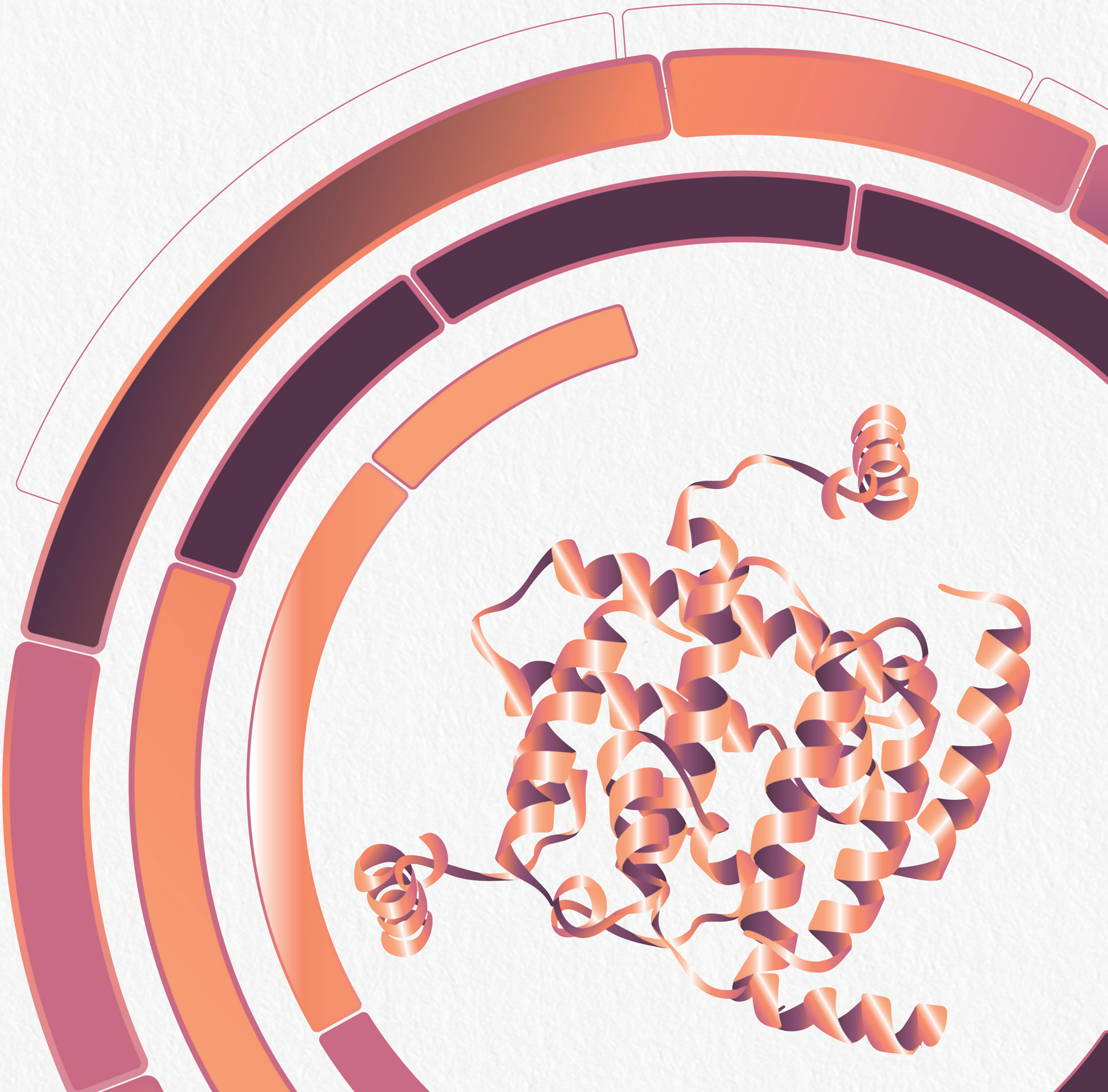
Methods and lessons in long-read proteogenomics

Renee Salz



Methods and lessons in long-read proteogenomics

Renee Salz

ISBN:	978-94-6506-261-7
Cover design:	Ridderprint | www.ridderprint.nl
Lay-out:	Ridderprint | www.ridderprint.nl
Print:	Ridderprint | www.ridderprint.nl

Methods and lessons in long-read proteogenomics

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 27 augustus 2024
om 12.30 uur precies

door

Renee Lynn Salz
geboren op 7 februari 1995
te Boston (Verenigde Staten)

Promotor:

Prof. dr. P.A.C. 't Hoen

Copromotor:

Prof. dr. ir. P.J. Volders (Universiteit Gent, België)

Manuscriptcommissie:

Prof. dr. C.F.H.A. Gilissen
Prof. dr. V. Guryev (Rijksuniversiteit Groningen)
Dr. J. Gloerich

Methods and lessons in long-read proteogenomics

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Tuesday, August 27, 2024
at 12:30 pm

by

Renee Lynn Salz
born on February 7, 1995
in Boston (United States)

PhD supervisor:

Prof. dr. P.A.C. 't Hoen

PhD co-supervisor:

Prof. dr. ir. P.J. Volders (Ghent University, Belgium)

Manuscript Committee:

Prof. dr. C.F.H.A. Gilissen
Prof. dr. V. Guryev (University of Groningen)
Dr. J. Gloerich

Table of Contents

1

Chapter 1

General introduction

Proteogenomics: what is it?

Proteogenomics is a recently-introduced methodology for obtaining new insights into basic biological processes controlling gene expression regulation and the discovery of novel disease mechanisms and biomarkers. Proteogenomics is a multi-omics technique that leverages data from the genome, transcriptome and proteome of a sample to inform one other. Proteogenomics methodology enables the comprehensive elucidation of complex, inter-dependent biological systems. The bioinformatics methodologies to extract insights from these data are error prone and an active area of research. This thesis contains a critical examination of the latest proteogenomics methodologies, original research to improve bioinformatics methods used in proteogenomics and a demonstration of their potential. In this introduction, I will outline the types of important biological questions that can be explored with the help of proteogenomics, the key experimental methods that produce the data used in proteogenomics, and the bioinformatics methods that make the resulting biological insights possible (Figure 1).

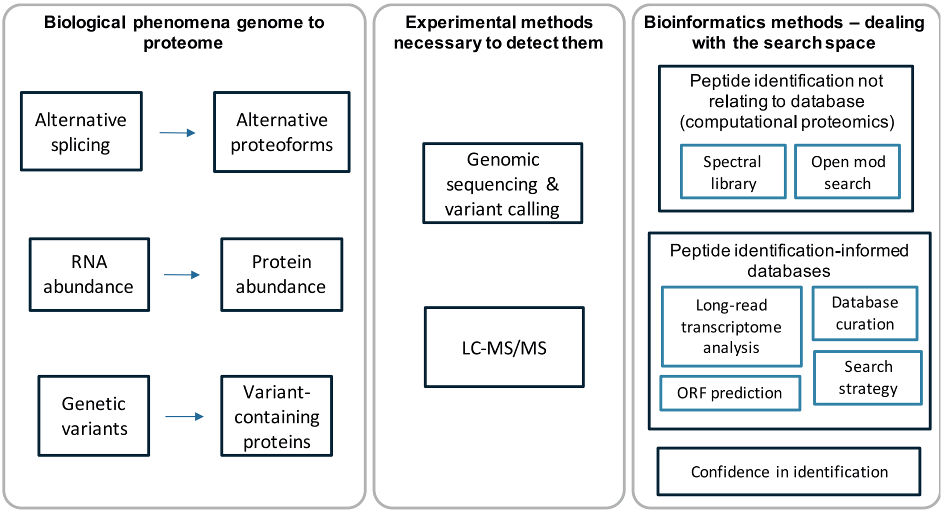

Biological phenomena genome to proteome
Alternative splicing
Alternative proteoforms
RNA abundance
Protein abundance
Genetic variants
Variant-containing proteins
Experimental methods necessary to detect them
Genomic sequencing & variant calling
LC-MS/MS
Bioinformatics methods – dealing with the search space
Peptide identification not relating to database (computational proteomics)
Spectral library
Open mod search
Peptide identification-informed databases
Long-read transcriptome analysis
Database curation
ORF prediction
Search strategy
Confidence in identification

Figure 1: Structure of Chapter 1. LC-MS/MS = liquid chromatography tandem mass spectrometry. ORF= open reading frame. Mod = modification.

Biological phenomena in the translation of genome to proteome

In any human cell, intricate regulatory pathways are at play. We have an estimated 20,000 genes that eventually become perfectly folded, biologically active proteins carrying out specific functions exactly when they are needed[1]. Ultimately, knowledge of the proteome translates closely to understanding of cellular mechanisms. There is, however, a lot of activity before a protein is produced. The paradigm one gene, one transcript, one protein does not hold in eukaryotes. There is an order of magnitude more proteins than genes, and an even greater number of transcripts that either do not produce functional protein or do so at wildly varying proportions. The path to protein from DNA is controlled by numerous factors. To understand the regulatory factors at play, we have to study the genome *and* all its products.

Alternative splicing

Diversity on the protein level can be largely attributed by alternative splicing (AS)[2,3]. AS is a mechanism by which exons from a gene are joined together in different combinations to make various transcripts and thus produce various different proteins. AS is an essential and widespread process in multi-exon genes, and an important contributor to genomic diversity as well as tissue specificity[4]. AS begins with transcription of DNA sequence into precursor RNA containing both exons and introns of a gene. The spliceosome assembles on the pre-RNA and catalyzes the removal of introns according to 5’ donor and 3’ acceptor splice sites (GU and AG respectively). Different combinations of introns are included and excluded from the mature mRNA, which is referred to as AS. There are several categories within AS that refer to the alternative use, skipping and/or retention of either exons or introns.

Transcript diversity is largely attributed to AS and is a reason that humans have approximately equal or sometimes fewer genes than some less complex eukaryotic organisms such as a freshwater crustacean (*Daphnia pulex*). This diversity is useful; AS allows us to react to a variety of certain cellular needs/environments. In immune functions, for example, there is a heavy reliance on AS for T-cell response to antigens[5]. In embryonic development, there is carefully coordinated AS that is location- and time-specific[6,7]. The diversity that AS creates subsequently affects downstream processes such as cell signaling and protein-protein interaction networks[8,9]. These functions are so essential that AS malfunction is cause for a variety of cancers and other diseases[10–12].

The assumption behind AS is that alternatively spliced transcripts result in changes in proteins that are expressed. The reality is more nuanced, since there are translational regulatory mechanisms that come in between a spliced, mature transcript and a resulting protein[13]. The signals for this type of regulation are often contained in the mRNA sequence itself. Lengths and sequences of the untranslated regions can affect interaction with translation machinery, slowing or stopping the production of protein from a given transcript[14]. Certain motifs in 3’ untranslated regions bind to microRNAs (miRNAs), which serve to degrade transcripts thereby inhibiting translation[15]. Some transcripts containing abnormal stop codons are simply destroyed before translation through nonsense-mediated decay as a mechanism to prevent the production of potentially aberrant proteins[16]. These are just a few examples of the many mechanisms of post-transcriptional regulation in humans. Even if translation inhibition is absent, the predicted open reading frame may not be the one that is being translated. Many non-canonical open reading frames have been discovered in recent years[17]. Considering the many factors at play, assuming knowledge of which proteins are present from the transcriptome alone is misguided.

RNA and protein abundance

Since mRNA presence is prerequisite for protein production, the abundance of RNA is often used as the proxy for protein. However, variance in protein abundance explained by mRNA abundance ranges from 30-80% depending on the biological system, experimental setup and statistical models[18]. Although the two abundances correlate poorly to one another, the abundance of mRNA is a good proxy for presence (versus absence) of detectable protein in cells[19]. Regulation at the level of mRNA abundance is thus setting the “on” or “off” state of the gene[20]. Other post-transcriptional regulation mechanisms such as RNA interference are responsible for the fine-tuning of protein levels[21]. Even taking into account post-translational regulatory processes would lead to an inaccurate estimation of protein levels, as there are additional regulatory feedback loops between transcription and translation that are not completely understood[22].

Genetic variation

Diversity in transcripts and proteins can also arise from genetic variation. Large-scale sequencing efforts have led to the creation of a reference human genome and a better understanding of the true extent of genetic variation[23–26]. There are many types of genetic variations, but the most common and interesting from a molecular function perspective are small-scale. Single nucleotide polymorphisms (SNPs) involve a single nucleotide change at a particular position on the DNA sequence that could be present in either coding or non-coding regions. SNPs are the most common human variants. There are 4-5 million in any individuals’ genome. SNPs are associated with a wide range of phenotypic traits[27,28]. While the effects of individual SNPs on disease are typically small, their collective effects, summarized in genetic risk scores, can be large. Rare variations in the genome tend to have even larger effects on disease risk and are responsible for the majority of monogenetic disease[29,30].

The reason that single base pair perturbations can cause disease is because of their ability to alter or disable a molecular function. A variant can cause multiple types of functional disruption with varying severity depending on where it is located. If the variants occurs in regions where proteins are coded, it can affect the functioning of a protein and disease phenotype[31,32]. Missense variants, a type of variant that causes a single amino acid substitution in a protein, are an important class of variants that are responsible for a reported 60% of Mendelian diseases[33]. A missense variant can alter/ destroy the functionality of the protein it occurs in, referred to as deleterious, or it could remain completely unchanged. A variant could also fall in intronic or intergenic regions. These variants are generally less impactful than those in coding regions, but there is evidence of non-coding variants having pathogenic effects[34,35]. In general, a more disruptive variant is more likely to be the cause of disease, but exact mechanisms must be experimentally verified to infer causation.

The number of catalogued rare variants is increasing, as is our need to understand their possible link to disease[36]. Finding the cause of a disease requires a deeper understanding of the specific way a variant disrupts the system in question. The latest high-throughput experimental methods to measure variants' effect on phenotype such as deep mutational scanning do not predict clinical phenotypes and still struggle to scale genome-wide, making computational prediction preferable[37,38]. It is possible to computationally predict variant effects by simply comparing the coordinate of the variant with the annotation at that location. However, some predictions require additional information to determine the extent of protein disruption. For missense variants, chemical properties of amino acids/protein structures as well as sequence conservation at the position of the substitution in an alignment with homologous sequences are two of the most important determining factors. Specific software such as PolyPhen2[39], SIFT[40] and CADD[41] use one or both of these factors in a classifier to determine the deleteriousness of a missense variant, and by extension, the likelihood of the variant in question to be associated with disease.

Experimental methods enabling proteogenomics

The biological questions that proteogenomics addresses requires a plethora of data which has only recently become practical to collect. Experimental methods to collect that data have improved rapidly in recent years, becoming more accurate, higher throughput and more easily accessible than ever before. The correct data is the indispensable basis for understanding the complex, individualized journey from DNA to protein.

Genomic sequencing and variant calling

Next generation sequencing (NGS) refers to post-Sanger high throughput sequencing of DNA and RNA, and is hallmarked by resolving short sequences from millions of fragments in parallel. State-of-the-art NGS platforms allow the elucidation of whole human genomes as well as comprehensive transcriptomes and genetic variants. Most technologies in NGS re-purpose DNA replication machinery to arrange fluorescent molecules for detection by a machine. The most widespread NGS technology on the market today is Illumina sequencing, a sequencing-by-synthesis (SBS) method utilizing reversible dye terminators[42]. In Illumina sequencing, millions of template DNA strands are bound to a glass slide and amplified in-situ. During sequencing, each template is extended one base at a time with fluorescent bases in subsequent cycles of reagent administration and washing. During each cycle, the intensity and positions of the fluorescent signals are captured by a microscope. After each cycle, a restoration step occurs wherein all modified bases are converted back to regular bases, priming the system for the next round of base extension. At completion of the run, the colors are matched back to their subsequent base in a process referred to as basecalling. The bases recorded from a single template position form a "read", which is typically up to 150 base pairs[43].

