteins[57], changes in DNA methylation patterns, or disruptions of enhancer and silencer elements. These processes can modulate the transcription levels of nearby or even distant genes[56,58].

The architecture of genes is further subdivided into *exons* and *introns* (see Figure 1.3). Exons are sequences that remain in mature RNA following transcription and RNA splicing[59,60]; in messenger RNA (mRNA), they directly encode the amino acid sequence of proteins. In contrast, introns are sequences located between exons that are transcribed into RNA but removed during splicing, and do not contribute directly to the final RNA or protein product. Since introns are typically several kilobases long, while exons usually span only a few hundred base pairs[51], the majority of variants mapping to gene regions are found within introns. These intronic variants can influence the regulation of transcription, affect mRNA processing, or contribute to alternative splicing (see Figure 1.4)[60]. While *whole-genome sequencing* (WGS) aims to cover the entire genome, *whole-exome sequencing* (WES) focuses specifically on capturing all exons of all known genes[61].

Within exons, a critical distinction exists between regions that are translated into protein and

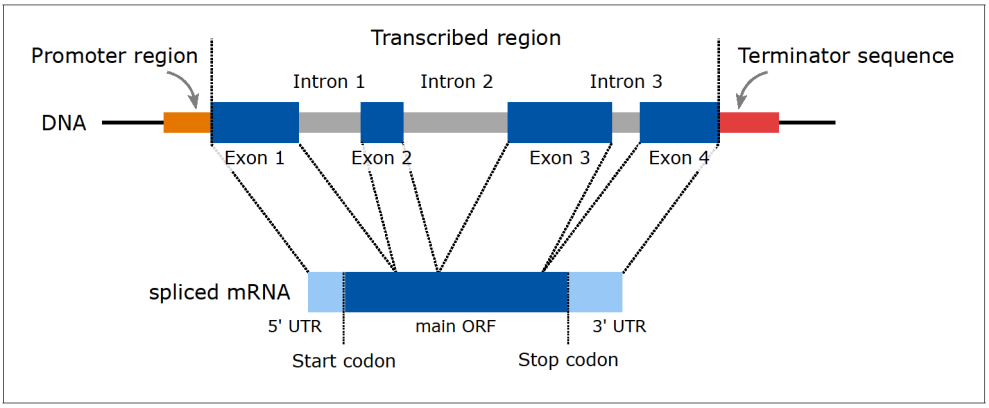


Figure 1.3: Canonical structure of the eukaryotic gene. Introns and exons are transcribed into RNA, followed by the removal of intron regions during splicing. Mature spliced mRNA is divided into the untranslated regions (UTRs) and the main open reading frame (ORF), which is translated into protein. Figure adapted from[62].

regions that are not. The portion of the spliced transcript that is translated into a functional protein is referred to as an *open reading frame* (ORF). Within the ORF, three nucleotides always code for one amino acid or signal the termination of translation; these triplets are known as codons. As there are 64 possible codons, but only 21 amino acids, multiple different codons can encode the same amino acid. Three codons, known as *stop codons*, do not encode an amino acid but signal the end of translation. The main ORF of a transcript defines the annotated protein-coding sequence, beginning with a start codon (AUG) and ending at the nearest stop codon[63].

Sequences outside the boundaries of the main ORF, but still part of the mature mRNA, are known as untranslated regions (UTRs). While traditionally considered non-coding, recent advances have led to the annotation of novel small ORFs within UTRs[64,65], some of which have been shown to be translated, suggesting that the functional output of these regions may be more complex than previously thought. Genetic variants located in UTRs can impact gene expression at the post-transcriptional level. Such variants may disrupt motifs required for ribosome recognition and binding, alter mRNA stability and localization, or affect regulatory elements within UTRs, such as microRNA binding sites and promoter regions[58]. Changes in these regions predominantly influence the efficiency of translation rather than the protein sequence itself.

In contrast, variants within the main ORF directly modify the canonical protein-coding sequence, potentially resulting in amino acid substitutions, insertions or deletions, premature termination, or frameshifts. While these variants can have profound consequences for the structure and function of the encoded protein, contributing to altered cellular processes or disease phenotypes, most of them are not pathogenic and are commonly observed both between and within human populations[66].

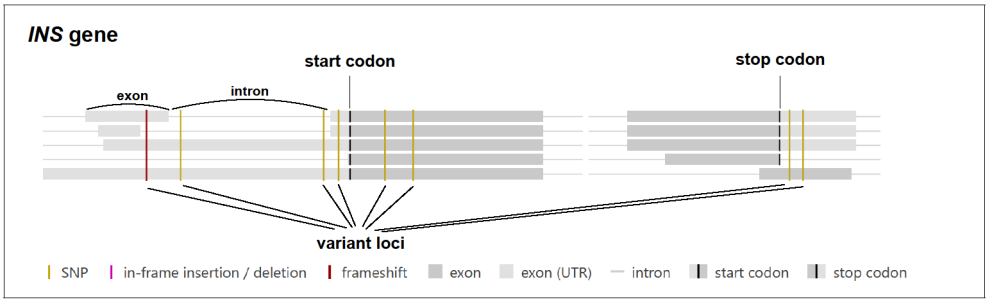


Figure 1.4: Five different splicing alternatives of the insulin gene (*INS*), with annotated genetic variants.

The variants inherited from our parents, called *germline* variants, are present in every cell of our body and can be passed on to future generations. Germline changes affect the entire organism's genome. Conversely, new genetic variants also continuously arise throughout an individual's lifetime via random events - these are known as *somatic* or *de novo* variants. Somatic variants appear in non-reproductive cells and are not inherited or passed to offspring; instead, they can result in cellular mosaicism within the body, and their accumulation is linked to diseases such as cancer[67,68]. Variants which newly arise either in the sperm or egg of the parent prior to fertilization are known as *prezygotic de novo* variants, and are found in every cell of the offspring[69].

1.2.3 Variant classification

Genetic variants can be classified in two principal ways: by their pathogenicity (clinical significance) and by their molecular consequence (effect on the transcript or protein sequence). The most widely adopted scheme for classifying pathogenicity was developed by the American College of Medical Genetics and Genomics (ACMG)[70]. This system categorizes genetic variants into five classes based on the likelihood that a variant causes or contributes to disease: 'pathogenic,' 'likely pathogenic,' 'uncertain significance,' 'likely benign,' and 'benign'[70]. Variants of uncertain significance (VUS) are those for which there is insufficient or conflicting evidence regarding their role in disease. As of 2023, 41% of the over 2 million variants in the ClinVar database were classified as VUS or conflicting[71], highlighting that the classification of variant pathogenicity remains a major challenge in modern medical research[72,73].

Variants, regardless of their pathogenicity, can also be classified based on their effect on the gene, transcript, or protein sequence. The nomenclature for sequence variants was developed by the Human Genome Variation Society (HGVS), providing an essential tool for consistency in reporting genetic variation[74,75]. HGVS offers standardized variant descriptions at three levels: DNA, RNA, and protein. In short, all three levels describe six common types of variants: substitution, deletion, duplication, insertion, inversion, and deletion-insertion (indel)[74]. Additionally, variants at the protein level can be described as frameshifts if the number of bases inserted or deleted from a coding sequence is not a multiple of three. Frameshift variants

completely alter the reading frame downstream and, in most cases, introduce an early stop codon[74,75].

The HGVS nomenclature differs from other commonly used terms by strictly focusing on sequence alterations and avoiding terminology that implies biological interpretation[74,75]. For example, substitutions in DNA sequences are commonly referred to as *single nucleotide polymorphisms* (SNPs)[76,77]. At the RNA level, substitutions are often described as *missense* variants, where the altered codon encodes a different amino acid[73,78] or *nonsense* variants, where a codon encoding an amino acid is changed to a stop codon, resulting in a truncated protein[78,79] (see Figure 1.5). Substitutions at the RNA level that do not change the encoded amino acid sequence are referred to as synonymous variants[80,81].

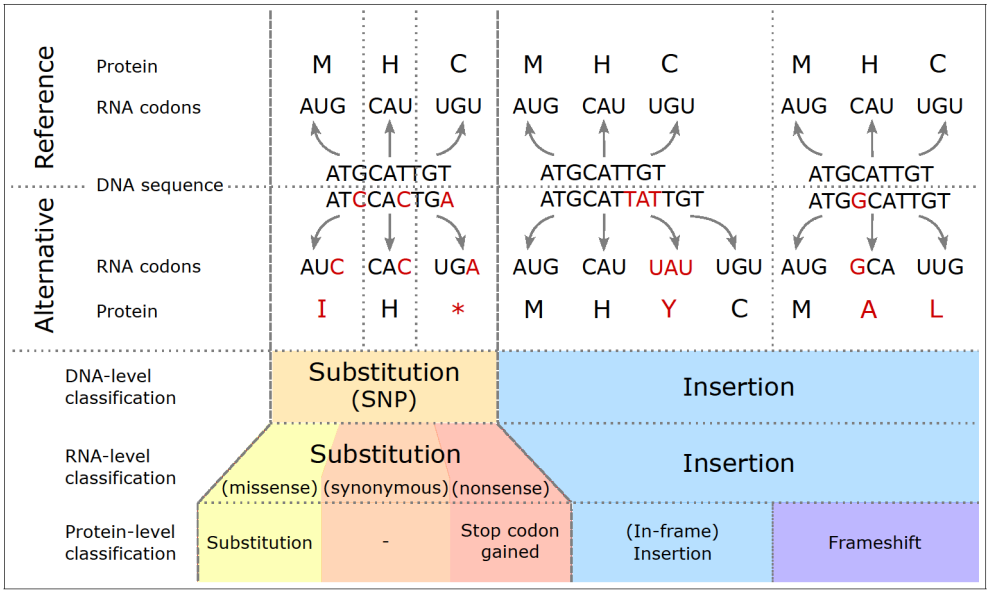


Figure 1.5: Classification of variants on the level of DNA, RNA, and protein. While HGVS classifies variants arising from a single nucleotide change as substitutions on all three levels[74], other commonly used terms for variant consequences on the RNA and protein sequence are illustrated here. The asterisk sign (*) denotes the stop codon.

Similarly to the term *polymorphism*, HGVS discourages the use of the term *mutation*, as it implies pathogenicity[75]. The nomenclature also recommends standardized numbering systems for nucleotides or amino acids within the reference sequence at multiple levels, each indicated by a specific prefix in the variant description (see Table 1.1)[74]. The process of mapping the genomic coordinates of features from one reference build to another (e.g. from GRCh37 to GRCh38) is known as *liftover*[82,83].

Level	Variant description prefix	Position relative to
Genomic DNA	g.	First nucleotide of the genomic reference sequence
Coding DNA	c.	First nucleotide of the translation start codon of the coding DNA reference sequence
RNA	r.	First nucleotide of the translation start codon of the RNA reference sequence or first nucleotide of the noncoding RNA reference sequence
Protein	p.	First amino acid of the protein sequence

Table 1.1: Numbering of positions in the reference sequence as defined by HGVS. Table adapted from[74]

1.2.4 Haplotypes and phasing

While genetic variants are often described as individual loci scattered across the genome, in reality, they are not inherited in isolation. Instead, alleles of germline variants are passed down in groups that originate from the same parental chromosome. This non-random distribution of alleles is known as *linkage disequilibrium* (LD)[84]. The specific combinations of alleles present on the same physical copy of a chromosome are called *haplotypes*[85,86]. In the regular human genome, there are two haplotypes for each of the 22 autosomes: one inherited from the mother and one from the father. For the sex chromosomes, females carry two haplotypes on chromosome X, while males carry one haplotype on chromosome X and one on chromosome Y. Importantly, the X and Y chromosomes share small homologous segments known as *pseudoautosomal regions* (PARs) at both ends (the 5' and 3' ends), where their sequences are identical and can recombine during meiosis[87,88].

When a genome is sequenced, the reads are aligned to a reference genome and variants are called, producing a list of loci where the sequenced genome differs from the reference. At this stage, it is possible to determine for each variant whether an individual is homozygous (both alleles identical) or heterozygous (two different alleles). However, this information alone does not reveal which alleles are located together on the same parental chromosome – in other words, we lack the assignment of variants to the paternal and maternal haplotypes[89]. To reconstruct this arrangement and identify which variants are inherited together, a process called *phasing* is required. Phasing resolves the chromosome-specific pattern of alleles, enabling a haplotype-level view of the genome that is essential for understanding inheritance, compound heterozygosity, and the combined effects of multiple variants within the same haplotype[89,90].

Phasing can be achieved directly using individual sequencing reads – a method known as read-based phasing – which relies on sequencing technology to provide physical evidence for haplotype structure[90,91] (Figure 1.6). This approach typically offers high accuracy and low er-

ror rates but is limited by read length and coverage gaps. Statistical phasing, also known as population-based phasing, is an indirect approach that uses large population datasets to impute haplotypes. While it can cover genome regions not connected by sequencing reads, it may perform poorly for rare variants or in genetically diverse populations[86,91].

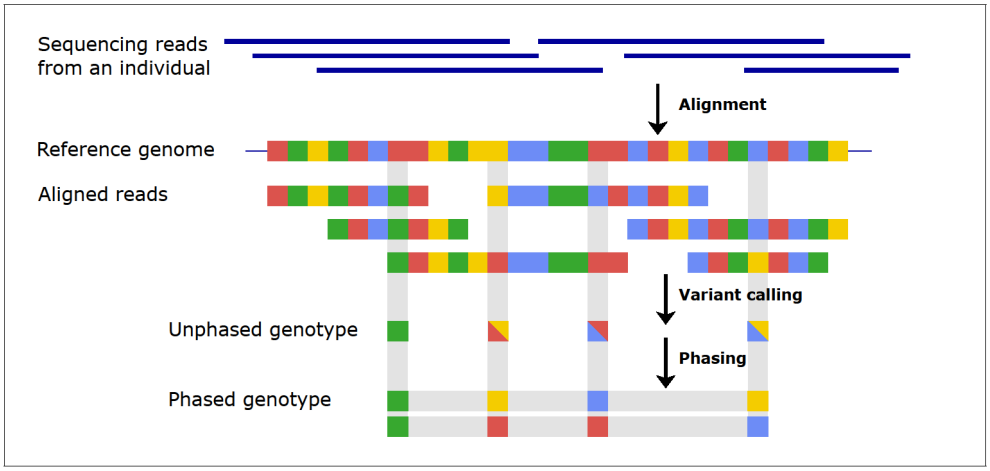


Figure 1.6: Overview of variant calling and read-based phasing. Figure adapted from[92].

1.2.5 Population-wide genomic data and genetic epidemiology

Large-scale genotype datasets have enabled new insights into human evolutionary history and admixture[93,94], and, when combined with clinical, phenotypic, or environmental data, provide unique opportunities to deepen our understanding of complex traits and disorders[95]. Genome-wide association studies (GWAS) are a central approach in genetics, investigating the entire genome of large groups of individuals to identify genetic variants statistically associated with specific traits or diseases[96]. Unlike traditional candidate gene studies, GWAS helps remove bias in gene selection by scanning the entire genome without prior assumptions, thereby targeting the actual genes associated with diseases[97]. The genetic makeup of a disease is known as its genetic architecture, which encompasses the number of genetic variants involved, and the magnitude of their effects on the phenotype.

So-called *polygenic* diseases arise from the combined influence of many genetic variants, each contributing a small effect to the overall phenotype[97]. On the other hand, in *monogenic* diseases, few risk variants in a single gene with a large effect disrupt a physiological pathway[98]. Such large-effect variants tend to be rare in populations due to the negative selection pressures they introduce (see Figure 1.7). While the genes having extreme effects on diseases, causing monogenic phenotypes, can be different from those with mild influence, causing complex traits; there can be some overlap or crosstalk between monogenic disease-associated genes and those that contribute to polygenic risk[97,98].

Each variant identified in GWAS is assigned an effect size value, which can be used to quantify the aggregated risk of developing a disease for a given individual. The sum of the number of risk alleles in an individual, weighted by their effect size, is referred to as a *polygenic risk score* (PRS)[99–101]. A valuable resource for GWAS analysis and PRS development is the UK Biobank, which includes genotyping data from over 490,000 participants, as well as extensive phenotypic data, environmental exposures, and electronic health records[95,102]. Additionally, the Genome Aggregation Database (gnomAD) presents a dataset of genetic variants aggregated from over 140,000 human exomes and genomes[72]. GWAS and PRS predictions enabled by such resources are now available for most non-communicable disorders with major public health impact[99,103].

Variants are often characterized by their frequency within a population – common variants typically have small individual effect sizes, while rare variants are much less prevalent but may have larger effects on traits and diseases. Detecting low-frequency variants in GWAS is challenging because most of them are not captured by genotyping arrays. As a result, standard GWAS approaches are better suited for identifying common variants of modest effect than rare variants of large effect[104] (Figure 1.7).

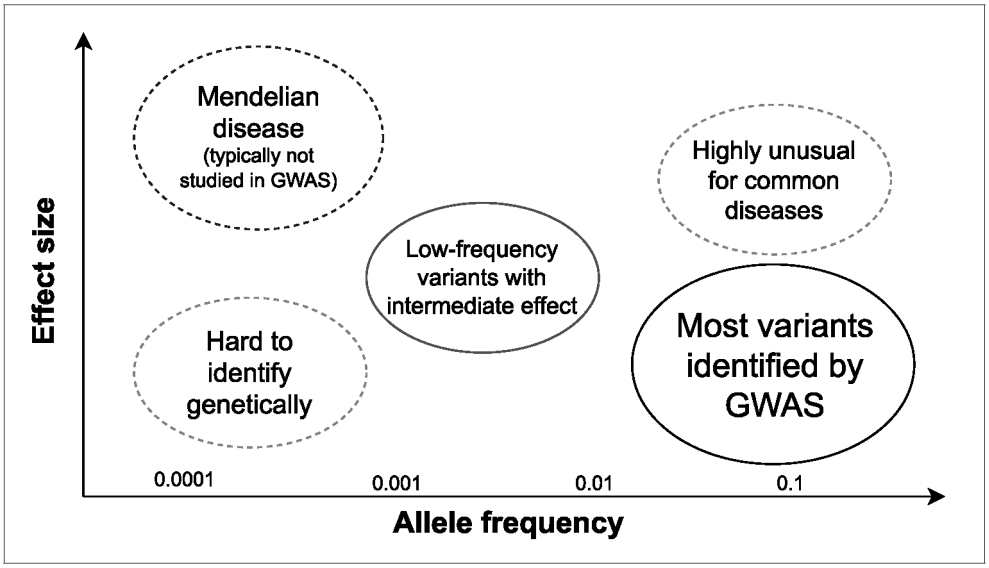


Figure 1.7: Comparison of common and rare genetic variants by their population frequency and effect size. Rare variants of low to modest effect size are hard to identify in GWAS, resulting in most significant associations found in alleles of frequency of 1% or higher of moderate to lower effect size. Figure adapted from[104].

1.2.6 Social and ethical implications of diversity in genomic research

As introduced in Section 1.1, allele frequencies and effects differ between populations. Large genetic panels are predominantly composed of individuals of European ancestry; over 93%

of UK Biobank participants are European[95], and as of 2017, approximately 79% of all GWAS participants were of European ancestry[105,106]. Additionally, UK Biobank exhibits a "healthy volunteer bias": participants are, on average, healthier, older, more highly educated, and of higher socio-economic status than the general UK population[107,108]. Such biases limit the generalizability of observed associations. For example, when using European-derived statistics to calculate PRS for 17 anthropometric and blood-panel traits in the UK Biobank, prediction accuracy dropped for non-European populations, with a 1.6-fold decrease for Hispanic American and South Asian ancestry, and up to a 4.5-fold decrease for African ancestry[105]. Similarly, gnomAD has reported an absence of representation from regions of Asia, as well as Oceania and the majority of the African continent[72].

Similarly to the example of Section 1.1, where several benign variants were misclassified as pathogenic - leading to a series of misdiagnoses in Black American individuals[10] – PRS derived from predominantly European-ancestry panels are less accurate for other populations[100,105]. While association studies and the calculation of risk scores are not the central topics of this thesis, they serve as well-established examples of how the lack of diversity in reference databases hinders our ability to generalize findings and predictions. These issues have also motivated the development of more diverse genomic panels, upon which this work heavily relies.

The 1000 Genomes Project (1kGP), launched in 2008 and conducted until around 2015, was among the first global efforts to create a diverse catalog of human genetic variation across populations. It involved over 2,500 individuals from 26 populations worldwide, using a combination of low-coverage whole-genome sequencing and higher-coverage whole-exome sequencing[109]. Anonymized, phased genotypes from the 1000 Genomes panel are publicly available with unrestricted access.

More recently, the ongoing All of Us Research Program aims to enroll over 1 million diverse participants from the United States, collecting genotype data, electronic health records, and other comprehensive phenotype data[110]. Similarly, other national genome projects, such as the PRECISE initiative in Singapore[111] and Biobank Japan[112], strive to capture the genetic diversity of their respective populations, making these datasets broadly available to researchers[95]. These projects share in common the commitment to data sharing and aggregation as a tool for innovation on a global scale, delivering more value and greater insight from already existing resources[113,114]. As with the early Human Genome Project, making data publicly available can also be seen as a way to counteract private companies' control over scientific knowledge[115].

Consortia that aggregate data from hundreds of thousands of individuals operate under the assumption that participants receive equitable and fair treatment, with an even distribution of risks and benefits[114]. However, numerous examples demonstrate that Indigenous Peoples, minority populations, and disadvantaged groups often bear greater risks while receiving fewer benefits[114,116]. Cases of inadequate informed consent and consultation, as well as misinterpretation or misuse of samples and data, highlight the need for improved community involvement and oversight in research projects[114,116–118].

Notably, failures in consultation with Native American tribes occurred within the All of Us project, prompting the National Congress of American Indians to pass a resolution urging the National Institutes of Health to immediately develop guidelines requiring individual tribal nations to provide consent and oversight for any samples or data collected from their community members[114,119]. Concerns regarding consent, ownership, and access to genetic data are increasingly addressed by data sovereignty approaches, which prioritize the rights of individuals and communities to control their genomic information. An example of this is the "DNA on loan" protocol, which emphasizes that biological samples are not owned by researchers but are considered "on loan" for the specific purpose and duration of the research agreed upon with the donor or community[120,121]. Such initiatives reflect a growing recognition within the genomics community of the importance of community participation in research[114,117,122].

A review by Hudson et al.[114] identifies three main principles for promoting better research practices: (1) building trust with communities through early consultation and participation in data governance[123,124]; (2) enhancing institutional accountability through community-specific review boards that oversee resources, databases, and biobanks[117,125]; and (3) improving equity by appropriately disclosing the origin of samples, ensuring that any value generated from the data is shared, and allowing secondary users to engage with the relevant communities[126,127].

Several projects already implement these principles. The Aotearoa New Zealand Variome is a Māori-led initiative developing the country's genomic catalogue[128]. The Silent Genomes study in British Columbia, Canada, consults and collaborates with Indigenous partners, developing a governance structure for a genomic database as a clinical tool and establishing policy guidelines for the oversight of biological samples and data[121]. Such projects illustrate good practices for the inclusion of diverse communities in genomic resources; however, their integration into larger consortia remains a challenge.

1.2.7 Pangenomes

Recently, a new approach has emerged as an alternative to using a single genome reference complemented by panels of diverse genotypes. A *pangenome* is a comprehensive collection of genome sequences from many individuals within the same species, designed to capture the full breadth of genetic variation present in populations. Pangenomes can be described in terms of their *core genome*, which contains sequences shared by all individuals, and the *accessory genome* (also known as *flexible* or *dispensable*), which comprises regions and alleles present only in some individuals[129].

Pangenomes are typically represented as graphs – data structures composed of vertices (nodes) connected by edges (links) – allowing for a compact and efficient representation (Figure 1.8)[129]. The first draft of the human pangenome reference was published in 2022 by the Human Pangenome Reference Consortium (HPRC)[130]. This initial draft was assembled using

high-quality long-read sequencing of 47 samples from the 1kGP, selected to best represent the individual subpopulations included in the original panel.

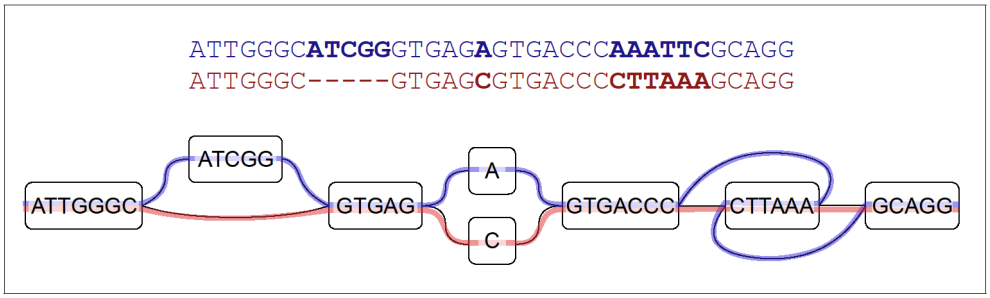


Figure 1.8: Example of a pangenome graph. Figure adapted from [130].

The project is ongoing, and future releases are expected to include a much larger and more diverse set of genomes. The current pangenome already provides much better coverage of the whole genome compared to the GRCh38 reference, and is comparable to the T2T-CHM13 build [130]. While the pangenome can be used for variant calling after alignment with the GRCh38 reference, producing phased genotype files [131], its most significant improvement is in the annotation of structural variants (i.e., variant loci longer than 50,000 bases) [130].

1.3 Mass spectrometry-based proteomics

While genomics provides a view of an organism's genetic material, it alone cannot capture the full complexity of biological systems [132]. The actual presence and abundance of proteins—the functional products of genes—vary greatly not only between tissues and cell types, but also in response to physiological changes. Factors such as disease, injury, infection, circadian rhythm, growth, and aging all influence which proteins are expressed and in what quantities at any given time [133–135].

Proteomics is the large-scale study of the entire set of proteins, known as the *proteome*, detected within a biological system such as a cell, tissue, or organism [16,136]. Since proteins, the blue-collar workers of biology [19], carry out most cellular functions and directly influence phenotype, proteomics has become a cornerstone of modern biomedical research [137].

The earliest proteomic methods, such as two-dimensional gel electrophoresis, relied on antibodies to detect and quantify proteins separated on gels [138]. Today, affinity-based methods remain widely used for protein quantification. These approaches use binding reagents – most commonly antibodies – developed specifically for each target protein [139,140]. Modern platforms such as OLINK and Somalogic [141] employ carefully designed arrays of binding reagents to probe broad or focused panels of proteins, offering a high-throughput and less labor-intensive means of quantification, especially in complex samples like blood plasma [142].

However, affinity-based methods are limited to detecting proteins included in pre-defined panels and depend heavily on the availability and specificity of high-quality binding reagents for each protein of interest[140]. Additionally, these methods do not provide sequence-level information, making it impossible to distinguish between proteins encoded by different haplotypes or alternatively spliced transcripts of the same gene[143]. For this reason, mass spectrometry (MS)-based analysis is often considered the gold standard in proteomics, as it provides both sequence-level and quantitative information for all detectable proteins. Despite technical limitations in sensitivity and coverage, MS enables global, unbiased characterization of proteomes, making it invaluable for both discovery and targeted analyses in biomedical research[137,144].

1.3.1 Laboratory methods for mass spectrometry-based proteomic experiments

Proteomic experiments can be performed on a wide variety of sample types. Typical samples include cell lines cultured in the laboratory[145] as well as frozen tissue sections obtained from biopsies or surgical specimens[134]. In addition to solid tissues, a broad range of body fluids – such as blood plasma, serum, cerebrospinal fluid, urine, and breast milk – are profiled to investigate both normal physiological processes and disease states[146].

Proteins are first extracted from the biological sample using a suitable lysis buffer. The buffer (e.g., 5% sodium dodecyl sulfate (SDS)) and additional treatments like sonication and heating help disrupt cells and solubilize proteins while minimizing degradation[147]. Pairs of cysteine residues within or between protein molecules often form disulfide bonds. While these bonds are important for maintaining the folded structure of the protein[148], in sample preparation, they are reduced using reagents such as tris(2-carboxyethyl)phosphine (TCEP), followed by alkylation (often with iodoacetamide) to prevent their reformation[147]. This step ensures proteins are unfolded and accessible for enzymatic digestion.

In so-called *bottom-up* proteomics, proteins are digested into peptides by proteases, with trypsin being the gold standard[147,149]. The resulting peptide mixture is purified to remove residual salts, detergents, and other contaminants, and the purified peptides are diluted to appropriate concentrations[147]. In liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS), the peptides generated from protein digestion first enter the liquid chromatography (LC) system. Typically, less hydrophobic peptides elute from the LC column faster, while more hydrophobic peptides interact longer and elute later[144], reducing the number of peptides entering the mass spectrometer simultaneously. The time at which a peptide elutes from the column is referred to as its *retention time*[150].

After separation by LC, peptides are ionized into charged ions to be analyzed by tandem mass spectrometry. The ionized peptides, referred to as precursors, enter the mass analyzer, where their mass-to-charge ratios (m/z) and intensities are measured (MS1). Selected precursors are then fragmented, often by higher-energy collisional dissociation (HCD), producing fragment

ions that are analyzed in a second mass analyzer (MS2)[151]. The MS/MS spectrum consists of the m/z and intensity of the fragment ions.

In data-dependent acquisition (DDA), only the most abundant precursor ions detected in the MS1 scan are selected for fragmentation and MS2 analysis. This approach aims to reduce redundant precursor selection and maximize the depth of proteome coverage[152]. Alternatively, data-independent acquisition (DIA) strategies select all precursors within a series of defined *m/z* windows for fragmentation and MS2. This reduces the bias introduced by filtering precursors[153]. However, since all precursors within an *m/z* window are fragmented simultaneously, the resulting spectra are more complex, posing challenges for data interpretation[153,154]. Recent advances in instrumentation – such as increased acquisition speed and reduced isolation width of m/z windows[154] – as well as improvements in software for DIA data analysis[155,156] are leading to DIA gradually replacing DDA-MS for many applications[154].

1.3.2 Standard proteomic data processing

Raw spectra from the instrument contain discrete intensity values for each *m/z* at high resolution. Initial preprocessing steps remove high-frequency noise by smoothing, low-frequency noise by baseline correction, and detect meaningful peaks using one of many peak picking algorithms[157,158], resulting in a sparse representation of the spectra.

A key step in interpreting MS spectra is the correct assignment of a peptide sequence. Most approaches use a database of protein sequences to generate a list of expected peptides, simulating enzymatic cleavage of each protein sequence (*in-silico* digestion). Such a list of all possible peptides in the context of mass spectrometry data processing may be referred to as the *search space*. For each peptide, a list of expected *m/z* values for its fragment ions is generated and compared to the experimental spectra to identify the closest match. Numerous software tools have been developed to match peptide sequences to mass spectra[159–162], commonly referred to as proteomic *search engines*. After searching, these tools report a list of *peptide-spectrum matches* (PSMs).

In the following, this thesis will focus on DDA mass spectrometry, and thus assume a single correct peptide assignment to each spectrum. However, this is not always the case, as even in DDA, several precursors may be selected for fragmentation simultaneously, resulting in chimeric spectra[163,164]. This, and the application to DIA, will be further touched upon in the discussion.