With this technology we can *observe* human variation, but *calling* it is harder. Simply put, positions that deviate from the reference sequence with enough read support are called as variants. There are plenty of technical challenges involved with calling variants from NGS-derived raw sequencing reads. Even when primers and low-quality bases are removed from the reads, artifacts may still be present in the data that could be confused for variation[44,45]. This post-processing can be more complicated, for instance if targeted sequencing protocols were used. Subsequently, the reads must be mapped to a reference genome, a non-trivial task for which over 60 mapping algorithms have been designed[46]. The performance of these mappers varies and is dependent on the sequencing quality, read length and sequencing error rate. After mapping, a variant-calling algorithm will *call* variants in the sample by iteratively assessing each position of interest and combining information from all the reads mapped to that position. The number of times a position is observed in the sequencing data (depth) is a major consideration as it influences the computational load of the algorithms but also the accuracy of the variant calling[47,48].

Despite cost-effectiveness and high accuracy of short-read sequence technologies such as Illumina SBS, these technologies are associated with intrinsic limitations. Large genomic regions with high inter-individual (structural) variation are problematic to assess. RNA sequencing can be performed when RNA is converted to cDNA, but isoform variation in RNA transcriptomes is difficult to observe. Long-read sequencing has been quite successful in addressing these[49–52]. Two distinct technologies dominate the long-read sequencing scene: Pacific Biosciences' single molecule real-time sequencing (SMRT) and Oxford Nanopore Technologies' nanopore sequencing (ONT) (Figure 2). SMRT sequencers detect fluorescence corresponding to nucleotides that are added by an immobilized polymerase on the bottom of a well. ONT sequencing measures ionic current fluctuations that result from single stranded nucleic acids passing through a biological nanopore; resistances in the pore vary with the different nucleotides. While sequencing accuracy of these technologies was initially much lower compared to short-read sequencing, it has improved dramatically since their introduction, and has shown to be further improved with consensus of multiple sequencing "passes" of the same read[53,54]. Having recently resolved the last unknown portions of the human genome[24,55], long-read sequencing has successfully established its indispensability in the genome sequencing space. There is now a thriving ecosystem of an estimated number of 350 bioinformatics tools to process long-read data[56].

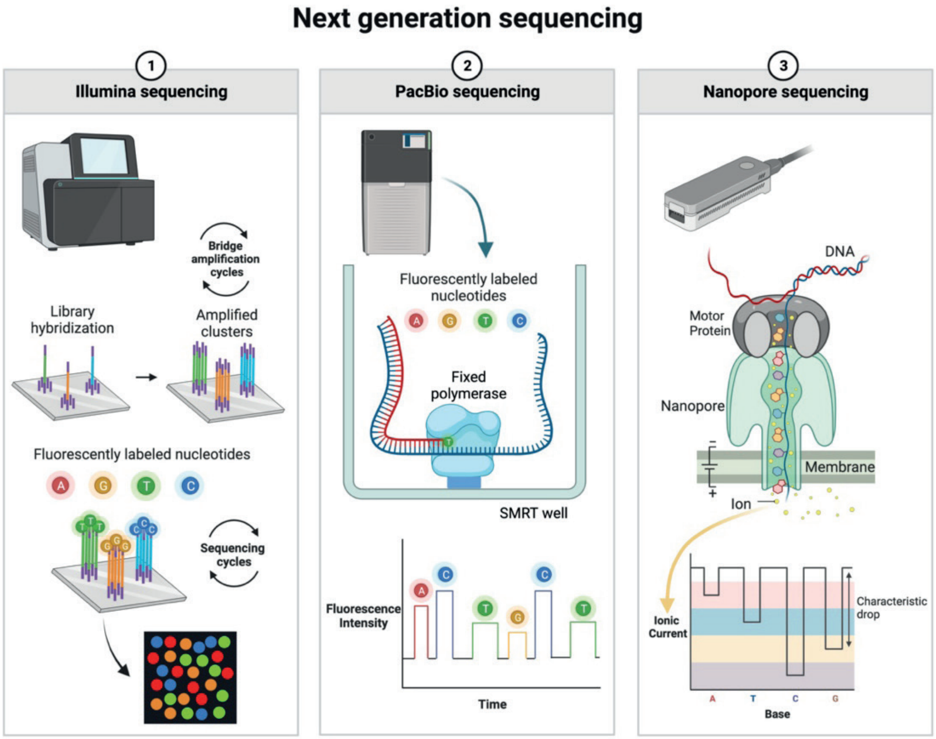


Figure 2: Under the hood of next generation sequencing approaches.

Long-read sequencing shows promise over short-read sequencing, particularly in transcript isoform detection[51,57,58]. It is difficult to use short reads to resolve full transcript structures (mRNAs), which are on average 2 kilobase (kb) in length[59,60], but can be as long as 20 kb. Performing *de novo* assembly of the short reads to detect new transcripts cannot overcome the limitation, because there are too many exon combination possibilities based on available reads, even sophisticated algorithms struggle[61,62]. Long read sequencing is addressing the limitations by outputting reads that are 3kb on average, clearing the length of the bulk of human transcripts. Similar to short read data, reads from long-read sequencing are typically first aligned to the genome before assembly. Specific alignment software/settings must be used with long reads since they more frequently have features such as sequencing errors and short exons[63–65]. Although long-reads reduce the ambiguity of the assembled transcriptome, there are still conflicting assembly methods/definitions of transcript novelty that cause considerable variation in output between assemblers[66–68]. Some definitions consider short exons, non-canonical splice junctions and/or alternative 3 and 5' ends on transcripts (with otherwise known splicing patterns) to be novel transcripts. Other more conservative tools assume that one or more of these events are artifactual and "correct" them away using reference junction coordinates. Some research suggests that the best way to resolve this ambiguity is to combine long and short read sequencing in a hybrid approach[69,70].

LC-MS/MS to observe genomic variation

Support for the extent of diversity in the transcriptome can be provided in the form of protein evidence. The most common way to provide this on a whole-proteome level is with mass spectrometry proteomics[71]. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) in proteomics is a commonly-used method that involves the separation of proteins by chemical properties and subsequent analysis to determine presence of proteins in a biological sample[72].

In a typical shotgun proteomics protocol, proteins are first extracted from the sample and fractionated to reduce sample complexity and increase sensitivity. Proteins are then broken down into peptides using enzymatic cleavage, typically with Trypsin. Peptides are then separated based on hydrophobicity in an LC step and then ionized, generating charged ions. In data-dependent acquisition (DDA) mode, selected precursor ions are fragmented with a fragmentation technique depending on the MS instrument. Resulting product ions are analyzed to determine the amino acid sequence of the peptide, which is then bioinformatically matched back to the protein of origin.

DDA has intrinsic detection limitations, however[73]. Since only the highest abundant precursor ions are selected for fragmentation, the detection rate of low abundant peptides is hampered[74]. Also, precursor selection has some randomness that limits reproducibility[75].

Since AS events and genetic variants can only be observed in a small number of peptides, relatively higher coverage is necessary to detect these events. Some details of the protocol can be optimized for the desired amount of coverage without the need for additional replicates. For instance, separating the isolate in more fractions can increase the amount of the sample going through the mass spectrometer and thus increase the number of peptide spectra produced[76]. The use of additional digestion enzymes other than trypsin can expand the variety of peptides, increasing the chance of covering the event with a optimally detectable peptide[77]. Precursor selection processes including parallel reaction monitoring (commonly referred to as targeted proteomics) can be utilized to increase sensitivity to certain variant peptides, but they require the use of synthetic peptides chosen beforehand[78].

Bioinformatics methods encompassing proteogenomics

Terabytes worth of biological data can be generated, but discovery only happens after making sense of the data using bioinformatics. The methodological improvements seen in the last years have begun to provide the necessary measurements, but bioinformatic post-processing is the last piece of the puzzle to provide answers to our aforementioned biological questions. Accurate peptide identification is the backbone of discovery in the field of proteogenomics, as the answers to all the biological questions can ultimately be found in the proteome.

Unfortunately, the accurate identification of peptide spectra is the biggest challenge at present. Even if we manage to experimentally acquire the spectra of non-standard peptides, there are difficulties in identifying them as such[79]. A major issue lies in the creation of the database used to identify the peptides. Peptides acquired in a DDA experiment are usually identified using a database search method wherein peptides are compared against all proteins that can be expected to be in the sample[80]. In human-derived samples, this would be all human proteins from Uniprot plus a “junk” database containing proteins from common lab reagents. A standard database search would thus at best yield only identifications from proteins that are already known, and yield incorrect identifications at worst[81]. *De novo* sequencing in proteomics eliminates the need for a peptide search database, but suffers from prediction accuracies around 35%, making database-search the more reliable identification method in general[82]. In a typical proteomics workflow, less than half of the acquired spectra go unidentified and are thus discarded[83]. In general, proteo(geno)mics aims to increase that yield by screening for peptides outside of the refence proteome. Computational proteomics focuses on a data-driven approach while proteogenomics uses a data-informed approach to address obstacles in the identification of unidentified spectra (Figure 3).

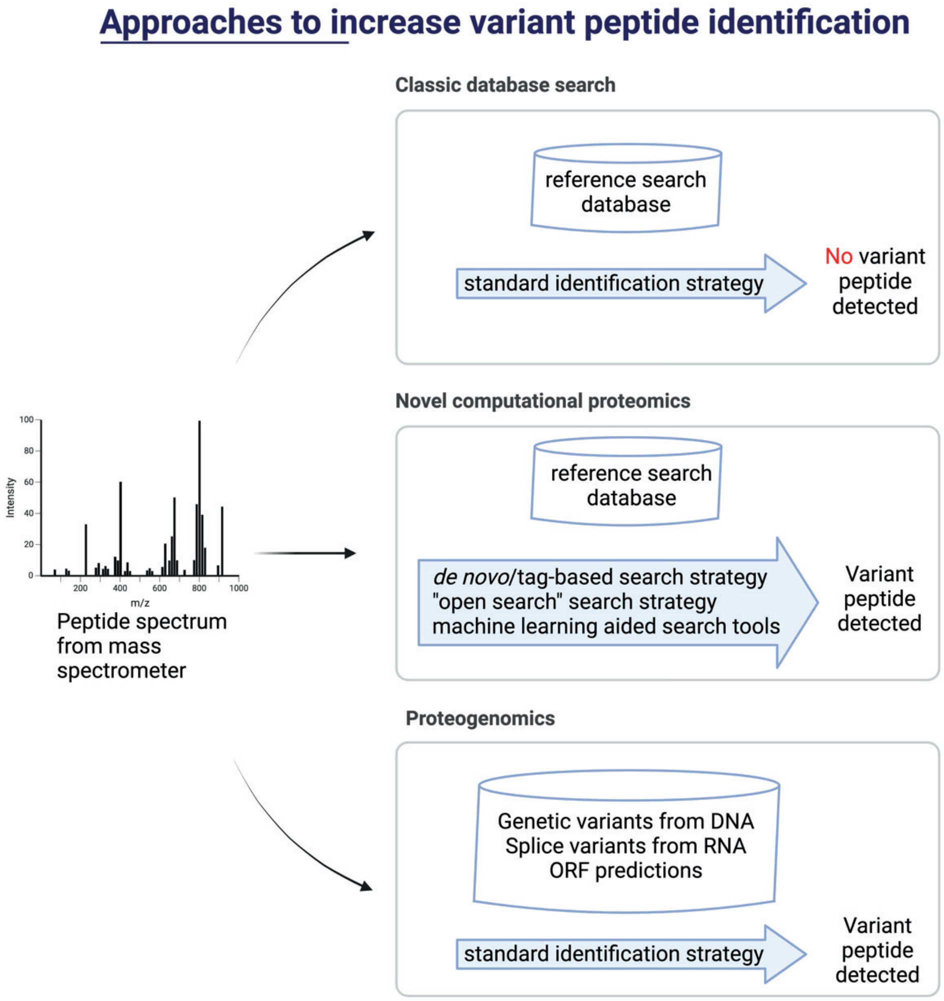

Approaches to increase variant peptide identification
Classic database search
reference search database
standard identification strategy
No variant peptide detected
Novel computational proteomics
reference search database
de novo/tag-based search strategy
"open search" search strategy
machine learning aided search tools
Variant peptide detected
Proteogenomics
Genetic variants from DNA
Splice variants from RNA
ORF predictions
standard identification strategy
Variant peptide detected
Intensity
100
80
60
40
20
0
200
400
600
800
1000
m/z
Peptide spectrum from mass spectrometer

Figure 3: Approaches to enable variant peptide identification

Computational proteomics methods to increase identifications

Although technically not proteogenomics, computational proteomics also seeks to increase the accurate identification of peptides not found in the database. Computational proteomics encompasses computational methods such as *de novo* sequencing to increase sensitivity. Improvement in this area focuses on resolving ambiguities in acquired spectra that underpin the challenges in peptide identification.

Some of these spectral ambiguities are technical in nature, originating from MS/MS protocol/machinery. When using search databases to identify peptides, generated spectra are matched to theoretical spectra and the quality of the match is scored. It is straightforward to predict peptide sequences from full protein sequences using an *in silico* digest, as enzymes such as Trypsin digest amino acid sequences in a predictable manner[84]. Predicting the expected spectral peak heights and m/z values from any known peptide sequence is more complex. B- and y-ions are the most abundant products from fragmentation since the peptide bond is most easily broken, so most theoretical spectra are based on these. There are, however, many other fragmentation products that can be produced and contribute to the end peptide spectrum, which results in imperfect matches in the real world[85]. Spectral library searching is a method used to address this issue. It utilizes a collection of previously identified full spectra, resulting in a gain of sensitivity and selectivity. However, it is limited to previously-observed spectra and, similar to database searching, does not aid the discovery of novel peptides/proteins[86].

There are also biological explanations for spectral ambiguities; m/z values may deviate due to post-translational modifications (PTMs), for instance. PTMs such as phosphorylation are common in proteins and essential for protein regulation and cellular signaling[87]. Spectra from PTM-containing peptides may have slightly shifted peak(s) as compared to their non-PTM containing counterparts, and they are thought to make up a third of unassigned spectra[88]. There are over 1,500 different PTMs which can produce mass shifts on any given spectra[89]. Despite their prevalence and importance, many popular peptide search engines fail to accurately identify most of these PTMs[90]. This issue extends to amino acid substitutions; peptides with single amino acid substitutions originating from (potentially disease-informative) genomic variation such as SNPs can manifest into a mass shift indistinguishable from that from a PTM.