1.3.3 Confidence scoring and error rate estimation

Ideally, the search engine would return a list containing all and only the peptides actually entering the instrument. However, the resulting list of PSMs includes peptides wrongly assigned

to spectra (false positives) and fails to identify peptides that were in fact measured (false negatives)[165]. While it is impossible to fully compensate for false negatives, there are established methods to control the rate of false positives.

Before searching, the protein sequence database can be appended with sequences known not to be present in the sample but that will generate theoretical peptides with identical precursor *m/z* and different fragmentation patterns. This is typically achieved by reversing or shuffling the *target* (real) peptide sequences from the database[166] resulting in a set of *decoy* sequences that are submitted to the search engine alongside the target sequences[167]. This is called the *target-decoy* approach[168].

The list of PSMs returned by the search engine will then include both target and decoy matches. For each PSM, a set of features is collected to describe the quality of the match, such as search engine scores, and other indicators. Since decoy matches are known to be false (i.e., "true negatives"), a machine learning classification model can be trained to discriminate between negative and positive PSMs[169]. The model then assigns a score to each PSM, and target PSMs scoring above a selected threshold are considered significant.

In almost all cases, thresholding will result in a proportion of decoy PSMs scoring above the threshold. Consequently, some accepted target PSMs will also be incorrect, and statistical measures help estimate the prevalence of such misclassified significant PSMs (Figure 1.9). Given a score threshold, the *false discovery rate* (FDR) quantifies the percentage of target PSMs scoring above the threshold that are expected to be incorrect[170]. This is the most commonly reported statistic in proteomic experiments, as it describes the overall quality of a group of reported PSMs. The *q-value* of a PSM is the minimal FDR threshold at which that PSM would be accepted[170]. For statistics describing the quality of individual PSMs, the *posterior error probability* (PEP) is reported, representing the probability that a target PSM, given its score, is incorrect[165]. Both PEPs and q-values can be estimated using the model trained for scoring PSMs, and are typically returned by software tools along with the score for each PSM. Further research on identification procedures in the context of the work of this thesis can be found in Additional Paper 1.

It is important to note that in DDA mass spectrometry, multiple PSMs are often reported for each peptide[171]. While the statistical methods described above consider sets of PSMs, in many scenarios only the highest-scoring PSM is selected per peptide. Applying the same methods for estimating PEPs and q-values after such filtering produces peptide-level statistics, and reporting the percentage of falsely identified peptides may be preferred over the percentage of incorrect PSMs[171].

Technical limitations, such as the resolution and sensitivity of mass spectrometry, account for only some of the errors. When matching mass spectra against a database which does not fully capture the range of peptides present in the sample, search engines will force matches to the closest available entries rather than the true underlying peptides, or these spectra may remain

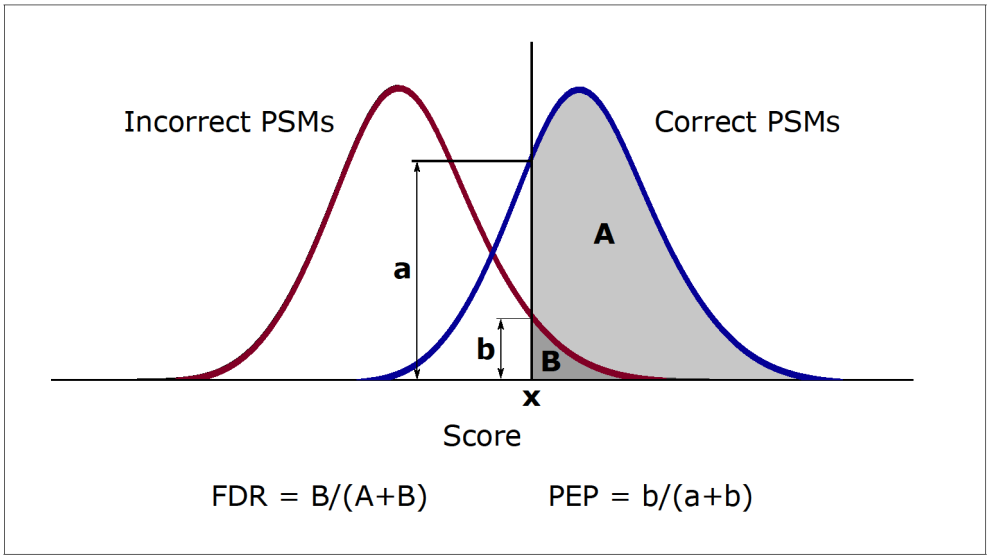


Figure 1.9: False discovery rates (FDR) and posterior error probabilities (PEP) can be estimated by modeling the distributions of correct and incorrect PSMs. Figure adapted from[165].

unassigned[172]. Conversely, expanding the search space excessively increases the chances of random PSMs at a given score, raising the false discovery rate[173]. False and missing peptide identifications may therefore be attributed to the limited representativeness of the search space, and choosing which sequences to include in the protein database is a critical and non-trivial decision.

1.4 Proteogenomics

While the fields of genomics and proteomics are often presented as two separate scientific disciplines, their objects of study are inherently interconnected. Genes are the units that encode proteins, and variants in the genome affect both their sequence and their abundance (see Section 1.2.2). Numerous methods have been developed to interconnect the genomic and proteomic analysis. For example, with recent advancements in affinity-based methods for quantifying protein abundances, large-scale studies investigating genome-wide associations between variants in non-coding regions and protein abundances are becoming increasingly popular[174]. Genomic regions where variation is statistically associated with the abundance of a protein are known as *protein quantitative trait loci* (pQTLs)[175]. For example, pQTL studies have identified novel loci associated with Parkinson's disease[176], as well as type 1 and type 2 diabetes[177,178]. In addition to affinity-based techniques, some pQTL studies use DIA mass spectrometry to quantify protein abundances. For instance, a recent study by Niu et al.[179] identified QTLs regulating at least one third of all detected proteins during pediatric development.

Traditionally, mass spectrometry-based methods rely on searching experimental spectra against a database of reference protein sequences. These sequences are typically derived from the reference genome sequence, which is composed of a limited number of arbitrarily selected individual genomes (see Section 1.2.1). This approach assumes that all peptides potentially present in the sample are represented in the reference database. However, this assumption can never be fully met, as genetic variants in protein-coding regions alter amino acid sequences, and novel protein sequences may arise from unannotated ORFs or alternatively spliced transcripts[173]. Similarly, pQTL studies employing antibody-based techniques typically quantify only the abundance of epitopes and are not designed for sequence-level analysis. This can lead to artifacts where isoforms or sequence variants are undetected, biasing protein quantification. If unaccounted for, these artifacts may be falsely identified as pQTLs[174,179,180].

While the proteomic search space can never be truly complete, it can be expanded by incorporating additional information from genomics and transcriptomics. The scientific field integrating genomics and proteomics is known as proteogenomics[173,181]. This approach enables systematic investigation of how genetic variation shapes the proteome and, conversely, provides experimental validation for the presence of variant peptides and novel protein-coding loci[173,182] (Figure 1.10).

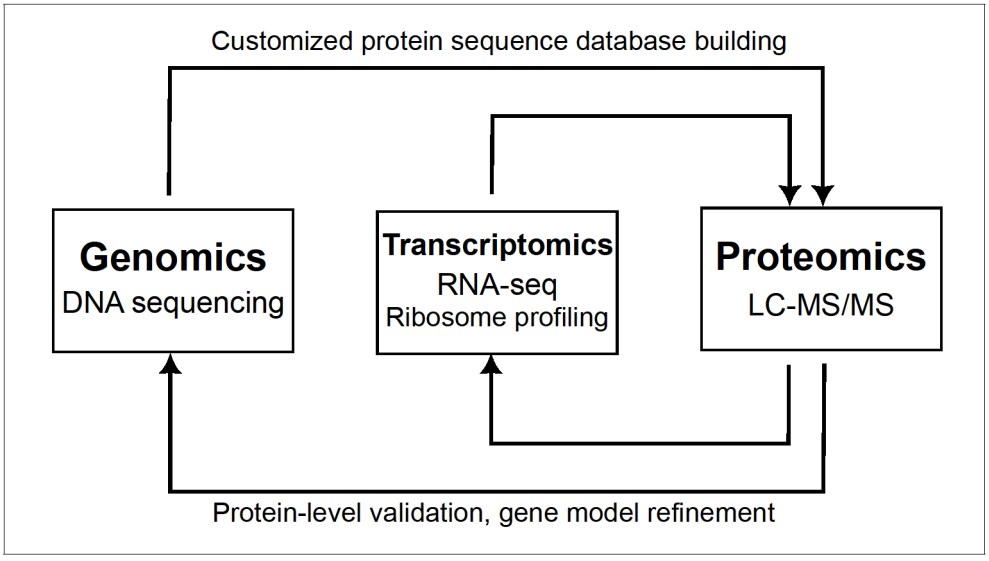


Figure 1.10: The concept of proteogenomics – a field integrating genomic, transcriptomic, and proteomic approaches. Figure adapted from[173].

Since their emergence, proteogenomic approaches have been widely used in biomarker discovery and in the investigation of disease-specific genetic variants[183–186]. However, while the field of genomics has extensively researched the prevalence of common, rare, and somatic variants and associated biases and limitations caused by a lack of diversity in reference data (see Section 1.2.6), knowledge is sparse on the overall impact of common genetic variation on the human proteome and the biases introduced by reference databases in proteomic analyses.

1.4.1 Reference and extended sequence databases

Reference protein sequence databases provide representative protein sequences for each known gene. These resources are typically optimized to present the most relevant set of sequences for download and use in search engines[187]. In this thesis, three major resources providing reference sequence databases for proteomics are considered.

The Universal Protein Knowledgebase (UniProtKB)[188] consists of a reviewed, non-redundant database (SwissProt) providing a single, representative protein sequence for each gene, as well as an unreviewed, computationally annotated set (TrEMBL) that includes protein sequences encoded by multiple alternatively spliced transcripts for each gene[188]. The Reference Sequence (RefSeq) project, maintained by the National Center for Biotechnology Information (NCBI), provides annotated sets of genomic, transcript, and protein sequence records[189]. Similar to UniProt, RefSeq offers both manually curated and computationally predicted and annotated databases.

Ensembl[55] is a platform that integrates available genomic data for species across the tree of life. For humans and common model organisms, Ensembl provides sequence and annotation information at the level of the genome, individual genes, spliced transcripts, and proteins[55]. Due to its comprehensive annotation – linking the locations of genes, exons, transcripts, and start codons – Ensembl is often used as the reference for proteogenomic analyses. Additionally, the recent Matched Annotation from NCBI and EMBL-EBI (MANE) project[54] has defined a set of representative transcripts (and corresponding proteins) for protein-coding genes in the human GRCh38 reference genome, ensuring identical annotation between Ensembl and RefSeq[54,55].

Proteogenomic approaches rely on software tools to build upon these resources and extend reference databases to include the consequences of genetic variants, as well as novel protein sequences encoded by non-coding RNAs or previously unannotated ORFs[173]. Proteogenomic database generation tools typically use *in-silico translation*, a computational process in which nucleotide sequences (DNA or RNA) are split into codons and converted into predicted protein sequences[181]. These sequences can be generated by applying genetic variants to modify reference cDNA sequences of spliced transcripts, or by using transcriptomic data such as RNA sequencing or ribosome profiling[181,190]. The list of genetic variants can either be provided as input to the tool[191] or automatically retrieved from public resources[192].

When the annotation of the ORF is known, in-silico translation is straightforward. However, when working with novel or non-coding transcripts, tools often rely on *six-frame* translation. In this approach, nucleotides can be divided into codons in three different ways on the forward strand and three ways on the reverse strand. While this method ensures that no potential protein-coding sequences are missed, it results in six different protein entries for each sequence, even though only one is likely to be correct[190].

The alternative approach of using RNA sequencing or ribosome profiling presents additional

challenges. mRNA and proteins are not necessarily simultaneously present in a sample: mRNA can be present in a cell at one time point, but the accumulation of the corresponding protein may occur later; similarly, proteins may be detected in other locations where the mRNA cannot be captured[20]. A classical example is the protein insulin, that can be detected in biofluids but is produced in pancreatic beta cells. As a result, even if transcriptomic and proteomic data are generated from the same sample at the same time point, the sequenced mRNAs may not accurately reflect the proteins present in the sample.

1.4.2 Sequence variants and proteomics

The problem of missing variant peptides in standard proteomic workflows gained attention in cancer research, where tumors are characterized by a high rate of somatic mutations[193,194]. Initiatives such as the Clinical Proteomic Tumor Analysis Consortium (CPTAC)[195] have demonstrated the impact of proteogenomics, particularly in oncology, by enabling the identification of cancer-associated variants and neoantigens at the protein level and enhancing our ability to link genetic changes to functional biology[184,195]. Rare variants have also been characterized by proteogenomic studies beyond cancer. For example, a *de novo* SNP introducing a premature stop codon in NAA30 was shown to disrupt the NatC-mediated acetylation of protein N-terminal, linked to developmental delay[196]. *De novo* duplications of the *ATAD3* gene locus were reported to cause perinatal mitochondrial cardiac failure[186].

Approaches to investigating sequence variants in proteins by mass spectrometry are often performed in a personalized manner, where individual whole-genome, whole-exome, or RNA sequencing data are used to construct protein sequence databases[134,185,191]. Furthermore, proteogenomic studies of sequence variants typically focus on disease-associated variants to identify biomarkers. Such methods have been applied in cancer[184,194] and in neurodegenerative diseases such as Alzheimer's disease[197].

While the expansion of the search space increases the likelihood of random matches, the challenges of identifying variant peptides reach beyond this issue. Variant peptides often differ from their canonical counterparts in a single amino acid. Matches of spectra arising from variant peptides to canonical sequences are therefore not entirely arbitrary, while still being false[173]. As the current methods of FDR estimation are designed to distinguish between random and non-random identifications, false matches to variant peptides are often overrepresented in the reported sets of confident peptide identifications[134,173]. Wang et al.[134] have characterized the transcriptome and proteome of 29 healthy human tissues, showing strong differences between the abundance of mRNA and protein in the same sample, and confirming known challenges associated with identifying protein sequence variants. Overall, this study observed 7.4% of non-synonymous variants uncovered by WES in confidently identified peptides, 32% of which were further experimentally verified[134].

1.4.3 Population-based studies of protein sequence variation

Population-based studies that account for amino acid variants in mass spectrometry-based proteomics remain relatively rare. Cao and Xing[191] performed a re-analysis of three publicly available mass spectrometry to account for common and population-specific variants retrieved from GENCODE, reporting an increase of 0.3% in the number of PSMs[191]. Rodrigues et al.[198] analyzed 1,064 tumor and matching blood samples from the CPTAC dataset using cohort-specific protein sequence databases, detecting over 18,000 germline variants in PSMs and relating these findings to GWAS and QTL studies to provide a comprehensive view of germline variation's impact on cancer patients' proteomes[198].

Additionally, Wang et al.[199] built a protein sequence database using 97 samples from the pangenome reference (some of which are not publicly available from HPRC) and re-analyzed two public mass spectrometry datasets, reporting 4,991 novel peptide sequences and 3,921 amino acid substitutions.

1.4.4 Protein haplotypes

Tools for the generation of proteogenomic sequence databases that rely on user-provided or public lists of genomic variants typically consider these variants independently, generating one protein sequence per alternative allele[92]. However, as described in Section 1.2.4, alleles of germline variants are inherited in groups from each parent, resulting in each individual carrying two unique combinations of alleles, known as haplotypes. In datasets with phased genotypes, the assignment of each allele to a specific haplotype can be determined.

This was first addressed by Spooner et al.[92], who introduced the concept of *protein haplotypes* – unique protein sequences encoded by specific combinations of alleles (Figure 1.11). *Haplosaurus*, a resource available through Ensembl both online and as a command-line tool, enables the inspection of protein haplotypes encoded by individual genes[92]. This thesis builds upon the concept of protein haplotypes in the context of mass spectrometry-based proteogenomics.

1.5 Algorithms and tools for high-throughput omics

The concepts introduced so far – genomic, proteomic, and proteogenomic analysis of human samples - have been established not only through the development of novel experimental methods and instruments. Equally, they rely on the existence of relevant computational methods and tools enabling data interpretation, and ultimately drive new scientific discoveries[22]. While not aiming to give an exhaustive review, this section introduces some of the fundamental software

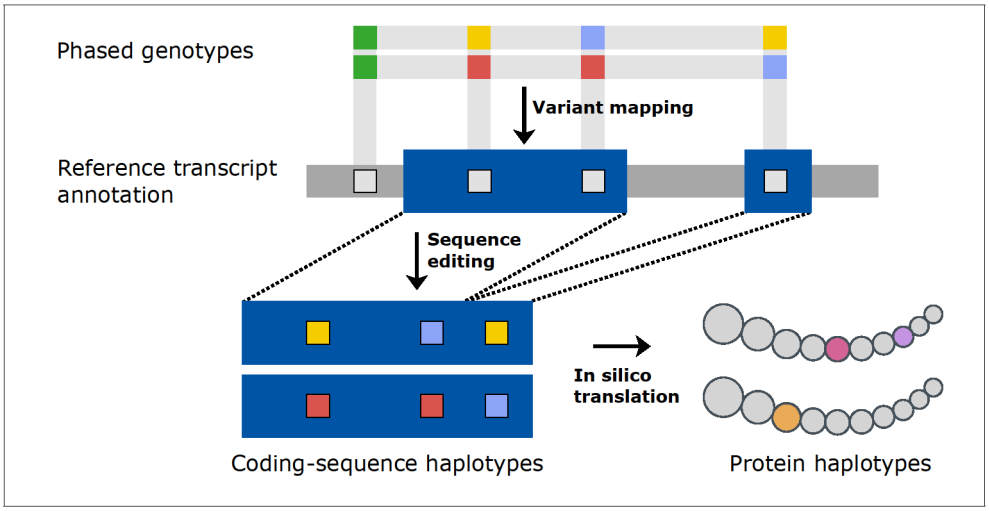


Figure 1.11: Protein haplotypes can be obtained by the alignment of phased genotype data with sequences of spliced transcript, and in-silico translation. Figure adapted from[92].

tools and discusses their relevance to the work presented in this thesis.

1.5.1 Tools for genomics

Sequence assembly methods play a fundamental role in genomics by enabling the reconstruction of complete genome sequences from the millions or billions of DNA fragments generated by sequencing technologies[200]. These algorithms made the initial assembly of the human genome reference possible[201] and remain essential in every genome sequencing project. Once a reference genome is assembled, reads from genome sequencing experiments can be aligned to the reference to identify variants. While NGS technologies enabled a breakthrough in sequencing throughput, the use of the Burrows-Wheeler Transform for indexing the reference genome led to more than a tenfold improvement in the speed of read alignments while simultaneously reducing memory footprint[27,28].

The 1kGP catalyzed many of the advances in tool development in genomics. Of high importance was the development of specialized data file formats for storing read alignments (SAM) and genetic variant calls (VCF)[202]. Accompanying these formats are toolkits such as SAMtools and BCFtools, which handle file format conversions, querying, statistics, variant calling, and variant effect analysis[202].

In addition to read alignment and variant calling, phasing also relies on efficient algorithms and software tools. A widely used method that improved the accuracy and speed of statistical phasing is the Segmented Haplotype Estimation and Imputation Tool (SHAPEIT)[203]. Fast and accurate read-based phasing, particularly beneficial for long-read sequencing techniques, can

be performed using WhatsHap[204]. These algorithms, which scale linearly with respect to the number of reference haplotypes (SHAPEIT) or the number of variant loci (WhatsHap), are essential for haplotype analysis.

These sets of tools have enabled the creation of genomic reference sequences and panels of phased genotypes, which are crucial for the proteogenomic analyses presented in this work.

1.5.2 Tools for proteomics

The transformation of mass spectrometry from a technique for characterizing individual, isolated proteins to a high-throughput approach for untargeted proteomic analysis was in part enabled by the development of search engine software that assigns peptide sequences to spectra using a protein sequence database. The pioneering SEQUEST algorithm[159] inspired the development of several other search engines, including X!Tandem[160], Crux[161], Mascot[205], or Tide[206]. After the database search, a set of features is collected for each PSM or peptide, and a machine learning model is used to estimate q-values and PEPs for each entry, allowing the selection of a set of identifications at a desired FDR (see Section 1.3.3). The Percolator algorithm[169] automatically constructs a training dataset from high-scoring target PSMs and decoy PSMs, and iteratively trains a support vector machine to perform the classification.

This approach presents an advantage over using a fixed, pre-trained model for each experiment, as the characteristics of spectra and the distributions of true and false PSMs can vary between experiments[169]. Another advantage of Percolator is that it does not require a predefined number of features to model the underlying score distribution, allowing the addition of novel features better describing the quality of PSMs[207].

In proteomic applications that require a large search space, such as proteogenomics or the identification of post-translational modifications, search engines typically report higher rates of false positive matches[208]. To better model the score distribution and distinguish between random and true identifications, novel approaches use machine learning-based predictors to extend the feature space used by Percolator or similar algorithms[209]. For example, MS2PIP[210] predicts the intensity of fragment ions in the MS2 spectrum given a peptide sequence, DeepLC[211] predicts the expected chromatographic retention time, and Prosit[212] is capable of predicting both the retention time and fragment ion intensities. Differences between these predictions and the observed values can serve as auxiliary confidence metrics, improving the accuracy of PSM scoring in Percolator[209].

The lack of advanced computational skills has often been a barrier preventing researchers from using search engine tools and interpreting proteomic data. User-friendly interfaces and software such as FragPipe[213], MaxQuant[214], or PeptideShaker[215] have lifted this barrier by providing accessible platforms for data analysis. These tools facilitate downstream interpretation by combining outputs from multiple search engines, inferring protein presence from

peptide identifications, and enabling intuitive filtering by FDR at the PSM, peptide, and protein levels[215]. PeptideShaker also supports automated reanalysis of selected publicly available proteomic datasets[215], and more recently, PepQuery[216] enables a fast and targeted identification of a peptide or protein sequence of interest in almost any public proteomic dataset, greatly improving possibilities of peptide validation and discovery.

1.5.3 Proteogenomic database-generation tools

Proteogenomic analyses use all of the above-mentioned methods to investigate the interplay between the genome and the proteome. The critical step that bridges genomic and proteomic toolsets is the creation of extended protein sequence databases. Early approaches to custom database construction for proteogenomics relied on exhaustive six-frame translation of the genome, gene prediction, or translation of expressed sequence tags. While comprehensive, these methods often resulted in databases containing extremely large numbers of protein sequences unlikely to be present in the sample, hence inflating the search space and making downstream analysis challenging[217].

A major advancement was the development of integrated toolsets such as customProDB[218], which allow users to create protein databases that incorporate splice variants, non-canonical ORFs, and sample-specific sets of variants. These tools work directly with aligned sequencing reads and VCF, enabling more precise and relevant database construction. More recently, Haplosaurus[92] allowed the inspection of protein haplotypes encoded by individual genes, and published a set of databases of protein haplotypes derived from the 1kGP, aligned with the GRCh37 reference genome. Specialized bioinformatics packages such as py-pgatk[192] and PrecisionProDB[191] have further automated the creation of customized protein databases. These tools can download reference data from resources such as Ensembl and use VCF files to align variant calls, generating a set of expected protein sequences with less manual intervention.

1.5.4 Open-source software and reproducible science

Whether we are gathering and annotating a set of variants from a sequencing project, searching a mass spectrometry dataset against a sequence database and filtering the results to a given FDR threshold, or creating a protein database from a list of variants, each of these tasks involves numerous steps. As novel methods require increasingly complex workflows, managing each step manually is becoming infeasible[219]. To address this, bioinformatic tools often aggregate different steps of the workflow into so-called *pipelines*. While custom scripts can be written to create pipelines, workflow management systems such as Snakemake and Nextflow are now commonly used[219]. In these systems, users define a series of analysis steps as distinct “rules”, each specifying its required input, output, and the commands or scripts to be executed. The tools automatically determine dependencies between steps and construct a Directed Acyclic

Graph to optimize the execution order.

Such workflow management tools are essential for the reproducibility of published scientific work (Figure 1.12). By automating and formalizing complex analysis steps into defined, version-controlled workflows, pipelines make it possible to re-run the same sequence of steps with the same parameters, software versions, and configurations - either on new data or by different researchers - and achieve consistent results[219,220].

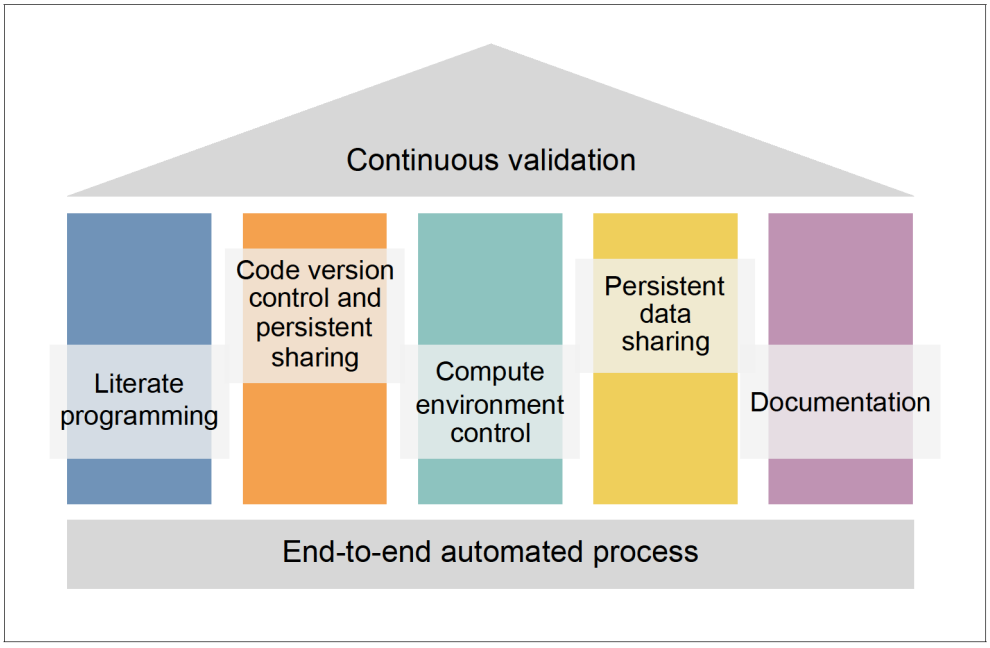


Figure 1.12: Main principles of computational reproducibility. Literate programming combines analytical code with human-readable text; version control allows tracking changes through the development process; environment control requires providing software in container images or environment packages; data sharing enables reproduction, auditing, and reuse; documentation should outline the contents of the software package, and give instruction for usage and reproduction. Figure and concept adapted from[220].

Consistency in software versions is achieved through the use of containers or environments. Containers are complete, portable software units that encapsulate an application and all its dependencies (e.g., libraries, binaries, and configuration files). Created with tools like Docker or Singularity, containers bundle everything needed for a program to function, isolating it from the host operating system and minimizing compatibility issues[221,222].

Environments, managed by tools such as Conda, specify a collection of software packages and their versions, enabling reproducible installation of dependencies, but they operate within the host operating system rather than providing full isolation like containers do[223]. By integrating a workflow management tool (e.g., Snakemake) with an environment manager (e.g., Conda), one can create pipelines where a single command will install all dependencies and execute the entire workflow from start to finish. The Nextflow community has taken this a step further

with *nf-core*[224], a set of tools to automate pipeline creation, testing, deployment, and synchronization. The project also maintains a curated collection of pipelines that can be readily used by scientists. For example, the proteogenomic database generation tool *py-pgatk* is available through the *pgdb* pipeline in *nf-core*[192].

A concept closely related to the principles of reproducibility is open-source software (OSS). In OSS, the source code is made freely available to the public, allowing anyone to view, use, modify, and distribute it for any purpose[225]. Naturally, OSS fosters sharing, community engagement, and dialogue[226] – qualities that are strongly encouraged by the guiding principles of Findability, Accessibility, Interoperability, and Reusability (FAIR)[214]. These principles have been further detailed and extended to research software applications by the FAIR4RS Working Group[227], as outlined in Table 1.2.

Findability	Software, and its associated metadata, is easy for both humans and machines to find.
Accessibility	Software, and its metadata, is retrievable via standardised protocols.
Interoperability	Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
Reusability	Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).

Table 1.2: The FAIR principles for research software. Table adapted from[227].

Reproducibility has been a central focus in the development of the software presented in this thesis. By adhering to open-source principles and using workflow management tools, all analysis steps, dependencies, and configurations are fully documented and reproducible. This ensures that other researchers can easily reproduce the results, extend the methods, or apply the software to new datasets.

1.5.5 Proteomic data sharing

Data sharing in proteomics accelerates scientific discovery, fostering collaboration across the global research community, and ensuring that results of proteomic studies are reproducible[228]. ProteomeXchange (PX)[229] is a consortium of proteomic resources with the primary goal of standardizing and disseminating public MS proteomic datasets. It integrates several key resources, including the Proteomics Identifications (PRIDE)[230] database maintained by the European Bioinformatics Institute, and the Mass Spectrometry Interactive Virtual Environment (MassIVE)[231] maintained by the University of California, San Diego. A notable feature of these systems is the Universal Spectrum Identifier (USI)[232], a unique identifier that can be used to retrieve a single spectrum from a specified deposited dataset. When appended with the assigned

peptide sequence, the spectrum can also be visualized with annotated fragment ion peaks.

In 2021, users downloaded an average of 210,555 GB of data from PX each month[229], and as of 21 August 2025, 21,442 datasets involving human samples had been deposited into PX. However, when dealing with human proteomic data, implementing FAIR principles presents challenges as such data may contain sensitive information that could compromise personal privacy, and patients might not have consented to sharing these data. Alternatively, such data is publishable only via controlled access repositories, accessible only to authorized researchers[228].

1.6 Data visualization

Long before the emergence of high-throughput technologies and bioinformatic software pipelines, visual methods were used to gain insight from complex medical information. In the 1850s, a century before the discovery of the DNA double-helix structure, the work of John Snow mapping cholera cases in London helped identify the source of the outbreak[233]. Similarly, during the Crimean War, Florence Nightingale's charts revealed that more soldiers were dying from preventable diseases than from battle wounds (Figure 1.13), persuading authorities to improve sanitary conditions in military hospitals[233,234].

In the ongoing effort to manage and interpret ever more complex datasets, both automated data analysis and effective data visualization are critical[235,236]. For example, Rittenbruch et al.[237] show that visual applications facilitate effective communication between experts from diverse fields such as biology, computer science, and medicine. Today, as medical and life sciences rely on data science more than ever, visualization plays a crucial role.

1.6.1 Principles of abstract data visualization

While there are experimental methods that produce visual output within the field of molecular biology, and even proteomics[238], the methods central to this thesis produce data that are numerical, and apart from a position within a sequence, do not have a spatial dimension. Abstract data visualization is founded on principles of clarity and efficiency, and is guided by several key frameworks. The core principles of clear data visualizations were laid out by Tufte[233], who introduced the so-called data-ink ratio, maximizing the proportion of space used to represent actual data relative to the total "ink" used in the graphic. Central to this is the directive to "show data variation, not design variation", which means reducing or eliminating all unnecessary visual decorations or redundant elements[233]. For instance, visualizations of genome annotations, as exemplified by Figure 1.14, show relatively dense tracks of elements, with every visual channel (position, color, length, height) encoding a property in the data.

While the principles by Tufte focus on the static visual clarity and integrity of the visualization

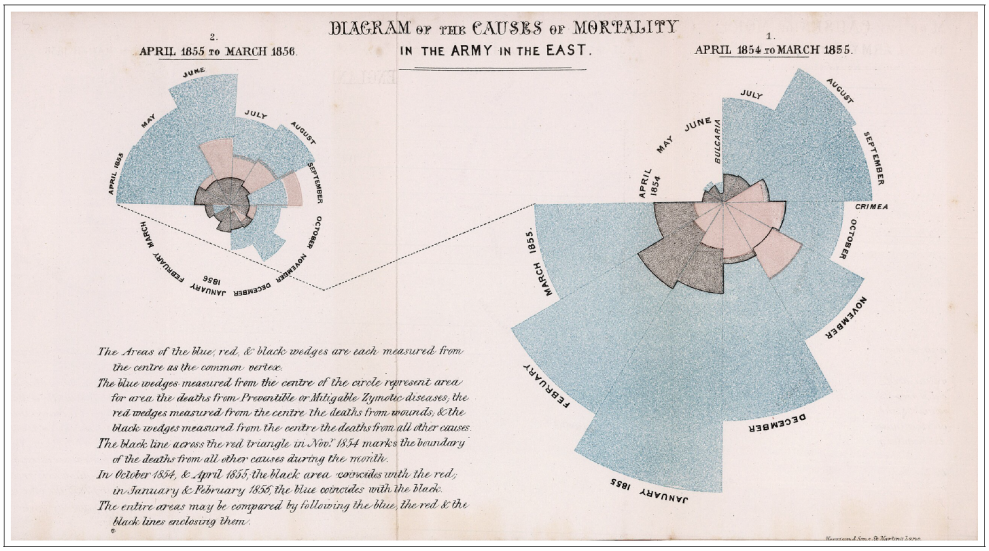


Figure 1.13: The diagram created by Florence Nightingale, showing the causes of death in the British Army during the Crimean war. Public domain via Wikimedia Commons.

itself, Shneiderman[239] addresses how users engage with data visually in an interactive manner. His Visual Information Seeking Mantra – *Overview first, zoom and filter, then details on demand* – advocates first presenting an accessible overview to give context and reveal large-scale patterns, then enabling zooming and filtering to focus on specific subsets without losing overall context, and finally offering details on demand to explore fine-grained information only when needed[239]. Genomic visualizations, especially genome browsers, follow this pattern, showing the overview of the genome, before the user selects a subsequence or searches for a specific gene, revealing a more detailed view, as shown in Figure 1.14.

Expanding on Tufte, Kindlmann and Scheidegger[240] argue that while useful, these directives should be understood as qualities of good visualizations, rather than guidance on how to create those. They highlight three aspects of visualizations: the structure of the data itself, how the data is represented in the computer, and how humans perceive the visual result. Based on these distinctions, they propose three main rules: (1) changing the data's digital format shouldn't change what viewers see; (2) if you change the actual data, the visualization should change to show it; and (3) important changes in the data should be clearly reflected in the visuals[240].

While Tufte emphasizes design clarity, Kindlmann formalizes systematic design processes, and Shneiderman guides user interaction, Munzner and Brehmer[241] offer a framework that bridges the gap between high-level user goals and low-level actions. This multi-level typology categorizes the motivation behind visualization tasks (why), the techniques and operations used (how), and the relevant data or outputs involved (what), providing concise descriptions for visualization tasks[241]. These foundational principles underpin all effective data visualization tools in bioinformatics, a selection of which will be discussed below.

1.6.2 Visualizing biological sequences

Unlike general abstract data, biological sequences (DNA, RNA, and protein) are naturally sequentially ordered, with the position of each element (e.g., nucleotide or amino acid) being important for biological interpretation. This strict ordering requires that visualizations preserve and emphasize the linear arrangement of sequence elements—unlike many general visualizations, which can often rearrange data without loss of meaning[242]. Another distinguishing feature is the sparse distribution of biologically relevant patterns across multiple scales[242]. Functionally important motifs, conserved regions, or variants may be scattered over very long sequences, sometimes spanning tens of thousands or millions of bases. As with the standardized numbering systems for variant loci along chromosomes, genes, transcripts, and proteins established by HGVS (see Section 1.2.3), elements in sequence visualizations can be mapped onto axes that may use different scales.

In genomic and transcriptomic visualization, specialized tools enable exploration and analysis of sequence data. Platforms such as the Integrative Genomics Viewer[243] and UCSC Genome Browser (Figure 1.14)[244] provide comprehensive environments for viewing genomic information, presenting multiple annotation tracks aligned along a unified genomic coordinate system. While the work presented in this thesis does not build upon these tools directly, the display of different splice alternatives of a single gene as separate rows, displaying exons as rectangles along a common scale, as well as highlighting variant loci by vertical lines, provides a well established way to display the genomic context when visualizing proteogenomic datasets.

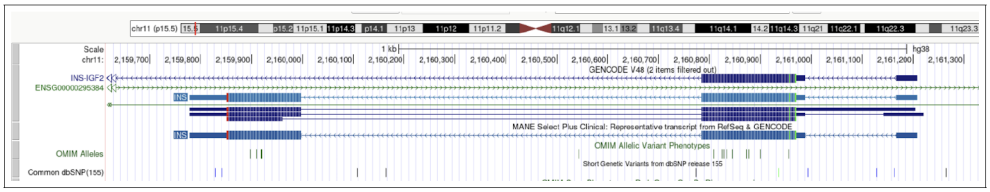


Figure 1.14: The *INS* gene locus shown in the UCSC Genome Browser[244].

Resources providing reference sequence databases like Ensembl[55] and UniProt[188] also offer interactive interfaces for browsing and exploring their data. Ensembl features a genome browser that integrates gene annotations, sequence variants, and transcript splicing, while UniProt provides visualization tools highlighting functional domains, post-translational modifications, annotated sequence features, and variant positions within protein sequences.

1.6.3 Visualizing proteomic data

Various visualization approaches address additional aspects of proteomic data beyond protein sequence analysis. Extensive research has focused on visualizing protein structures[245], including protein-protein interactions and complexes[246], and highlighting key structural fea-

tures[247]. These structural visualizations are essential for investigating protein function and molecular mechanisms within the cellular context. Protein interaction networks are commonly represented using interactive graph-based visualizations, as implemented in several Cytoscape plug-ins[248], often using data from the STRING database of protein interactions[249].

Other visualization tools are designed to represent raw MS proteomic data for quality assessment purposes. Applications like PeptideShaker[215], PDV[250], and Skyline[251] offer graphics to evaluate the quality and reliability of proteomics experiments, including PSMs. Common displays in these tools include annotated mass spectra, where peaks correspond to peptide fragments, and line charts illustrating peptide elution profiles - representing the quantity of a peptide eluted from the chromatographic column over time. The updated version of PeptideShaker Online[33] extends this functionality by integrating visualizations of protein structures and interaction networks into a single comprehensive interface. Large mass spectrometry data repositories, such as PRIDE[29] and MassIVE[231], provide similar interactive visualizations of mass spectra annotated with corresponding fragment ions (Figure 1.15). Any of the billions of deposited spectra can be accessed using the USI[232], making it a valuable resource for proteomics data analysis and verification.

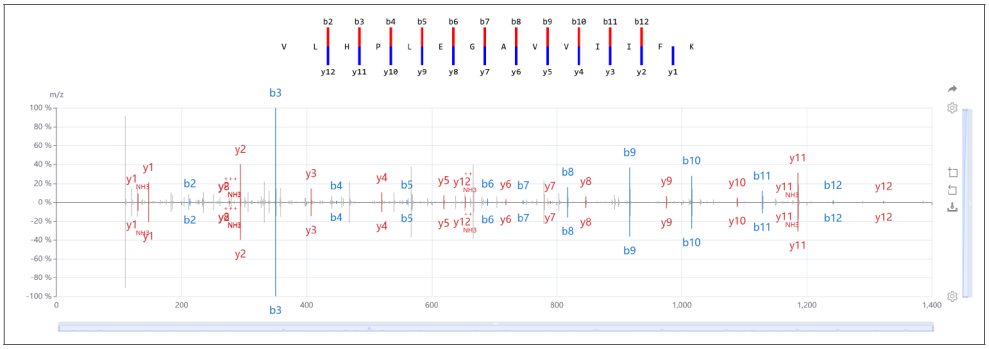


Figure 1.15: An example of a PSM visualization, showing peaks in the spectrum annotated with corresponding fragment ions, accessed via USI[232] in the PRIDE[29] repository.

1.6.4 Evaluation of visualisation design

While the concepts outlined in Section 1.6.1 provide a framework for creating high-quality visualization tools, validating their design is essential to ensure that these tools genuinely enhance understanding, support accurate interpretation, and meet the needs of their intended users[252]. Munzner[253] introduced a nested model that offers a structured approach to visualization design and validation (Figure 1.16). This model divides the design process into four interconnected levels: understanding the context and user needs (domain problem characterization), determining the most relevant data and tasks (data abstraction), selecting effective visual representations and interactions (visual encoding and interaction), and ensuring that computational methods efficiently support these choices (algorithm design)[253]. The nested

structure of the model means that decisions or errors at higher levels propagate through the system, making early-stage clarity and validation especially important. This framework was considered during the development of a visualization tool presented in this thesis, which will be discussed further.

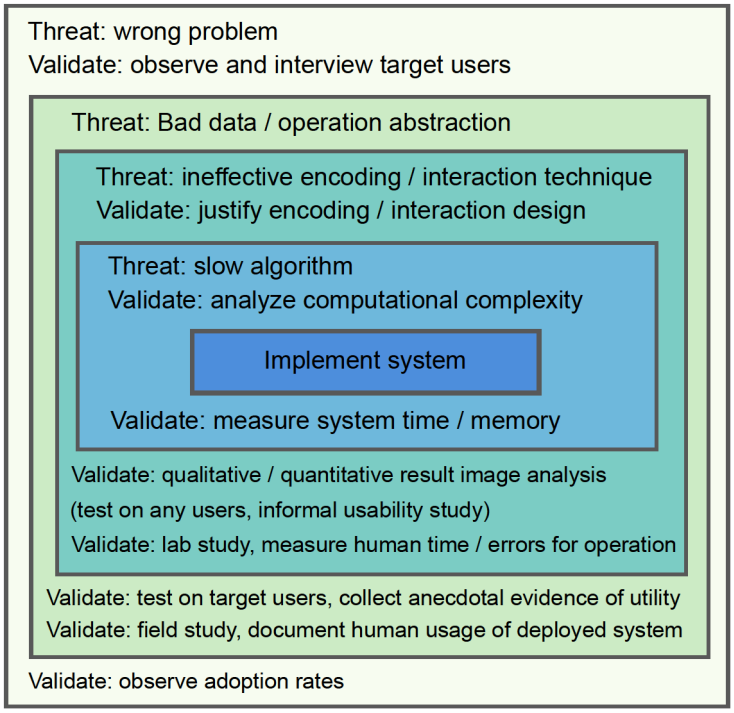


Figure 1.16: A nested model for visualization design and validation. Figure adapted from[253].

Various methods have been developed to validate visualizations at different levels of this model. While many evaluations focus on how well a visualization conveys explicit data "facts," such as values, relationships, and trends[254], others assess broader qualities like insight generation and user engagement. In the following, I will discuss selected examples of these validation approaches.

Among the simplest and most commonly used validation tools is the System Usability Scale (SUS)[255], which uses a structured questionnaire to quickly gauge overall user satisfaction and usability. SUS provides a single score based on responses to 10 questions and can be easily applied to a wide range of tools or products, making it useful for benchmarking[255]. However, SUS lacks diagnostic depth. It indicates whether usability is a concern but does not reveal specific issues or their underlying causes. The ICE-T (Insight, Confidence, Engagement, Tasks) framework[254] evaluates visualizations by assessing the insights users gain, their confidence during exploration, engagement level, and effectiveness in task completion. This framework quantifies the value of a visualization through a series of low-level heuristics, rated by evaluators using a standardized scale. As a more detailed and visualization-focused method, ICE-T is pri-

marily intended for use by visualization experts and can help identify specific issues within a visualization system[254].

A variety of other approaches are also available, including user studies with think-aloud protocols[256,257], cognitive walkthroughs[258], or formal task-based performance assessments measuring speed, accuracy, and error rates[259,260]. Ideally, a multi-method approach—combining quantitative surveys, qualitative interviews, and task performance metrics—would provide a comprehensive view of visualization effectiveness and support iterative design improvements. However, implementing such approaches can be challenging, as it requires substantial resources, time, and access to a sufficiently large and diverse group of users. Recruiting participants and designing and coordinating evaluation methods can be particularly demanding for specialized tools with a limited user base.

2 Aims of the study

The work presented in this thesis aims to advance our understanding of how common haplotypes influence protein sequences in humans, and to improve the integration of genetic information into proteomic analysis, with a focus on applicability in precision medicine.

1. The aim of the study presented in **Paper 1** is to utilize the concept of protein haplotypes, unique protein sequences encoded by sets of genetic alleles in linkage disequilibrium, to investigate the discoverability of novel variant peptides encoded by haplotypes by mass spectrometry.

2. The aim of the study presented in **Paper 2** is to provide an open-source software pipeline for the creation of sequence databases of protein haplotypes, based on phased genotypes of population panels.

3. The aim of the study presented in **Paper 3** is to present a visual and interactive web-based platform to explore the protein haplotype sequences possibly encoded by human genes, and to browse the identifications of variant peptides in public mass spectrometry data.

3 Main results

3.1 Paper 1

Vašíček, J., Skiadopoulou, D., Kuznetsova, K. G., Wen, B., Johansson, S., Njølstad, P. R., Bruckner, S., Käll, L., Vaudel, M. **Finding Haplotypic Signatures in Proteins**. *GigaScience* 2023, 12, giad093.

The first study investigates the signatures left by common haplotypes in protein sequences and their discoverability using mass spectrometry. Following enzymatic digestion, protein haplotypes produce peptides classified into three categories: (1) **canonical peptides**, encoded by gene regions without alternative alleles that change the amino acid sequence; (2) **single-variant peptides**, where one alternative allele produces a non-synonymous change; and (3) **multivariant peptides**, in which two or more alleles in LD encode non-synonymous changes within the same peptide.

To demonstrate the utility of this approach, we used a database of common protein haplotypes, built from the 1000 Genomes Project (Phase 3, GRCh37) by Spooner et al.[92]. Through in-silico tryptic digestion and peptide alignment, we found that 7.82% of the human proteome can potentially be mapped to variant peptides. Specifically, of 102,595 amino acid substitutions encoded by common haplotypes, 12.42% could be discoverable in multivariant peptides (Figure 3.1).

Reanalyzing raw mass spectrometry files from healthy tonsil tissue published by Wang et al.[134], we assessed the prevalence and quality of spectra matching multivariant peptides. After database search, we detected alternative alleles for 4,582 variants, 6.37% of which were found in multivariant peptides. Overall, only 0.6% of all reported PSMs corresponded to variant peptides - a proportion lower than the expected 1% error rate. To address reliability, we present further evidence supporting selected multivariant peptide matches through predicted fragment ion intensities (MS2PIP[210]) and additional validation by PepQuery[216].

As an example, we highlight a multivariant peptide encoded by the actin-like domain of the *POTEI* gene, differing from canonical actin in a single amino acid residue. We emphasize

that without including the relevant POTEI protein haplotypes in the search database, spectra arising from these variant peptides would likely be arbitrarily assigned to actin, demonstrating the importance of haplotype-aware proteomic analyses.

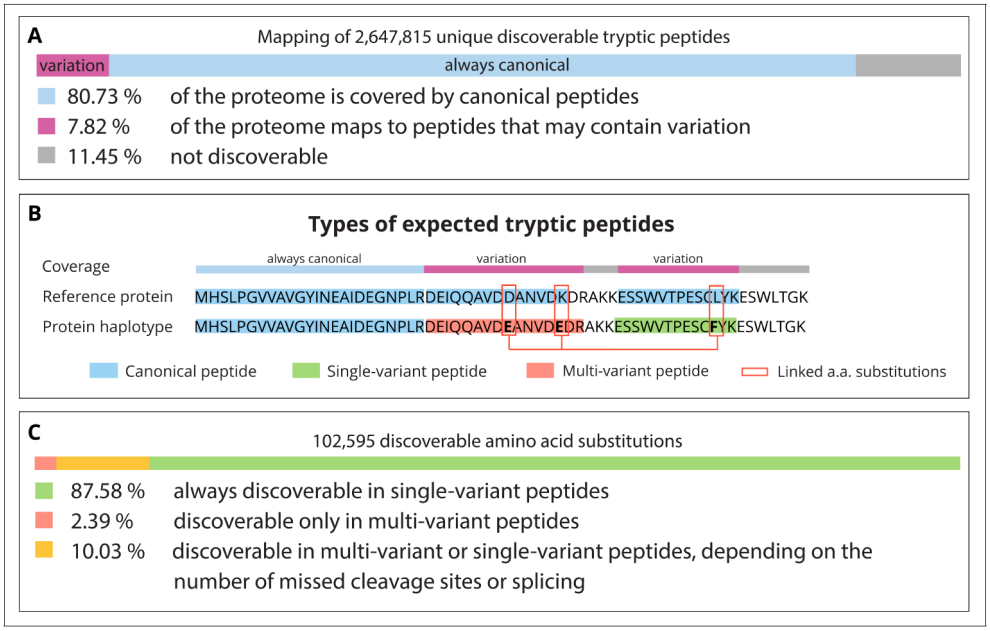


Figure 3.1: Coverage of the human proteome by variant tryptic peptides, and discoverability of amino acid substitutions. Proteome coverage is expressed as the proportion of residues that may map to at least one variant peptide, as illustrated in part B.

3.2 Paper 2

Vašíček, J., Kuznetsova, K. G., Skiadopoulou, D., Unger, L., Chera, S.;, Ghila, L. M., Bandeira, N., Njølstad, P. R., Johansson, S., Bruckner, S., Käll, L., Vaudel, M. **ProHap Enables Human Proteomic Database Generation Accounting for Population Diversity**. *Nature Methods* 2025, 12, 273–277.

In Paper 1, we used a protein haplotype database generated by Haplosaurus[92] based on the GRCh37 genome build, which only included amino acid substitutions, excluding insertions, deletions, and loss of stop codons. As Haplosaurus is meant to investigate specific proteins with a focus on drug design, a haplotype-aware database-generation tool suitable for mass spectrometry-based proteomics was missing.

To overcome these limitations, the Paper 2 introduces **ProHap**, a Python-based software pipeline designed to efficiently create comprehensive protein haplotype sequence databases from phased genotype datasets (Figure 3.2A). Alongside ProHap's release, we provide three sets of protein databases derived from major population genomics resources: the 1000 Genomes Project Phase 3 aligned to the GRCh38 reference genome[261], the HRC Release 1.1[86], and the Human Pangenome Reference Consortium Release 1[130].

In-silico digestion of the 1000 Genomes-derived databases revealed that the set of protein sequences derived from individuals of African ancestry exhibits a larger fraction of the proteome mapping to variant peptides, yet all five superpopulations showed at least 9% of the proteome potentially attributable to variant peptides (Figure 3.2B). This suggests that similarly to the biases reported in genomic association studies[105], using only canonical protein databases in mass spectrometry created biases against populations with different haplotypic structures.

The analysis of in-silico digested peptides also suggested that using haplotype-aware databases is advantageous even when working with individuals of European or East Asian ancestry. To validate their practical utility, we searched a blood plasma proteomics[262] dataset against the protein sequence database created by ProHap from the European superpopulation in the 1000 Genomes dataset. Among the identified variant peptides, we highlight four specific examples, identified respectively in 44, 39, 13, and 12 of the 52 individuals included in the study. Additionally, we analyzed mass spectrometry data from a stem cell experiment for which donor genotypes are publicly available. Using ProHap, we constructed a personalized protein database reflecting the donor's haplotypes. This analysis identified peptides encoded by both the reference and the alternative allele for 87 loci in the genome where the donor is known to be heterozygous. Additionally, a multivariant peptide mapping to a haplotype of the CAP1 protein was identified, containing five amino acid substitutions.

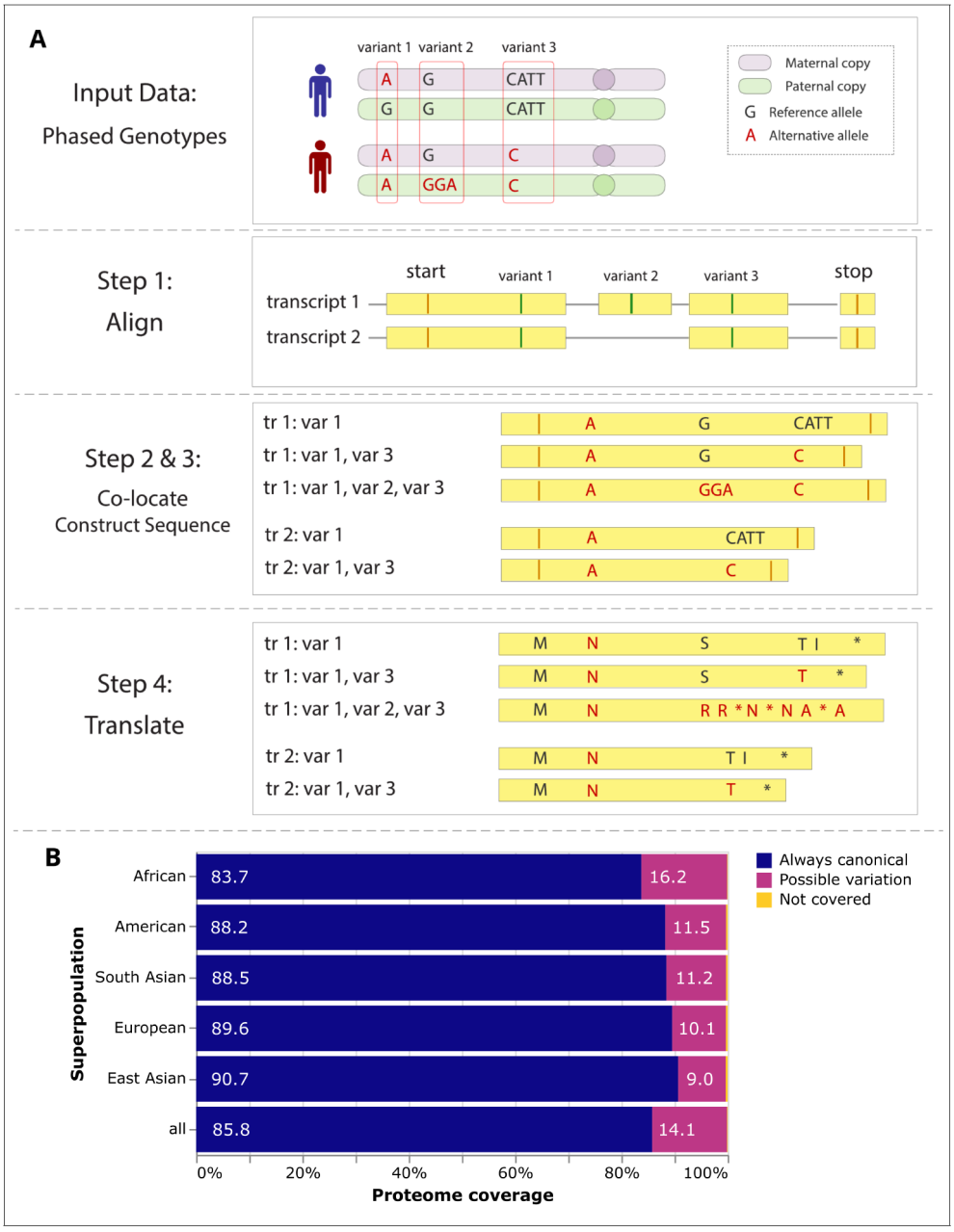


Figure 3.2: A: Overview of the ProHap pipeline to create protein haplotype databases from phased genotype datasets. B: Coverage of the proteome by variant peptides in the databases derived from the five superpopulations included in the 1000 Genomes panel shows a higher percentage of the proteome mapping to variant peptides in the African superpopulation.

3.3 Paper 3

Vašíček, J., Skiadopoulou, D., Kuznetsova, K. G., Käll, L., Vaudel, M., Bruckner, S. **ProHap Explorer: Visualizing Haplotypes in Proteogenomic Datasets**. *IEEE Computer Graphics and Applications* 2025, 45(5), 64-77.

In both Paper 1 and Paper 2, we have presented novel approaches to bridge the interpretation of proteomic and genomic datasets. Genetic variants (combined into haplotypes), annotations of exons and open reading frames, peptide identifications in mass spectrometry datasets, and the associated confidence metrics are considered together, enabling a comprehensive exploratory analysis. Until now, visualization tools have primarily focused on either genomics or proteomics independently, lacking an integrative platform to explore the complex relationships between genetic variation and proteomic evidence.

This work introduces **ProHap Explorer**, a web-based interactive visualization tool designed to bridge this gap by enabling researchers to investigate the influence of common human haplotypes on protein sequences in breadth and depth. The presented approach models proteogenomic data as a graph across four primary layers: gene, transcript, protein, and peptide. To contextualize experimental data, additional nodes in the graph represent mass spectra and biological samples. ProHap Explorer features a dual-view interface facilitating transitions between broad exploration and in-depth gene-level analysis. The Explore View offers an overview of all human genes, displaying coverage and variation in mass spectrometry datasets, while the Detail View allows detailed examination of protein sequences encoded by one gene, corresponding transcripts, splice variants, and peptide identifications. Both views include interactive visualizations and exportable data tables, enhancing user accessibility.

ProHap Explorer's practical utility is exemplified through case studies where variant peptides from the CPN2 (Figure 3.3) and CAP1 proteins, previously identified in Paper 2, have been validated within the interface. Users can easily visualize these peptides in the broader proteogenomic context, viewing the frequency of the respective haplotypes within the 1000 Genomes panel. Notably, we find that all participants in the 1000 Genomes panel carry the haplotype encoding the multivariant peptide in CAP1, suggesting that the reference sequence in GRCh38 at this locus is not expected to be present in the general population.

This platform represents an important advance building upon the previous development of ProHap in Paper 2, and the foundational observation in both Paper 1 and 2 that a substantial fraction of the human proteome maps to variant peptides. The three studies presented in this thesis together enable the proteogenomic community to more accurately characterize and explore protein diversity across and within human populations.

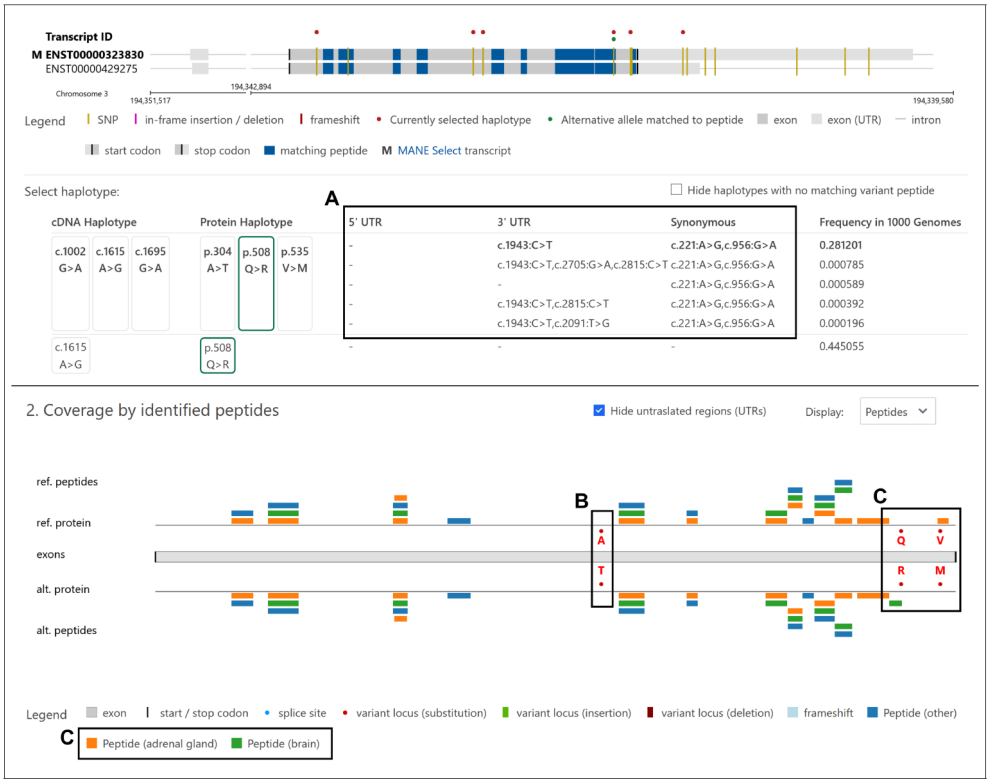


Figure 3.3: Detail of the *CPN2* gene in ProHap Explorer. A: Several haplotypes encode the same protein sequence. B: Presence of the reference or alternative allele cannot be estimated due to lacking coverage. C: We only see the alternative allele for the second substitution, and the reference allele for the third substitution. However, the respective peptides are observed in different tissue samples.

4 Methodological considerations

4.1 Data sources

This thesis is based on three principal sources of data: genome annotations, phased genotypes, and raw mass spectrometry proteomic data. First, reference genome annotations were systematically retrieved from the Ensembl[55] database, which provides a consistently updated and version-controlled resource for both gene models and reference sequences. This approach ensures that the analytical pipelines can flexibly specify the desired Ensembl release, automatically downloading the relevant annotations and sequences as the foundational layer for all downstream analyses.

For genotype data, the 1kGP Phase 3[109] dataset served as the primary resource for phased human genotypes. In Paper 1, we used the 1kGP data as processed by Spooner et al.[92], aligned with the older GRCh37 genome build, while Papers 2 and 3 utilized the same project's genotypes aligned to the more recent GRCh38 assembly[261]. Additionally, Paper 2 incorporated the HRC Release 1.1[86] dataset, which was restricted to variant calls mapped to GRCh37; to use these with up-to-date genome builds, we performed a liftover of coordinates using GeneBe[263]. The tool was chosen for its convenient Python implementation, and support for variant description in various formats. Noting that recent reviews assessed the most commonly used liftover tools to be performing similarly[264,265], we did not compare multiple liftover tools specifically for this dataset. Further, as part of Paper 2, a publicly available healthy donor genotype—also aligned with GRCh37—was employed to generate a personalized protein database.

While the liftover of variant coordinates to GRCh38 was required, such that these results are comparable to the databases created using 1kGP, this process is not optimal: differences between genome builds can cause conflicts, such as instances where an alternative allele in GRCh37 becomes the reference sequence in GRCh38 or vice versa[266]. In such cases, individuals carrying the 'reference' allele in GRCh37 will not have that variant called at all, and after liftover to GRCh38, the same allele – now classified as 'alternative' – will be missing from the updated genotype dataset.