In computational proteomics, machine learning techniques have been instrumental efforts to resolve the ambiguity around PTM and variant mass shifts. Improvements of identification methods using such can help reduce the size of search space, such as in the so-called “open modification search” [91]. Sequence tag-based peptide identification is one such method that utilizes short stretches of *de novo* sequencing (tags) combined with sequence database search to allow for flexibility in identification by removing all candidates without the *de novo* sequenced tag from the database[92,93]. The mass differences between pairs of peptides with the same sequence tag can improve differentiation between PTMs. The use of additional information from the LC-MS/MS protocol itself can also be used to aid identification efforts. For example, retention time, or the time it takes for a peptide to be eluted from a chromatography column, is a measure that has been successfully used to identify false positive identifications[94,95]. Better modeling of the real-world effects of MS protocol leads to improved theoretical spectra with which to match. Peak intensity prediction for matching was improved by training an ML model on large spectral library sets, allowing for the benefits of spectral library matching on unseen peptides[96,97].

Peptide identification using informed sequence databases

The basis of all proteogenomics approaches involves the inclusion of DNA and/or RNA sequences and their predicted protein products, but the balance between the power of additional genomic information and the costs in proteomic detectability is a delicate one. In order to correctly identify a peptide, it must be present in the search database (completeness), but have an identification strategy that is both sensitive and specific. Unfortunately, the more sequences included in a search database, the higher the likelihood that the best scoring match is incorrect[98]. Using a more complete database can ironically result in fewer identifications than using a more limited consensus database[99,100]. Prudent usage of transcriptome and SNP information from the sample in question is advisable to harness the increasing availability of nucleotide sequencing data to inform peptide matches.

Short versus long read sequencing data in the search database

As the peptide search database is populated with RNA sequencing information, the quality of the methods used to assemble the transcriptome directly impacts the quality of the proteomic findings. Short-read sequencing is unable to confidently resolve full isoforms, particularly with more lowly expressed transcripts[101]. Therefore, complex computational methods are necessary to identify novel transcripts and their corresponding protein products using short-read data[102–104]. For example, splice-graph based databases are used to include all possible protein products of short-read-detected spice junctions in the search database. Such methods, however, are not ideal since the corresponding amino acid sequence from these splice junctions cannot be confirmed among the multiple possibilities. In practice, this means choosing between the enforcement of one reading frame that may not be correct[105] or the addition of many redundant entries in the database (6 frame translation) [106]. Long read RNA sequencing is better suited to aid discovery in the proteome by reducing the amount of noise in the search database as compared to short read sequencing. However, it is important to define what constitutes novelty in the transcriptome (see genomics sequencing section above). The use of a looser definition including *e.g.* non-canonical junctions and alternative 3/5' ends results in more 'novel' transcripts but likely more false positive transcripts and a larger search space if used in the peptide search database.

Open reading frame prediction

Novel transcripts discovered by long-read RNA sequencing must be converted into (predicted) protein sequences before being incorporated into the search database. There are several options for ORF prediction. The most comprehensive solution is a full six-frame translation (6FT), which creates six predicted protein sequences per transcript taking into account all possible frames of translation that could occur[107–109]. The correct ORF will be present in the six, but one or multiple incorrect amino acid fragments are also included, reducing specificity[99]. One simple way to reduce six frames to three is to include strand information. A few other common strategies to eliminate unlikely ORFs in 6FT include imposition of a minimum length threshold for the predicted protein product, selection based on homology to known coding sequences and the use of ORF prediction software[110–113]. Identification of certain nucleotide features that point to codingness has led to statistical models that output accurate ORF predictions[114–116]. CPAT is one such example that uses a logistic regression model to predict transcript codingness and likely ORF sequence using sequence features such as hexamer (6NT window) and codon usage biases.

Options for adding variants to search database

Ultimately, taking all the above considerations in account makes it more likely that the non-reference sequences that are added to the proteogenomic search database are accurate. Depending on the goals of the study, the search database can be further refined by including genetic variants (based on WGS/WES) or alternative isoforms from an individual. In cases where the variants of the individual are not known, it is still possible to identify them by incorporating known variants from popular databases such as dbSNP and COSMIC [117–119]. In the same vain, alternative isoforms can be inferred from publicly available Expressed Sequence Tags [120–122]. An attractive option to reduce the database size is to only include protein sequences that correspond to transcripts that were found to be expressed in the sample. It is however still common practice to also include all canonical proteins when adding variants to the search database, regardless of whether their transcripts were expressed because RNA sequencing in a typical experiment does not have full coverage.

Confidence in peptide identification

In peptide identification, pairings are made with the closest match between theoretical spectra derived from peptide sequences and observed spectra (peptide-spectrum matches, PSMs). The resemblance between the theoretical and observed spectrum is reflected in a score. As many PSMs are incorrect, a scoring system combined with a confidence threshold will help remove the majority of the false positive matched. False discovery rate (FDR) estimation is a commonly used procedure to help determine a statistically sound decision when removing false positives[123]. The most common method is the target-decoy model; reversed or shuffled peptide sequences called decoys are added alongside the true (target) peptides in the search database[124,125]. The underlying assumption is that the score distribution of decoy matches and incorrect target matches will be similar. The list of reported discoveries is sorted by score and filtered such that the proportion of decoy matches in the final list is less than a desired threshold, thus removing the lower-scoring incorrect target matches. FDR itself is estimated by dividing the number of decoy identifications by the number of target identifications above a certain threshold.

This method of FDR estimation has been widely accepted but is imperfect. Larger databases, which are common in proteogenomics, yield fewer peptide identifications at the same FDR threshold[126,127]. It can also unfavor certain subcategories of peptides, such as variant peptides[128]. Global FDR estimation fails in those settings due to the relatively smaller size and heterogeneous nature of identifications in this subcategory[98,129]. Other refined methods of calculating FDR for variant/subgroup peptides have been proposed to increase the number of identifications, but they have not been independently benchmarked[119,130]. Some studies attempt to circumvent the problem altogether using so-called multi-pass strategies whereby multiple searches are performed on subsets of the total search database[119,131,132]. They yield more identifications, but error rates are impossible to quantify, calling their validity into question [133].

Scope of this thesis

In this thesis, the power of long-read transcriptome sequencing is exploited to further understanding of transcriptomic and proteomic variation. This thesis explores the potential of novel proteogenomics methodology with long-read sequencing, suggests a way forward in the field, and provides an open-source tool of potential use in diagnostics. The thesis also aims to assess the gains of proteogenomics over standard proteomics to identify variant proteins. In Chapter 2, we examine in detail how and to what extent long-read transcriptome information can lead to discovery of protein variants using proteogenomics. An application of these methodologies to uncover novel biology was performed in Chapter 3, where long-read proteogenomics was applied to profile the host-pathogen interaction in response to multiple pathogenic stimuli. We study the use of alternative isoform usage in certain cellular conditions. A new method to leverage the increased transcriptional diversity (as a result of long-read sequencing) to re-annotate patient variants is described in Chapter 4. This new method, called SUsPECT, utilizes a more accurate set of transcripts and proteoforms to better estimate the impact of variants on protein function than using standard reference databases. The method may be useful in aiding diagnostics for patients with rare monogenic disease as long-read transcript data (and corresponding proteome data) becomes more widely available. I conclude with the applications, limitations and future outlook on the proteogenomics field in Chapter 5.

2

Chapter 2

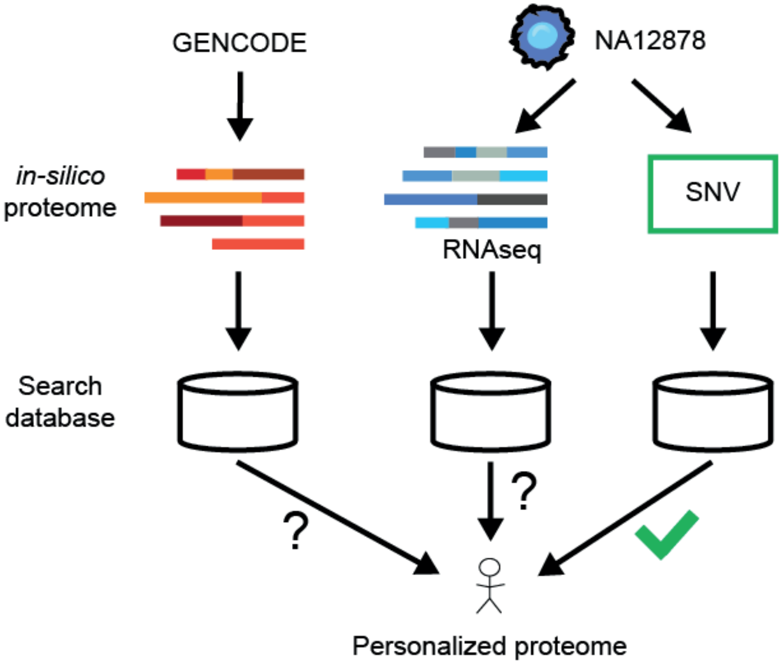
The Personalized Proteome: Comparing Proteogenomics and Open Variant Search Approaches for Single Amino Acid Variant Detection

Renee Salz, Robbin Bouwmeester, Ralf Gabriels, Sven Degroeve, Lennart Martens,
Pieter-Jan Volders, Peter A.C. ’t Hoen

Published in Journal of Proteome Research (2021)

Abstract

Discovery of variant peptides such as single amino acid variant (SAAV) in shotgun proteomics data is essential for personalized proteomics. Both the resolution of shotgun proteomics methods and the search engines have improved dramatically, allowing for confident identification of SAAV peptides. However, it is not yet known if these methods are truly successful in accurately identifying SAAV peptides without prior genomic information in the search database. We studied this in unprecedented detail by exploiting publicly available long-read RNA seq and shotgun proteomics data from the gold standard reference cell line NA12878. Searching spectra from this cell line with the state-of-the-art open modification search engine *ionbot*™ against carefully curated search databases resulted in 96.7% false positive SAAVs and an 85% lower true positive rate than searching with peptide search databases that incorporate prior genetic information. While adding genetic variants to the search database remains indispensable for correct peptide identification, inclusion of long-read RNA sequences in the search database contributes only 0.3% new peptide identifications. These findings reveal the differences in SAAV detection that result from various approaches, providing guidance to researchers studying SAAV peptides and developers of peptide spectrum identification tools.



Introduction

Proteomes display significant inter-individual variability [134,135] and personal proteomes may delineate disease risk and pave the way for personalized disease prevention and treatment. Personalized cancer treatment, for instance, is already instigated based on the detection of peptides containing single amino acid variants (SAAVs) that often serve as excellent biomarkers [136–141]. Detecting these SAAV peptides reliably, however, is a formidable challenge. Previously, scientists looked for protein evidence of a small number of variants in particular and resorted to targeted proteomics approaches such as selected reaction monitoring (SRM) [142–145]. Alternatively, BLAST-like query tools such as peptimapper and PepQuery [146,147] or database tools like XMAn v2 [148] and dbSAP [149] can be used to investigate single events [150,151]. Proteogenomics, the integration of genome and transcriptome information, is a more holistic and higher-throughput form of mass spectrometry- (MS-) based detection of variant peptides.

A main limiting factor of SAAV peptide (called 'variant peptide' for the remainder of the manuscript) detection with shotgun proteomics is the tandem mass spectrometry (MS/MS) technology itself. Since MS/MS spectra are generally too noisy to call a peptide sequence de novo, current MS/MS analysis methods rely on a database of known peptides. This limits the ability to detect unknown peptides such as variant peptides. The most flexible way to detect variant peptides is an exhaustive search; allowing any possible amino acid substitution at any position in the peptide sequence [152,153]. However, this strategy increases the search space immensely to a point where it is no longer useful in practice. The larger search space leads to ambiguity in peptide identification and thus a high number of false positive hits [154,155]. Therefore, more careful curation of sequences in the search database pays off.

Databases of peptides containing variants from dbSNP have been created to facilitate the search for SAAVs [149,156], and simply adding these variant peptides to the database showed promise early on [156,157]. Not all dbSNP variants however, are expected to be found in every sample, and including them all may lead to false identifications [106]. In addition, rare and unique variants may be overlooked. A proteogenomics approach where only those variant peptides predicted from genome or transcriptome information are added to the peptide search databases, can improve their detection. Proteogenomics pipelines have streamlined this process of incorporating personal genome information into a proteomic search database [158–162]. In addition, there is evidence that including correct sequence variant information, including often-overlooked sample-specific indels and frameshifts, improves variant peptide identification workflows [163]. Yet, false discovery rate (FDR) correction is needed to compensate for the increase of database size and complexity [154,155]. When searching for evidence of specific peptides such as variant peptides, an additional subset specific FDR correction should be made [164].

In addition to SAAVs, alternative splicing may also introduce sample specific peptides. Alternative splicing is commonplace as 90% of genes undergo alternative splicing [165]. Since protein reference databases do not cover all protein isoforms produced by alternative splicing, sample-specific transcriptome information is advantageous. Typically, the information on alternatively spliced sequences comes from RNA sequencing. Short-read RNA seq is however not ideal for properly capturing the complete splicing patterns and the resulting open reading frames (ORFs). Traditionally, this is circumvented by including 3- or 6-frame translations of the sample's transcriptome. However, this approach was found to expand the database far too much for eukaryotic organisms, leaving few remaining hits after FDR correction [166]. Studies utilizing long-read RNA seq frequently discover previously unannotated transcript structures. Thus, full-length transcripts may add essential information for correct ORF prediction and peptide identification.

An emerging alternative to proteogenomics methods for the detection of variant peptides is the 'open search' method. This allows unexpected post-translational modifications and amino acid substitutions in the peptide spectrum match, while maintaining accurate FDR and a workable computation time. Using sequence tag-based approaches, the search space is narrowed with de novo sequence tags, which makes room for the addition of all possible SAAV peptides in the search space [167–171]. These methods were historically not as effective as classical proteogenomics searches in finding variant peptides, since there is difficulty in discerning between post-translationally modified and SAAV peptides. However, this situation has recently improved with the inclusion of optimized probabilistic models [172]. One implementation of the tag-based method improved with such models is *ionbotTM* (manuscript in preparation; compomics.com/ionbot), which is a machine learning search engine that uses MS2PIP [173] and ReSCore [174] to significantly improve the accuracy of peptide match scoring.

The main objective of this study is to compare a previously established proteogenomics approach based on long-read sequencing with a recently-developed open search method for the detection of true variant peptides. In simpler terms, we compare a genome-informed search space with typical spectrum identification settings to a genome-uninformed search space with advanced identification settings. We aim to understand the power of, and potential biases associated with, using an open search method without prior information about the genome. For this, we make use of high-confidence nucleotide sequencing and (ultra)-deep proteomics data from a gold standard cell line NA12878. Using correct ORFs from the long-read transcriptome and high-confidence phased variants belonging to this cell line, we gain a unique perspective on exactly what advantages can be gained by each approach.

Experimental section

NA12878 Data sources

Variant information was obtained from Illumina platinum genomes (ftp://platgene_ro@ussd-ftp.illumina.com/2017-1.0/hg38/small_variants/NA12878/). The reference genome used was GRCh38, which can be downloaded from the pre-computed 1000 genomes GRCh38 BWA database at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/ (with decoys). Transcript structures for NA12878 were sourced from the ONT consortium [175]. In the consortium, Workman et al sequenced 9.9 million reads corresponding to 33,894 transcripts and 20,289 genes. The reference transcriptome and proteome are from GENCODE v29.

Shotgun proteomics data came from the [176] study, downloaded from Peptide Atlas (http://www.peptideatlas.org/PASS/PASS00230). This dataset consists of 417 TMT6plex runs from 54 samples, with the reference tag (126.77) on NA12878 in every case.

Creation of the search databases

In total, four search databases were created (see Table 1 below). 1) Database based on ONT transcriptome sequences only (referred to as 'ONT'), 2) database based on GENCODE coding transcriptome only (referred to as 'Ref'), 3) a database that is the union of 1) and 2) and contains no NA12878 specific variants (referred to as variant-free or VF), and 4) the same sequences as database 3), but contains NA12878 specific variants (referred to as variant-containing or VC). A simple depiction can be found in Figure 1A and Table S1, while the detailed full workflow can be found in Figure S1. Each database had MaxQuant [177] contaminant sequences appended before search.

The Ref search database was made by filtering GENCODE v29 predicted ORFs for those that were complete (no 5' or 3' missingness). The ONT database was created using transcript structures provided by the NA12878 consortium (https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md). The coordinates in the junction file (PSL format) provided were converted to BED with BEDOPS [178] and used to fetch the corresponding stretch of sequence from the GRCh38 genome with bedtools [179] getfasta. The exons were assembled using in-house scripts to form the full transcripts, and those that were non-identical to transcript sequences in GENCODE (“novel”) were then submitted to 2 ORF prediction software; ANGEL v2.4 (“dumb” ORF prediction on default settings) and SQANTI2 v2.7(https://github.com/Magdoll/SQANTI2). The translation of transcripts predicted by both prediction programs were added to the search database. ORFs from GENCODE were used for transcript sequences in ONT identical to transcript sequences in GENCODE.

The VF database was simply the union of the Ref and ONT databases. The VC database was created by first creating full-length coding sequences (CDS) with variants included by replacing reference nucleotides according to the VCF file per CDS fragment, for every CDS fragment. If only homozygous variant(s) were present in a CDS fragment, only one variant CDS fragment was generated. If a CDS contained at least one heterozygous variant, two variant CDS sequences were generated corresponding to the different alleles. Fragments were then assembled to full CDS. If a full CDS contained at least one CDS fragment with a heterozygous variant, two full CDS were generated corresponding to each allele. For those full CDS that contained at least one variant, the variant version(s) of the sequences replaced the non-variant versions in the VF database to create the VC database.

Spectral search and post processing

Each run from Wu et al 2003 was first converted to the Mascot Generic Format(MGF) format using msconvert [180] with MS2 peak picking enabled. Each dataset was then searched against the four search databases described in the previous section, using *ionbot*™ version 0.5. Fixed and variable modifications were set according to the protocol in Wu et al. Open modification settings were enabled for all four runs, while open variant settings (for SAAV detection) were enabled for all runs except for on the VC database. Searches allowed for up to two missed cleavages. When parsing the search results, only spectra with an observed TMT6plex reporter ion 126.77 (corresponds to cell line NA12878) were retained.

Since sub-setting PSMs into groups such as variant peptides requires separate FDR correction [164], both VC and VF underwent a separate FDR correction for the variant peptide subset. Successful FDR correction requires the modeling of potential false positive peptide identifications using appropriate decoy peptides. In the case of variant peptides, this means a sufficient number of decoy variant peptide identifications must be present to accurately model the population of false positive peptides. Reversed sequences thus underwent the same processing steps as the true sequences in order to create the appropriate decoys. The distributions were checked for successful modeling (Figure S2).

A variant peptide list was created to compare with *ionbot*™ identifications from searches of the VC and VF. The list was created with an in-house Python script that performs an in-silico trypsin digest (allowing for up to 2 missed cleavages) with the pyteomics v 4.2 [181] package and checks per protein for peptides that differ by only one amino acid between the VF and VC database. I and L were treated as identical, and a potential variant peptide was disqualified if it appears in any other reference protein sequence.

ionbot™ identifications presumed to be variant peptides (and variant peptide decoys) underwent subset-specific FDR correction for both combination databases, but the exact subset of variant peptides differed between the two searches due to different assumptions. The assumption in the VF database is that variants in the genome are unknown, so all predicted variant peptides (and predicted variant decoy peptides) were pooled for FDR correction. In the VC database, only known variant peptides (and corresponding decoy peptides) are pooled for FDR correction. We expect the different approaches to subset FDR to be comparable, as *ionbot*™ does not include duplicate peptides in the search database. This means that the databases being compared are of similar size on the peptide level, which is the level at which the FDR correction is performed. Q value calculation and cutoff ($q < 0.01$) were performed with an in-house python script (distribution can be seen in Figure S2). Retention time predictions were calculated with DeepLC [182]. All scripts referred to in this manuscript can be found in the GitHub repository (https://github.com/cmbi/NA12878-saav-detection).

Results

Search database makeup

The main goal of this study is to evaluate the added value of transcriptomics data for SAAV identification in proteomics data. In this evaluation, SAAV identification with and without transcriptomics prior knowledge is compared for a state-of-the-art open search engine. To this end, we searched the NA12878 deep shotgun proteomics data set with four distinct search databases corresponding to two comparisons, as outlined in Figure 1A. The first comparison was between databases based on the Oxford Nanopore (ONT) long-read transcriptome, the GENCODE reference proteome (referred to as Ref) or the combination of the two (referred to as combi, Figure 1B). In this comparison, all searches were run with open modification settings that allow for one mutation in the peptide match. The second comparison was between a regular and an open variant search using databases that did and did not include NA12878 genome sequencing-derived variants, respectively. This comparison was performed for the combi databases only. The analysis with the variant-free combi database will be referred to as the VF method and the analysis with the variant-containing combi database will be referred to as the VC method. In this comparison, open modification search was enabled for both methods, but open variant search was only enabled in the VF method to allow for the detection of SAAVs. Open variant search is disabled in the VC method, because the variants were already incorporated in the VC search database.

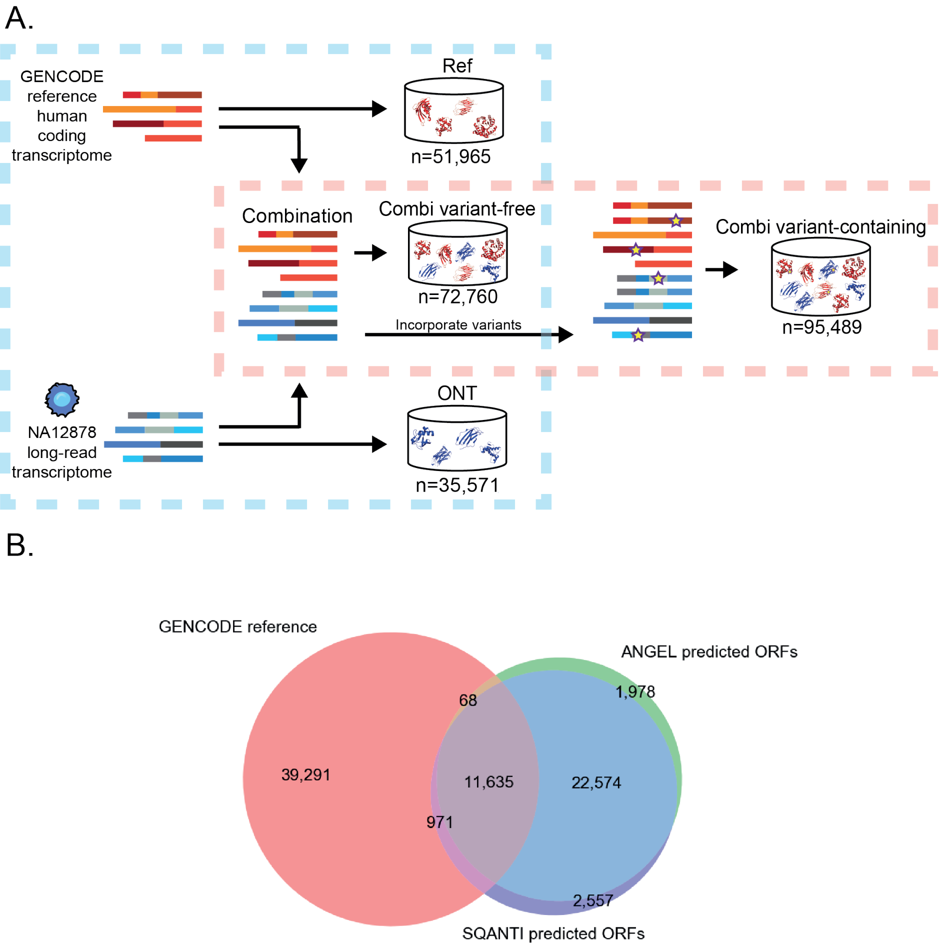


Figure 1: Creation of the search databases**.** **(A)** Three databases were made to make comparison between use of different sources of sequences. One with only translations of transcriptome sequences (ONT), one with only the reference proteome (GENCODE), and one with the union of the two. This comparison is denoted with a blue square. Variants from NA12878 were incorporated into the combination database from A and compared to the combination database without variants. This comparison is denoted with a red square. (B) The number of (predicted) ORF in the different sources used to construct the VF search database and their overlap. The sources included the GENCODE v29 reference ORFs and the predicted ORFs from ONT RNAseq. Two ORF prediction softwares (ANGEL and SQANTI) were used to determine candidate ORFs and the intersection was included in the final search database.

Adding the long-read transcriptome for the cell line does not contribute to additional peptide identifications in practice

Reliable peptide identification normally requires a comprehensive search database. We first investigated whether novel transcripts from long-read transcriptome sequencing would contribute to peptide identifications in the NA12878 shotgun proteomics data. The ONT database contained 35,248 full-length transcript sequences, 64% of which were novel. Although the combi database containing these novel predicted ORFs was 42% larger than the Ref database (Figure 1B), the number of unique peptides from these sequences made up a mere 2.3% of the search database (Figure 2, top panel). The addition of ONT-derived ORFs to the Ref ORFs thus translated to an only modest increase in the number of unique peptides in the search database (Figure 2, lower panel). A likely explanation for this is the fact that many of the novel ONT transcripts demonstrate high similarity to existing reference sequences. The sequences usually only differed in the length of the 3’ or 5’ UTR or the in use of alternative exon junctions rather than completely novel exons. The exact frequencies of these events are difficult to estimate, but when looking at the set of novel ORFs from the ONT transcriptome, 73% of them can be attributed to known GENCODE coding genes. Conversely, the GENCODE genes that had novel isoforms in the ONT set corresponded to 27% of all GENCODE coding genes. In terms of observed peptide identifications, 67% of the ORFs in ONT set had at least one peptide match (when including PSMs that also matched to peptides present in GENCODE). However, the number of unique peptide matches to the novel ONT transcripts was much smaller: only 0.3% of unique peptides identified to the combi databases mapped exclusively to novel ONT transcripts. This indicates that the transcriptome database does not contribute significantly to the proteomic search results and suggests that alternative splicing and mRNA processing events do not contribute much to the diversity of the MS-detectable fraction of the proteome.

Aside from the contributions from the ONT-only sequences, it is also interesting to investigate protein identifications that were not found in the ONT transcriptome. While these should theoretically not be present, roughly 20% of identified peptides are exclusively matched with the ENCODE transcripts (Figure 2). As expected, this percentage is smaller than the 42% of peptides in the search database that are exclusive to GENCODE transcripts, but still a significant fraction. This suggests that it is best to still use a reference transcript database, even if there is full transcriptome sequencing data available.

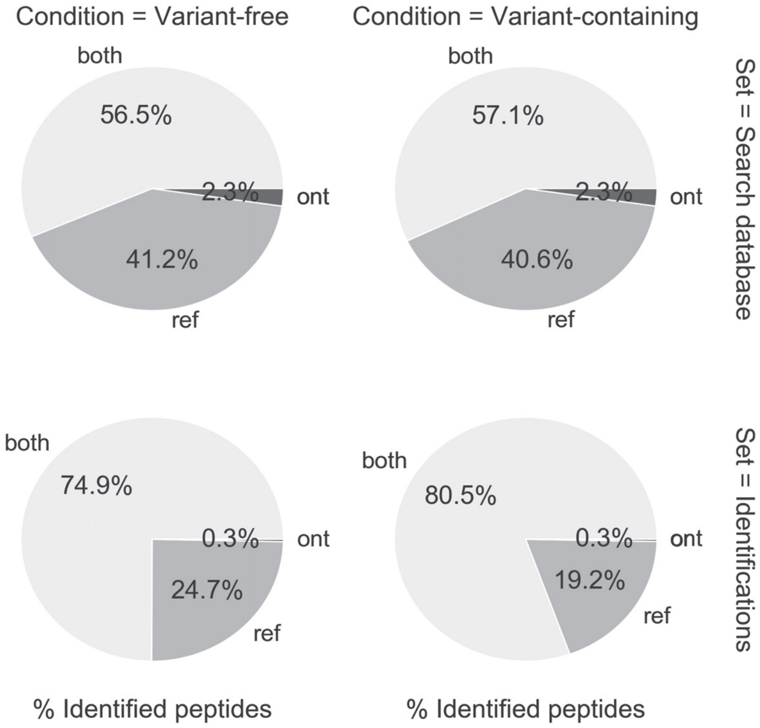
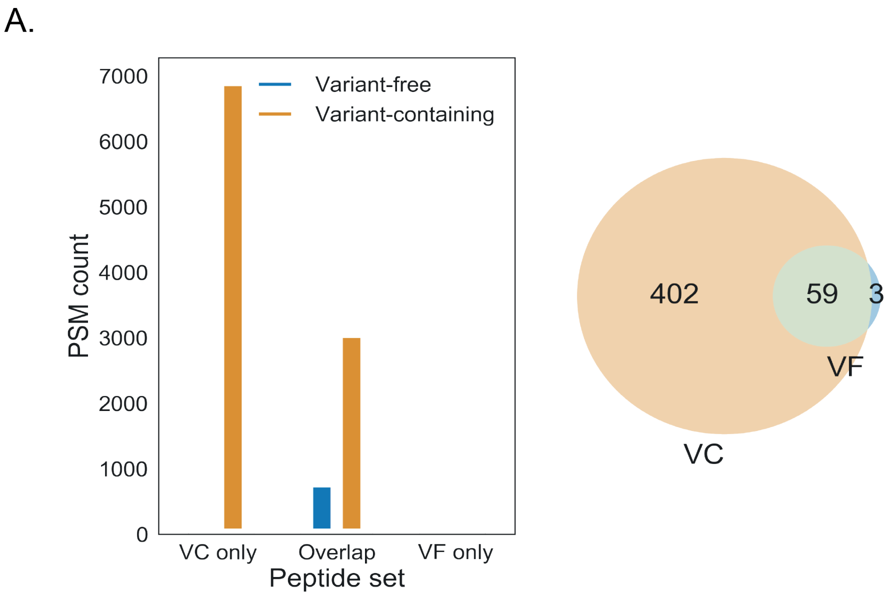


Figure 2: Detectible peptides per method. Theoretical (upper pie charts) and observed (lower pie charts) proportions of peptides when searching against VC (right) or VF (left) search databases. This shows percentages of matched peptides attributed only to GENCODE proteins, only ONT proteins, and those that match to proteins in both databases.