Thirdly, public MS proteomic datasets were used to exemplify the usage of protein haplotype

databases in proteomic studies, and take the first steps towards charting the influence of common haplotypes on the human proteome. In Papers 1 and 3, a subset of a reference dataset of 29 healthy human tissues[134] was used, allowing to chart the presence of variant peptides across different tissues. Specifically, Paper 1 focused on tonsil tissue samples, matching the original study's aim to detect non-canonical peptides and alternative splicing signatures, while Paper 3 expanded this comparison to multiple tissues, using our interface for highlighting cross-tissue peptide identification. Paper 2 pivoted to blood plasma samples, reflecting a common clinical research scenario, to assess the robustness and applicability of our methods in biobank-scale analyses. In addition, we incorporated MS data from an in-house experiment on healthy-donor stem cells. Having the donor genome enabled linking personal-level genomic and proteomic analyses.

4.2 Mass spectrometry data analysis

We developed an MS data analysis pipeline using Snakemake to identify and score peptides, as detailed in Paper 1 and Additional Paper 1, with the coordination and bioinformatic implementation led by Dafni Skiadopoulou. The workflow begins by converting raw spectra into the mzML format using ThermoRawFileParser[158]. Peptide identification is performed via SearchGUI's command-line interface[162] with the X!Tandem[160] algorithm (employed in Paper 1 and Paper 2), and the Tide[161] algorithm (in Paper 2 and Paper 3). PeptideShaker[215] is used for downstream processing and the export of PSM lists for feature prediction and rescoring with Percolator[169]. Fragment ion intensities are predicted using MS2PIP[210], while chromatographic retention times (RT) are predicted using DeepLC[211]. Calculation of similarity metrics between the observed and predicted RT and fragment spectra is detailed in Additional Paper 1.

In Paper 1, Percolator[169] was initially run with only the default features, and similarity metrics were analyzed independently. In Paper 1, PSMs corresponding to variant peptides are further analyzed using PepQuery[216], and were rejected if a canonical peptide match (with or without post-translational modifications) scored better or equal. In Additional Paper 1 we showed that including prediction-based metrics in Percolator's feature space can improve PSM rescoring, leading to Paper 2's adoption of extended features as the standard approach. For stem cell experiments in Paper 2, peptide identification used the MSFragger algorithm within the FragPipe pipeline[267,268], and was paired with retention time and fragmentation predictors implemented in MSBooster[269].

4.3 Publishing research software

Publishing research software requires adherence to best practices that ensure the software is reliable, maintainable, and reusable across different research contexts. The FAIR principles adapted for research software (FAIR4RS)[227] provide a set of guidelines which help maximize its impact and facilitate reproducibility. Here, qualities of the published tools are discussed within this framework.

4.3.1 Application of the FAIR principles

The tools developed and utilized throughout this thesis broadly align with the FAIR4RS principles. Each tool is accompanied by a unique and descriptive name as well as a public GitHub repository, facilitating discoverability and accessibility (FAIR4RS: F1, A1). The associated codebases are maintained, and software is clearly versioned upon publication (F1.2, R1.2). Data input and output formats are standardized whenever possible, with adherence to domain-specific norms promoting interoperability (I1). Input files for ProHap are expected in the VCF format, developed within the 1kGP. Where phasing information is available, a standard format for phased VCF file is accepted. With phasing unavailable, we have developed a simplified version of ProHap (called ProVar), considering variants independently, and accepting any VCF file.

Protein sequence data are typically exchanged in the FASTA format, a widely accepted text-based file type containing header lines with information such as protein identifiers and gene or transcript annotations. However, the FASTA files accepted by common proteomic search engines are conventionally structured for canonical proteomes. In ProHap-generated databases, it is necessary to supplement sequence headers with additional information such as the presumed translation start site, and identifiers of the encoding haplotypes (accounting for synonymous variants). To maintain compatibility, users may export this extended metadata separately and generate simplified FASTA files suitable for standard search engines, effectively bridging advanced dataset richness with practical downstream use.

Furthermore, to close the gap between standard protein search engine outputs and insights gained from viewing peptide sequences in their genomic context, the ProHap Peptide Annotator pipeline was developed. This tool takes as input PSM or peptide reports from search engines along with genome annotations from Ensembl and ProHap database files, and produces richly annotated lists of peptides. These annotations include matched gene names and identifiers, genomic and protein coordinates of reference and alternative alleles encoding the peptide, and other biologically relevant features such as splice junction overlaps (I1, I2).

Metadata, licences, and documentation accompany all tools and databases through GitHub and Zenodo repositories (F2, F3, R1.1). Reproducibility is ensured via Snakemake workflows

coupled with conda environment specifications; pipelines and a series of steps are available to automatically download necessary data and reproduce the main results reported in Paper 1 and 2. Similarly, the software and pipeline constructing the graph database for Paper 3, called ProHap Graph, is openly shared, along with instructions to build the ProHap Explorer web frontend from source. However, due to the complexity in deploying graph databases and web applications, full replication of such infrastructure remains challenging, limiting perfect reproducibility in practical terms.

4.3.2 Validating visualization design

Visualization tools require validation beyond adherence to general software principles such as FAIR. Designing effective visual representations follows well-established rules and methodologies discussed in the literature (see Section 1.6). One influential framework is Tamara Munzner's Nested Model for Visualization Design and Validation[253], which guides the verification of visualization systems across four interrelated levels (see Figure 1.16). This section discusses the validation of the visualization tool presented in Paper 3 within this framework.

The design process begins at the outermost level with a characterization of the domain problem - ensuring that the visualization's purpose, domain-specific tasks, and users' needs are well understood. This foundational understanding was primarily developed in Paper 1, as the initial designs for an interactive visual interface were drafted during this study. The second level involves data and task abstraction, which translates domain-specific vocabulary and challenges into computer science concepts, data structures, and analytical tasks. Much of this mapping was realized during the development of ProHap and the ProHap Peptide Annotator pipeline in Paper 2. This work required an in-depth examination of all the data properties, from handling genome annotations, phased genotypes, protein sequence data, to generating databases of protein haplotypes. The implementation of these tools provided a purely computational perspective on the data, beneath domain expertise.

At the third level, the design of visual encoding and interaction is justified. Paper 3 provides a detailed rationale for chosen visual representations, interactive features, and interface layouts, drawing from best practices and standards in the visualization community. At the innermost level, algorithm design ensures efficient implementation of visual and interaction techniques. While no formal computational complexity analysis was performed, initial performance testing did not reveal major bottlenecks. This aspect will continue to be monitored as the software evolves to maintain responsiveness and scalability.

A critical component of the nested model is downstream validation at each level after the system's implementation. Paper 3 reports an informal, qualitative usability study involving domain experts external to the development team, addressing the downstream portion of the third level. This study generated early feedback on intuitiveness and practical applicability,

highlighting the tool's potential and areas for enhancement. As the tool gains wider adoption, community input is expected to fuel iterative refinements, effectively addressing the outermost levels of domain characterization and abstraction in a real-world context.

5 Discussion

5.1 Reference genome builds and sequencing quality

Many datasets of phased genotypes remain aligned to the older GRCh37 reference genome, including key resources such as the HRC Release 1.1[86]. In Paper 2, when working with the HRC dataset, as well as the personal genotypes of the stem cell donor, a liftover of variant coordinates to GRCh38 was performed, for comparability, as discussed in Section 4.1. As there can be conflicts or ambiguity between genome builds, liftover of variant calls poses challenges. Ideally, new variant calling should be performed on original sequencing reads, as was done for the realignment of the 1kGP to GRCh38[261], but this approach requires access to raw read data, which is often unavailable. The reliance on Ensembl-based annotations further reinforces the use of GRCh38, since the Ensembl annotations for new human genome assemblies are only available to a limited extent in the Variant Effect Predictor tool[270,271]. Since the new T2T-CHM13 assembly mainly expands noncoding and repetitive regions, its impact on annotated protein-coding regions is expected to be minimal. Several new protein-coding genes were suggested in the novel sequences added in the assembly, however, their presence is yet to be verified[272].

5.1.1 Using pangenomes

The initial release of the HPRC dataset was published as the work on Paper 2 was already near completion, and therefore, the tools developed did not consider the graph-based data structures for genomic reference data. Still, the adoption of pangenomes would present advantages compared to traditional VCF-based representations. Notably, pangenome graphs model structural variants and repeated regions more effectively[129]. Using a pangenome as the reference instead of a single, linear sequence brings conceptual changes: the distinction between 'reference' and 'variant' peptides is blurred, as all observed peptides could be incorporated as part of a comprehensive reference. While this shift would not hinder proteomic analysis, the concepts of variant peptides remain essential for identifying the limitations of canonical proteome-centric searches and for clinical communication about pathogenicity and trait associations. Integrat-

ing variant descriptions into a pangenome framework is possible, but methods for this remain under active development.

Constructing protein sequence databases directly from pangenomes - without annotation resources like Ensembl - remains challenging. It requires alternative strategies for open reading frame (ORF) selection, such as six-frame translation, which dramatically increases the search space and introduces a high proportion of biologically irrelevant sequences. While Wang et al.[199], have implemented pangenome-derived protein databases, details about ORF selection in this study remain unclear. For these reasons, in Paper 2, we used the HPRC dataset decomposed into a VCF file and aligned with GRCh38, rather than building databases directly from an unannotated pangenome.

The HPRC dataset was generated using long-read sequencing technologies, such as PacBio HiFi and Oxford Nanopore[273], which offer substantial advantages in assembling highly variable and structurally complex genomic regions. Even after alignment to the GRCh38 reference, the HPRC provides markedly improved coverage of genes that are difficult to resolve with short-read data, exemplified by the human leukocyte antigen (HLA) loci.

5.1.2 ProHap and the HLA

The HLA genes encode the major histocompatibility complex proteins, which play a central role in immune system function by presenting antigens to T cells and initiating immune responses. HLA proteins are among the most polymorphic in the human genome, with thousands of different alleles known[274], driving substantial inter-population diversity and conferring genetic susceptibility or resistance to various diseases, from autoimmune conditions to infectious disease responses[275]. Notably, the frequency and combinations of HLA alleles vary dramatically between global populations, reflecting adaptation to different pathogen pressures and population histories[276,277].

What are commonly referred to as HLA alleles are, in fact, protein haplotypes – distinct combinations of nucleotide substitutions that occur together on the same copy of the protein-coding regions of HLA genes. These substitutions in HLA are systematically annotated in databases such as UniProt, providing curated sequences for a wide variety of common alleles. While a detailed analysis of the HLA region is beyond the scope of this thesis, comparing the protein haplotypes encoded by the *HLA-A* gene, generated by ProHap from the 1kGP and HPRC datasets, to the sequences of the most common HLA-A protein sequences as annotated in UniProt, revealed that the common HLA-A protein haplotypes obtained using the 1kGP did not reproduce the common alleles annotated in UniProt. However, the protein haplotypes obtained using the HPRC data resembled more closely the known groups of HLA-A sequences (so-called *supertypes*[274]), indicating a large improvement in resolution.

While the HPRC shows a better ability to resolve complex genetic loci, the panel in its first re-

lease included sequences of only 47 individuals. In genomics, studies to inspect the association of HLA alleles with diseases are typically performed across populations with large numbers of samples (e.g., over 450,000 participants in the UK Biobank[278], or over 1,200 participants in a focused study in First Nations peoples of Oceania[277]). With the upcoming releases of HPRC, our ability to model the distribution of common protein haplotypes encoded by complex genetic loci will improve, enabling new avenues for investigating their role in human traits and diseases.

5.2 Looking beyond genetic information

The thesis focuses on the presence and discoverability of peptides encoded by germline genetic variants in proteomic datasets, which represents an important first step in characterizing proteomic diversity linked to genetic variation. However, beyond identification, the ultimate goal often lies in accurately quantifying the abundance of both reference and variant peptides to understand their biological and functional impact. To date, no fully established methods exist for reliably estimating the relative abundance of variant versus canonical peptides in complex samples.

One promising avenue involves the use of DIA proteomics, which has recently seen substantial advances in depth of coverage and quantification accuracy[154]. Unlike DDA, which employs a spectrum-centric approach – reporting PSMs and inferring peptide presence from these assignments – DIA often utilizes a peptide-centric approach, reporting a single quantitative value per peptide[251]. Reporting a single value per peptide reduces the complexity associated with multiple precursors eluting within the same retention time window and being represented by the same spectrum. Tools such as DIA-NN[156], which use deep learning to quantify peptides from DIA MS data, allow the use of predicted spectral libraries.

Such libraries are obtained by applying retention time and fragment ion intensity predictors onto a peptide sequence database, and the libraries are then matched against the observed spectra to identify and quantify peptides. Conversely, a spectrum-centric approach can also be applied to DIA data by first deconvoluting spectra to represent individual precursors[279]; similarly, chimeric spectra in DDA, which contain fragments from multiple precursors, can be deconvoluted to improve peptide assignment[230]. These approaches often employ prediction tools in a similar manner as DIA-NN. The impact of expanding the search space with protein haplotypes on DIA analysis remains to be explored. Extending haplotype-aware methods to DIA proteomics presents a critical step towards the quantification of relative protein haplotype abundances in large cohorts.

A major challenge in both DDA and DIA is distinguishing true variant peptides from those with post-translational modifications (PTMs). An observed mass shift in a spectrum can result either from an amino acid substitution or the presence of a PTM. “Open” modification searches,

looking for any possible PTMs, are often not feasible as they substantially inflate the FDR despite rescoring strategies[280]. Multi-pass searches as implemented in PepQuery[216], as demonstrated in Paper 1, can help differentiate some cases, but further refinement and automation are needed for reliable large-scale application. Notably, PepQuery analysis revealed that false positives are overrepresented among variant peptide identifications compared to those encoded by canonical sequences. This suggests that current FDR control methods, designed to distinguish between random and non-random peptide identifications, may not adequately account for the presence of variant peptides, even when using orthogonal feature sets based on predictors, as falsely assigned variant peptides still may be partially correct. Addressing these challenges is critical for accurate interpretation and quantification of variant peptides in proteomic studies.

Compounding these difficulties, some of the most variable and medically significant peptides, such as HLA-bound peptides[281] and hypervariable regions of antibodies[282], are particularly difficult to capture and quantify. For instance, HLA-bound peptides are cleaved in the cytosol and trimmed in the endoplasmic reticulum[281], and therefore do not follow any known enzymatic cleavage patterns. Being low in abundance as compared with their source protein, specialized experimental workflows beyond standard shotgun proteomics are needed for these peptides to be captured[281]. The methods investigating the entire repertoire of HLA-bound peptides in a sample are called *immunopeptidomics*[281,283]. Regular proteomic search engines can be used to match peptide sequences to spectra in immunopeptidomic datasets, however, the need to consider all possible peptide sequences of a given length range (typically 8–11 residues for HLA class I, 6–24 residues for HLA class II) results in a search space two orders of magnitude larger than in proteomics[209]. This again results in inflated FDR, which can be compensated through rescoring and using machine learning-based predictors[209]. Still, the effect of haplotype-aware analysis on immunopeptidomic studies is yet to be explored.

This work has focused in detail on the discoverability of peptides encoded by different haplotypes and deriving peptides from genes and proteins. Conversely, inferring proteins and genes from peptides remains challenging, and even more so when accounting for protein haplotypes. Peptide identifications often cover only a limited section of the entire protein, and their combinations do not uniquely discriminate between sequences encoded by different splicing alternatives and haplotypes of the same gene. Therefore, inferring the presence of a particular protein sequence from a set of identified peptides within a sample is often impossible[284].

To improve the coverage of proteome by identified peptides, previous studies have used multiple enzymes for protein digestion in addition to trypsin[134]. Similarly, in Paper 2, we have included the cleavage pattern of six enzymes in the in-silico digestion, improving the proteome coverage by discoverable peptides to over 99% (see Figure 3.2), as compared to 89% reported in Paper 1 (see Figure 3.1), which only used trypsin. Although the digestion using multiple enzymes, the annotation of peptides with their encoding genomic loci allowed by the ProHap Peptide Annotator in Paper 2, or the visual alignment of peptides to proteins as shown in Paper 3 may enhance our ability to distinguish between unique protein haplotypes, a systematic

analysis of the implications of haplotype-aware methods on protein inference is an important avenue for future research.

5.3 Potential for further software development

The ProHap workflow currently relies on Ensembl gene annotations, which poses challenges when comparing peptide identifications obtained using ProHap-derived databases with those from standard proteomic workflows that often depend on UniProt. This discrepancy in reference sources can complicate the conversion of protein identifiers between Ensembl and UniProt, making cross-platform analyses cumbersome. Ideally, the pipeline should support the generation of protein haplotype databases based entirely on UniProt, thereby facilitating direct comparisons. This approach has previously been adopted by PrecisionProDB[191] and would be a valuable enhancement to the tools presented here, improving the interoperability and flexibility of the workflow.

Integrating UniProt protein identifiers into ProHap Explorer would also enhance user experience, as participants in the tool evaluation highlighted UniProt IDs as an intuitive and familiar way to search for proteins of interest. Such integration would make the platform more accessible to a broader community of researchers and align it even better with widely used resources in proteomics. Beyond mapping identifiers, an integration with UniProt may facilitate access to protein structure data available from the Protein Data Bank (PDB)[21], and variant loci may be visualized within the retrieved protein structures.

While the interaction design implemented in Paper 2 generally adheres to Shneiderman's Visual Information Seeking Mantra, there are additional recommended tasks that have yet to be fully realized. Shneiderman's expanded framework includes seven tasks: Overview, Zoom, Filter, Details-on-demand, Relate, History, and Extract[239]. Of these, the Relate task is only partially supported; users can compare protein haplotype and canonical sequences encoded by the same gene, and examine peptide identifications within these pairs across tissues or datasets. However, relating protein haplotypes or peptide identifications between different genes is not yet implemented. The History task, which supports actions such as "undo" and progressive refinement, is also limited—currently, users can switch between different genes, each opened in a dedicated tab, but more robust history tracking may be considered. Finally, a zoom feature in the interactive visualizations representing a particular gene would be beneficial to allow users a close inspection of gene regions with a high level of variation.

Validation of ProHap Explorer has so far been informal, involving a small panel of experts who were not involved in its development. Formal validation methods, such as the System Usability Scale (SUS), require a large number of users, while frameworks like ICE-T demand significant time investment and expert input. Due to these constraints, formal validation was not prioritized within the available development timeframe. Nevertheless, rigorous validation

will be considered in future iterations of ProHap Explorer, especially as the tool evolves and is potentially integrated with other resources. This will ensure that the platform meets the needs of the broader scientific community and maintains high standards of usability and reliability.

5.4 Impact of haplotype-aware proteomics

Proteogenomic studies in medical research that use mass spectrometry most commonly focus on personalized analyses, investigating the effects of rare variants. In contrast, population-based proteomic studies have primarily examined the impact of variants in non-coding regions on protein abundance. The work presented here enables the inclusion of protein sequence-altering variants in both individualized and population-level studies. In Paper 2, we demonstrate this by searching for common variants in blood serum samples from individuals after weight loss, as well as by investigating signatures of personal haplotypes in a stem cell experiment using the donor's genotype.

We also show in Paper 2 that a substantial portion of the proteome may map to variant peptides across all superpopulations included in the 1kGP. These findings reflect both differences between individual genomes and the reference sequence, as well as the level of diversity within defined groups. It is well established that the African superpopulation exhibits greater diversity between individuals than other groups[109,285], which explains the larger proportion of the proteome potentially mapping to variant peptides in this population. Furthermore, even at the individual level, people of African ancestry differ from the reference proteome to a greater extent than most other 1kGP participants (Figure 5.1).

Looking ahead, ProHap could be used with mass spectrometry-based proteomics to supplement GWAS and pQTL studies for polygenic traits. For example, after identifying variants significantly associated with a trait or disease, can we observe the effects of these variants—or others in linkage disequilibrium—on the protein sequence? Are the corresponding variant peptides detectable? The ability to perform biobank-scale proteomic analysis accounting for protein haplotypes will enable a better understanding of polygenic disease risk and the associated molecular mechanisms.

With the adaptation of haplotype-aware approaches to DIA proteomics, the ability to reliably quantify variant peptides will enable us to compare their abundance to that of reference peptides, providing insights into protein stability, secretion, or degradation. For instance, the CEL protein truncated by a frameshift variant tends to aggregate in the pancreas, and is secreted into the blood at a lower rate, causing maturity-onset diabetes of the young (MODY) type 8[286,287]. The alternative protein allele would therefore be observed in a higher abundance in the pancreatic tissue than the reference allele. Would a systematic comparison of the abundance of different alleles in proteins help explain their role in polygenic traits?

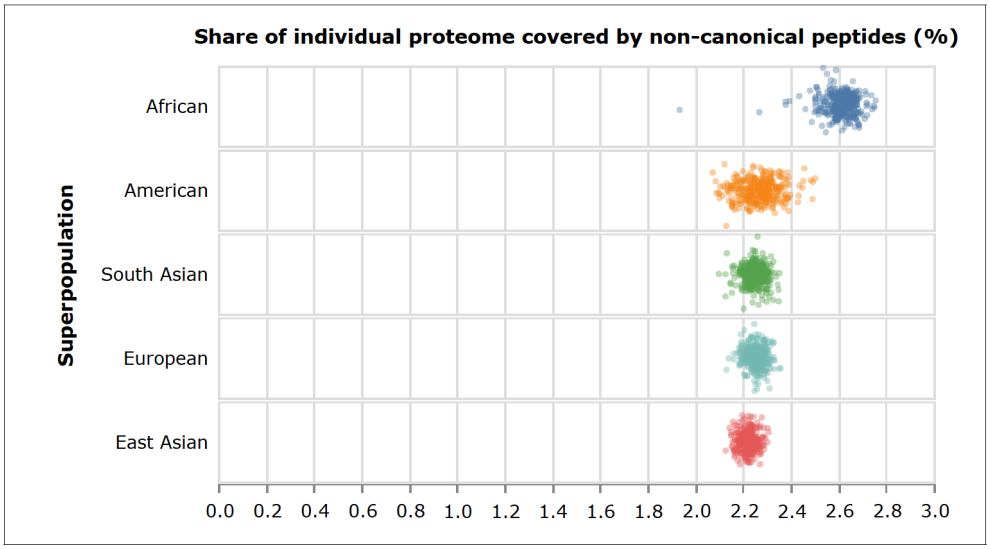


Figure 5.1: Percentage of the individual proteome covered by variant peptides among the participants of the 1kGP, grouped by superpopulation, where each dot represents an individual.

We have also briefly addressed the application of protein haplotype analysis in cell line studies. As demonstrated in Paper 2, when working with stem cells derived from a specific donor, incorporating the donor's genotype into the proteomic search space enables the detection of variant peptides in a manner similar to that observed in blood or tissue samples. However, HeLa cells, and many other established cell lines, exhibit extensive genomic abnormalities compared to the normal human genome, including aneuploidy (changes in chromosome number) and complex structural rearrangements[288]. In HeLa, these features are partly a reflection of its cancerous origin but may also be the result of long-term adaptation to cell culture conditions[288]. While accounting for the specific genomic landscape of cell lines could improve the accuracy of such proteomic studies, the presence of chromosomal abnormalities and large structural variants would require alternative strategies to fully capture the proteomic complexity of these cell lines.

Furthermore, the concept of protein haplotypes has so far been explored only in human studies. While there is potential to extend these approaches to model organisms with well-annotated diploid genomes, such as mice, the benefit of accounting for common haplotypes in animal studies has yet to be described.

5.5 Ethical considerations

A growing concern in proteogenomics is that proteomic data may require the same level of confidentiality as genomic data if enough rare variant peptides are detected and linked to specific

individual haplotypes[289]. In the context of multi-omics association studies, where genetic information is already included, this development should not introduce new ethical challenges, as participants provide informed consent for secure handling of their genomic data. While we have shown that it is possible to confidently map hundreds of variant peptides in an individual without accessing their genetic information, in our case, these variants are common and cannot uniquely identify an individual within the general population. Nevertheless, research in forensic science has demonstrated that even this number of variant peptides can help distinguish one individual from another[290]. Moreover, as ProHap can now generate protein haplotype databases from large genomic datasets, such as the All of Us Research Program, there is a realistic possibility that increasingly detailed proteomic references will reflect rare haplotypes. These developments will have implications for informed consent regarding the use of proteomic data, as well as for data publishing and protection practices[289].

Beyond data confidentiality, other challenges remain. Genomic studies have highlighted biases introduced by the overrepresentation of European populations in reference datasets. GWAS is commonly performed exclusively on individuals of European ancestry (see Section 1.2.6), and the percentage of participants of European ancestry in GWAS has increased since 2017, reaching over 88% as of August 2025, as obtained from the list of studies available in the GWAS Catalog[291]. The extent of European overrepresentation in proteomic studies has not been well described.

Our approach enables haplotype-aware analysis in proteomics using any available panel of haplotypes – including underrepresented populations. However, when obtaining informed consent from individuals in vulnerable groups, special consideration of risks, benefits, and community involvement is essential[122]. These steps are crucial, and while the methods developed here can be applied in line with the best practices, and the software tools executed in secure environments without breaching data privacy, it is likely that haplotype-aware proteomic analysis will first be applied in populations already well represented in genomic studies.

6 Conclusion

In summary, this thesis advances the field of proteogenomics by enabling the systematic identification and analysis of common protein sequence variants at both individual and population levels. By integrating haplotype-aware databases, robust data processing pipelines, and an interactive visualization platform, this work provides a new framework for exploring proteomic diversity encoded by common genetic variation. The results demonstrate that based on the 1000 Genomes Project data, a substantial share of the human proteome may be obscured to mass spectrometry-based proteomic analysis unless common haplotypes are accounted for. Moreover, individuals of the African superpopulation present the highest share of the proteome covered by variant peptides both individually and combined, indicating that proteomic analyses not accounting for haplotype diversity will be biased against these populations.

We also show that haplotype-aware approaches are capable of identifying peptides encoded by common haplotypes in mass spectrometry-based proteomic datasets. These findings may lay the groundwork for more nuanced studies of the biological and clinical significance of protein haplotypes, and for the integration of haplotype-aware proteomic workflows into population-scale multi-omics studies. To better chart the overall impact of common haplotypes on the human proteome, we developed a visual web-based tool to allow researchers to explore the protein haplotypes encoded by any gene. Identifications of peptides in public mass spectrometry-based datasets are mapped to the protein haplotype sequences, and peptides identified in specific tissues, or in specific datasets can be distinguished. This tool is available online and will be updated to further encompass more public mass spectrometry data.

The examples illustrating the principles of precision medicine – the identification of the causal mechanisms of an individual autoimmune condition, and the improvements of treatment options for several subclasses of diabetes – have shown how the close inspection of candidate genes, or of the genetic makeup of a single individual, can allow for life-changing medical interventions. This work was completed with the hope that the developed approaches and resources will open new avenues for investigating the functional consequences of genetic variation in an unbiased manner at scale, improving applications in risk scores, disease association studies, and precision medicine.

7 Future perspectives

So far, this work has demonstrated that including common protein haplotypes in the proteomic search space can enhance peptide identification in healthy tissues and plasma. By accounting for the actual combinations of genetic variants present in individuals, we address a key limitation of traditional reference-based searches, which may miss variant peptides or misassign spectra, especially in genetically diverse populations. The initial findings in healthy samples lay a foundation for extending these methods to more complex biomedical contexts.

Most applications of these proteogenomic methods will focus on understanding disease phenotypes. In particular, the study of polygenic non-communicable diseases stands to benefit from a direct investigation of variant proteins themselves, rather than solely relying on genetic association studies. By identifying and quantifying variant peptides, we can begin to systematically answer new questions about the functional consequences of genetic variation. For example, how do variant proteins interact with other proteins, or form functional complexes and polymers? How does the presence of variant peptides influence the results of pQTL studies, especially in regions of high linkage disequilibrium where multiple variants may co-occur and jointly alter protein sequences?

Additionally, the systematic inclusion of haplotype information allows us to investigate the significance of variant combinations that only occur together. This enables us to address important questions, such as whether certain combinations of variants are required for protein stability or function. Furthermore, certain rare variants may only become pathogenic in the context of specific common variants located in the same gene or elsewhere. These approaches also provide new opportunities to explore evolutionary processes by revealing patterns of variant co-occurrence that may reflect selective pressures or functional constraints over time. By integrating proteomic data with haplotype information, we open new avenues for understanding the molecular mechanisms underlying complex traits and for improving the interpretation of genetic association studies in biomedical research.

Today, many biobanks have paired exome or genome sequencing data with biological samples such as blood serum, providing a rich resource for integrative proteogenomic analysis. Norway and the Nordic countries, for example, offer a unique setting with a moderately diverse population, and benefit from a unified, high-quality healthcare system and excellent biobank

infrastructure[292]. Applying the proteogenomic approaches developed in this thesis to such datasets could greatly enhance our understanding of disease mechanisms, explain regional differences in disease incidence, and enable more nuanced stratification or clustering of patients based on both genetic and proteomic profiles.

Finally, this thesis presents two novel software tools, whose utility has been demonstrated on selected example datasets. However, the true impact of these tools will ultimately depend on their adoption and sustained use by the broader research community. To ensure long-term relevance, it is essential that these tools are actively maintained and updated over time. The most effective strategy will likely involve merging them with similar resources into a collaboratively maintained toolset. Moreover, incorporating peptide identifications from multiple proteomic datasets into ProHap Explorer, spanning human tissues and blood plasma samples, is essential to provide a reliable resource, potentially supplementing well-established platforms such as gnomAD[72]. With this in mind, we are currently making plans to secure the future development and integration of the tools introduced in this thesis, aiming to maximize their value and usability for researchers in proteogenomics and related fields.

Bibliography

[1] E. A. Ashley. Towards precision medicine. *Nature Reviews Genetics*, 17(9):507–522, Sept. 2016. doi: 10.1038/nrg.2016.86.

[2] E. A. Worthey, A. N. Mayer, G. D. Syverson, et al. Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine*, 13(3):255–262, Mar. 2011. doi: 10.1097/GIM.0b013e3182088158.

[3] A. L. Gloyn, E. R. Pearson, J. F. Antcliff, et al. Activating Mutations in the Gene Encoding the ATP-Sensitive Potassium-Channel Subunit Kir6.2 and Permanent Neonatal Diabetes. *New England Journal of Medicine*, 350(18):1838–1849, Apr. 2004. doi: 10.1056/NEJMoa032922.

[4] E. R. Pearson, I. Flechtner, P. R. Njølstad, et al. Switching from Insulin to Oral Sulfonylureas in Patients with Diabetes Due to Kir6.2 Mutations. *New England Journal of Medicine*, 355(5):467–477, Aug. 2006. doi: 10.1056/NEJMoa061759.

[5] O. Søvik, P. Njølstad, I. Følling, et al. Hyperexcitability to sulphonylurea in MODY3. *Diabetologia*, 41(5):607–608, Apr. 1998. doi: 10.1007/s001250050956.

[6] E. R. Pearson, S. Pruhova, C. J. Tack, et al. Molecular genetics and phenotypic characteristics of MODY caused by hepatocyte nuclear factor 4a mutations in a large European collection. *Diabetologia*, 48(5), May 2005. doi: 10.1007/s00125-005-1738-y.

[7] A. Molven and P. R. Njølstad. Role of molecular genetics in transforming diagnosis of diabetes mellitus. *Expert Review of Molecular Diagnostics*, 11(3):313–320, Apr. 2011. doi: 10.1586/erm.10.123.

[8] S. Johansson, H. Irgens, K. K. Chudasama, et al. Exome Sequencing and Genetic Testing for MODY. *PLOS ONE*, 7(5):e38050, May 2012. doi: 10.1371/journal.pone.0038050.