Variant-containing method allows detection of many more genome-supported variant peptides

We subsequently studied the effect of the inclusion of sample-specific variants in the search database. In the VF method, the data is analyzed with an open variant search, thus letting the search engine predict single amino acid substitutions. This is in contrast to the VC method, where no variants are predicted and only genetically supported variant are present in the search database. We detected 461 variant peptides by the VC method and 62 by the VF method, with 59 overlapping between the two methods (Figure 3A). The greater majority of variant peptides that were detected by the VF method only (n=1,805), were not supported by the genome and are likely false positives (Figure 3B). In addition, one third of variant peptide matches that appeared to be supported by the genome, actually contained an incorrect amino acid substitution. Thus, the inclusion of variant peptides derived from personal genomes in search databases is far superior to the use of a variant free database combined with an open variant search. Some examples of identified variant peptides can be found in Figure S3.



B.

	VC	VF genome-unsupported	VF genome-supported	VF genome-supported and correctly identified SAAV
PSM	10,015	41,340	807	592
Peptide	461	1,805	61	50

Figure 3: Detection of variant peptides using (combination) VF and VC databases.
(A) variant PSMs (left) and unique peptides (right) attributed to genome-supported variant peptides. (B) PSM and peptide counts found by each method.

Detectible variant peptides have attributes that differ from expected variant peptides

Out of the 34,968 peptides in the genome-supported variant peptide list, only 462 were detected by either or both the VC and VF methods (Figure 4A). They are not a random sample of all possible variant peptides. Namely, some variant peptides are easier to detect than others depending on their abundance and/or properties, and that differs even between methods. For instance, the VF method tends to find longer variant peptides (in a range of 16-27 aa) and misses the shorter variant peptides (Figure

4B). This highlights the larger amount of ambiguity in variant peptide identification proportional to the lower number of peaks in the spectra. The VC method does not suffer from this ambiguity and allows for detection of a wider range of variant peptide lengths than variant-free, especially shorter variant peptides (p=0.0017 K2 samp). While there is a bias in variant peptide length, we did not find clear evidence that the position of the variant within the peptide affects detection of the variant peptide in either of the methods. In addition, the amino acid substitution itself affects detectability, since the corresponding mass shift in the MS/MS spectrum needs to be separated from noise or similar mass shifts corresponding to other modifications in order to be identified. There are some predefined limitations to SAAV detection with the VF method that lead to certain amino acid substitutions getting detected less than expected (Figure 4C). Amino acids on which there are fixed modifications can't have variants in the open variant search, meaning substitutions at K and C are not detected. Substitutions affecting the trypsin digest, such as those involving R, can also not be detected.

Erroneous variant peptide identifications are difficult to discern from true variant peptide identifications

The misidentifications from the open variant-free approach can be separated into false negatives and false positives. False negative identification is where the VC method identifies variant peptides, but those same spectra are identified by the VF method as non-variant peptides. False positive misidentification is where the VF method identified variant peptides that were not supported by the genome.

There were 402 unique false negative peptides observed (Figure 5A). These false negatives peptides were classified as variant peptides by VC method but not by VF, although they were contained in the VF search space. Identifying causes of false negatives requires investigation of how the VC peptides were identified with the VF method. There was no particular length peptide that was mis-identified more than others in general, despite the difference in detectible peptide length (Figure S4). The peptide identifications were similar between the VF and VC methods. In general, length correlated highly between the identifications of the two methods (R^2=0.9071, p=0). When comparing individual peptide identifications per method for mismatches and length difference, the largest source of error was a 1 aa length difference. Nonvariant peptides with a 1 aa length difference from the variant peptide were being identified instead of the correct variant peptide in >30% of the false negatives (Figure S4).

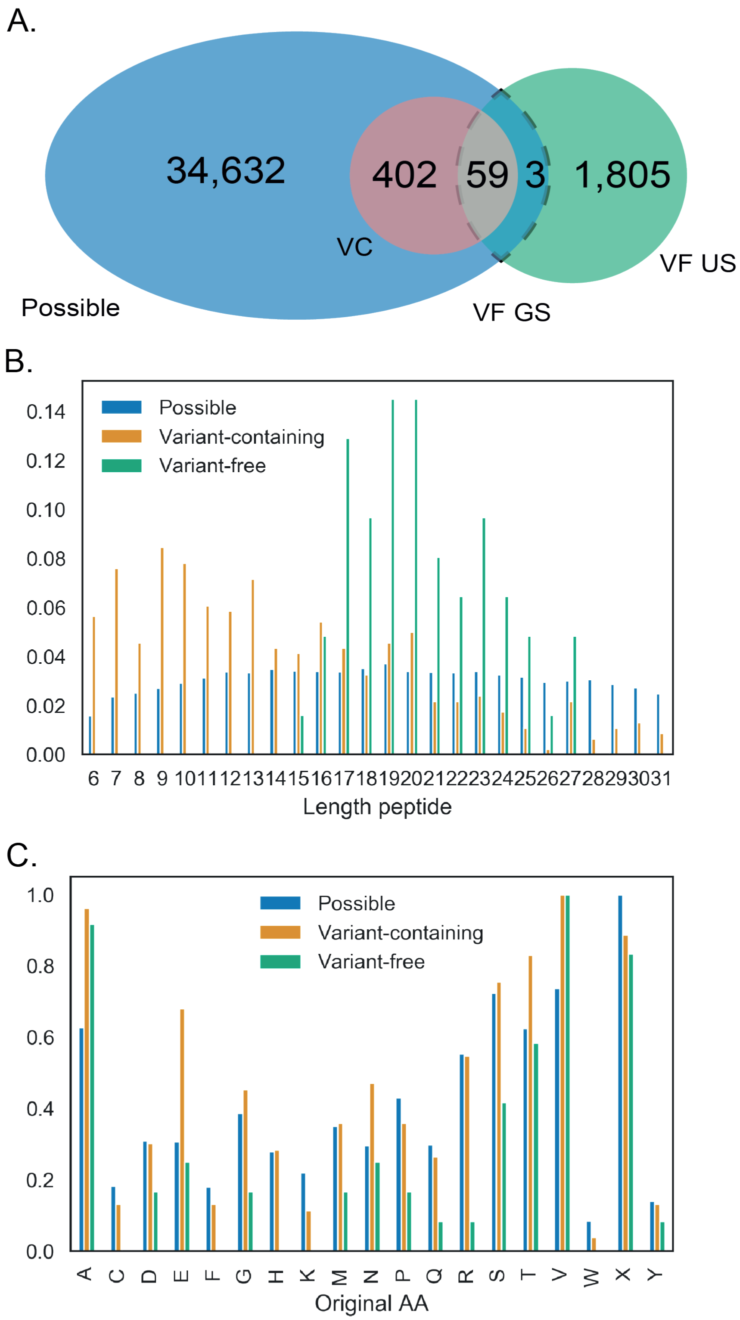


Figure 4. Properties of detected variants compared to expected. (A) The groups of variant peptides being compared. Each circle, including all overlaps, are being compared to each other. (B) Length distribution differences between detected variant peptides by the different variant detection methods. (C) Normalized (divided by max) frequency of variation per original (reference) amino acid.

Another possible source of false negative errors that was investigated is SAAVs being mistaken for unexpected post-translational modifications. In the false negative set, this did not appear to be an issue. The false negative VF identifications had approximately the same rate of unexpected PTMs (Figure S4).

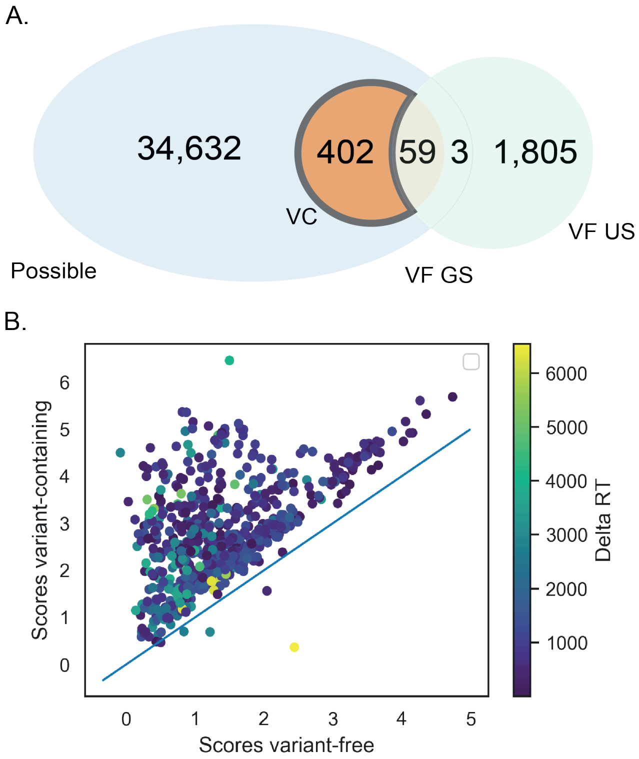


Figure 5: False negative variant misidentifications. (A) Investigation of causes of mis-identification of peptides in the variant-free set. (B) Scores of those mis-identified peptides in VF vs VC set. Each point corresponds to one false negative variant peptide. Percolator PSM score is used. Color corresponds to delta retention time.

To further understand how false negatives could occur, we compared the peptide matching scores of the false negative spectra for the VF and VC search methods (Figure 5B). Higher scores indicate higher confidence in assignment of spectra. VC scores for false negative peptides were generally higher than the VF scores (mean score ratio VC/VF = 1.31). However, a large fraction of the false negatives received comparable scores in the VC and VF search methods. This could indicate a ranking problem: the variant peptide received a score equal to another peptide, to which the peptide spectrum was ultimately assigned. Delta retention time can often be a useful independent validator when score disagrees between the different search methods. Despite high retention time discrepancies in this particular data set, observed retention time aligns relatively well with predicted retention time for those spectra that received higher scores in VC.

The genome-supported variants are a tiny fraction of the high confidence variant peptide predictions from the VF database, indicating a high false positive rate (Figure 6A). We investigated whether there are distinguishing features between genome-supported and genome-unsupported variants. Reassuringly, scores of true positives were slightly higher than false positives (Figure 6B, p=1.34e-26, ANOVA).

A closer inspection of genome-unsupported variants reveals potential sources of confusion for variant prediction algorithms, leading to false positive identifications. There was a high level of concordance of peptides matched to these spectra in general. Two thirds of spectra that corresponded to genome-unsupported variant peptide identifications by VF had the same base peptide identifications in both the VF and VC searches. Mass shifts predicted to be SAAV in VF were commonly predicted to be ‘unexpected’ PTMs by the VC method (Figure 6C). A common PTM mistaken as a SAAV in VF was threonine oxidation, but many PTMs contributed to this mix-up. There was no clear trend to the identification errors, underlining the difficulty of correctly classifying minor mass shifts corresponding to PTMs and SAAVs.

Evaluation of the variant peptides’ SNPs of origin

The detection of variant peptides is ultimately a means to understanding which single nucleotide variants (SNVs) are expressed on the protein level. By incorporating SNVs into predicted ORFs, we ended up with a theoretical set of 34,968 variant peptides originating from 9,298 SNVs from all chromosomes, of which 5,989 are heterozygous variants.

In the case of a heterozygous variant, both variant peptides and their reference counterparts can be identified in some ratio. A ratio different from 0.5 may be indicative of preferred expression of one of the alleles on the protein level, otherwise known as ASPE (allele specific protein expression). Presence and magnitude of ASPE is potentially key information that can be used to understand biological mechanisms. However, technical biases of search methodology may invalidate potential findings by distorting these ratios. For the VF method, the reference peptide was identified more frequently than the variant peptide (p=0.013, one-way ANOVA). The opposite was true for the VC method.

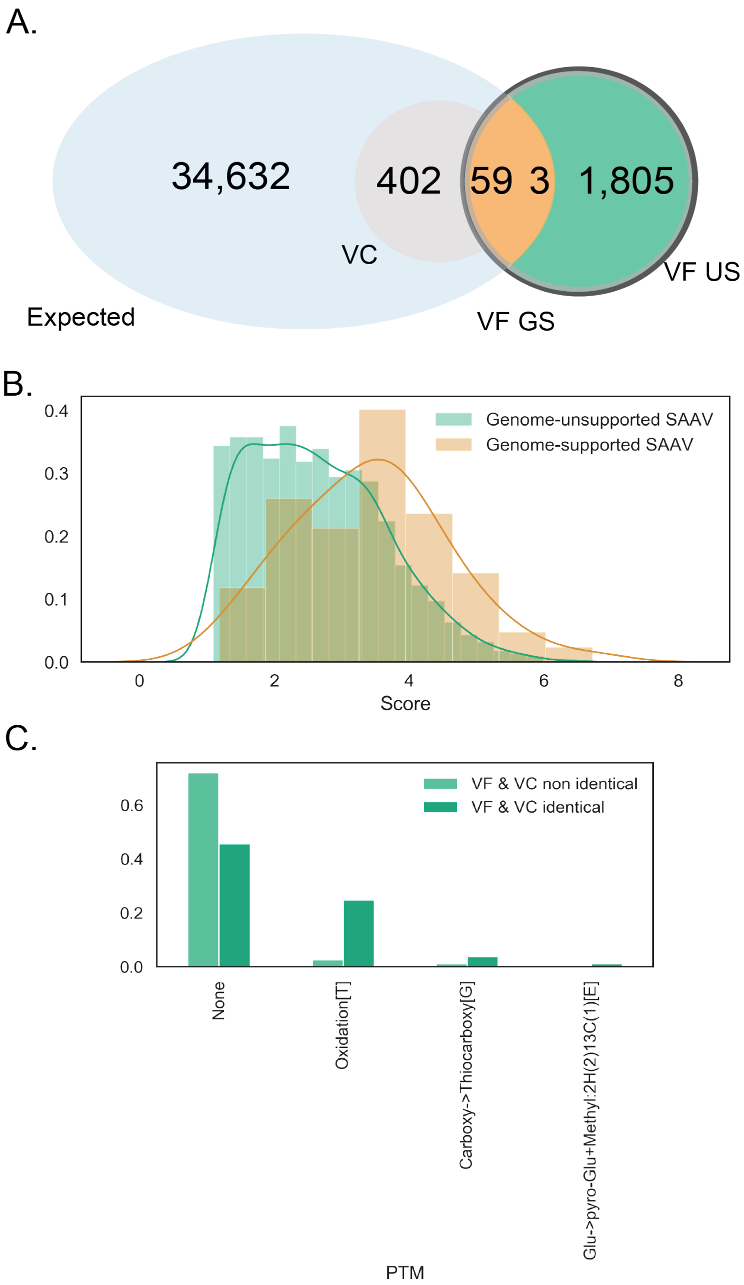


Figure 6: False positive misidentifications. (A) False positive misidentifications are genome-unsupported (US) variants predicted by the variant-free method (VF). The venn diagram highlights the subset of variants that are being investigated in this figure. These 2,998 variants were predicted by *ionbot*™ to be variant peptides, but were not found with the variant containing set. All but 7 were variants unsupported by genome information. (B) Relative score distributions between genome supported vs unsupported variants in the variant-free set. (C) Unexpected modifications by the VC set corresponding to all 'false positive' predicted variant PSMs in the VF set.

Homozygous variants can be used as a type of control to understand the bias in search methods, since we know that only one of the two alleles can be expressed. In case of homozygous variants, the variant peptide is expected to be present in all cases – with no reference counterpart. This was observed for the VC but not for the VF method (Figure 7A). Thus, without prior information about zygosity, the VF method tends to be conservative in identifying SAAV peptides, resulting in a higher likelihood of the reference peptide than its variant counterpart.