[9] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2011. ISBN 978-0-309-22222-8.

[10] A. K. Manrai, B. H. Funke, H. L. Rehm, et al. Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*, 375(7):655–665, Aug. 2016. doi: 10.1056/NEJMsa1507092.

[11] G. Sirugo, S. M. Williams, and S. A. Tishkoff. The Missing Diversity in Human Genetic Studies. *Cell*, 177(1):26–31, Mar. 2019. doi: 10.1016/j.cell.2019.02.048.

[12] B. Berger, J. Peng, and M. Singh. Computational solutions for omics data. *Nature Reviews Genetics*, 14(5):333–346, May 2013. doi: 10.1038/nrg3433.

[13] K. J. Karczewski and M. P. Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299–310, May 2018. doi: 10.1038/nrg.2018.4.

[14] N. G. Anderson, A. Matheson, and N. L. Anderson. Back to the future: The human protein index (HPI) and the agenda for post-proteomic biology. *PROTEOMICS*, 1(1): 3–12, 2001.

[15] R. Aebersold, L. E. Hood, and J. D. Watts. Equipping scientists for the new biology. *Nature Biotechnology*, 18(4):359–359, Apr. 2000. doi: 10.1038/74325.

[16] S. D. Patterson and R. H. Aebersold. Proteomics: the first decade and beyond. *Nature Genetics*, 33(3):311–323, Mar. 2003. doi: 10.1038/ng1106.

[17] M. A. Gillette, C. R. Jimenez, and S. A. Carr. Clinical Proteomics: A Promise Becoming Reality. *Molecular & Cellular Proteomics*, 23(2), Feb. 2024. doi: 10.1016/j.mcpro.2023.100688.

[18] R. Mayeux. Biomarkers: Potential uses and limitations. *NeuroRX*, 1(2):182–188, Apr. 2004. doi: 10.1602/neurorx.1.2.182.

[19] V. Özdemir, E. S. Dove, U. K. Gürsoy, et al. Personalized medicine beyond genomics: alternative futures in big data—proteomics, environtome and the social proteome. *Journal of Neural Transmission*, 124(1):25–32, Jan. 2017. doi: 10.1007/s00702-015-1489-y.

[20] C. Buccitelli and M. Selbach. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, Oct. 2020. doi: 10.1038/s41576-020-0258-4.

[21] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, Jan. 2019. doi: 10.1093/nar/gky949.

[22] Y. Perez-Riverol, W. Bittremieux, W. S. Noble, et al. Open-Source and FAIR Research Software for Proteomics. *Journal of Proteome Research*, Apr. 2025. doi: 10.1021/acs.jproteome.4c01079.

[23] E. S. Lander, L. M. Linton, B. Birren, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001. doi: 10.1038/35057062.

[24] H.-B. Zhang and C. Wu. BAC as tools for genome sequencing. *Plant Physiology and Biochemistry*, 39(3):195–209, Mar. 2001. doi: 10.1016/S0981-9428(00)01236-5.

[25] M. Gross. Riding the wave of biological data. *Current Biology*, 21(6):R204–R206, Mar. 2011. doi: 10.1016/j.cub.2011.03.009.

[26] W. W. Soon, M. Hariharan, and M. P. Snyder. High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, 9(1):640, Jan. 2013. doi: 10.1038/msb.2012.61.

[27] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, Mar. 2009. doi: 10.1186/gb-2009-10-3-r25.

[28] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. doi: 10.1093/bioinformatics/btp324.

[29] Y. Perez-Riverol, C. Bandla, D. Kundu, et al. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Research*, 53(D1):D543–D553, Jan. 2025. doi: 10.1093/nar/gkae1011.

[30] C. Dai, J. Pfeuffer, H. Wang, et al. quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. *Nature Methods*, 21(9):1603–1607, Sept. 2024. doi: 10.1038/s41592-024-02343-1.

[31] M. Vaudel, A. S. Venne, F. S. Berven, et al. Shedding light on black boxes in protein identification. *PROTEOMICS*, 14(9):1001–1005, 2014. doi: 10.1002/pmic.201300488.

[32] V. Jalili, E. Afgan, Q. Gu, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research*, 48(W1):W395–W402, July 2020. doi: 10.1093/nar/gkaa434.

[33] Y. M. Farag, C. Horro, M. Vaudel, and H. Barsnes. PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data. *Journal of Proteome Research*, 20(12):5419–5423, Dec. 2021. doi: 10.1021/acs.jproteome.1c00678.

[34] A. Campbell. Genetics (as a Discipline). In S. Maloy and K. Hughes, editors, *Brenner's Encyclopedia of Genetics (Second Edition)*, page 284. Academic Press, San Diego, Jan. 2013. ISBN 978-0-08-096156-9. doi: 10.1016/B978-0-12-374984-0.00632-X.

[35] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, Apr. 1953. doi: 10.1038/171737a0.

[36] R. Wu and A. D. Kaiser. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, 35(3):523–537, Jan. 1968. doi: 10.1016/S0022-2836(68)80012-9.

[37] A. M. Giani, G. R. Gallo, L. Gianfranceschi, and G. Formenti. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18:9–19, Jan. 2020. doi: 10.1016/j.csbj.2019.11.002.

[38] R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, 1979. doi: 10.1093/nar/6.7.2601.

[39] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct. 2004. doi: 10.1038/nature03001.

[40] Y. Guo, Y. Dai, H. Yu, et al. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, Mar. 2017. doi: 10.1016/j.ygeno.2017.01.005.

[41] S. Ballouz, A. Dobin, and J. A. Gillis. Is it time to change the reference genome? *Genome Biology*, 20(1):159, Aug. 2019. doi: 10.1186/s13059-019-1774-4.

[42] R. Chen and A. J. Butte. The reference human genome demonstrates high risk of type 1 diabetes and other disorders. In *Biocomputing 2011*, pages 231–242. WORLD SCIENTIFIC, Nov. 2011. ISBN 978-981-4335-04-1. doi: 10.1142/9789814335058_0025.

[43] Human Genome Overview - Genome Reference Consortium. URL `https://www.ncbi.nlm.nih.gov/grc/human`.

[44] D. M. Church, V. A. Schneider, T. Graves, et al. Modernizing Reference Genome Assemblies. *PLOS Biology*, 9(7):e1001091, July 2011. doi: 10.1371/journal.pbio.1001091.

[45] P. Flicek, B. L. Aken, B. Ballester, et al. Ensembl's 10th year. *Nucleic Acids Research*, 38 (suppl_1):D557–D562, Jan. 2010. doi: 10.1093/nar/gkp972.

[46] E pluribus unum. *Nature Methods*, 7(5):331–331, May 2010. doi: 10.1038/nmeth0510-331.

[47] S. Nurk, S. Koren, A. Rhie, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, Apr. 2022. doi: 10.1126/science.abj6987.

[48] Z. D. Stephens, S. Y. Lee, F. Faghri, et al. Big Data: Astronomical or Genomical? *PLOS Biology*, 13(7):e1002195, July 2015. doi: 10.1371/journal.pbio.1002195.

[49] N. D. Olson, J. Wagner, N. Dwarshuis, et al. Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics*, 24(7):464–483, July 2023. doi: 10.1038/s41576-023-00590-0.

[50] I. Kockum, J. Huang, and P. Stridh. Overview of Genotyping Technologies and Methods. *Current Protocols*, 3(4):e727, Apr. 2023. doi: 10.1002/cpz1.727.

[51] L. Stein. Genome annotation: from sequence to biology. *Nature Reviews Genetics*, 2(7): 493–503, July 2001. doi: 10.1038/35080529.

[52] I. Ezkurdia, D. Juan, J. M. Rodriguez, et al. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23 (22):5866–5878, Nov. 2014. doi: 10.1093/hmg/ddu309.

[53] S. L. Salzberg. Open questions: How many genes do we have? *BMC Biology*, 16(1):94, Aug. 2018. doi: 10.1186/s12915-018-0564-x.

[54] J. Morales, S. Pujar, J. E. Loveland, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604(7905):310–315, Apr. 2022. doi: 10.1038/s41586-022-04558-8.

[55] S. C. Dyer, O. Austine-Orimoloye, A. G. Azov, et al. Ensembl 2025. *Nucleic Acids Research*, 53(D1):D948–D957, Jan. 2025. doi: 10.1093/nar/gkae1071.

[56] K. R. Chi. The dark side of the human genome. *Nature*, 538(7624):275–277, Oct. 2016. doi: 10.1038/538275a.

[57] G. McVicker, B. van de Geijn, J. F. Degner, et al. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science*, 342(6159):747–749, Nov. 2013. doi: 10.1126/science.1242429.

[58] F. Zhang and J. R. Lupski. Non-coding genetic variants in human disease. *Human Molecular Genetics*, 24(R1):R102–R110, Oct. 2015. doi: 10.1093/hmg/ddv259.

[59] J. Rogers and R. Wall. A mechanism for RNA splicing. *Proceedings of the National Academy of Sciences*, 77(4):1877–1879, Apr. 1980. doi: 10.1073/pnas.77.4.1877.

[60] A. Newman. RNA splicing. *Current Biology*, 8(25):R903–R905, Dec. 1998. doi: 10.1016/S0960-9822(98)00005-0.

[61] B. Rabbani, M. Tekin, and N. Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, 59(1):5–15, Jan. 2014. doi: 10.1038/jhg.2013.114.

[62] A. Zien, B. Schölkopf, K. Tsuda, and J. Vert. A primer on molecular biology. *Kernel Methods in Computational Biology, 3-34 (2004)*, Jan. 2004.

[63] P. Sieber, M. Platzer, and S. Schuster. The Definition of Open Reading Frame Revisited. *Trends in Genetics*, 34(3):167–170, Mar. 2018. doi: 10.1016/j.tig.2017.12.009.

[64] J.-P. Couso and P. Patraquim. Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology*, 18(9):575–589, Sept. 2017. doi: 10.1038/nrm.2017.58.

[65] J. M. Mudge, J. Ruiz-Orera, J. R. Prensner, et al. Standardized annotation of translated open reading frames. *Nature Biotechnology*, 40(7):994–999, July 2022. doi: 10.1038/s41587-022-01369-0.

[66] M. Lek, K. J. Karczewski, E. V. Minikel, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, Aug. 2016. doi: 10.1038/nature19057.

[67] J. G. Hall. Review and hypotheses: somatic mosaicism: observations related to clinical genetics. *American Journal of Human Genetics*, 43(4):355–363, Oct. 1988.

[68] L. V. Horebeek, B. Dubois, and A. Goris. Somatic Variants: New Kids on the Block in Human Immunogenetics. *Trends in Genetics*, 35(12):935–947, Dec. 2019. doi: 10.1016/j.tig.2019.09.005.

[69] M. Mohiuddin, R. F. Kooy, and C. E. Pearson. De novo mutations, genetic mosaicism and human disease. *Frontiers in Genetics*, 13, Sept. 2022. doi: 10.3389/fgene.2022.983668.

[70] S. Richards, N. Aziz, S. Bale, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–424, May 2015. doi: 10.1038/gim.2015.30.

[71] D. M. Fowler and H. L. Rehm. Will variants of uncertain significance still exist in 2030? *The American Journal of Human Genetics*, 111(1):5–10, Jan. 2024. doi: 10.1016/j.ajhg.2023.11.005.

[72] K. J. Karczewski, L. C. Francioli, G. Tiao, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, May 2020. doi: 10.1038/s41586-020-2308-7.

[73] J. Cheng, G. Novati, J. Pan, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 0(0):eadg7492, Sept. 2023. doi: 10.1126/science.adg7492.

[74] J. T. den Dunnen, R. Dalgleish, D. R. Maglott, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*, 37(6):564–569, 2016. doi: 10.1002/humu.22981.

[75] J. T. den Dunnen. Sequence Variant Descriptions: HGVS Nomenclature and Mutalyzer. *Current Protocols in Human Genetics*, 90(1):7.13.1–7.13.19, 2016. doi: 10.1002/cphg.2.

[76] M. Cargill, D. Altshuler, J. Ireland, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3):231–238, July 1999. doi: 10.1038/10290.

[77] N. J. Schork, D. Fallin, and J. S. Lanchbury. Single nucleotide polymorphisms and the future of genetic epidemiology. *Clinical Genetics*, 58(4):250–264, 2000. doi: 10.1034/j.1399-0004.2000.580402.x.

[78] M. Vihinen. Variation Ontology for annotation of variation effects and mechanisms. *Genome Research*, 24(2):356–364, Feb. 2014. doi: 10.1101/gr.157495.113.

[79] M. Mort, D. Ivanov, D. N. Cooper, and N. A. Chuzhanova. A meta-analysis of nonsense mutations causing human genetic disease. *Human Mutation*, 29(8):1037–1047, 2008. doi: 10.1002/humu.20763.

[80] R. C. Hunt, V. L. Simhadri, M. Iandoli, et al. Exposing synonymous mutations. *Trends in Genetics*, 30(7):308–321, July 2014. doi: 10.1016/j.tig.2014.04.006.

[81] M. Vihinen. When a Synonymous Variant Is Nonsynonymous. *Genes*, 13(8):1485, Aug. 2022. doi: 10.3390/genes13081485.

[82] N.-C. Chen, L. F. Paulin, F. J. Sedlazeck, et al. Improved sequence mapping using a complete reference genome and lift-over. *Nature Methods*, 21(1):41–49, Jan. 2024. doi: 10.1038/s41592-023-02069-6.

[83] C. Ormond, N. M. Ryan, A. Corvin, and E. A. Heron. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Briefings in Bioinformatics*, 22(5):bbab069, Sept. 2021. doi: 10.1093/bib/bbab069.

[84] D. E. Reich, M. Cargill, S. Bolk, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, May 2001. doi: 10.1038/35075590.

[85] D. F. Conrad, M. Jakobsson, G. Coop, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, 38(11):1251–1260, Nov. 2006. doi: 10.1038/ng1911.

[86] S. McCarthy, S. Das, W. Kretzschmar, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279–1283, Oct. 2016. doi: 10.1038/ng.3643.

[87] D. Freije, C. Helms, M. S. Watson, and H. Donis-Keller. Identification of a Second Pseudoautosomal Region Near the Xq and Yq Telomeres. *Science*, 258(5089):1784–1787, Dec. 1992. doi: 10.1126/science.1465614.

[88] A. Flaquer, G. A. Rappold, T. F. Wienker, and C. Fischer. The human pseudoautosomal regions: a review for genetic epidemiologists. *European Journal of Human Genetics*, 16 (7):771–779, July 2008. doi: 10.1038/ejhg.2008.63.

[89] Y. Choi, A. P. Chan, E. Kirkness, et al. Comparison of phasing strategies for whole human genomes. *PLOS Genetics*, 14(4):e1007308, Apr. 2018. doi: 10.1371/journal.pgen.1007308.

[90] M. Martin, P. Ebert, and T. Marschall. Read-Based Phasing and Analysis of Phased Variants with WhatsHap. In B. A. Peters and R. Drmanac, editors, *Haplotyping: Methods and Protocols*, pages 127–138. Springer US, New York, NY, 2023. ISBN 978-1-0716-2819-5. doi: 10.1007/978-1-0716-2819-5_8.

[91] V. Bansal. Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes. *Bioinformatics*, 35(14):i242–i248, July 2019. doi: 10.1093/bioinformatics/btz329.

[92] W. Spooner, W. McLaren, T. Slidel, et al. Haplosaurus computes protein haplotypes for use in precision drug design. *Nature Communications*, 9(1):4128, Oct. 2018. doi: 10.1038/s41467-018-06542-1.

[93] G. Hellenthal, G. B. J. Busby, G. Band, et al. A Genetic Atlas of Human Admixture History. *Science*, 343(6172):747–751, Feb. 2014. doi: 10.1126/science.1243518.

[94] A. Bergström, S. A. McCarthy, R. Hui, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484):eaay5012, Mar. 2020. doi: 10.1126/science.aay5012.

[95] H. Lee, W. Kim, N. Kwon, et al. Lessons from national biobank projects utilizing whole-genome sequencing for population-scale genomics. *Genomics & Informatics*, 23(1):8, Mar. 2025. doi: 10.1186/s44342-025-00040-9.

[96] E. Uffelmann, Q. Q. Huang, N. S. Munung, et al. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, Aug. 2021. doi: 10.1038/s43586-021-00056-9.

[97] D. J. M. Crouch and W. F. Bodmer. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences*, 117(32):18924–18933, Aug. 2020. doi: 10.1073/pnas.2005634117.

[98] A. C. Fahed, M. Wang, J. R. Homburger, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications*, 11(1): 3635, Aug. 2020. doi: 10.1038/s41467-020-17374-3.

[99] C. M. Lewis and E. Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, 12(1):44, May 2020. doi: 10.1186/s13073-020-00742-5.

[100] M. L. Page, E. L. Vance, M. E. Cloward, et al. The Polygenic Risk Score Knowledge Base offers a centralized online repository for calculating and contextualizing polygenic risk scores. *Communications Biology*, 5(1):899, Sept. 2022. doi: 10.1038/s42003-022-03795-x.

[101] N. J. Lennon, L. C. Kottyan, C. Kachulis, et al. Selection, optimization and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse US populations. *Nature Medicine*, 30(2):480–487, Feb. 2024. doi: 10.1038/s41591-024-02796-z.

[102] C. Bycroft, C. Freeman, D. Petkova, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, Oct. 2018. doi: 10.1038/s41586-018-0579-z.

[103] C. Chatelain, S. Lessard, K. Klinger, et al. Building a human genetic data lake to scale up insights for drug discovery. *Drug Discovery Today*, 30(6):104385, June 2025. doi: 10.1016/j.drudis.2025.104385.

[104] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, May 2008. doi: 10.1038/nrg2344.

[105] A. R. Martin, M. Kanai, Y. Kamatani, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, Apr. 2019. doi: 10.1038/s41588-019-0379-x.

[106] J. Morales, D. Welter, E. H. Bowler, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biology*, 19(1):21, Feb. 2018. doi: 10.1186/s13059-018-1396-2.

[107] E. Stamatakis, K. B. Owen, L. Shepherd, et al. Is Cohort Representativeness Passé? Poststratified Associations of Lifestyle Risk Factors with Mortality in the UK Biobank. *Epidemiology*, 32(2):179, Mar. 2021. doi: 10.1097/EDE.0000000000001316.

[108] T. Schoeler, D. Speed, E. Porcu, et al. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, 7(7):1216–1227, July 2023. doi: 10.1038/s41562-023-01579-9.

[109] A. Auton, G. R. Abecasis, D. M. Altshuler, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct. 2015. doi: 10.1038/nature15393.

[110] T. A. of Us Research Program Investigators. The “All of Us” Research Program. *New England Journal of Medicine*, 381(7):668–676, Aug. 2019. doi: 10.1056/NEJMsr1809937.

[111] E. Wong, N. Bertin, M. Hebrard, et al. The Singapore National Precision Medicine Strategy. *Nature Genetics*, 55(2):178–186, Feb. 2023. doi: 10.1038/s41588-022-01274-x.

[112] A. Nagai, M. Hirata, Y. Kamatani, et al. Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, 27(Supplement_III):S2–S8, 2017. doi: 10.1016/j.je.2016.12.005.

[113] R. I. Amann, S. Baichoo, B. J. Blencowe, et al. Toward unrestricted use of public genomic data. *Science*, 363(6425):350–352, Jan. 2019. doi: 10.1126/science.aaw1280.

[114] M. Hudson, N. A. Garrison, R. Sterling, et al. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nature Reviews Genetics*, 21 (6):377–384, June 2020. doi: 10.1038/s41576-020-0228-x.

[115] B. Salter and C. Salter. Controlling new knowledge: Genomic science, governance and the politics of bioinformatics. *Social Studies of Science*, 47(2):263–287, Apr. 2017. doi: 10.1177/0306312716681210.

[116] R. R. McInnes. 2010 Presidential Address: Culture: The Silent Language Geneticists Must Learn— Genetic Research with Indigenous Populations. *The American Journal of Human Genetics*, 88(3):254–261, Mar. 2011. doi: 10.1016/j.ajhg.2011.02.014.

[117] N. A. Garrison, M. Hudson, L. L. Ballantyne, et al. Genomic Research Through an Indigenous Lens: Understanding the Expectations. *Annual Review of Genomics and Human Genetics*, 20(Volume 20, 2019):495–517, Aug. 2019. doi: 10.1146/annurev-genom-083118-015434.

[118] E. G. Cohn, M. Husamudeen, E. L. Larson, and J. K. Williams. Increasing Participation in Genomic Research and Biobanking Through Community-Based Capacity Building. *Journal of Genetic Counseling*, 24(3):491–502, June 2015. doi: 10.1007/s10897-014-9768-6.

[119] K. Fox. The Illusion of Inclusion — The "All of Us" Research Program and Indigenous Peoples' DNA. *New England Journal of Medicine*, 383(5):411–413, July 2020. doi: 10.1056/NEJMp1915987.

[120] L. Arbour and D. Cook. DNA on Loan: Issues to Consider when Carrying Out Genetic Research with Aboriginal Families and Communities. *Community Genetics*, 9(3):153–160, June 2006. doi: 10.1159/000092651.

[121] A. A. Valiani. Frontiers of Bio-Decolonization: Indigenous Data Sovereignty as a Possible Model for Community-Based Participatory Genomic Health Research for Racialized Peoples in Postgenomic Canada. *Genealogy*, 6(3):68, Sept. 2022. doi: 10.3390/genealogy6030068.

[122] K. G. Claw, M. Z. Anderson, R. L. Begay, et al. A framework for enhancing ethical genomic research with Indigenous communities. *Nature Communications*, 9(1):2957, July 2018. doi: 10.1038/s41467-018-05188-3.

[123] M. Tauali'i, E. L. Davis, K. L. Braun, et al. Native Hawaiian Views on Biobanking. *Journal of Cancer Education*, 29(3):570–576, Sept. 2014. doi: 10.1007/s13187-014-0638-6.

[124] M. Hudson, A. Beaton, M. Milne, et al. *Te Mata Ira: Guidelines for Genomic Research with Maori*. Te Mata Hautū Taketake – Māori & Indigenous Governance Centre, University of Waikato, Hamilton, New Zealand, 2016. ISBN 978-0-473-36937-8.

[125] N. R. Caron, M. Chongo, M. Hudson, et al. Indigenous Genomic Databases: Pragmatic Considerations and Cultural Contexts. *Frontiers in Public Health*, 8, Apr. 2020. doi: 10.3389/fpubh.2020.00111.

[126] J. Golan, K. Riddle, M. Hudson, et al. Benefit sharing: Why inclusive provenance metadata matter. *Frontiers in Genetics*, 13, Sept. 2022. doi: 10.3389/fgene.2022.1014044.

[127] J. Ambler, A. A. Diallo, P. K. Dearden, et al. Including Digital Sequence Data in the Nagoya Protocol Can Promote Data Sharing. *Trends in Biotechnology*, 39(2):116–125, Feb. 2021. doi: 10.1016/j.tibtech.2020.06.009.

[128] S. P. Robertson, J. H. Hindmarsh, S. Berry, et al. Genomic medicine must reduce, not compound, health inequities: the case for hauora-enhancing genomic resources for New Zealand. *The New Zealand Medical Journal (Online)*, 131(1480):81–89, Aug. 2018.

[129] The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, page bbw089, Oct. 2016. doi: 10.1093/bib/bbw089.

[130] W.-W. Liao, M. Asri, J. Ebler, et al. A draft human pangenome reference. *Nature*, 617 (7960):312–324, May 2023. doi: 10.1038/s41586-023-05896-x.

[131] G. Hickey, J. Monlong, J. Ebler, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, 42(4):663–673, Apr. 2024. doi: 10.1038/s41587-023-01793-w.

[132] T. E. Angel, U. K. Aryal, S. M. Hengel, et al. Mass spectrometry based proteomics: existing capabilities and future directions. *Chemical Society Reviews*, 41(10):3912–3928, May 2012. doi: 10.1039/c2cs15331a.

[133] M.-S. Kim, S. M. Pinto, D. Getnet, et al. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014. doi: 10.1038/nature13302.

[134] D. Wang, B. Eraslan, T. Wieland, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, 15(2):e8503, Feb. 2019. doi: 10.15252/msb.20188503.

[135] K. Tsuo, M. A. Argentieri, D. Gadd, et al. Proteomic prediction of disease largely reflects environmental risk exposure, Aug. 2025.

[136] S. Adhikari, E. C. Nice, E. W. Deutsch, et al. A high-stringency blueprint of the human proteome. *Nature Communications*, 11(1):5301, Oct. 2020. doi: 10.1038/s41467-020-19045-9.

[137] M. Su, Z. Zhang, L. Zhou, et al. Proteomics, Personalized Medicine and Cancer. *Cancers*, 13(11):2512, Jan. 2021. doi: 10.3390/cancers13112512.

[138] P. Y. Lee, N. Saraygord-Afshari, and T. Y. Low. The evolution of two-dimensional gel electrophoresis - from proteomics to emerging alternative applications. *Journal of Chromatography A*, 1615:460763, Mar. 2020. doi: 10.1016/j.chroma.2019.460763.

[139] S. Xie, M. , Colby, B. , Betul, et al. Emerging affinity-based techniques in proteomics. *Expert Review of Proteomics*, 6(5):573–583, Oct. 2009. doi: 10.1586/epr.09.74.

[140] O. Stoevesandt, , and M. J. Taussig. Affinity proteomics: the role of specific binding reagents in human proteome analysis. *Expert Review of Proteomics*, 9(4):401–414, Aug. 2012. doi: 10.1586/epr.12.34.

[141] H. Wang, T. Zhao, J. Zeng, et al. Methods and clinical biomarker discovery for targeted proteomics using Olink technology. *PROTEOMICS – Clinical Applications*, 18(5): 2300233, 2024. doi: 10.1002/prca.202300233.

[142] B. Lehallier, D. Gate, N. Schaum, et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nature Medicine*, 25(12):1843–1850, Dec. 2019. doi: 10.1038/s41591-019-0673-2.

[143] D. E. Haslam, J. Li, S. T. Dillon, et al. Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAscan platforms. *PROTEOMICS*, 22(13-14):2100170, 2022. doi: 10.1002/pmic.202100170.

[144] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928): 198–207, Mar. 2003. doi: 10.1038/nature01511.

[145] T. Geiger, A. Wehner, C. Schaab, et al. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins *. *Molecular & Cellular Proteomics*, 11(3), Mar. 2012. doi: 10.1074/mcp.M111.014050.

[146] J. M. Bader, V. Albrecht, and M. Mann. MS-Based Proteomics of Body Fluids: The End of the Beginning. *Molecular & Cellular Proteomics*, 22(7), July 2023. doi: 10.1016/j.mcpro.2023.100577.

[147] M. Wojtkiewicz, L. Berg Luecke, M. I. Kelly, and R. L. Gundry. Facile Preparation of Peptides for Mass Spectrometry Analysis in Bottom-Up Proteomics Workflows. *Current Protocols*, 1(3):e85, 2021. doi: 10.1002/cpz1.85.

[148] T. J. Bechtel and E. Weerapana. From structure to redox: The diverse functional roles of disulfides and implications in disease. *PROTEOMICS*, 17(6):1600391, 2017. doi: 10.1002/pmic.201600391.

[149] E. J. Dupree, M. Jayathirtha, H. Yorkey, et al. A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes*, 8(3):14, Sept. 2020. doi: 10.3390/proteomes8030014.

[150] L. Moruz and L. Käll. Peptide retention time prediction. *Mass Spectrometry Reviews*, 36 (5):615–623, 2017. doi: 10.1002/mas.21488.

[151] G. C. McAlister, D. H. Phanstiel, J. Brumbaugh, et al. Higher-energy Collision-activated Dissociation Without a Dedicated Collision Cell*. *Molecular & Cellular Proteomics*, 10 (5):O111.009456, May 2011. doi: 10.1074/mcp.O111.009456.

[152] N. W. Bateman, S. P. Goulding, N. J. Shulman, et al. Maximizing Peptide Identification Events in Proteomic Workflows Using Data-Dependent Acquisition (DDA). *Molecular & Cellular Proteomics*, 13(1):329–338, Jan. 2014. doi: 10.1074/mcp.M112.026500.

[153] A. Doerr. DIA mass spectrometry. *Nature Methods*, 12(1):35–35, Jan. 2015. doi: 10.1038/nmeth.3234.

[154] K. Fröhlich, M. Fahrner, E. Brombacher, et al. Data-Independent Acquisition: A Milestone and Prospect in Clinical Mass Spectrometry–Based Proteomics. *Molecular & Cellular Proteomics*, 23(8):100800, Aug. 2024. doi: 10.1016/j.mcpro.2024.100800.

[155] R. Bruderer, O. M. Bernhardt, T. Gandhi, et al. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues *[S]. *Molecular & Cellular Proteomics*, 14 (5):1400–1410, May 2015. doi: 10.1074/mcp.M114.044305.

[156] V. Demichev, C. B. Messner, S. I. Vernardis, et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1):41–44, Jan. 2020. doi: 10.1038/s41592-019-0638-x.

[157] C. Bauer, R. Cramer, and J. Schuchhardt. Evaluation of Peak-Picking Algorithms for Protein Mass Spectrometry. In M. Hamacher, M. Eisenacher, and C. Stephan, editors, *Data Mining in Proteomics: From Standards to Applications*, pages 341–352. Humana Press, Totowa, NJ, 2011. ISBN 978-1-60761-987-1. doi: 10.1007/978-1-60761-987-1_22.

[158] N. Hulstaert, J. Shofstahl, T. Sachsenberg, et al. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *Journal of Proteome Research*, 19(1): 537–542, Jan. 2020. doi: 10.1021/acs.jproteome.9b00328.

[159] J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, Nov. 1994. doi: 10.1016/1044-0305(94)80016-2.

[160] R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, June 2004. doi: 10.1093/bioinformatics/bth092.

[161] C. Y. Park, A. A. Klammer, L. Käll, et al. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *Journal of Proteome Research*, 7(7):3022–3027, July 2008. doi: 10.1021/pr800127y.

[162] H. Barsnes and M. Vaudel. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *Journal of Proteome Research*, 17(7):2552–2555, July 2018. doi: 10.1021/acs.jproteome.8b00175.

[163] A. Michalski, J. Cox, and M. Mann. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC–MS/MS. *Journal of Proteome Research*, 10(4):1785–1793, Apr. 2011. doi: 10.1021/pr101060v.

[164] V. Dorfer, S. Maltsev, S. Winkler, and K. Mechtler. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *Journal of Proteome Research*, 17(8):2581–2589, Aug. 2018. doi: 10.1021/acs.jproteome.7b00836.

[165] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research*, 7(1):40–44, Jan. 2008. doi: 10.1021/pr700739d.

[166] R. E. Moore, M. K. Young, and T. D. Lee. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, Apr. 2002. doi: 10.1016/S1044-0305(02)00352-5.

[167] J. C. Wright and J. S. Choudhary. DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. *Journal of proteomics & bioinformatics*, 9(6):176–180, June 2016. doi: 10.4172/jpb.1000404.

[168] E. Debrie, M. Malfait, R. Gabriels, et al. Quality Control for the Target Decoy Approach for Peptide Identification. *Journal of Proteome Research*, 22(2):350–358, Feb. 2023. doi: 10.1021/acs.jproteome.2c00423.