It is evident that some variant peptides were observed much more often than their reference counterparts or vice versa. The VC heterozygous variant peptide identifications should not suffer from the technical reference bias and allow for detection of allele specific expression on the protein level. The VC-detected heterozygous variants were divided in two groups; one group with more counts for the reference peptide (reference-biased, N=78) and one group with more counts for the alternative peptide (alternative-biased, N=123). The two groups demonstrated a clear and significant difference in population allele frequency (p=6.45e-08. Figure 7B). Those with lower allele frequencies displayed a stronger reference bias. This could be explained by the fact that rare variants in coding regions have a higher likelihood of causing undesirable effects on the resulting protein. Any deleterious effects resulting from the variant on protein stability would be visible as depletion of the alternative allele.

One significant subgroup of heterozygous variants was particularly biased towards the alternative allele. Fourty-four out of 183 variant peptides supported by more than two PSMs did not have any detected reference counterparts. One third of these variants had a substitution involving arginine or lysine (tryptic cleavage sites). One gene, HLA-DBQ1, had two alternative alleles instead of one reference and one alternative. In general, the score distribution for these highly biased group was lower than the score distribution for all VC detected variant peptides. The allele frequencies of this group were not different to those of the overall alternative-biased group (p=0.5, ANOVA). There was also no correlation between the RNA expression of these genes to the variant peptide expression (R^2=0.01). Also, a comparison the list of genes displaying ASE on the RNA level from [175] to the heterozygous genes with variant peptides detected on the protein level yielded negligible overlap (2 genes).

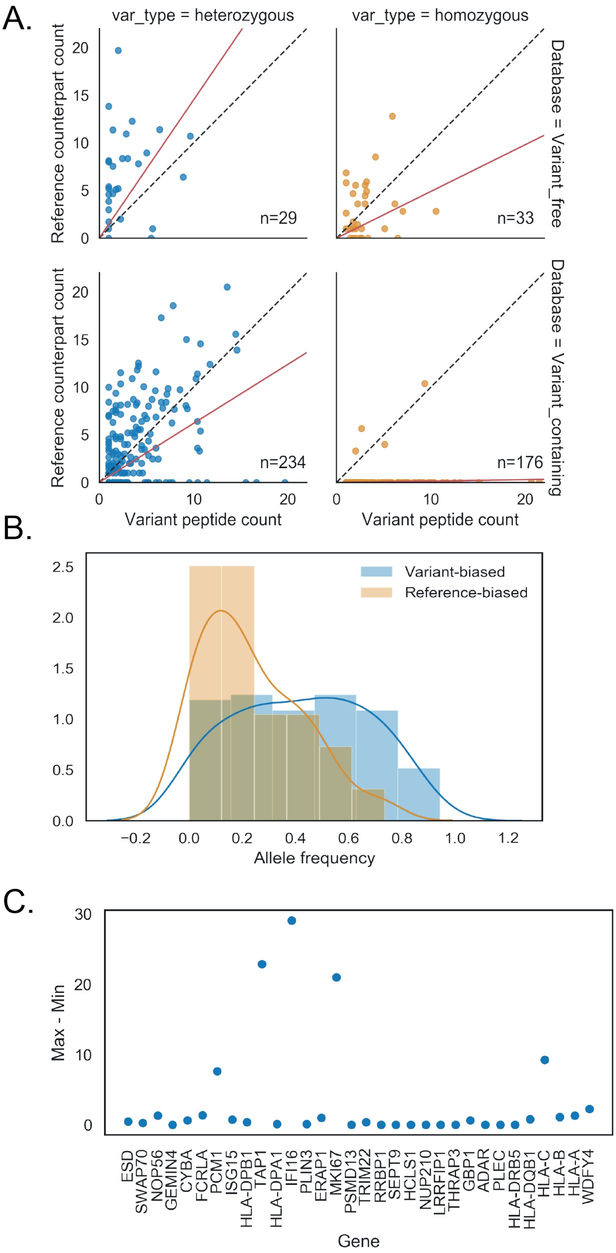


Figure 7: Underlying SNPs detected on the protein level. (A) Variant peptide abundance vs reference counterpart split by zygosity and search database, square root transformed. (B) Separating heterozygous variants in the variant-containing database by whether more variant peptide was found (variant-biased) or more of the reference counterpart was found (reference-biased) revealed differences in allele frequency distributions. (C) Ratio variability of genes with 2 or more variant peptides. Ratio is defined by the variant counterpart abundance divided by variant peptide abundance. Y axis shows max – min per gene.

A total of 33 genes were detected through two or more unique variant peptides. For variant peptides within a gene, the reference peptide to variant peptide ratio should be consistent, unless there are different protein isoforms as a consequence of alternative splicing. This was the case for the majority of genes with multiple variant peptides belonging to the same gene (Figure 7C). Five of these genes were represented by multiple variant peptides with inconsistent ratios. HLA-C, IFI16 and MKI67 had peptides matching to non-identical (sets of) isoforms within the gene. PCM1 had peptides matched to 24 isoforms. That is four times the average number of isoforms matched by a variant peptide in the VC search. Thus, inconsistent variant to reference peptide ratios within a gene can generally by attributed to differing abundances of protein isoforms.

Discussion

Here, we have carried out an investigation of the effects of proteogenomic additions to a proteomics search database. To this end, we compared a typical proteomics approach to a purely proteomics method utilizing state-of-the-art open search. We observed that the addition of transcriptomic sequences to the search database did not have significant effects on the overall peptide identification rate. There was a roughly equal number PSMs from the three databases, despite the long-read transcriptome search database being 40% smaller than that of the union of it and the reference. At the same time, the matches to reference-only sequences in the combination database imply that >20% of peptide identifications are missed. This suggests a large portion of false identifications when using a database comprised of only ONT sequences.

The fact that around a quarter of peptide identifications cannot be attributed to the transcriptomics data is rather surprising. There are a couple possible explanations. Using transcriptomics data from different cells than the proteomics data (different labs and different year) will unavoidably cause some discrepancies [183]. This could also be attributed to protein stability in the cell, as proteins are detectable for some time after RNA have already been degraded [184]. Also notable is the fact that including the transcriptome sequences did not seem to add significantly to the peptide detections; the proportion of novel peptides found was lower than the proportion of novel transcripts found. As this cell line/organism is so well studied, it is likely that the vast majority of present proteins have already been characterized. For other cell types and organisms with more novel transcripts, adding (full length) transcriptomes may lead to more peptide identifications.

Two different search methods were used to identify non-reference peptides derived from SNVs: a proteogenomics approach, in which all variants known from the genome sequence were added to the search database, and an ‘open variant search’, where only reference peptides were included in the search database and one amino acid differences were allowed by the search engine. The proteogenomics approach was clearly superior, as it detected 7-times more variant peptides , whereas the open variant search suffered from many false positive identifications that were not supported by the genome sequence, and from large numbers of false negatives. Nevertheless, also the proteogenomics search method detected only a minor fraction of the variant peptides predicted to be present in the genome. It has been estimated before that maximum ~70% of variants in protein coding regions are theoretically detectible in an ideal shotgun proteomics experiment considering peptide lengths 7-40 aa [185]. The number of variants found with a proteogenomics method in practice is much lower, depending on method details. Some studies either use a statistically dubious 'multi-tier' method [186,187] or skip FDR sub-setting altogether [188] and report the number of variants detected to be in the region of 10%. We detect only 1% of the theoretically present variant peptides, despite the ~4M spectra present in this dataset, making it one of the deepest proteomics datasets currently available. This is partly due to the careful control of FDRs in our study. Also other conservative efforts to detect variant peptides using FDR sub-setting or targeted proteomics validation detect <1% of all theoretically present variant peptides [157,187,189].

While open search lags behind the proteogenomics approach for the moment, it has promise. Algorithms are being continuously improved to better differentiate signal from noise, which will reduce the false positives and false negatives in variant peptide detection [190]. There are several upcoming methodologies to further refine the open search to increase accuracy, either adding to existing peptide identification tools or standalone with promising results such as Open-pfind [191], TagGraph [172], MSFragger [192], Crystal-C [193]. There are considerable challenges still to face in their detection, particularly in noise/signal differentiation. This is especially complicated as variants often co-occur with other PTMs such as phosphorylation [163,187]. Current detection methods including *ionbot*™ cannot handle the complexity of two modifications on one site. However, deep neural networks show great promise with difficult peptide identifications [194]. Using methods of machine learning along with orthogonal information such as peptide retention time should result in significant improvements in open search [195]. This in combination with rapidly improving data-independent acquisition removes detection limitations of low-abundance or otherwise difficult to detect peptides [196], which is currently a considerable hurdle in SAAV peptide detection [188]. Including open search is clearly useful and bound to get more accurate. This study used *ionbot*™ as the sole predictor of unexpected modifications/SAAVs, and comparison between identification tools was difficult as no other identification software tested reported the precise reporter ions per matched spectra (to be able to separate TMT tags corresponding to different cell lines). A study to compare methods given these updates is certainly warranted and ensemble methods may eventually be used to even more accurately predict these unexpected modifications/SAAVs.

One important implication of correctly detecting SAAVs is the ability to observe allele specific expression on the protein level. A targeted proteomics approach has recently been described to study ASPE (allele specific protein expression) with high confidence [197]. It found no correlation between RNA and protein level ASE for the few variants studied, highlighting the utility of having higher throughput methods to study this phenomenon. One simple way to measure ASPE when using a proteogenomics approach is by comparing the spectral counts for the SAAV and its reference counterpart, since a reference counterpart usually has equal detectability by MS/MS [185]. Here we found low correlation between the abundance of the variant and reference counterparts regardless of VF or VC method. This is potentially indicative for a high level of ASPE. In contrast, [187] demonstrated a high correlation between variant and reference peptides. This may be attributed to the low stringency associated with using the multi-tier search strategy for SAAV detection. We found no correlation between ASE and ASPE was found in this study which is consistent with the findings of Shi et al.

Conclusions

Our study provides guidance for the detection of variant peptides that shape the personal proteome. While personal genomes currently seem indispensable for the characterization of personal proteomes, new computational and analytical tools and new file formats to accommodate personal proteome information will allow us to get the fullest picture possible of the individual proteome, even without personal genome information.

Acknowledgements

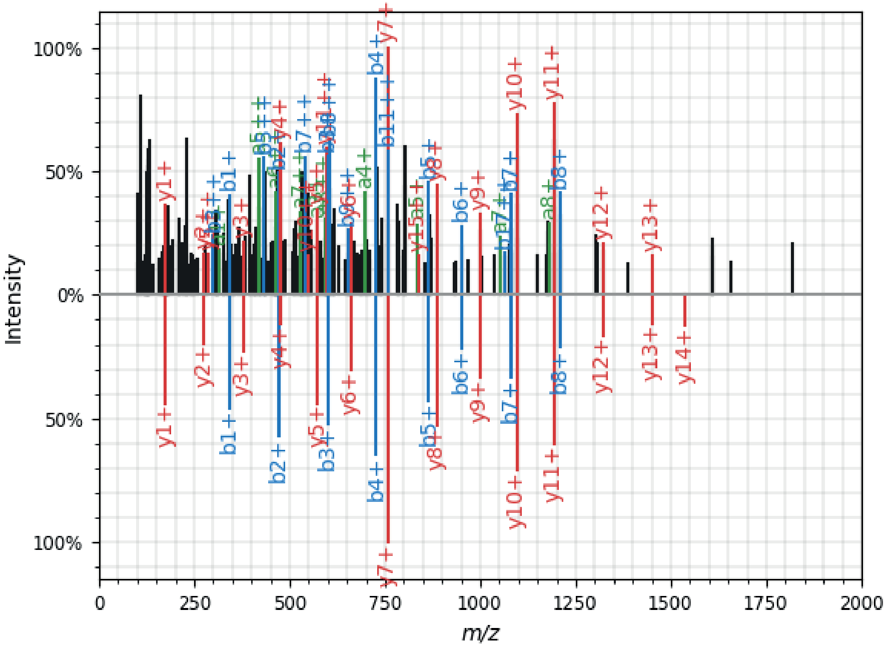
This work is funded by the Radboud Universitair Medisch Centrum.

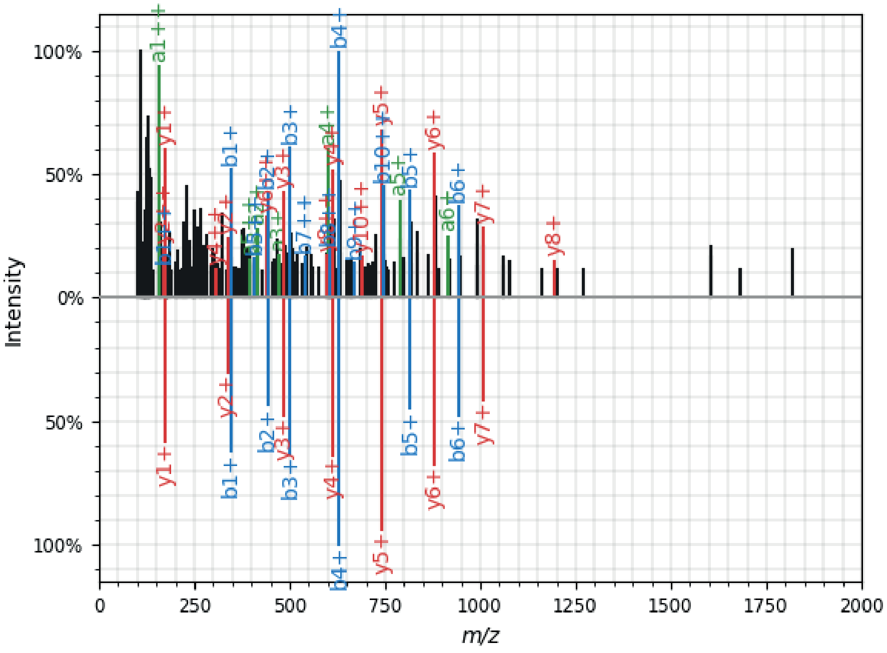
Authors' contributions

RS wrote the analysis scripts, performed the statistical analysis and drafted the manuscript. LM and P'tH conceived of the study. PV and P'tH participated in its design and coordination, and helped to draft the manuscript. SD, RG and RB assisted with the running of the spectra identification tools. All authors read and approved the final manuscript.

PV and P'tH are shared last authors.

Supporting information





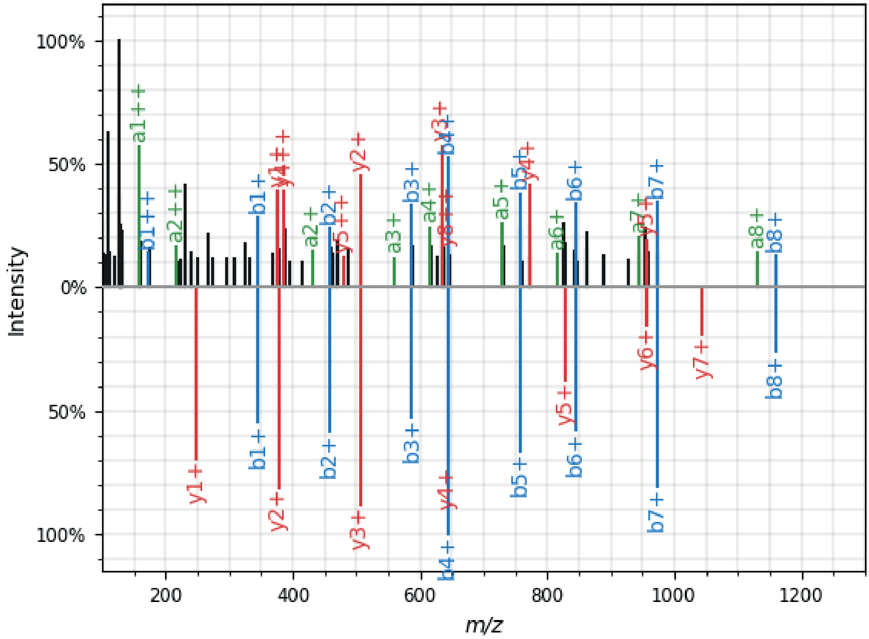


Figure S3: Annotated variant peptide spectra in mirror plots, with theoretical spectra (as predicted by MS2PIP) in the bottom half for reference. Plots made with spectrum_utils python package. Top: variant peptide LQQQHSEQPPLQPSPVTTR, substitution M → T, on chromosome 1 pos 179882939, scan id Linfeng_012511_HapMap39_6.8739.8739.3. Middle: variant peptide DVGEWQHEEFYR, substitution R → G, on chromosome 16 pos 3674464, scan id Linfeng_030911_HapMap46_2.12742.12742.3. This is one of the peptides where no reference counterparts were detected (while 90 variant peptides were identified). Bottom: variant peptide DLEGLSQWHEEK, substitution W → R, on chromosome 22 pos 36292132, scan id Linfeng_080711_HapMap59_5.15580.15580.3. This is one of the rare variant peptide identifications (AF = 0.001).

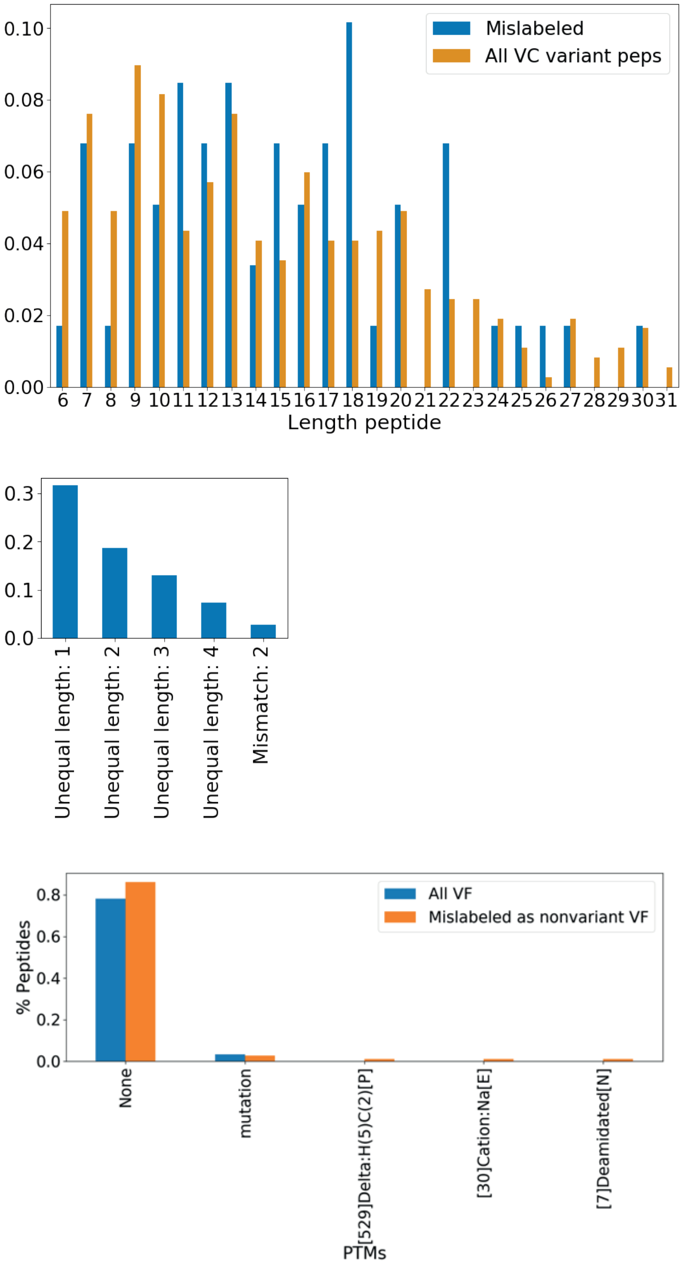
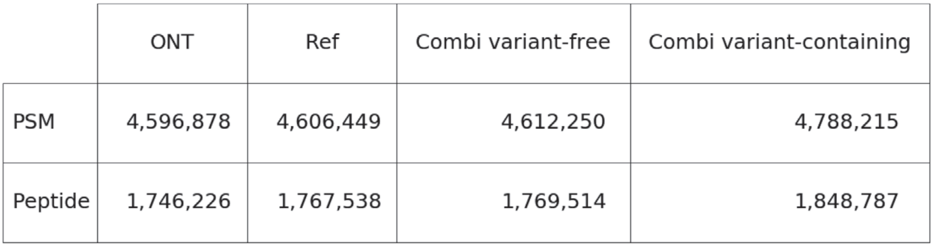


Figure S4: Investigation of false negative ('mislabeled') identifications by *ionbot*™. Top figure shows the density of mislabeled peptides per length, as compared to lengths of all variant peptides identified by the VC method. Middle figure shows the 5 most common causes of misidentification of variant peptides by *ionbot*™. Bottom figure shows unexpected modifications of the false negatives versus the unexpected modifications by all VF identifications. Unlabeled y axises refer to density.

Table S1: Side-by-side comparison of the contents of the search database

Search database contents	Sequences in GENCODE	Sequences in the ONT transcriptome	NA12878-specific variants
ONT	No	Yes	No
Ref	Yes	No	No
VF	Yes	Yes	No
VC	Yes	Yes	Yes

Table S2: Absolute numbers of PSMs and peptides detected per method.



	ONT	Ref	Combi variant-free	Combi variant-containing
PSM	4,596,878	4,606,449	4,612,250	4,788,215
Peptide	1,746,226	1,767,538	1,769,514	1,848,787

https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00264

3

Chapter 3

Multi-omic profiling of pathogen-stimulated primary immune cells

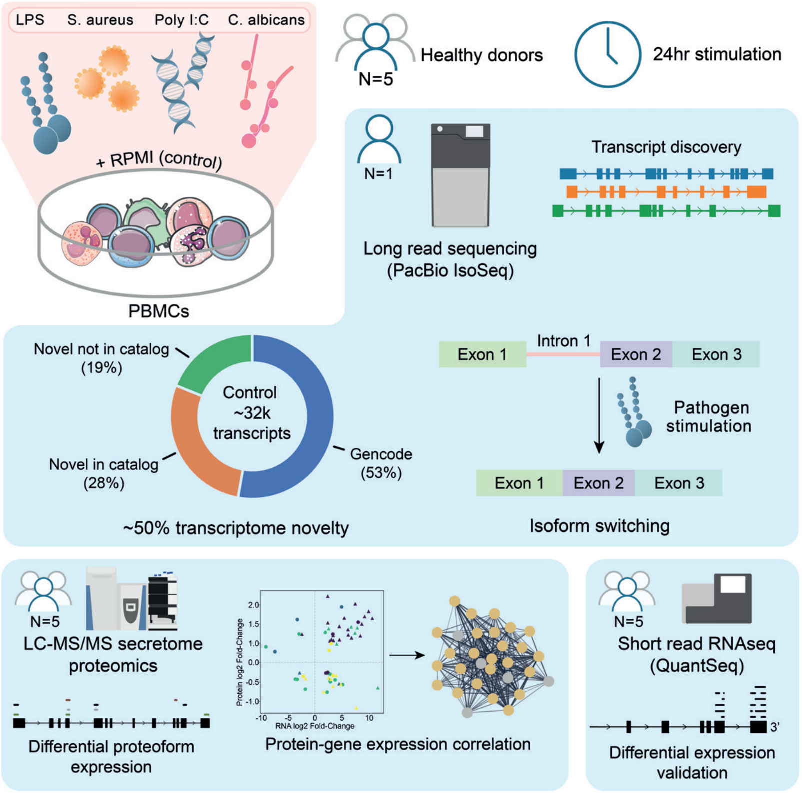
Renee Salz*, Emil E. Vorsteveld*, Caspar I. van der Made, Simone Kersten, Merel Stemerdink, Tabea V. Riepe, Tsung-han Hsieh, Musa Mhlanga, Mihai G. Netea, Pieter-Jan Volders, Alexander Hoischen#, Peter A.C. 't Hoen#

*these authors contributed equally
#these authors contributed equally

Accepted for publication in iScience (2024)

Abstract

We performed long-read transcriptome and proteome profiling of pathogen-stimulated peripheral blood mononuclear cells (PBMCs) from healthy donors to discover new transcript and protein isoforms expressed during immune responses to diverse pathogens. Long-read transcriptome profiling reveals novel sequences and isoform switching induced upon pathogen stimulation, including transcripts that are difficult to detect using traditional short-read sequencing. Widespread loss of intron retention occurs as a common result of all pathogen stimulations. We highlight novel transcripts of *NFKB1 and CASP1* that may indicate novel immunological mechanisms. RNA expression differences did not result in differences in the amounts of secreted proteins. Clustering analysis of secreted proteins revealed a correlation between chemokine (receptor) expression on the RNA and protein levels in *C. albicans*- and Poly(I:C)-stimulated PBMCs. Isoform aware long-read sequencing of pathogen-stimulated immune cells highlights the potential of these methods to identify novel transcripts, revealing a more complex transcriptome landscape than previously appreciated.



Introduction

Immune system responses within the context of an infection are shaped by the nature of the infection and by inter-individual variability, contributing to differential susceptibility to infections and to various diseases with an inflammatory component. Dynamic expression of transcripts and proteins in a range of cells responsible for the innate immune response is important to shape the first line of defense against a wide variety of pathogens[198]. Pattern recognition receptors (PRR) initiate acute inflammatory responses, activating signaling cascades that converge on various transcription factors. Multiple levels of regulation orchestrate the dynamic expression of transcripts and proteins, including transcriptional and post-transcriptional checkpoints such as mRNA splicing and protein translation. Examples of this include the regulatory role of alternative splicing of Toll-like receptors (TLR) and their downstream signaling factors[199,200].

Methods for investigating innate immune responses include *in vitro* stimulation of primary immune cells with pathogens or microbial components. These methods allow for specific investigation of host-pathogen interactions that shape the immune response elicited by specific cell types and have been under extensive investigation in research on the innate immune system[201,202]. Stimuli that are commonly used include molecules that stimulate a specific TLR, such as *E. coli* lipopolysaccharide (LPS) for TLR4[203], dsRNA mimicking Poly(I:C) for TLR3[204] and imidazoquinolines for TLR7/8[205]. Stimulation is also often elicited by live or heat-killed pathogens[206,207], which stimulate at the same time a broader range of PRRs[208,209].

Transcriptome characterization is traditionally performed using short read RNA sequencing. These sequencing approaches are limited by their short-read length (approximately 150-300 bp), necessitating the computational reconstruction of whole transcripts and making detection of different transcripts of the same gene inaccurate. This limitation is especially pronounced in immune biology, where tight regulation of isoform expression has previously been described to play a major role in processes, such as the expression of multiple IL-32 transcripts with different inflammatory potency[210] and alternative splicing of *CD45* in T cell activation[211]. Recent long-read sequencing approaches provide a more complete and accurate reflection of the transcriptome. Sequencing technologies provided by PacBio and Oxford Nanopore allow for the sequencing of mRNA (or cDNA) molecules from the ultimate 3’-end to the ultimate 5’-end, which have given a more comprehensive view into the complexity of the transcriptome. A number of studies have indicated that the isoform landscape is much more complex than previously appreciated[212–214]. Long-read mRNA sequencing has provided insight into regulatory mechanisms of immune responses, for instance in alternative splicing in macrophages[215] , and allows for accurate sequencing of complex transcripts in immune cells[216]. Further work on cell-type specific long-read transcriptomes have shown preferential expression of transcripts in specific cell types[217].

The impact of the newly discovered transcripts as well as post-transcriptional processes can only be fully understood by observing the proteome. Many studies have characterized the transcriptomic landscape of the human immune response, but a multi-omics view of immunity is necessary as mRNA profiles are not enough to understand immune activation[218,219]. Transcript information can be leveraged to study the proteome, including identification of novel proteoforms resulting from alternative splicing. Proteoforms discovered by proteogenomics methodologies have already been found to have a role in immunological processes, for instance in immune-regulating micropeptides[220] and tumor neoantigen production[221].

Here, we stimulated peripheral blood mononuclear cells (PBMCs) with multiple microbial stimuli *in vitro/ex vivo* and performed long- and short-read RNA sequencing and secretome proteomics to gain insight into potential differences in immune response. We aim to provide insight into the immune transcriptome and proteome of immune cells during innate immune responses against a variety of pathogens.

Material & methods

Ex vivo PBMC experiments

Venous blood was drawn from five healthy donors[202] and collected in 10mL EDTA tubes. Isolation of peripheral blood mononuclear cells (PBMCs) was conducted as described elsewhere[222]. In brief, PBMCs were obtained from blood by differential density centrifugation over Ficoll gradient (Cytiva, Ficoll-Paque Plus, Sigma-Aldrich) after 1:1 dilution in PBS. Cells were washed twice in saline and re-suspended in serum-free cell culture medium (Roswell Park Memorial Institute (RPMI) 1640, Gibco) supplemented with 50 mg/mL gentamicin, 2 mM L-glutamine and 1 mM pyruvate. Cells were counted using a particle counter (Beckmann Coulter, Woerden, The Netherlands) after which the concentration was adjusted to 5×10^6/mL. *Ex vivo* PBMC stimulations were performed with $5{\times}10^5$ cells/well in round-bottom 96-well plates (Greiner Bio-One, Kremsmünster, Austria) for 24 hours at 37°C and 5% carbon dioxide. Cells were treated with lipopolysaccharide (*E. coli* LPS, 10 ng/mL), *Staphylococcus aureus* (ATCC25923 heat-killed, $1{\times}10^6$/mL), TLR3 ligand Poly I:C (10 µg/mL), *Candida albicans* yeast (UC820 heat-killed, $1{\times}10^6$/mL), or left untreated in regular RPMI medium as normal control. After the incubation period of 24h and centrifugation, supernatants were collected and stored at -80°C until further processing. For the RNA isolation, cells were stored in 350 µL RNeasy Lysis Buffer (Qiagen, Rneasy Mini Kit, Cat nr. 74104) at −80°C until further processing.

RNA and protein isolation

RNA was isolated from the samples using the RNeasy RNA isolation kit (Qiagen) according to the protocol supplied by the manufacturer. The RNA integrity of the isolated RNA was examined using the TapeStation HS D1000 (Agilent), and was found to be ≥7.5 for all samples. Accurate determination of the RNA concentration was performed using the Qubit (ThermoFisher).