[169] L. Käll, J. D. Canterbury, J. Weston, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, Nov. 2007. doi: 10.1038/nmeth1113.

[170] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7(1):29–34, Jan. 2008. doi: 10.1021/pr700600n.

[171] A. Lin, T. Short, W. S. Noble, and U. Keich. Improving Peptide-Level Mass Spectrometry Analysis via Double Competition. *Journal of Proteome Research*, 21(10):2412–2420, Oct. 2022. doi: 10.1021/acs.jproteome.2c00282.

[172] J. Griss, Y. Perez-Riverol, S. Lewis, et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature Methods*, 13(8): 651–656, Aug. 2016. doi: 10.1038/nmeth.3902.

[173] A. I. Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11):1114–1125, Nov. 2014. doi: 10.1038/nmeth.3144.

[174] K. Suhre, M. I. McCarthy, and J. M. Schwenk. Genetics meets proteomics: perspectives for large population-based studies. *Nature Reviews Genetics*, 22(1):19–37, Jan. 2021. doi: 10.1038/s41576-020-0268-2.

[175] F. Aguet, K. Alasoo, Y. I. Li, et al. Molecular quantitative trait loci. *Nature Reviews Methods Primers*, 3(1):4, Jan. 2023. doi: 10.1038/s43586-022-00188-6.

[176] S. Kaiser, L. Zhang, B. Mollenhauer, et al. A proteogenomic view of Parkinson's disease causality and heterogeneity. *npj Parkinson's Disease*, 9(1):24, Feb. 2023. doi: 10.1038/s41531-023-00461-9.

[177] N. Yazdanpanah, M. Yazdanpanah, Y. Wang, et al. Clinically Relevant Circulating Protein Biomarkers for Type 1 Diabetes: Evidence From a Two-Sample Mendelian Randomization Study. *Diabetes Care*, 45(1):169–177, Nov. 2021. doi: 10.2337/dc21-1049.

[178] F. Ghanbari, N. Yazdanpanah, M. Yazdanpanah, et al. Connecting Genomics and Proteomics to Identify Protein Biomarkers for Adult and Youth-Onset Type 2 Diabetes: A Two-Sample Mendelian Randomization Study. *Diabetes*, 71(6):1324–1337, Mar. 2022. doi: 10.2337/db21-1046.

[179] L. Niu, S. E. Stinson, L. A. Holm, et al. Plasma proteome variation and its genetic determinants in children and adolescents. *Nature Genetics*, pages 1–12, Feb. 2025. doi: 10.1038/s41588-025-02089-2.

[180] K. Suhre, Q. Chen, A. Halama, et al. A genome-wide association study of mass spectrometry proteomics using the Seer Proteograph platform, June 2024.

[181] G. Menschaert and D. Fenyö. Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrometry Reviews*, 36(5):584–599, 2017. doi: 10.1002/mas.21483.

[182] J. D. Jaffe, H. C. Berg, and G. M. Church. Proteogenomic mapping as a complementary method to perform genome annotation. *PROTEOMICS*, 4(1):59–77, 2004. doi: 10.1002/pmic.200300511.

[183] M. Fejzo, N. Rocha, I. Cimino, et al. GDF15 linked to maternal risk of nausea and vomiting during pregnancy. *Nature*, 625(7996):760–767, Jan. 2024. doi: 10.1038/s41586-023-06921-9.

[184] H. Zhang, T. Liu, Z. Zhang, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, 166(3):755–765, July 2016. doi: 10.1016/j.cell.2016.05.069.

[185] L. Reilly, S. Seddighi, A. B. Singleton, et al. Variant biomarker discovery using mass spectrometry-based proteogenomics. *Frontiers in Aging*, 4, Apr. 2023. doi: 10.3389/fragi.2023.1191993.

[186] A. E. Frazier, A. G. Compton, Y. Kishita, et al. Fatal Perinatal Mitochondrial Cardiac Failure Caused by Recurrent De Novo Duplications in the ATAD3 Locus. *Med*, 2(1):49–73.e10, Jan. 2021. doi: 10.1016/j.medj.2020.06.004.

[187] E. H. Bowler-Barnett, J. Fan, J. Luo, et al. UniProt and Mass Spectrometry-Based Proteomics—A 2-Way Working Relationship. *Molecular & Cellular Proteomics*, 22(8), Aug. 2023. doi: 10.1016/j.mcpro.2023.100591.

[188] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, Jan. 2025. doi: 10.1093/nar/gkae1010.

[189] N. A. O'Leary, M. W. Wright, J. R. Brister, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, Jan. 2016. doi: 10.1093/nar/gkv1189.

[190] F. Zickmann and B. Y. Renard. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics*, 31(12):i106–i115, June 2015. doi: 10.1093/bioinformatics/btv236.

[191] X. Cao and J. Xing. PrecisionProDB: improving the proteomics performance for precision medicine. *Bioinformatics*, 37(19):3361–3363, Oct. 2021. doi: 10.1093/bioinformatics/btab218.

[192] H. M. Umer, E. Audain, Y. Zhu, et al. Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides. *Bioinformatics*, 38(5):1470–1472, Mar. 2022. doi: 10.1093/bioinformatics/btab838.

[193] S. Faulkner, M. D. Dun, and H. Hondermarck. Proteogenomics: emergence and promise. *Cellular and Molecular Life Sciences*, 72(5):953–957, Mar. 2015. doi: 10.1007/s00018-015-1837-y.

[194] J. A. Alfaro, A. Sinha, T. Kislinger, and P. C. Boutros. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature Methods*, 11(11):1107–1113, Nov. 2014. doi: 10.1038/nmeth.3138.

[195] M. J. Ellis, M. Gillette, S. A. Carr, et al. Connecting Genomic Alterations to Cancer Biology with Proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery*, 3(10):1108–1112, Oct. 2013. doi: 10.1158/2159-8290.CD-13-0219.

[196] S. Varland, K. M. Brønstad, S. J. Skinner, and T. Arnesen. A nonsense variant in the N-terminal acetyltransferase NAA30 may be associated with global developmental delay and tracheal cleft. *American Journal of Medical Genetics Part A*, 191(9):2402–2410, 2023. doi: 10.1002/ajmg.a.63338.

[197] Y. Li, X. Wang, J.-H. Cho, et al. JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. *Journal of Proteome Research*, 15(7):2309–2320, July 2016. doi: 10.1021/acs.jproteome.6b00344.

[198] F. Martins Rodrigues, N. V. Terekhanova, K. J. Imbach, et al. Precision proteogenomics reveals pan-cancer impact of germline variants. *Cell*, 188(9):2312–2335.e26, May 2025. doi: 10.1016/j.cell.2025.03.026.

[199] D. Wang, R. Bouwmeester, P. Zheng, et al. Proteogenomics analysis of human tissues using pangenomes, May 2024.

[200] Y. Meng, Y. Lei, J. Gao, et al. Genome sequence assembly algorithms and misassembly identification methods. *Molecular Biology Reports*, 49(11):11133–11148, Nov. 2022. doi: 10.1007/s11033-022-07919-8.

[201] H. Chial. DNA Sequencing Technologies Key to the Human Genome Project. *Nature Education*, 1(1):219, Jan. 2008.

[202] P. Danecek, J. K. Bonfield, J. Liddle, et al. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, Feb. 2021. doi: 10.1093/gigascience/giab008.

[203] O. Delaneau, J. Marchini, and J.-F. Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, Feb. 2012. doi: 10.1038/nmeth.1785.

[204] M. Patterson, T. Marschall, N. Pisanti, et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology*, 22(6):498–509, June 2015. doi: 10.1089/cmb.2014.0157.

[205] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS*, 20(18):3551–3567, 1999. doi: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

[206] B. J. Diament and W. S. Noble. Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra. *Journal of Proteome Research*, 10(9):3871–3879, Sept. 2011. doi: 10.1021/pr101196n.

[207] L. Käll, J. D. Storey, and W. S. Noble. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*, 24(16):i42–i48, Aug. 2008. doi: 10.1093/bioinformatics/btn294.

[208] A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, 73(11):2092–2123, Oct. 2010. doi: 10.1016/j.jprot.2010.08.009.

[209] A. Declercq, R. Bouwmeester, A. Hirschler, et al. MS2Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *Molecular & Cellular Proteomics*, page 100266, July 2022. doi: 10.1016/j.mcpro.2022.100266.

[210] S. Degroeve and L. Martens. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, 29(24):3199–3203, Dec. 2013. doi: 10.1093/bioinformatics/btt544.

[211] R. Bouwmeester, R. Gabriels, N. Hulstaert, et al. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods*, 18(11):1363–1369, Nov. 2021. doi: 10.1038/s41592-021-01301-5.

[212] S. Gessulat, T. Schmidt, D. P. Zolg, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509–518, June 2019. doi: 10.1038/s41592-019-0426-7.

[213] F. Yu, G. C. Teo, A. T. Kong, et al. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nature Communications*, 14(1):4154, July 2023. doi: 10.1038/s41467-023-39869-5.

[214] S. Tyanova, T. Temu, and J. Cox. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, 11(12):2301–2319, Dec. 2016. doi: 10.1038/nprot.2016.136.

[215] M. Vaudel, J. M. Burkhart, R. P. Zahedi, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology*, 33(1):22–24, Jan. 2015. doi: 10.1038/nbt.3109.

[216] B. Wen and B. Zhang. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. *Nature Communications*, 14(1):2213, Apr. 2023. doi: 10.1038/s41467-023-37462-4.

[217] K. V. Ruggles, K. Krug, X. Wang, et al. Methods, Tools and Current Perspectives in Proteogenomics. *Molecular & Cellular Proteomics*, 16(6):959–981, June 2017. doi: 10.1074/mcp.MR117.000024.

[218] X. Wang and B. Zhang. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics*, 29(24):3235–3237, Dec. 2013. doi: 10.1093/bioinformatics/btt543.

[219] J. Leipzig. A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, 18(3):530–536, May 2017. doi: 10.1093/bib/bbw020.

[220] M. Ziemann, P. Poulain, and A. Bora. The five pillars of computational reproducibility: bioinformatics and beyond. *Briefings in Bioinformatics*, 24(6):bbad375, Nov. 2023. doi: 10.1093/bib/bbad375.

[221] J. M. Perkel. Workflow systems turn raw data into scientific knowledge. *Nature*, 573 (7772):149–150, Sept. 2019. doi: 10.1038/d41586-019-02619-z.

[222] F. da Veiga Leprevost, B. A. Grüning, S. Alves Aflitos, et al. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, 33(16):2580–2582, Aug. 2017. doi: 10.1093/bioinformatics/btx192.

[223] B. Grüning, R. Dale, A. Sjödin, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, July 2018. doi: 10.1038/s41592-018-0046-7.

[224] P. A. Ewels, A. Peltzer, S. Fillinger, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3):276–278, Mar. 2020. doi: 10.1038/s41587-020-0439-x.

[225] E. Raymond. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3): 23–49, Sept. 1999. doi: 10.1007/s12130-999-1026-0.

[226] G. von Krogh and S. Spaeth. The open source software phenomenon: Characteristics that promote research. *The Journal of Strategic Information Systems*, 16(3):236–253, Sept. 2007. doi: 10.1016/j.jsis.2007.06.001.

[227] M. Barker, N. P. Chue Hong, D. S. Katz, et al. Introducing the FAIR Principles for research software. *Scientific Data*, 9(1):622, Oct. 2022. doi: 10.1038/s41597-022-01710-x.

[228] M. Shome, T. M. MacKenzie, S. R. Subbareddy, and M. P. Snyder. The Importance, Challenges, and Possible Solutions for Sharing Proteomics Data While Safeguarding Individuals' Privacy. *Molecular & Cellular Proteomics : MCP*, 23(3):100731, Feb. 2024. doi: 10.1016/j.mcpro.2024.100731.

[229] E. W. Deutsch, N. Bandeira, Y. Perez-Riverol, et al. The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Research*, 51(D1):D1539–D1548, Jan. 2023. doi: 10.1093/nar/gkac1040.

[230] M. Frejno, M. T. Berger, J. Tüshaus, et al. Unifying the analysis of bottom-up proteomics data with CHIMERYS. *Nature Methods*, 22(5):1017–1027, May 2025. doi: 10.1038/s41592-025-02663-w.

[231] M. Wang, J. Wang, J. Carver, et al. Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems*, 7(4):412–421.e5, Oct. 2018. doi: 10.1016/j.cels.2018.08.004.

[232] E. W. Deutsch, Y. Perez-Riverol, J. Carver, et al. Universal Spectrum Identifier for mass spectra. *Nature Methods*, 18(7):768–770, July 2021. doi: 10.1038/s41592-021-01184-6.

[233] E. R. Tufte and P. R. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[234] L. McDonald. Florence Nightingale, statistics and the Crimean War. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 177(3):569–586, June 2014. doi: 10.1111/rssa.12026.

[235] F. P. Brooks. The computer scientist as toolsmith II. *Commun. ACM*, 39(3):61–68, Mar. 1996. doi: 10.1145/227234.227243.

[236] S. I. O'Donoghue. Grand Challenges in Bioinformatics Data Visualization. *Frontiers in Bioinformatics*, 1, June 2021. doi: 10.3389/fbinf.2021.669186.

[237] M. Rittenbruch, K. Vella, M. Brereton, et al. Collaborative Sense-Making in Genomic Research: The Role of Visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4477–4489, Dec. 2022. doi: 10.1109/TVCG.2021.3090746.

[238] A. Mund, F. Coscia, A. Kriston, et al. Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nature Biotechnology*, pages 1–10, May 2022. doi: 10.1038/s41587-022-01302-5.

[239] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In B. B. Bederson and B. Shneiderman, editors, *The Craft of Information Visualization*, Interactive Technologies, pages 364–371. Morgan Kaufmann, San Francisco, Jan. 2003. ISBN 978-1-55860-915-0. doi: 10.1016/B978-155860915-0/50046-9.

[240] G. Kindlmann and C. Scheidegger. An Algebraic Process for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2181–2190, Dec. 2014. doi: 10.1109/TVCG.2014.2346325.

[241] M. Brehmer and T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, Dec. 2013. doi: 10.1109/TVCG.2013.124.

[242] S. Nusrat, T. Harbig, and N. Gehlenborg. Tasks, Techniques, and Tools for Genomic Data Visualization. *Computer Graphics Forum*, 38(3):781–805, 2019. doi: 10.1111/cgf.13727.

[243] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, et al. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, Jan. 2011. doi: 10.1038/nbt.1754.

[244] G. Perez, G. Barber, A. Benet-Pages, et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Research*, 53(D1):D1243–D1249, Jan. 2025. doi: 10.1093/nar/gkae974.

[245] S. I. O'Donoghue, D. S. Goodsell, A. S. Frangakis, et al. Visualization of macromolecular structures. *Nature Methods*, 7(3):S42–S55, Mar. 2010. doi: 10.1038/nmeth.1427.

[246] K. Furmanová, A. Jurčík, B. Kozlíková, et al. Multiscale Visual Drilldown for the Analysis of Large Ensembles of Multi-Body Protein Complexes. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):843–852, Jan. 2020. doi: 10.1109/TVCG.2019.2934333.

[247] M. Shi, J. Gao, and M. Q. Zhang. Web3DMol: interactive protein structure visualization based on WebGL. *Nucleic Acids Research*, 45(W1):W523–W527, July 2017. doi: 10.1093/nar/gkx383.

[248] N. T. Doncheva, Y. Assenov, F. S. Domingues, and M. Albrecht. Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, 7(4):670–685, Apr. 2012. doi: 10.1038/nprot.2012.004.

[249] D. Szklarczyk, K. Nastou, M. Koutrouli, et al. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, Jan. 2025. doi: 10.1093/nar/gkae1113.

[250] K. Li, M. Vaudel, B. Zhang, et al. PDV: an integrative proteomics data viewer. *Bioinformatics*, 35(7):1249–1251, Apr. 2019. doi: 10.1093/bioinformatics/bty770.

[251] L. K. Pino, B. C. Searle, J. G. Bollinger, et al. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrometry Reviews*, 39(3):229–244, 2020. doi: 10.1002/mas.21540.

[252] H. Lam, E. Bertini, P. Isenberg, et al. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, Sept. 2012. doi: 10.1109/TVCG.2011.279.

[253] T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, Nov. 2009. doi: 10.1109/TVCG.2009.111.

[254] E. Wall, M. Agnihotri, L. Matzen, et al. A Heuristic Approach to Value-Driven Evaluation of Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):491–500, Jan. 2019. doi: 10.1109/TVCG.2018.2865146.

[255] A. Bangor, P. T. Kortum, and J. T. Miller. An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, 24(6):574–594, July 2008. doi: 10.1080/10447310802205776.

[256] S. Elling, L. Lentz, and M. de Jong. Combining Concurrent Think-Aloud Protocols and Eye-Tracking Observations: An Analysis of Verbalizations and Silences. *IEEE Transactions on Professional Communication*, 55(3):206–220, Sept. 2012. doi: 10.1109/TPC.2012.2206190.

[257] A. Gegenfurtner and M. Seppänen. Transfer of expertise: An eye tracking and think aloud study using dynamic medical visualizations. *Computers & Education*, 63:393–403, Apr. 2013. doi: 10.1016/j.compedu.2012.12.021.

[258] K. Allendoerfer, S. Aluker, G. Panjwani, et al. Adapting the cognitive walkthrough method to assess the usability of a knowledge domain visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 195–202, Oct. 2005. doi: 10.1109/INFVIS.2005.1532147.

[259] C. North, P. Saraiya, and K. Duca. A comparison of benchmark task and insight evaluation methods for information visualization. *Information Visualization*, 10(3):162–181, July 2011. doi: 10.1177/1473871611415989.

[260] B. Saket, A. Endert, and C. Demiralp. Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2505–2512, July 2019. doi: 10.1109/TVCG.2018.2829750.

[261] E. Lowy-Gallego, S. Fairley, X. Zheng-Bradley, et al. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Research*, 4:50, Dec. 2019. doi: 10.12688/wellcomeopenres.15126.2.

[262] P. E. Geyer, N. J. Wewer Albrechtsen, S. Tyanova, et al. Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Molecular Systems Biology*, 12 (12):901, Dec. 2016. doi: 10.15252/msb.20167357.

[263] P. Stawiński and R. Płoski. Genebe.net: Implementation and validation of an automatic ACMG variant pathogenicity criteria assignment. *Clinical Genetics*, 106(2):119–126, 2024. doi: 10.1111/cge.14516.

[264] P.-L. Luu, P.-T. Ong, T.-P. Dinh, and S. J. Clark. Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. *NAR Genomics and Bioinformatics*, 2(3):lqaa054, Sept. 2020. doi: 10.1093/nargab/lqaa054.

[265] K.-J. Park, Y. A. Yoon, and J.-H. Park. Evaluation of Liftover Tools for the Conversion of Genome Reference Consortium Human Build 37 to Build 38 Using ClinVar Variants. *Genes*, 14(10):1875, Oct. 2023. doi: 10.3390/genes14101875.

[266] H. Li, M. Dawood, M. M. Khayat, et al. Exome variant discrepancies due to reference-genome differences. *The American Journal of Human Genetics*, 108(7):1239–1250, July 2021. doi: 10.1016/j.ajhg.2021.05.011.

[267] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, et al. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods*, 14(5):513–520, May 2017. doi: 10.1038/nmeth.4256.

[268] G. C. Teo, D. A. Polasky, F. Yu, and A. I. Nesvizhskii. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *Journal of Proteome Research*, 20(1): 498–505, Jan. 2021. doi: 10.1021/acs.jproteome.0c00544.

[269] K. L. Yang, F. Yu, G. C. Teo, et al. MSBooster: improving peptide identification rates using deep learning-based features. *Nature Communications*, 14(1):4539, July 2023. doi: 10.1038/s41467-023-40129-9.

[270] A. Mushtaq. Cool stuff Ensembl VEP can do: supporting alternative human assemblies – Ensembl Blog, Aug. 2024. URL `https://www.ensembl.info/2024/08/09/cool-stuff-ensembl-vep-can-do-supporting-alternative-human-assemblies/`.

[271] F. J. Martin, M. R. Amode, A. Aneja, et al. Ensembl 2023. *Nucleic Acids Research*, 51 (D1):D933–D941, Jan. 2023. doi: 10.1093/nar/gkac958.

[272] C. Alkan, L. Carbone, M. Y. Dennis, et al. Implications of the first complete human genome assembly. *Genome Research*, 32(4):595–598, Apr. 2022. doi: 10.1101/gr.276723.122.

[273] V. Marx. Method of the year: long-read sequencing. *Nature Methods*, 20(1):6–11, Jan. 2023. doi: 10.1038/s41592-022-01730-w.

[274] J. Sidney, B. Peters, N. Frahm, et al. HLA class I supertypes: a revised and updated classification. *BMC Immunology*, 9(1):1, Jan. 2008. doi: 10.1186/1471-2172-9-1.

[275] C. A. Dendrou, J. Petersen, J. Rossjohn, and L. Fugger. HLA variation and disease. *Nature Reviews Immunology*, 18(5):325–339, May 2018. doi: 10.1038/nri.2017.143.

[276] P. Parham and T. Ohta. Population Biology of Antigen Presentation by MHC Class I Molecules. *Science*, 272(5258):67–74, Apr. 1996. doi: 10.1126/science.272.5258.67.

[277] L. Loh, P. M. Saunders, C. Faoro, et al. An archaic HLA class I receptor allele diversifies natural killer cell-driven immunity in First Nations peoples of Oceania. *Cell*, 187(24): 7008–7024.e19, Nov. 2024. doi: 10.1016/j.cell.2024.10.005.

[278] G. Butler-Laporte, J. Farjoun, T. Nakanishi, et al. HLA allele-calling using multi-ancestry whole-exome sequencing from the UK Biobank identifies 129 novel associations in 11 autoimmune diseases. *Communications Biology*, 6(1):1–17, Nov. 2023. doi: 10.1038/s42003-023-05496-5.

[279] R. Peckner, S. A. Myers, A. S. V. Jacome, et al. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nature Methods*, 15(5):371–378, May 2018. doi: 10.1038/nmeth.4643.

[280] E. K. Keenan, D. K. Zachman, and M. D. Hirschey. Discovering the landscape of protein modifications. *Molecular Cell*, 81(9):1868–1878, May 2021. doi: 10.1016/j.molcel.2021.03.015.

[281] A. W. Purcell, S. H. Ramarathinam, and N. Ternette. Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. *Nature Protocols*, 14(6):1687–1707, June 2019. doi: 10.1038/s41596-019-0133-y.

[282] J. Fiala, D. Schuster, S. Ollivier, et al. Protein-Centric Analysis of Personalized Antibody Repertoires Using LC-MS-Based Fab-Profiling on a timsTOF. *Journal of the American Society for Mass Spectrometry*, 35(6):1292–1300, June 2024. doi: 10.1021/jasms.4c00076.

[283] J. A. Vizcaíno, P. Kubiniok, K. A. Kovalchik, et al. The Human Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases. *Molecular & cellular proteomics: MCP*, 19(1):31–49, Jan. 2020. doi: 10.1074/mcp.R119.001743.

[284] A. I. Nesvizhskii and R. Aebersold. Interpretation of Shotgun Proteomic Data. *Molecular & Cellular Proteomics*, 4(10):1419–1440, Oct. 2005. doi: 10.1074/mcp.R500012-MCP200.

[285] L. Pereira, L. Mutesa, P. Tindana, and M. Ramsay. African genetic diversity and adaptation inform a precision medicine agenda. *Nature Reviews Genetics*, 22(5):284–306, May 2021. doi: 10.1038/s41576-020-00306-8.

[286] H. Ræder, S. Johansson, P. I. Holm, et al. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nature Genetics*, 38(1):54–62, Jan. 2006. doi: 10.1038/ng1708.

[287] A. Gravdal, X. Xiao, M. Cnop, et al. The position of single-base deletions in the VNTR sequence of the carboxyl ester lipase (CEL) gene determines proteotoxicity. *Journal of Biological Chemistry*, 296, Jan. 2021. doi: 10.1016/j.jbc.2021.100661.

[288] J. J. M. Landry, P. T. Pyl, T. Rausch, et al. The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3 Genes|Genomes|Genetics*, 3(8):1213–1224, Aug. 2013. doi: 10.1534/g3.113.005777.

[289] K. Boonen, K. Hens, G. Menschaert, et al. Beyond Genes: Re-Identifiability of Proteomic Data and Its Implications for Personalized Medicine. *Genes*, 10(9):682, Sept. 2019. doi: 10.3390/genes10090682.

[290] G. J. Parker, T. Leppert, D. S. Anex, et al. Demonstration of Protein-Based Human Identification Using the Hair Shaft Proteome. *PLOS ONE*, 11(9):e0160653, Sept. 2016. doi: 10.1371/journal.pone.0160653.

[291] GWAS Catalog. URL `https://www.ebi.ac.uk/gwas/docs/file-downloads`.

[292] P. R. Njølstad, O. A. Andreassen, S. Brunak, et al. Roadmap for a precision-medicine initiative in the Nordic region. *Nature Genetics*, 51(6):924–930, June 2019. doi: 10.1038/s41588-019-0391-1.

8 Scientific results

8.1 Finding Haplotypic Signatures in Proteins

Vašíček, J., Skiadopoulou, D., Kuznetsova, K. G., Wen, B., Johansson, S., Njølstad, P. R., Bruckner, S., Käll, L., Vaudel, M.

GigaScience, **12**, giad093 (2023)

The reprint of Paper 1 is permitted under the CC-BY 4.0 license.

GigaScience, 2023, **12**, 1–13
DOI: 10.1093/gigascience/giad093
Research

Finding haplotypic signatures in proteins

Jakub Vašíček [1,2,†], Dafni Skiadopoulou [1,2,†], Ksenia G. Kuznetsova [1,2], Bo Wen [3], Stefan Johansson [1,4], Pål R. Njølstad [1,5], Stefan Bruckner [6,‡], Lukas Käll [7,‡] and Marc Vaudel [1,2,8,*,‡]

[1]Mohn Center for Diabetes Precision Medicine, Department of Clinical Science, University of Bergen, Bergen 5021, Norway
[2]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen 5008, Norway
[3]Department of Genome Sciences, University of Washington, Seattle, WA 98195, United States
[4]Department of Medical Genetics, Haukeland University Hospital, Bergen 5021, Norway
[5]Children and Youth Clinic, Haukeland University Hospital, Bergen 5021, Norway
[6]Chair of Visual Analytics, Institute for Visual and Analytic Computing, University of Rostock, Rostock 18051, Germany
[7]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH–Royal Institute of Technology, Solna 17121, Sweden
[8]Department of Genetics and Bioinformatics, Health Data and Digitalization, Norwegian Institute of Public Health, Oslo 0473, Norway
*Correspondence address. Marc Vaudel, Universitetet i Bergen, Klinisk institutt 2, Postboks 7804, NO-5020 BERGEN, NORWAY. E-mail: Marc.Vaudel@uib.no
†These authors contributed to the work equally.
‡These authors jointly supervised the work.

Abstract

Background: The nonrandom distribution of alleles of common genomic variants produces haplotypes, which are fundamental in medical and population genetic studies. Consequently, protein-coding genes with different co-occurring sets of alleles can encode different amino acid sequences: protein haplotypes. These protein haplotypes are present in biological samples and detectable by mass spectrometry, but they are not accounted for in proteomic searches. Consequently, the impact of haplotypic variation on the results of proteomic searches and the discoverability of peptides specific to haplotypes remain unknown.

Findings: Here, we study how common genetic haplotypes influence the proteomic search space and investigate the possibility to match peptides containing multiple amino acid substitutions to a publicly available data set of mass spectra. We found that for 12.42% of the discoverable amino acid substitutions encoded by common haplotypes, 2 or more substitutions may co-occur in the same peptide after tryptic digestion of the protein haplotypes. We identified 352 spectra that matched to such multivariant peptides, and out of the 4,582 amino acid substitutions identified, 6.37% were covered by multivariant peptides. However, the evaluation of the reliability of these matches remains challenging, suggesting that refined error rate estimation procedures are needed for such complex proteomic searches.

Conclusions: As these procedures become available and the ability to analyze protein haplotypes increases, we anticipate that proteomics will provide new information on the consequences of common variation, across tissues and time.

Keywords: proteogenomics, haplotype, protein, bioinformatics, post-translational modification

Background

Linkage disequilibrium (LD) describes the nonrandom correlation between alleles at different positions in the genome in a population. LD arises when alleles at nearby sites co-occur on the same haplotype more often than expected by chance. When haplotypes are located in protein-coding portions of the genome and include nonsynonymous changes, they can alter protein sequences, forming so-called protein haplotypes, as defined by Spooner et al. [1]. Based on the co-occurrence of alleles in the 1000 Genomes Project [2] and their *in silico* translation, Spooner et al. [1] created a list of possible protein haplotype sequences. Notably, they stress that for 1 in 7 genes, the most frequent protein haplotype differs from the reference sequence in Ensembl [3]. In precision medicine, probing the proteotype—the actual state of the proteome—adds valuable information concerning the relationship between the genotype and the phenotype [4]. Therefore, it is important that genetic information, including LD, is taken into account in proteomics searches.

Proteins in biological samples can be identified by liquid chromatography coupled to mass spectrometry (LC-MS), usually after digestion into peptides [5]. Then, the measured spectra are matched to a database of expected protein sequences using a search engine [6]. The identified peptides are used to infer the presence of proteins [7] along with potential posttranslational modifications (PTMs) [8]. When the peptides cover the relevant parts of the protein sequences, it is also possible to discover the product of alternative splicing or genetic variation [9]. In precision medicine, proteomic searches need to be adapted to individual patient profiles by extending the search space to include noncanonical sequences [10].

This challenge is addressed by proteogenomics—the scientific field integrating genomics and proteomics into a joint approach [9, 11]. Recent work, mainly in the domain of cancer research, has shown that accounting for genetic variation in proteomic analyses provides the means to discover noncanonical proteins. Umer et al. [12] have developed a tool to generate databases of variant proteins derived from single-nucleotide polymorphisms (SNPs), insertions and deletions, and the 3-frame translation of pseudogenes and noncanonical transcripts, appended with a database of canonical proteins [12]. Levitsky et al. [13] use measures of

Received: April 13, 2023. **Revised:** September 24, 2023. **Accepted:** October 8, 2023
© The Author(s) 2023. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

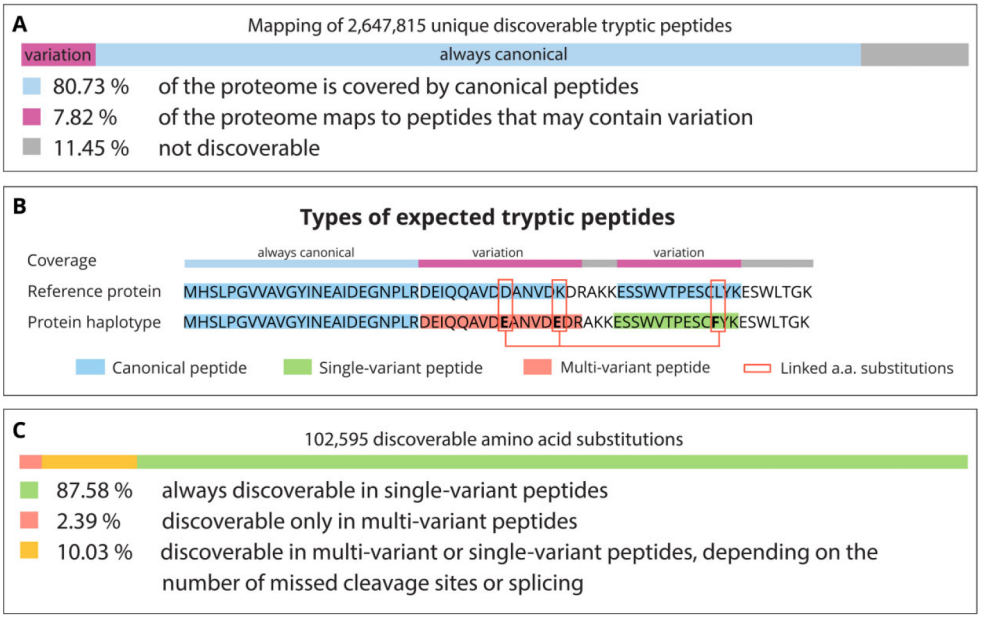


Figure 1: (A) Proteome coverage expressed in terms of the percentage of amino acids; that is, if 7 out of 100 residues belong to at least 1 discoverable peptide containing the product of a substitution, we say that 7% of the proteome maps to peptides containing variation. See main text for details and Materials and Methods for the handling of shared peptides. (B) Example of a reference sequence aligned to another haplotype. The classes of peptides following the cleavage pattern of trypsin are highlighted by a colored background. Three amino acid substitutions encoded by this haplotype are marked by red rectangles. The "coverage" layer indicates the alignment applied to obtain numbers shown in section A. (C) Distribution of variation in discoverable peptides. Amino acid variants are stratified based on the category of peptide in which the substitution caused by the respective variant can be identified.

proteome coverage, including variant peptides, to verify the presence of single amino acid variants. Choong et al. [14] proposed an algorithm to generate the optimal number of protein sequences containing combinations of amino acid substitutions possibly occurring in the same tryptic peptide. In their approach, the database includes not only the combinations of alleles encoded by haplotypes but all combinations possible per peptide. Lobas et al. [15, 16] showed that peptides containing variation were 2.5 to 3 times less likely to be identified than canonical peptides. Wang et al. [17] have analyzed data for 29 paired healthy human tissues from the Human Proteome Atlas project to detect amino acid variants at the protein level. However, the majority of amino acid variants predicted from exome sequencing could not be detected [17], suggesting that proteogenomics remains highly challenging and methods for discovering noncanonical proteins need further development.

Here, we used the protein haplotypes generated by Spooner et al. [1] to evaluate the ability of mass spectrometry–based proteomics to identify peptides encoded by combinations of variants in LD. We show that in some protein haplotypes, multiple amino acid substitutions affect the same peptide after digestion. Those protein haplotypes can only be identified if the combinations of amino acid variants are included in the search space, and several of these protein haplotypes are predicted to be more common than the reference sequence. Then, we mined the publicly available data from Wang et al. [17] for peptides including a combination of amino acid variants, demonstrating how such peptides can be identified according to the standards of the field but also how the quality control of the results remains challenging.

Results

The consequence of haplotypes on the proteomics search space

We digested *in silico* the protein sequences translated from haplotypes obtained from Spooner et al. [1] using the canonical cleavage pattern of trypsin, allowing for up to 2 missed cleavages. Note that indels were not considered, and we focused only on common variants with a minor allele frequency >1% in any population of the 1000 Genomes Project [2]; see Materials and Methods for details. After excluding contaminants, this yielded 2,647,815 unique tryptic peptide sequences of length between 8 and 40 amino acids (Fig. 1A). The coverage of protein sequences from Ensembl can be partitioned as follows: 80.73% can only be covered by canonical peptides, 7.82% map to peptides that may contain 1 or multiple amino acid substitutions, and the remaining yields sequences that are either too short or too long to be identified. Most peptides discoverable in proteomic studies therefore map to canonical sequences, making it challenging for nontargeted approaches to assess the allelic status of a common genetic variant using proteomics, in agreement with [15, 16].

We classify the obtained peptide sequences in 3 types (Fig. 1B): (i) canonical, no haplotype is known to yield an amino acid substitution in the sequence of this peptide; (ii) single-variant, a haplotype encodes an amino acid substitution in the sequence of this peptide; and (iii) multivariant, a haplotype encodes a set of 2 or more amino acid substitutions in the sequence of this peptide. In total, common haplotypes encode 102,595 amino acid substitutions, with 87.58% of them found only in single-variant

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

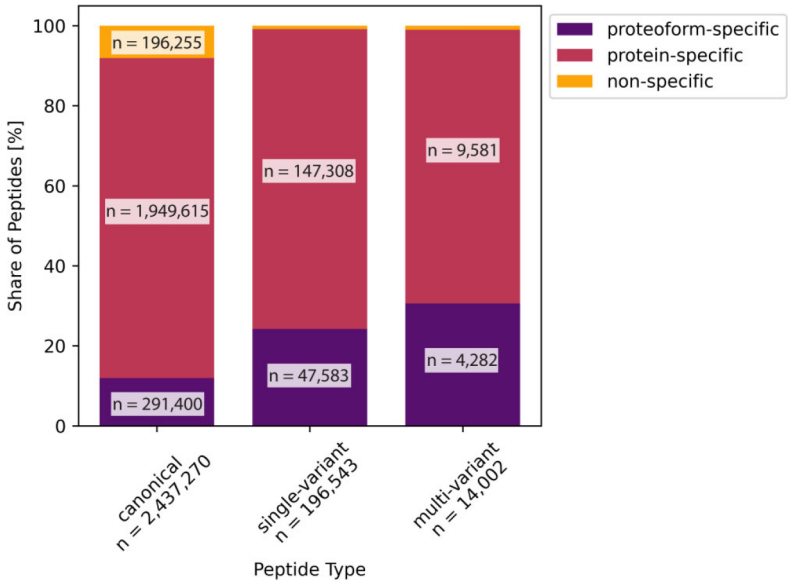


Figure 2: Classification of peptides based on their ability to distinguish between protein sequences (bar color) and to identify amino acid substitutions (position on x-axis). The height of the bars represents the distribution of categories (nonspecific, protein-specific, proteoform-specific) among the peptide types (canonical, single-variant, multivariant).

peptides, 2.39% in multivariant peptides, and 10.03% in either single- or multivariant peptides depending on the number of missed cleavages (Fig. 1C). Note that substitutions in different isoforms of the same protein are reported separately by Spooner et al. [1], creating multiple consequences for the same genetic variant. The total number of amino acid substitutions is consequently higher than the number of genetic variants. Interestingly, based on the frequencies among all participants in the 1000 Genomes project, 22.3% and 32.4% of the amino acid substitutions discoverable in single-variant and multivariant peptides, respectively, occur in protein haplotypes that are predicted to be more frequent than the Ensembl reference sequence. If these alleles are not accounted for, proteomics analyses will, therefore, not be able to identify these parts of the genome for the majority of individuals.

Peptides can be classified based on their ability to distinguish between protein sequences. We propose the following categories: (i) nonspecific peptides map to the products of different genes; (ii) protein-specific peptides map to multiple sequences, which are all products of the same gene; and (iii) proteoform-specific peptides map uniquely to a single form of a protein (i.e., single splice variant and haplotype), referred to as proteoform [18]. In this classification, based on the identification of a proteoform-specific peptide, one can uniquely identify products of a given gene. A protein-specific peptide allows for discriminating certain groups of proteoforms but does not yield a single candidate sequence (e.g., it determines which amino acid substitution is present but maps to multiple splicing variants). Nonspecific peptide sequences map to multiple genes, where the sequence of 1 gene matches the sequence of another, making it challenging to infer which protein is covered. We found 198,046 distinct nonspecific peptide sequences, covering up to 17.53% of the proteome. The prevalence of canonical, single-variant, and multivariant peptides among the above introduced types is displayed in Fig. 2, with exact numbers provided in Table 1. As expected intuitively, peptides containing the product of 1 or multiple variants present a higher ability to distinguish between protein products of different genes and between proteoforms of the same gene.

Table 1: Classification of peptides types and the number of *in silico* digested peptides in each of the categories

Peptide type (variation)	Peptide type (specificity)	Number of possible peptides
Canonical	Proteoform-specific	291,400
Canonical	Protein-specific	1,949,615
Canonical	Nonspecific	196,255
Single-variant	Proteoform-specific	47,583
Single-variant	Protein-specific	147,308
Single-variant	Nonspecific	1,652
Multivariant	Proteoform-specific	4,282
Multivariant	Protein-specific	9,581
Multivariant	Nonspecific	139

Matching multivariant peptides to mass spectra

To investigate the prevalence of spectra matching multivariant peptides encoded by common haplotypes and the quality of the obtained matches, we searched the deep proteomics data of healthy tonsil tissue made available by Wang et al. [17] against the sequences of common protein haplotypes using X!Tandem [19] as a search engine without refinement procedure and Percolator [20] with standard features for the evaluation of the confidence in all peptide-to-spectrum matches (PSMs). The resulting PSMs were thresholded at a 1% PSM-level false discovery rate (FDR). Note that because our study focuses on evaluating the quality of the spectrum matches, a PSM-level FDR was therefore preferred to peptide-level statistics. After thresholding, 1,318,152 target PSMs

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

remained (13,467 decoy PSMs would have passed the threshold), representing 176,193 unique peptide sequences (8,047 decoy peptide sequences would have passed the threshold), covering the alternative amino acid of 4,582 substitutions. The distribution of alternative alleles among single- and multivariant peptides (Fig. 3A) mirrored the values obtained from the *in silico* digestion of protein haplotypes (Fig. 1C). On average, the products of 2,249.67 substitutions were found per sample (2,360, 2,165, and 2,224 in samples 1, 2, and 3, respectively). The matched peptide sequences cover 21.56% of the proteome predicted to map exclusively to canonical peptides and 16.89% of the proteome possibly mapping to peptides with substitutions (Fig. 3B). Note, however, that 19,678 peptide sequences (identified in 231,181 PSMs) map to the products of multiple genes that cannot be distinguished, hence affecting the coverage estimates.

Out of the 1,318,152 spectra matched to peptides, 0.57% were matched to single-variant peptides and 0.03% were matched to multivariant peptides. The share of spectra matched to variant peptides is thus lower than the expected error rate, and currently, no method allows the evaluation of error rates in these subgroups of matches specifically. We thus investigated whether these classes of peptides showed signs of an overrepresentation of false-positive matches. No substantial difference was noticeable in the density of the posterior error probabilities (PEPs) and *q*-values for all 3 classes of PSMs (Fig. 3C, D), indicating that a more stringent FDR threshold would not alter the prevalence of variant peptides. We also compared the observed peptide retention time and fragmentation with predictions from DeepLC [21] and MS2PIP [22], respectively. Overall, the density of the distance to prediction in both retention time and fragmentation was very similar for all 3 classes of peptides (Fig. 3E, F), displaying no obvious shift in the distribution, which would have been indicative of a strong overrepresentation of false positives. Yet the distributions of variant and multivariant peptides showed stronger tails toward high distance to prediction compared to nonvariant peptides, indicative of the presence of false-positive matches. In comparison, the distance to prediction for decoys showed high retention time difference and low spectrum similarity.

Quality metrics on all matches are available as supplementary material. Three examples, sampled from the multivariant matches passing the FDR threshold at low, medium, and high PEP, representing high, medium, and low confidence, respectively, are displayed in Fig. 4 along with the predicted spectra. As expected, the share of peaks matching predicted fragment ions decreases as the PEP increases: (A) the highly confident match presents an excellent coverage of the spectrum with fragment ion masses, with an extensive mapping of the peptide y-ion series; the retention time distance to prediction of 320.8 seconds represents only a fraction of the gradient (approximately 2.5 hours); and the spectrum similarity to prediction, 0.79, shows good but not perfect agreement, which is in the lower range of the distribution of similarity scores for the canonical matches. (B) The medium confidence match presents a good coverage of the spectrum lacking prediction for many peaks, and the agreement scores with retention time and fragmentation predictions are excellent. (C) The low confidence match presents a poor coverage of the spectrum with poor agreement with retention time and fragmentation predictions. In addition to passing commonly accepted statistical thresholds, the matches in Fig. 4A and B would pass expert quality control. On the other hand, the match in 4C is most probably a false positive. Together, while these 3 sampled PSMs represent only a limited set of examples, they are very representative of the difficulty to confidently assess the presence of individual peptides from large proteomic experiments. This task is, however, important given that chimeric spectrum matches [23–26] and partial matches are known to be difficult to account for in error rate estimation [27, 28].

As highlighted by Spooner et al. [1], depending on the population studied, specific haplotypes often have higher frequencies than the canonical haplotype by Ensembl. For example, there are 5 haplotypes of the IQ motif containing the GTPase activating protein 2 (IQGAP2, ENSP00000274364) gene that have higher predicted frequencies than the canonical haplotype in the European population (with combined frequency of 84.9% according to Spooner et al. [1]). These haplotypes encode a tryptic peptide containing 2 amino acid substitutions when compared to the canonical sequence in Ensembl: VLWLDEIQQAVD**E**ANVD**E**DR (amino acid substitutions in bold). At position 527 of the protein sequence, aspartic acid is changed to glutamic acid (527D>E, rs2431352), and at position 532, lysine is changed to glutamic acid (532K>E, rs2909888), preventing cleavage by trypsin. In our results, 2 peptides overlapped with this sequence, 1 featuring a missed cleavage, supported by 13 and 10 spectra, respectively. Fig. 4D and E display 2 examples of highly confident matches, and Supplementary Table S1 lists PEP, *q*-value, and agreement with predictors for all PSMs. Altogether, the PEPs and agreement with predictors for these PSMs support the identification of this sequence and thus the presence of these haplotypes in the data reported by Wang et al. [17], consistent with the frequencies of these haplotypes in the European population. The sequence encoded by these haplotypes cannot be detected using canonical databases.

For diploid chromosomes, in the absence of deletion or copy number alteration, each individual carries 2 versions of a given gene—1 paternally and 1 maternally inherited—which can represent different haplotypes. We thus expect to find evidence for heterozygosity in some of the identified variants. We have come across such cases in 26, 21, and 19 genes in samples 1, 2, and 3, respectively. For example, the protein CR1 (complement component [3b/4b] receptor 1, ENSP00000356016) is commonly affected by multiple SNPs. First, at the position 2060, threonine is commonly changed to serine (2060T>S, rs4844609). Haplotypes including serine at position 2060 are expected in the European population with the combined frequency of 98%. Second, at the position 2065, isoleucine can be changed to valine (2065I>V, rs6691117); valine is expected in the European population with a frequency of 22.57%. However, a valine at position 2065 is only expected when preceded by a serine at position 2060. In one of the samples, we identified spectra matching confidently to both a multivariant peptide encoded by both alternative alleles (SFF**S**LTEI**V**R, substitutions in bold) and to a single-variant peptide encoded by the alternative allele of the first SNP (SFF**S**LTEIIR, substitution in bold). Mirrored spectra and associated quality metrics are shown in Fig. 5. In this case, including haplotypes in the protein database enables the identification of not only the alternative but also the reference allele of a variant.

While including the sequences from different haplotypes offers the ability to detect new protein haplotypes, it also increases the likelihood of similar peptides to map to different proteins. For example, the protein POTE ankyrin domain family member I (POTEI, ENSP00000392718) contains in its most frequent haplotype 8 amino acid substitutions, 2 of which fall into the same tryptic peptide. In the actin-like domain of POTEI at position 918, tyrosine changes to phenylalanine (918Y>F, rs147268452), and at position 929, methionine changes into threonine (929M>T, rs201878083), thus encoding the peptide LCFVALDFEQEMATAASSSSLEK

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

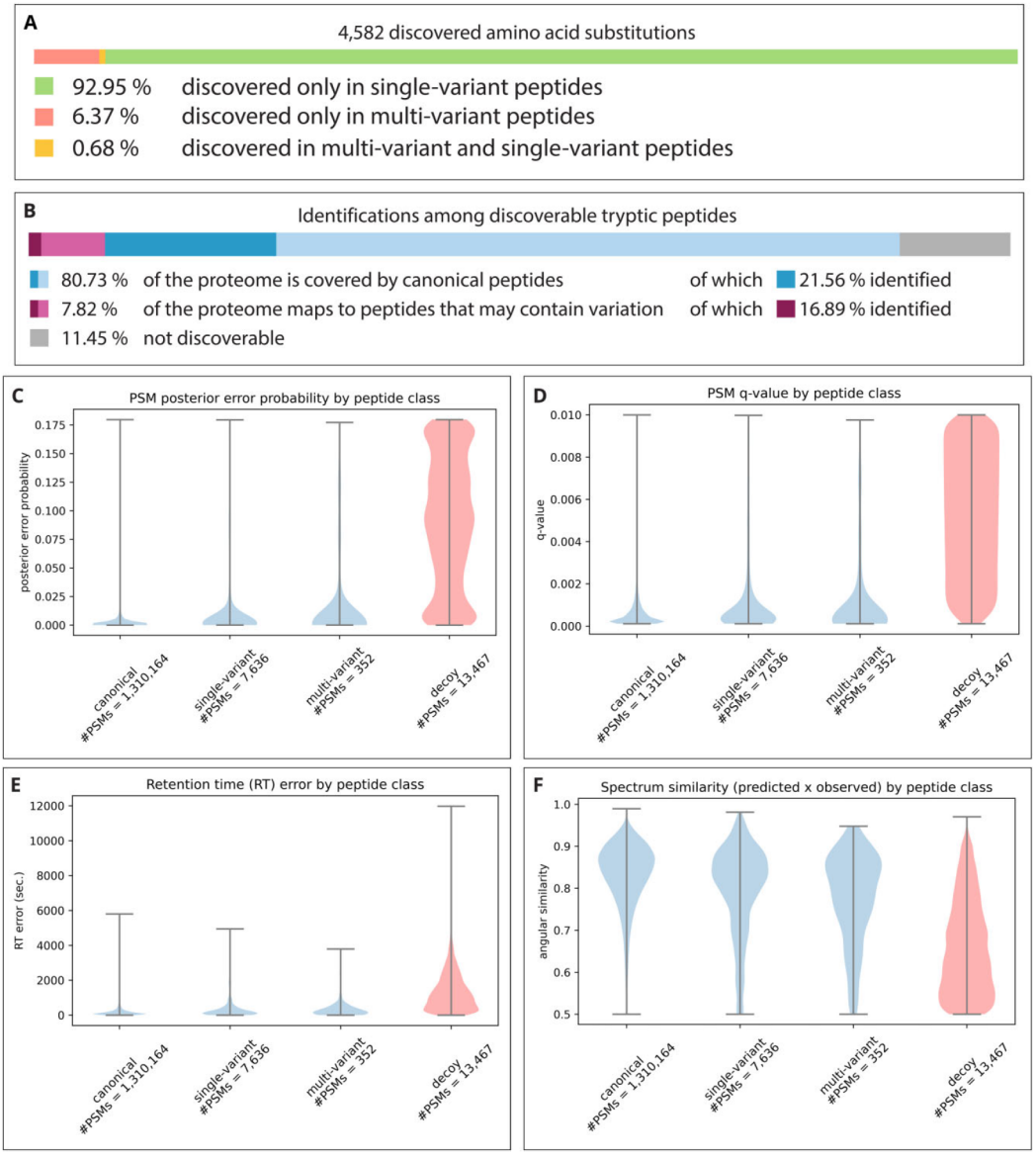


Figure 3: A: Coverage of the proteome by identified peptides, stratified by the possibility to contain variation. Lighter shades indicate the coverage by predicted peptides, darker shades represent the actual coverage by identified peptides. B: Distribution of variation in identified peptides. Amino acid variants are stratified based on the category of peptide in which the substitution caused by the respective variant can be identified. C-F: Distribution of four confidence measures among PSMs for peptide categories: posterior error probability (PEP), q-value, difference between observed and predicted retention time, and angular similarity between the observed and predicted spectrum. Decoy PSMs for this comparison were thresholded to 1% PSM-level FDR.

(Table 2). The frequency of this haplotype among participants of the European population in the 1000 Genomes project is 46%, while in this population, the aggregated frequency of all haplotypes not containing any of these substitutions is 1.98%. However, the sequence of the corresponding region of POTEI is highly similar to the sequence of actin beta (ACTB), actin gamma 1 (ACTG1), and actin alpha 1 (ACTA1), differing in 1, 1, and 2 residues, respectively. Such highly similar sequences represent peptides differing in their composition by only a few atoms, a mass difference that can be indistinguishable from a chemical or posttranslational modification (e.g., a chemical modification of methionine can be mistaken for a substitution of methionine to threonine [29]).

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

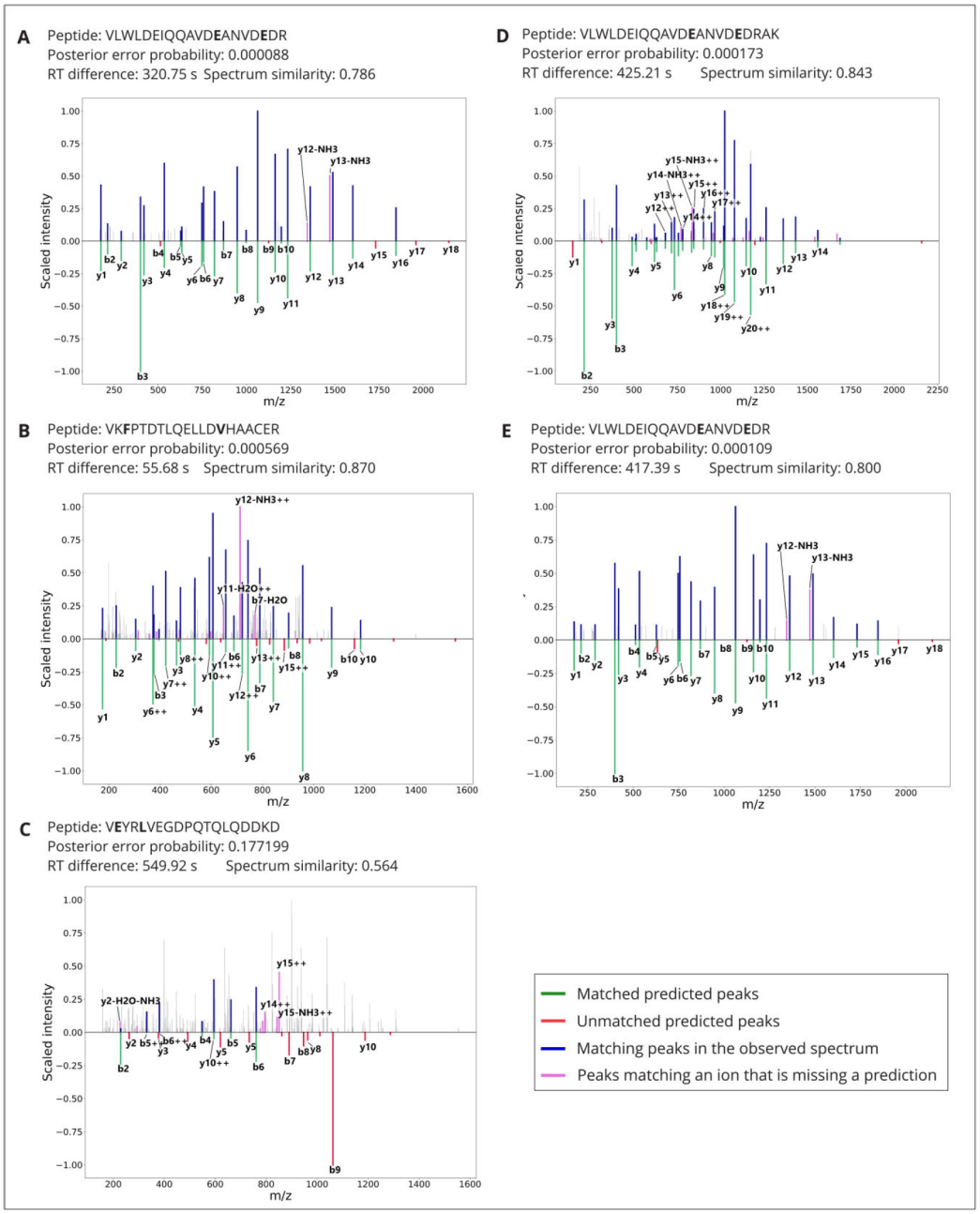


Figure 4: Quality control metrics and spectra of 5 multivariant PSMs. Amino acid substitutions are marked in bold. PSM A is among the 10% top-scoring matches to multivariant peptides by posterior error probability, B scores as the median value, and C is the lowest-scoring match to a multivariant peptide. PSMs D and E are within the 5 top-scoring matches for the most common haplotype of IQGAP2. The posterior error probability as obtained from Percolator is listed along with retention time difference to prediction as obtained from DeepLC and spectrum similarity with prediction as obtained from MS2PIP. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Note that in this representation, peaks matching a fragment ion with a predicted intensity of zero will not be annotated.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

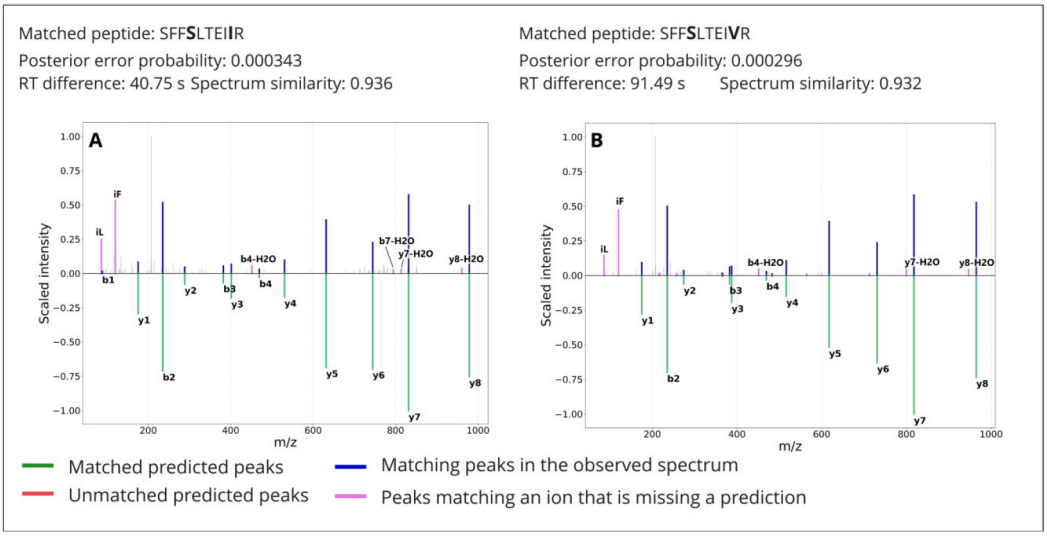


Figure 5: Quality control metrics and spectra for PSMs matching both the reference (A) and the alternative (B) allele in the same sample. The posterior error probability as obtained from Percolator is listed along with retention time difference to prediction as obtained from DeepLC and spectrum similarity with prediction as obtained from MS2PIP. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Note that in this representation, peaks matching a fragment ion with a predicted intensity of zero will not be annotated.

Table 2: PSMs mapping to the 5 highly similar protein sequences: actin gamma 1 (ACTG1)/actin beta (ACTB), actin alpha 1 (ACTA1), and 3 haplotypes of POTEI. REF marks the canonical sequence. We specify the number of confident PSMs matching the sequence of interest and number of samples with any spectra matching to these peptides.

Protein haplotype	Peptide sequence	No. of confident PSMs
ACTG1: REF ACTB: REF	LC**Y**VALDFE**Q**EMA**T**AASSSSLEK	3,298
ACTA1: REF	LC**Y**VALDFE**N**EMA**T**AASSSSLEK	385
POTEI: REF	LC**Y**VALDFE**Q**EMA**M**AASSSSLEK	103
POTEI: 918Y>F,929M>T	LC**F**VALDFE**Q**EMA**T**AASSSSLEK	19
POTEI: 918Y>F	LC**F**VALDFE**Q**EMA**M**AASSSSLEK	18

Therefore, telling these 2 proteins apart can be extremely challenging when accounting for variants. Conversely, if only canonical sequences are included in the database, the spectra obtained from POTEI will be arbitrarily assigned to actin. Numbers of spectra matching the corresponding regions of these proteins are listed in Table 2. Matching spectra to each of the peptide sequences in Table 2 have been identified in all 3 samples.

The need to distinguish very similar sequences makes the use of haplotype-specific databases particularly sensitive to the spectrum identification strategy. As an example, we conducted the search again after enabling the refinement procedure of X!Tandem. This procedure is a multistep approach that selects a limited set of proteins for a secondary search with different search parameters, including more modifications and relaxing thresholds (e.g., in terms of missed cleavages). While this procedure presents the advantage to quickly scan for new peptides, it makes the evaluation of matches challenging [30] and increases the likelihood to encounter cases where a modification can be mistaken for an amino acid substitution and vice versa. Fig. 6 shows such an example of 2 matches to the same spectrum, obtained using the refinement procedure: 1 peptide contains the product of the alternative allele of 2 variants (Fig. 6A) while the other has the product of the reference allele for 1 of the variants with a modification on the N-terminus compensating the mass difference (Fig. 6B). Both matches show a good matching of the higher-mass peaks and good agreements with the predictors but a high prevalence of unmatched peaks. Based on their scores, both matches would pass a 1% FDR threshold, but the similarity between the sequences makes it challenging to assess whether 1 or the other haplotype is a better match. This example shows the difficulty to distinguish variant peptides when the amino acid substitution has a mass difference equal or very similar to a modification. Overall, we observed inflated identification rates for multivariant peptides using the refinement procedure (1,060 PSMs with refinement vs. 342 PSMs without). For example, without the refinement procedure, 19 spectra matched the multivariant sequence of POTEI among the PSMs passing a 1% FDR threshold (Table 1); with the refinement procedure, the results contained 113 matching spectra. From the 94 additional matches, we suspect that many correspond to other sequences that were artifactually matched to this sequence, possibly through the addition of modifications.

Error rates derived from the target-decoy strategy rely on the modeling of the null distribution of scores using random matches. Distinguishing a variant peptide from a modified one, however, requires telling apart 2 matches that are very similar and both better than random. In such cases, it is expected that modeling the null distribution using random matches provides underestimated

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

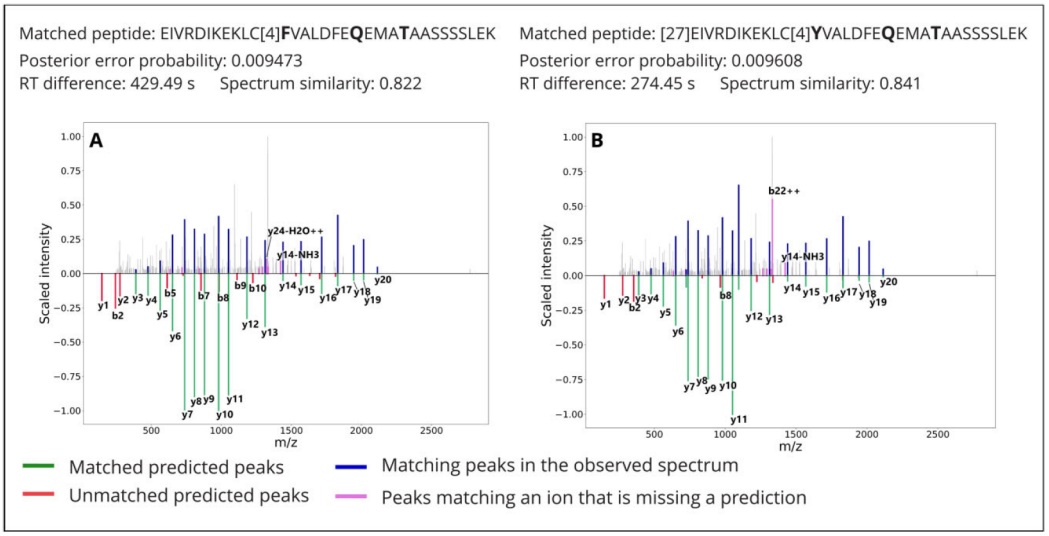


Figure 6: Comparison between predicted spectra for 2 different peptides matched to the same observed spectrum. The posterior error probability as obtained from Percolator is listed along with retention time difference to prediction as obtained from DeepLC and spectrum similarity with prediction as obtained from MS2PIP. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Numbers in the peptide sequence are identifiers of posttranslational modifications in UniMod [31].

error rates, and additional quality control measures can be applied to assess the quality of the matches [32]. We submitted the variant matches passing the target-decoy 1% FDR threshold in the X!Tandem search without the refinement procedure to PepQuery, a targeted peptide search engine providing additional validation for variant peptides identified using proteomics [33]. PepQuery found that a substantial share of the matches were low scoring or could also match another peptide (10% and 11% of the matches, respectively), and the prevalence of these matches decreased with the PEP (Fig. 7A). Conversely, 47% of the matches were labeled as confident, and the prevalence of confident matches increased with the PEP. The remaining matches were labeled as possibly matching a modification not considered in the original search—a rare posttranslational modification or an artifact introduced during sample preparation. Interestingly, the prevalence of such ambiguous matches was stable around 30% across PEP bins. These results highlight the difficulty posed by modifications in the confident identification of variant peptides. In the case of highly similar expected spectra between a variant and a modified peptide, analysts need to rely on prior knowledge on the likelihood of finding a given allele or modification in the sample studied or on the presence of diagnostic ions (Fig. 8). In the example of Fig. 8B, the detection of y29++ would advocate in favor of the variant peptide rather than the modified peptide, but this peak is of low intensity and therefore represents only thin evidence.

Moreover, we searched all spectra again using the search engine Tide [34], using the same parameters. Out of the 7,988 confident variant matches given by X!Tandem, 3,604 (45.12%) were confirmed by Tide. For 4,314 (54%) variant PSMs reported by X!Tandem, the spectra were not confidently matched to any peptide by Tide. The remaining 70 spectra were confidently matched to another peptide by Tide—in 51 cases to a canonical peptide, in 12 cases to a decoy peptide sequence, in 4 cases to a variant peptide encoded by a different haplotype but coming from the same gene, and in 3 cases to a contaminant.

Conclusion and Discussion

In this study, we propose to take advantage of the correlation between alleles through linkage disequilibrium to allow for the identification of peptides containing multiple linked amino acid substitutions, hence avoiding the computation of all possible combinations of alleles [14]. Co-occurring alleles in the protein-coding regions of a gene yield specific protein sequences—protein haplotypes. Building upon previous work in proteogenomics, we created a search space of protein haplotypes. We observe that 7.82% of the whole proteome maps to peptides that can contain an amino acid substitution, and up to 12.42% of all discoverable substitutions are located in peptides where multiple substitutions co-occur (multivariant peptides). These cases suggest that linkage disequilibrium between alleles resulting in amino acid substitutions should be included in a proteomics search space when identifying common variation. Subsequently, we performed a reanalysis of 3 samples of healthy tonsil tissue provided by Wang et al. [17]. We identified peptides encoded by haplotypes containing 4,582 unique amino acid substitutions compared to the reference sequences of Ensembl, 6.37% of which were found only in multivariant peptides.

Although searching haplotype-specific sequences allows for the discovery of novel peptides that match to protein haplotypes, numerous challenges still remain. Of the predicted haplotypes, 78.23% contain only substitutions, and the remaining haplotypes contain other types of polymorphisms (insertions, deletions, or polymorphisms introducing or removing a stop codon). These cannot be detected using the sequences obtained from Haplosaurus. Moreover, with the introduction of haplotypes, the search space

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

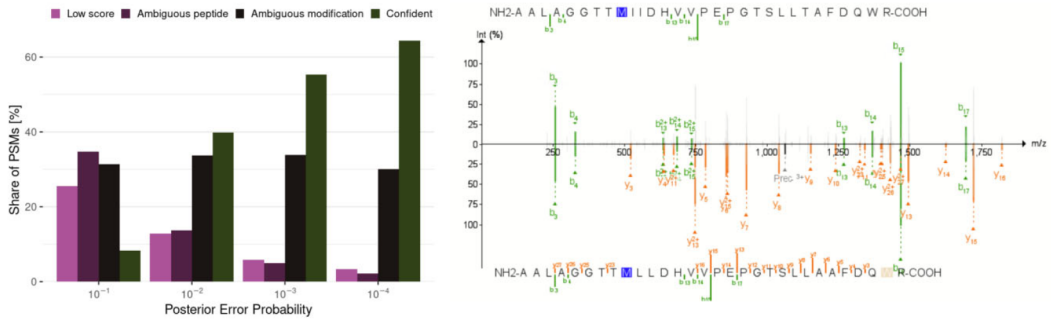


Figure 7: Analysis of variant peptides passing a target-decoy 1% FDR threshold using PepQuery. (A) Histogram of the PSM type according to PepQuery. Low score: the match was not further investigated by PepQuery due to a low score; Ambiguous peptide: the spectrum could be matched to a reference peptide at a similar score; Ambiguous modification: the spectrum could be matched to a reference peptide at a similar score when accounting for a modification that was not included in the original search; Confident: the match passed all PepQuery validation filters. (B) Mirrored annotated spectra obtained using PDV [35] of a variant PSM with better match when accounting for a modification not included in the search, here a dioxydation of tryptophan.

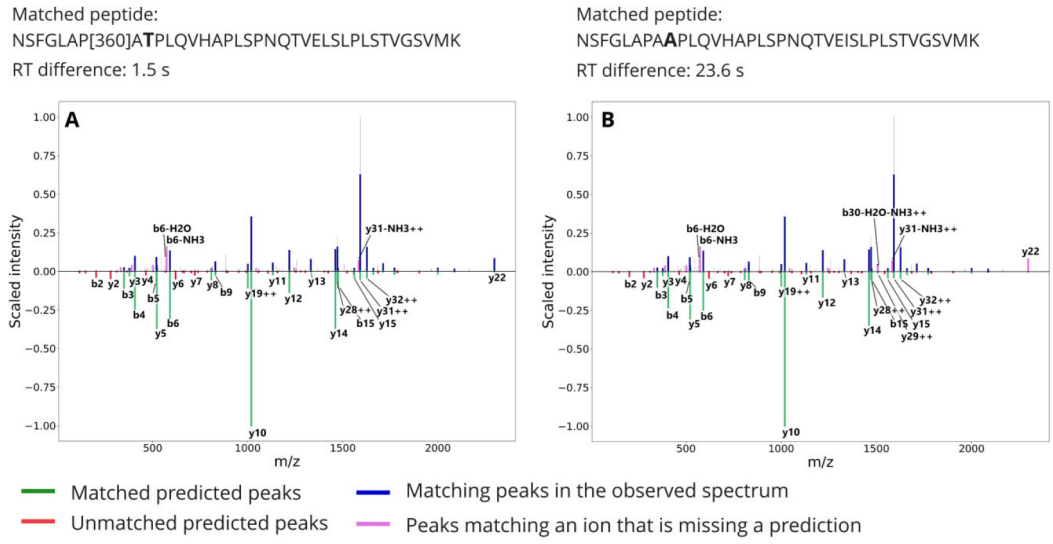


Figure 8: Comparison between predicted spectra as obtained from MS2PIP for 2 different peptides matched to the same observed spectrum. (A) Peptide candidate suggested by PepQuery is a canonical sequence with a modification. (B) Peptide candidate suggested in our search is a variant peptide. We list the retention time difference to prediction as obtained from DeepLC. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Numbers in the peptide sequence are identifiers of posttranslational modifications in UniMod [31].

consists of a large number of proteoforms with a high degree of similarity, making it challenging to infer which proteoform has been identified. Amino acid substitutions of a mass difference equal to a posttranslational or chemical modification are particularly challenging, as their distinction relies on the detection of few specific ions. This implies that searching without the correct haplotype or modification will generate incorrect sequences or modifications that are not caught by current error rate estimation strategies. Even worse, using the wrong haplotype on a protein sequence can result in a match in another protein. The prevalence of such errors in published proteomic datasets is currently unknown.

The dataset of protein haplotypes provided by Spooner et al. [1] was created using the genome assembly version GRCh37, which is now deprecated by Ensembl. During PepQuery analysis, we noted that a substantial share of variant peptides in GRCh37 would be canonical in GRCh38. For results that are fully up to date, a reanalysis of the data provided by the 1000 Genomes project on the current genome assembly is necessary. Limitations also come with the dataset of phased genotypes, as phasing may be inaccurate in regions with low linkage disequilibrium or in repetitive regions, resulting in an overestimation of haplotype frequencies [1]. Finally, the methods for the scoring of confidence of peptide-spectrum matches are not well suited to distinguishing between multiple

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

candidate sequences with a high degree of similarity. In the literature, the identification of variant peptides is validated by generating reference spectra using synthetic peptides [36, 37], but such an approach presents a substantial cost and low throughput. In the present work, we used retention time and fragmentation predictors to generate the reference spectra *in silico* and used these to evaluate the matches. Predictors can instead be directly coupled to Percolator, as implemented in MS2Rescore [38], and hence provide features that can improve the discrimination power between very similar peptides.

In conclusion, accounting for protein haplotypes in the search space for mass spectrometry–based proteomic identification improves the ability to cover relevant regions of the proteome and holds the potential to be utilized in the medical context, given that the database of protein haplotypes is complete and up to date, and novel methods of quality control are developed.

Materials and Methods

Database of protein sequences

The sequence database used for the search was built using data provided by Spooner et al. [1], who generated a database of protein haplotypes using their tool Haplosaurus, available as a part of the Ensembl Variant Effect Predictor [39]. The haplotypes were generated using phased genotype data from the 1000 Genomes project Phase 3, obtained using methods described in [2]. The haplotype analysis was performed using the transcript database Ensembl version 83 [40], human reference genome assembly version GRCh37 [1]. The data provided by Spooner et al. [1] can be found at [41]. For this work, we selected only protein haplotypes generated from minor alleles with frequency at least 1% worldwide. This database was appended with the list of canonical protein sequences in the corresponding version of Ensembl and a list of common sample contaminants, obtained from [42]. The resulting search space contains 104,736 reference sequences, assembly version GRCh37, 290,080 protein haplotype sequences obtained as described above, and 116 sequences of sample contaminants. In total, 394,959 decoy sequences were generated using the algorithm DecoyPyrat [43], provided by the tool py-pgatk [12]. The final protein sequence database in the FASTA format is available as supplementary material (SD1, SD2).

Classification of peptides

We classified peptide sequences as canonical, single-variant, or multivariant based on the number of amino acid substitutions they contain. If a peptide is canonical with respect to one protein sequence and single-variant or multivariant with respect to another protein sequence, it is classified as canonical. Similarly, if a peptide is a single-variant peptide with respect to one protein sequence and multivariant with respect to another protein sequence, it is classified as a single-variant peptide. Substitutions mapping to a peptide that has been "downgraded" in such manner are not considered as discovered, or discoverable.

Public data reanalysis

We used this database to perform a reanalysis on a subset of data published and initially analyzed by Wang et al. [17]—108 fractions from 3 samples of healthy tonsil tissue digested by trypsin, fragmented using higher-energy collisional dissociation (HCD) (MS experiment IDs: P013107, P010694, P010747).

The search was performed using the command-line interface of SearchGUI v. 4.1.16 [44], employing the X!Tandem search algorithm [19], allowing for the oxidation of methionine as a variable modification and carbamidomethylation of cysteine as a fixed modification, with the "quick acetyl" and "quick pyrolidone" options of X!Tandem enabled. PeptideShaker v. 2.2.20 [45] was used for postprocessing of the search results and export of the PSMs to Percolator v. 3.5 [20], which was used to evaluate the confidence of the matches and threshold using an FDR analysis [46]. The list of PSMs was filtered to retain matches with a q-value below 0.01 (i.e., FDR is lower than 1%). If a peptide matched to a contaminant sequence, it was removed from further analysis. As some of the canonical protein sequences in Ensembl contain multiple stop codons, the stop codon symbols were removed from their sequences for compatibility with X!Tandem. Peptides that would contain a stop codon were removed from further analysis.

Quality control

To provide supporting evidence for the confidence of the PSMs, chromatographic retention times were predicted by DeepLC v. 1.0.0 [21], and expected peptide fragment ion intensities were predicted using MS2PIP v. 3.6.3 [22]. Peptides passing the 1% FDR threshold were used for calibration of the DeepLC predictions. The absolute distance between the centered and scaled predicted and observed retention times was computed. The MS2PIP predictions were used to measure the distance between the predicted and observed spectrum. The peaks are scaled so that the median intensity in the observed spectrum corresponds to the median intensity in the prediction. A peak in the observed spectrum is considered matching to a peak in the prediction if it differs in m/z by no more than 10 ppm. The distance between the matched predicted peaks and the observed ones is expressed as their angular similarity, calculated by the formula in Equations 1 and 2:

$$C(M,P) = \frac{\sum_{i=1}^{n} m_i p_i}{\sqrt{\sum_{i=1}^{n} m_i^2}\sqrt{\sum_{i=1}^{n} p_i^2}} \quad (1)$$

$$A(M,P) = 1 - \frac{\arccos(C(M,P))}{\pi} \quad (2)$$

where $M = (m_1, \ldots, m_n)$ is the set of intensities for the matched measured peaks, and $P = (p_1, \ldots, p_n)$ is the set of intensities for the matched predicted peaks, and n is the number of matched peaks in the spectrum. $C(M, P)$ denotes the cosine similarity between M and P, and $A(M, P)$ denotes the angular similarity between M and P.

Predicted and observed spectra were also displayed as mirror plots for visual comparison in selected PSMs. The peaks in the observed spectrum matching to a predicted peak are highlighted in blue. As the intensity prediction for certain ion fragments by MS2PIP is missing, peaks matching those ions are highlighted in pink. The remaining measured peaks are displayed in gray. Peaks of the predicted spectrum are shown as negative values and labeled by the corresponding fragment ion. The predicted peaks that match a measured peak are displayed in green, and unmatched predicted peaks are displayed in red.

PepQuery analysis

The variant PSMs passing 1% FDR at PSM level using X!Tandem were further validated using PepQuery (v2.0.3) [33]. The following parameters were used: fixed modifications, carbamidomethylation of C; variable modifications, oxidation of M, ammonia loss of C, Glu→pyro-Glu of E, Gln→pyro-Glu of Q, acetylation of peptide N-term; precursor ion mass tolerance, 20 ppm; MS/MS mass tolerance, 0.05 Da; enzyme specificity, trypsin; maximum missed cleavages, 2; allowed isotope range: −1,0,1,2. The parameter "-hc" was also set in the analysis. The human protein database from

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

GENCODE Release 43 (GRCh37) was used as the reference protein database in the validation. The PSMs that passed all the filtering steps in PepQuery were considered confident. The filtering process is described in detail in [33]. Amino acid substitution modifications were not considered in the filtering process. PSMs classified as low scoring were assigned a score above the threshold of 12 by the Hyperscore algorithm, as is the default; see [33] for details. A complete list of variant PSMs with possible alternative peptide candidates suggested by PepQuery is available as supplementary material (SD5).

Source Code and Requirements

The pipeline to reproduce the postprocessing steps and a further description of the resulting files are provided in https://github.com/ProGenNo/FindingHaploSignatures [47].

- Project name: Finding Haplotypic Signatures in Proteins
- Project homepage: https://github.com/ProGenNo/FindingHaploSignatures
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: Snakemake v. 7.0.0 or higher, Anaconda 2022.10 or newer
- License: MIT

Additional Files

Supplementary Table S1. PSMs matching to the multivariant peptide covering a region of the most common haplotype of the IGQAP2 protein and their respective confidence measures. The posterior error probability and *q*-value as obtained from Percolator are listed along with retention time difference to prediction as obtained from DeepLC, as well as spectrum similarity with prediction as obtained from MS2PIP.

Supplementary Table S2. Search parameters used for the X!Tandem implementation in SearchGUI.

Data Availability

Supplementary data can be downloaded from figshare [47]. Other data further supporting this work are openly available in the *GigaScience* repository, GigaDB [48].

We provide the following files:

Supplementary Data 1: FASTA file including all target protein sequences (Ensembl reference proteome, protein haplotype sequences, contaminant sequences), excluding decoys.

Supplementary Data 2: FASTA file including all target and decoy sequences.

Supplementary Data 3: List of all peptide-to-spectrum matches (PSMs), resulting from the first run of X!Tandem without the refinement procedure, with all related metadata and quality control measures.

Supplementary Data 4: List of substitutions identified, along with IDs of corresponding PSMs.

Supplementary Data 5: List of variant PSMs and peptide candidates suggested by PepQuery, along with confidence scores for each peptide candidate.

Abbreviations

FDR: false discovery rate; HCD: higher-energy collisional dissociation; LC: liquid chromatography; LD: linkage disequilibrium; MS: mass spectrometry; MS/MS: tandem mass spectrometry; PEP: posterior error probability; PSM: peptide-to-spectrum match; PTM: post-translational modification.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the Research Council of Norway (project #301178 to MV), the University of Bergen, and the Novo Nordisk Foundation (project NNF20OC0063872 to S.J.). The funding bodies had no influence on the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

This research was funded, in whole or in part, by the Research Council of Norway 301178. A CC BY or equivalent license is applied to any Author Accepted Manuscript (AAM) version arising from this submission, in accordance with the grant's open access conditions.

Authors' Contributions

J.V. and D.S. developed the software. J.V., D.S., K.G.K., B.W., L.K., and M.V. analyzed the data. J.V. and B.W. quality controlled the results. J.V. wrote the initial draft of the manuscript. S.J., P.R.N., S.B., L.K., and M.V. oversaw the project and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The computations were performed on the Norwegian Research and Education Cloud (NREC), using resources provided by the University of Bergen and the University of Oslo (https://www.nrec.no).

References

1. Spooner W, McLaren W, Slidel T, et al. Haplosaurus computes protein haplotypes for use in precision drug design. Nat Commun 2018;9:4128. https://doi.org/10.1038/s41467-018-06542-1.
2. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015;526:68–74.
3. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res 2022;50:D988–95. https://doi.org/10.1093/nar/gkab1049.
4. Xuan Y, Bateman NW, Gallien S, et al. Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies. Nat Commun 2020;11:5248. https://doi.org/10.1038/s41467-020-18904-9.
5. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207. https://doi.org/10.1038/nature01511.
6. Verheggen K, Raeder H, Berven FS, et al. Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. Mass Spectrom Rev 2020;39:292–306. https://doi.org/10.1002/mas.21543.
7. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics MCP 2005;4:1419–40. https://doi.org/10.1074/mcp.R500012-MCP200.
8. Pagel O, Loroch S, Sickmann A, et al. Current strategies and findings in clinically relevant post-translational modification-

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

specific proteomics. Expert Rev Proteomics 2015;12:235–53. https://doi.org/10.1586/14789450.2015.1042867.
9. Menschaert G, Fenyö D. Proteogenomics from a bioinformatics angle: a growing field. Mass Spectrom Rev 2017;36:584–99. https://doi.org/10.1002/mas.21483.
10. Vizcaíno JA, Kubiniok P, Kovalchik KA, et al. The Human Immunopeptidome Project: a roadmap to predict and treat immune diseases. Mol Cell Proteomics MCP 2020;19:31–49. https://doi.org/10.1074/mcp.R119.001743.
11. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11:1114–25. https://doi.org/10.1038/nmeth.3144.
12. Umer HM, Audain E, Zhu Y, et al. Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides. Bioinformatics 2022;38:1470–2. https://doi.org/10.1093/bioinformatics/btab838.
13. Levitsky LI, Kuznetsova KG, Kliuchnikova AA, et al. Validating amino acid variants in proteogenomics using sequence coverage by multiple reads. J Proteome Res 2022;21:1438–48. https://doi.org/10.1021/acs.jproteome.2c00033.
14. Choong W-K, Wang J-H, Sung T-Y. MinProtMaxVP: generating a minimized number of protein variant sequences containing all possible variant peptides for proteogenomic analysis. J Proteomics 2020;223:103819. https://doi.org/10.1016/j.jprot.2020.103819.
15. Lobas AA, Karpov DS, Kopylov AT, et al. Exome-based proteogenomics of HEK-293 human cell line: coding genomic variants identified at the level of shotgun proteome. Proteomics 2016;16:1980–91. https://doi.org/10.1002/pmic.201500349.
16. Lobas AA, Pyatnitskiy MA, Chernobrovkin AL, et al. Proteogenomics of malignant melanoma cell lines: the effect of stringency of exome data filtering on variant peptide identification in shotgun proteomics. J Proteome Res 2018;17:1801–11. https://doi.org/10.1021/acs.jproteome.7b00841.
17. Wang D, Eraslan B, Wieland T, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol 2019;15:e8503. https://doi.org/10.15252/msb.20188503.
18. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. Nat Methods 2013;10:186–7. https://doi.org/10.1038/nmeth.2369.
19. Fenyö D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 2003;75:768–74. https://doi.org/10.1021/ac0258709.
20. Käll L, Canterbury JD, Weston J, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 2007;4:923–5. https://doi.org/10.1038/nmeth1113.
21. Bouwmeester R, Gabriels R, Hulstaert N, et al. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat Methods 2021;18:1363–9. https://doi.org/10.1038/s41592-021-01301-5.
22. Degroeve S, Martens L. MS2PIP: a tool for MS/MS peak intensity prediction. Bioinformatics 2013;29:3199–203. https://doi.org/10.1093/bioinformatics/btt544.
23. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC–MS/MS. J Proteome Res 2011;10:1785–93. https://doi.org/10.1021/pr101060v.
24. Houel S, Abernathy R, Renganathan K, et al. Quantifying the impact of Chimera MS/MS Spectra on peptide identification in large-scale proteomics studies. J Proteome Res 2010;9:4152–60. https://doi.org/10.1021/pr1003856.
25. Alves G, Ogurtsov AY, Kwok S, et al. Detection of co-eluted peptides using database search methods. Biol Direct 2008;3:27. https://doi.org/10.1186/1745-6150-3-27.
26. Dorfer V, Maltsev S, Winkler S, et al. Boosting peptide identifications by chimeric spectra identification and retention time prediction. J Proteome Res 2018;17:2581–9. https://doi.org/10.1021/acs.jproteome.7b00836.
27. Cifani P, Li Z, Luo D, et al. Discovery of protein modifications using differential tandem mass spectrometry proteomics. J Proteome Res 2021;20:1835–48. https://doi.org/10.1021/acs.jproteome.0c00638.
28. O'Bryon I, Jenson SC, Merkley ED. Flying blind, or just flying under the radar? The underappreciated power of de novo methods of mass spectrometric peptide identification. Protein Sci 2020;29:1864–78. https://doi.org/10.1002/pro.3919.
29. Chernobrovkin AL, Kopylov AT, Zgoda VG, et al. Methionine to isothreonine conversion as a source of false discovery identifications of genetically encoded variants in proteogenomics. J Proteomics 2015;120:169–78. https://doi.org/10.1016/j.jprot.2015.03.003.
30. Everett LJ, Bierl C, Master SR. Unbiased statistical analysis for multi-stage proteomic search strategies. J Proteome Res 2010;9:700–7. https://doi.org/10.1021/pr900256v.
31. Creasy DM, Cottrell JSU. Protein modifications for mass spectrometry. Proteomics 2004;4:1534–6. https://doi.org/10.1002/pmic.200300744.
32. Helsens K, Timmerman E, Vandekerckhove J, et al. Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. Mol Cell Proteomics 2008;7:2364–72. https://doi.org/10.1074/mcp.M800082-MCP200.
33. Wen B, Zhang B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. Nat Commun 2023;14:2213. https://doi.org/10.1038/s41467-023-37462-4.
34. Diament BJ, Noble WS. Faster SEQUEST searching for peptide identification from tandem mass spectra. J Proteome Res 2011;10:3871–9. https://doi.org/10.1021/pr101196n.
35. Li K, Vaudel M, Zhang B, et al. PDV: an integrative proteomics data viewer. Bioinformatics 2019;35:1249–51. https://doi.org/10.1093/bioinformatics/bty770.
36. Johansson HJ, Socciarelli F, Vacanti NM, et al. Breast cancer quantitative proteome and proteogenomic landscape. Nat Commun 2019;10:1600. https://doi.org/10.1038/s41467-019-09018-y.
37. Kuznetsova KG, Kliuchnikova AA, Ilina IU, et al. Proteogenomics of adenosine-to-inosine RNA editing in the fruit fly. J Proteome Res 2018;17:3889–903. https://doi.org/10.1021/acs.jproteome.8b00553.
38. Declercq A, Bouwmeester R, Hirschler A, et al. MS2Rescore: data-driven rescoring dramatically boosts immunopeptide identification rates. Mol Cell Proteomics 2022;21:100266. https://doi.org/10.1016/j.mcpro.2022.100266.
39. ensembl-vep. GitHub. 2023. https://github.com/Ensembl/ensembl-vep
40. Yates A, Akanni W, Amode MR, et al. Ensembl 2016. Nucleic Acids Res 2016;44:D710–6. https://doi.org/10.1093/nar/gkv1157.
41. McLaren W, Spooner W. 2017. https://doi.org/10.6084/m9.figshare.5545084.v1.
42. cRAP protein sequences. https://www.thegpm.org/crap/. Accessed 20 October 2022.
43. Wright JC, Choudhary JS. DecoyPyrat: fast non-redundant hybrid decoy sequence generation for large scale proteomics. J Proteomics Bioinform 2016;9:176–80. https://doi.org/10.4172/jpb.1000404.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

44. Vaudel M, Barsnes H, Berven FS, et al. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!tandem searches. Proteomics 2011;11:996–9. https://doi.org/10.1002/pmic.201000595.
45. Vaudel M, Burkhart JM, Zahedi RP, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotechnol 2015;33:22–24. https://doi.org/10.1038/nbt.3109.
46. Käll L, Storey JD, Noble WS. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics 2008;24:i42–8. https://doi.org/10.1093/bioinformatics/btn294.
47. Vasicek J, Skiadopoulou D, Kuznetsova KG et al., Supplementary data: Finding haplotypic signatures in proteins. 2022. https://doi.org/10.6084/m9.figshare.21408117.v3.
48. Vašíček J, Skiadopoulou D, Kuznetsova KG, et al. Supporting data for "Finding Haplotypic Signatures in Proteins." GigaScience Database. 2023. https://doi.org/10.5524/102458.

Downloaded from https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad093/7334270 by guest on 03 November 2023

Received: April 13, 2023. **Revised:** September 24, 2023. **Accepted:** October 8, 2023
© The Author(s) 2023. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

8.2 ProHap Enables Human Proteomic Database Generation Accounting for Population Diversity

Vašíček, J., Kuznetsova, K. G., Skiadopoulou, D., Unger, L., Chera, S.;, Ghila, L. M., Bandeira, N., Njølstad, P. R., Johansson, S., Bruckner, S., Käll, L., Vaudel, M.

Nature Methods, **12**, 273–277 (2025)

The reprint of Paper 2 is permitted by Springer Nature. All rights reserved.

8.3 ProHap Explorer: Visualizing Haplotypes in Proteogenomic Datasets

Vašíček, J., Skiadopoulou, D., Kuznetsova, K. G., Käll, L., Vaudel, M., Bruckner, S.

IEEE Computer Graphics and Applications, **45(5)**, 64-77 (2025)

The reprint of Paper 3 is permitted by IEEE. All rights reserved.

9 Errata

Errata for
Enabling Haplotype-Aware Proteomics to Better Connect Human Genomes and Proteomes

Jakub Vašíček

Thesis for the degree philosophiae doctor (PhD)

at the University of Bergen

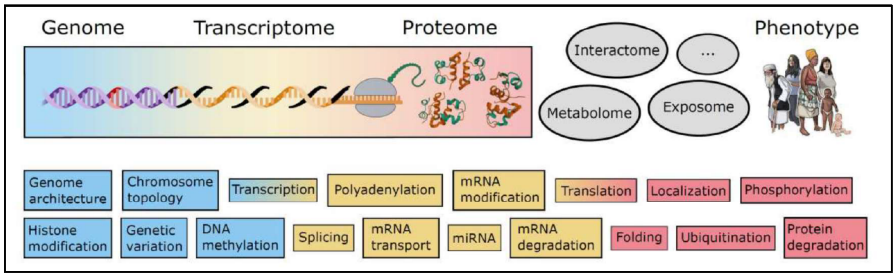
21. 12. 2025 _____ 22.12.25 _____

(date and sign. of candidate) (date and sign. of faculty)

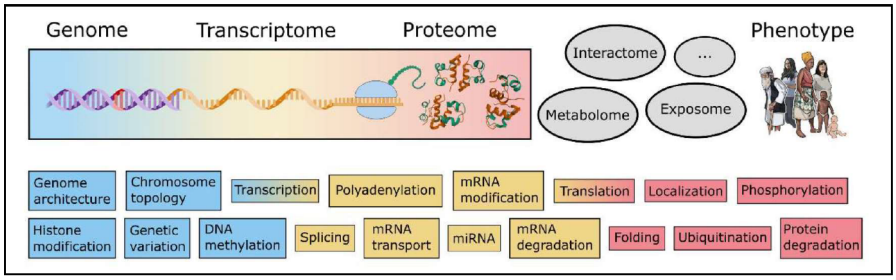
Errata

Page 2: Misplaced comma: "The initial completion of the Human Genome Project, catalyzed a new research paradigm" – corrected to "The initial completion of the Human Genome Project catalyzed a new research paradigm".

Page 3: Error in the rendering of Figure 1.1:



Corrected to:



Page 6: Missing parenthesis: "contribute to alternative splicing (see Figure 1.4 [60]." - corrected to "contribute to alternative splicing (see Figure 1.4) [60]."

Page 12: Missing word: "As introduced in Section 1.1, frequencies and effects differ between populations." – corrected to "As introduced in Section 1.1, allele frequencies and effects differ between populations."

Page 17: Misplaced word: "this thesis will focus on DDA mass spectrometry as, and thus assume (...)" – corrected to: "this thesis will focus on DDA mass spectrometry, and thus assume (...)".

Page 23: Typo: "protein haplotypes encoded by individual genesa" – corrected to: "protein haplotypes encoded by individual genes".

Page 29: Incorrect reference and formatting: "the work of John Snow mapping cholera cases in London helped identify the source of the outbreak [232]." – corrected to: "the work of John Snow mapping cholera cases in London helped identify the source of the outbreak[233]."

Graphic design: Communication Division, UiB / Print: Skipnes AS

uib.no

ISBN: 9788230874448 (print)
9788230883044 (PDF)