We extracted the secretome of the 24-hour stimulated PBMCs. To 250 µL of supernatant, 250 µL buffer containing 10% sodium dodecyl sulfate (SDS) and 100 mM triethylammonium bicarbonate (TEAB), pH 8.5 was added. Proteins were reduced by addition of 5 mM dithiothreitol and incubation for 30 minutes at 55˚C and then alkylated by addition of 10 mM iodoacetamide and incubation for 15 minutes at RT in the dark. Phosphoric acid was added to a final concentration of 1.2% and subsequently samples were diluted 7-fold with binding buffer containing 90% methanol in 100 mM TEAB, pH 7.55. The samples were loaded on a 96-well S-Trap^{TM} plate (Protifi) in parts of 400 µL, placed on top of a deepwell plate, and centrifuged for 2 min at 1,500 x g at RT. After protein binding, the S-trap^{TM} plate was washed three times by adding 200 µl binding buffer and centrifugation for 2 min at 1,500 x g at RT. A new deepwell receiver plate was placed below the 96-well S-Trap^{TM} plate and 125 µL 50 mM TEAB containing 1 µg of trypsin was added for digestion overnight at 37°C. Using centrifugation for 2 min at 1,500 x g, peptides were eluted in three times, first with 80 µL 50 mM TEAB, then with 80 µL 0.2% formic acid (FA) in water and finally with 80 µL 0.2% FA in water/acetonitrile (can) (50/50, v/v). Eluted peptides were dried completely by vacuum centrifugation.

Long-read library preparation and sequencing

Libraries were generated from one donor using the Iso-Seq-Express-Template-Preparation protocol according to the manufacturer's recommendations (PacBio, Menlo Parc, CA, USA). We followed the recommendation for 2-2.5kb libraries, using the 2.0 binding kit, on-plate loading concentrations of final IsoSeq libraries was 90pM (*C. albicans*, *S. aureus*, Poly(I:C), RPMI) and 100pM (LPS) respectively. We used a 30h movie time for sequencing.

The five samples were analyzed using the isoseq3 v3.4.0 pipeline. Each sample underwent the same analysis procedure. First CCS1 v6.3.0 was run with min accuracy set to 0.9. IsoSeq lima v2.5.0 was run in IsoSeq mode as recommended. IsoSeq refine was run with '--require-polya'. The output of IsoSeq refine was used as input for IsoQuant v3.1.2[223] with GRCh38.p13 v43 primary assembly from GENCODE. The settings were set for full length PacBio data, and quantification included ambiguous reads. In IsoQuant, transcripts were considered novel if their intron chains did not match intron chains found in GENCODE annotation version 39. Transcripts with fewer than 5 reads across all samples were excluded from further analyses (Supplemental table 1).

We sought to validate the novel transcripts identified using long-read sequencing using FANTOM5 CAGE data of CD14 monocytes (https://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.primary_cell.CAGEScan/CD14%2b%20monocyte%20derived%20endothelial%20progenitor%20cells%2c%20donor1.NCig10041.11229-116C5.hg19.GCTATA.clusters.bed.gz) that allows for the identification of transcripts with a matching TSS from this 5’ sequencing data. Transcripts with novel 5’ were considered to be supported with a CAGE peak if within 150 basepairs from the TSS.

Short-read library preparation and sequencing

RNA input was normalized to 200 ng for all samples/donors and libraries were generated using the QuantSeq 3’ mRNA-Seq Library Prep Kit-FWD from Lexogen (Lexogen) in accordance with the manufacturers’ protocol. In order to ensure high quality libraries, two separate preparations were performed, limiting the number of samples to 30 per preparation. End-point PCR was performed with 19 – 22 cycles, as indicated by a quantitative PCR on a 1:10 aliquot of a subset of double stranded cDNA libraries. Accurate quantification and quality assessment of the generated libraries was performed using Qubit dsDNA High Sensitivity assay (Thermo Fisher Scientific) and Agilent 2200 TapeStation (High Sensitivity D1000 ScreenTape, Agilent). Molarity of individual libraries was calculated using the cDNA concentration (Qubit) and average fragment size (TapeStation). Safeguarding sufficient read-depth for each sample, libraries were split in two separate runs. In each run, the baseline RPMI condition across all donors and time-points was included, in turn allowing sequencing bias assessment. The cDNA libraries of 35 samples were pooled equimolarly to 100 fmol. After a final dilution of both pools to a concentration of 4 nM, they were sequenced on a NextSeq 500 instrument (Illumina) with a final loading concentration of 1.4 pM.

FastQC v0.11.5 (Babraham Bioinformatics) was used to assess the quality of the obtained sequencing data, followed by removal of adapter sequences and poly(A) tails by Trim Galore! V.0.4.4_dev (Babraham Bioinformatics) and Cutadapt v1.18[224]. Since QuantSeq reads only provide coverage of the 3’ end of transcripts, we generated a set of transcripts representative of the full transcriptome by grouping transcripts based on unique 3’ sequences. Therefore, we separately mapped the filtered and trimmed reads to the long read transcriptome with Salmon v1.9.0 in mapping-based mode with decoys[225].

Differential expression analyses

To measure differential gene expression from long-read RNA sequencing, low abundance genes were filtered using a 10 CPM threshold with the conform package in python. Differentially expressed genes (DEGs) and transcripts were calculated for each condition versus control using the NOISeq R package[226] from the abundances generated with isoquant. TMM normalization was chosen and q-value threshold for DE was set at 0.95.

DEGs were generated from the salmon-mapped short-read RNA sequencing data using the samples from the same donor using NOISeq[226]. The two control samples (RPMI) per donor were treated as technical replicates. TMM normalization was chosen and q-value threshold for DE was set at 0.95. We validated the DEGs detected from long-read sequencing with those generated with the short-read data by comparing the linear correlation of the log2fold change values for each condition combination between both datasets using the lm() R function.

The up- and downregulated DEGs per condition-control pair were analyzed for pathway enrichment separately using gProfiler[29]. We used Gene Ontology biological process and molecular function and TRANSFAC transcription factor motifs gene sets[227,228]. A term size filter of between 100-500 was used to generate the final enrichment profiles.

Isoform switching

A first-pass isoform switching analysis was performed using swanvis v2.0[229]. For a second-pass isoform switching analysis, the resulting gene-level isoform switch p-values were imported into IsoformSwitchAnalyzeR v1.16.0 package in R[230]. Thresholds for isoform switching were set at 10 DPI (differential percent isoform use) and nominal p-value <0.05. Sequences corresponding to the significant isoform switches were analyzed with CPAT v1.2.4[114], hmmscan v3.3.2 with Pfam[231], and SignalP5[232] as a part of the IsoformSwitchAnalyzeR package.

Pathway analysis and gene network analysis of genes that were found to undergo isoform switching was performed in Cytoscape[233]. Default pathway analysis was performed, filtering for Gene Ontology Biological Process gene sets. An Enrichment Map was built from the enriched gene sets with a Jaccard similarity cutoff of 0.4[234].

Genes found to undergo intron retention gains/losses and genes with domain gains/ losses were separately analyzed using gProfiler. We used Gene Ontology Biological Process gene sets with a with a term size filter between 100-500 genes. We separately analyzed genes with domain gains or losses were using dcGOR[235]. We used the gene ontology molecular function gene sets with a term size filter between 100-500 genes.

LC-MS/MS analysis

Peptides were re-dissolved in 20 µL loading solvent A (0.1% trifluoroacetic acid in water/ acetonitrile) (98:2, v/v)) of which 4 µL was injected for LC-MS/MS analysis on an Ultimate 3000 RSLCnano system in-line connected to a Q Exactive HF mass spectrometer (Thermo). Trapping was performed at 10 µL/min for 4 min in loading solvent A on a 20 mm trapping column (made in-house, 100 µm internal diameter (I.D.), 5 µm beads, C18 Reprosil-HD, Dr. Maisch, Germany). The peptides were separated on a 250 mm Waters nanoEase M/Z HSS T3 Column, 100Å, 1.8 µm, 75 µm inner diameter (Waters Corporation) kept at a constant temperature of 45°C. Peptides were eluted by a non-linear gradient starting at 1% MS solvent B reaching 33% MS solvent B (0.1% formic acid (FA) in water/acetonitrile (2:8, v/v)) in 100 min, 55% MS solvent B (0.1% FA in water/acetonitrile (2:8, v/v)) in 135 min, 97% MS solvent B in 145 minutes followed by a 5-minute wash at 97% MS solvent B and re-equilibration with MS solvent A (0.1% FA in water).

The mass spectrometer was operated in data-dependent acquisition mode, automatically switching between MS and MS/MS acquisition for the 16 most abundant ion peaks per MS spectrum. Full-scan MS spectra (375-1500 m/z) were acquired at a resolution of 60,000 in the Orbitrap analyzer after accumulation to a target value of 3,000,000. The 16 most intense ions above a threshold value of 15,000 were isolated with a width of 1.5 m/z for fragmentation at a normalized collision energy of 28% after filling the trap at a target value of 100,000 for maximum 80 ms. MS/MS spectra (200-2000 m/z) were acquired at a resolution of 15,000 in the Orbitrap analyzer.

Protein identification and quantification

Two search databases were constructed; one database for proteoform detection and one database for quantification. The database used for sensitive detection of proteoforms was generated using a slightly adapted version of the Long Read Proteogenomics pipeline by Miller *et al*[236]. Since the pipeline uses a different long-read transcriptomics tool, small syntax adjustments were made to accommodate the use of Isoquant output. Additionally, a custom script was written to have Isoquant output mimic the required input format. The pipeline generated a GENCODE-PacBio hybrid database. The proteome from *C. albicans* (taxon ID 5476) and *S. aureus* (taxon ID 1280) were downloaded from UniProt and added to the search database. The search database used for quantification was created by downloading the proteome from *H. sapiens* (taxon ID 9609), *C. albicans* (taxon ID 5476) and *S. aureus* (taxon ID 1280) from UniProt. Metamorpheus default contaminants were added to both search databases.

Mass spectra were identified using Metamorpheus v1.0.0[237]. The Human Proteome Project Mass Spectrometry Data Interpretation Guidelines version 3.0 were applied[238]. Quantification was performed using FlashLFQ v 1.2.4.294[239] with all five individuals set as biological replicates and the two control (RPMI) samples per individual set as technical replicates. The following options enabled: normalization, shared peptide quantification, Bayesian fold change analysis, and match between runs (Supplemental table 2). An adapted version of SQANTI protein was used to search for novel peptides in the Metamorpheus identifications. Enrichment of secreted proteins was determined using the predicted secreted proteins from Human protein atlas[240] as reference.

Protein clustering

FlashLFQ raw protein expression values originating from the quantification database search were first square root transformed. To normalize for donor effects, the mean protein expression value per gene/individual was subtracted from all the expression values from the same gene/individual. Then z-score normalization was performed across all individuals per gene. K-means clustering was then performed using the kmeans() function in R with seed #82 and default parameters. We found four clusters to optimally represent the data according to the elbow plots (Supplemental figure 1). A heatmap was constructed with those clusters using the ComplexHeatmap package[241]. The proteins identifiers assigned to cluster #4 were converted to gene names and analyzed using gProfiler for enrichment analysis using both Gene Ontology Biological Process and Molecular Function gene sets. We further analyzed the protein found to form cluster 4 through a protein network analysis in Cytoscape[233].

Results

We stimulated PBMCs from five donors with four different microbial stimuli, mimicking bacterial (*E. coli* LPS, *S. aureus*), viral (Poly(I:C)) and fungal (*C. albicans*) infections. PBMCs were stimulated for 24 hours. RPMI incubation was used as a negative control (Figure 1A). To characterize full-length transcript structures, we performed long-read sequencing on PBMCs from one donor (Figure 1B). Additionally, shotgun proteomics data was generated from supernatants of the samples from all five donors. The proteomics data serves to corroborate differential gene/transcript expression and provide evidence of the protein-coding potential of novel transcripts identified through long-read RNA sequencing (Figure 1C). Short-read 3’ sequencing data of all five donors was generated to validate differential gene expression data generated from long-read RNA sequencing (Figure 1D).

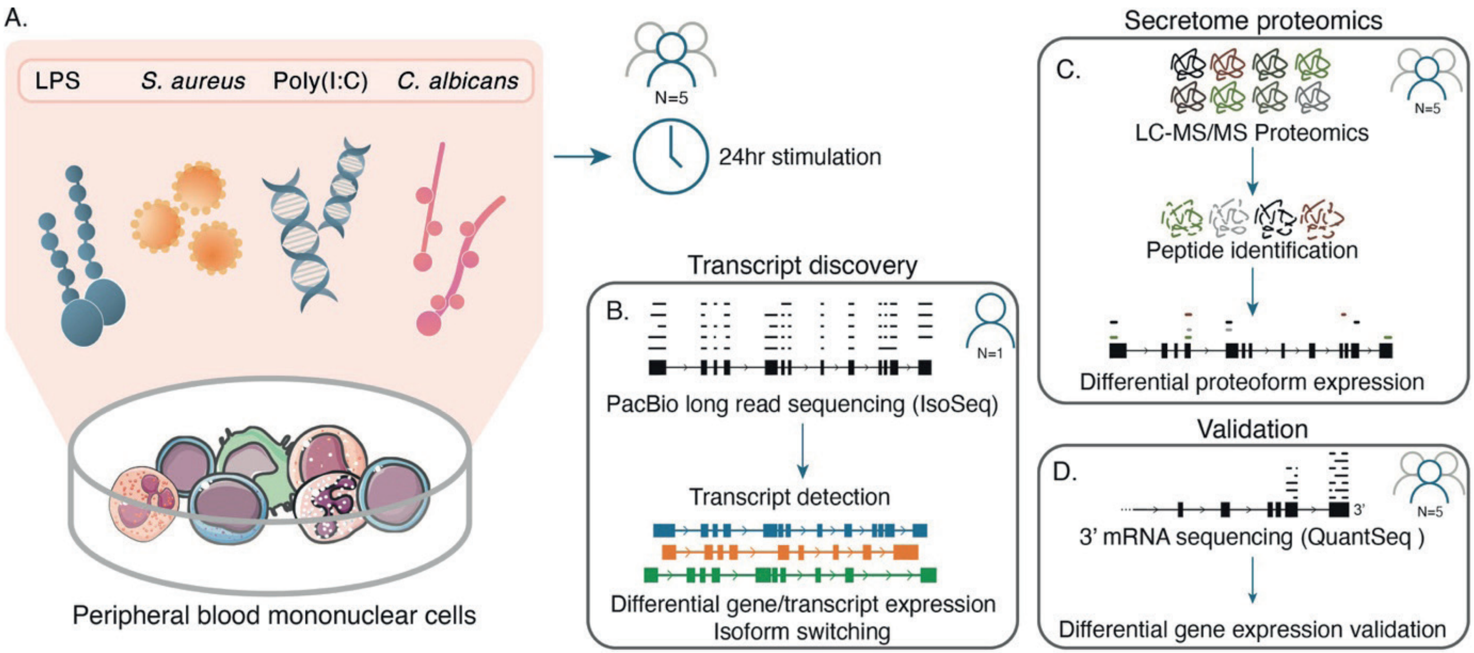

A.
LPS
S. aureus
Poly(I:C)
C. albicans
Peripheral blood mononuclear cells
N=5
24hr stimulation
Transcript discovery
B.
N=1
PacBio long read sequencing (IsoSeq)
Transcript detection
Differential gene/transcript expression
Isoform switching
Secretome proteomics
C.
N=5
LC-MS/MS Proteomics
Peptide identification
Differential proteoform expression
Validation
D.
N=5
3'
3' mRNA sequencing (QuantSeq)
Differential gene expression validation

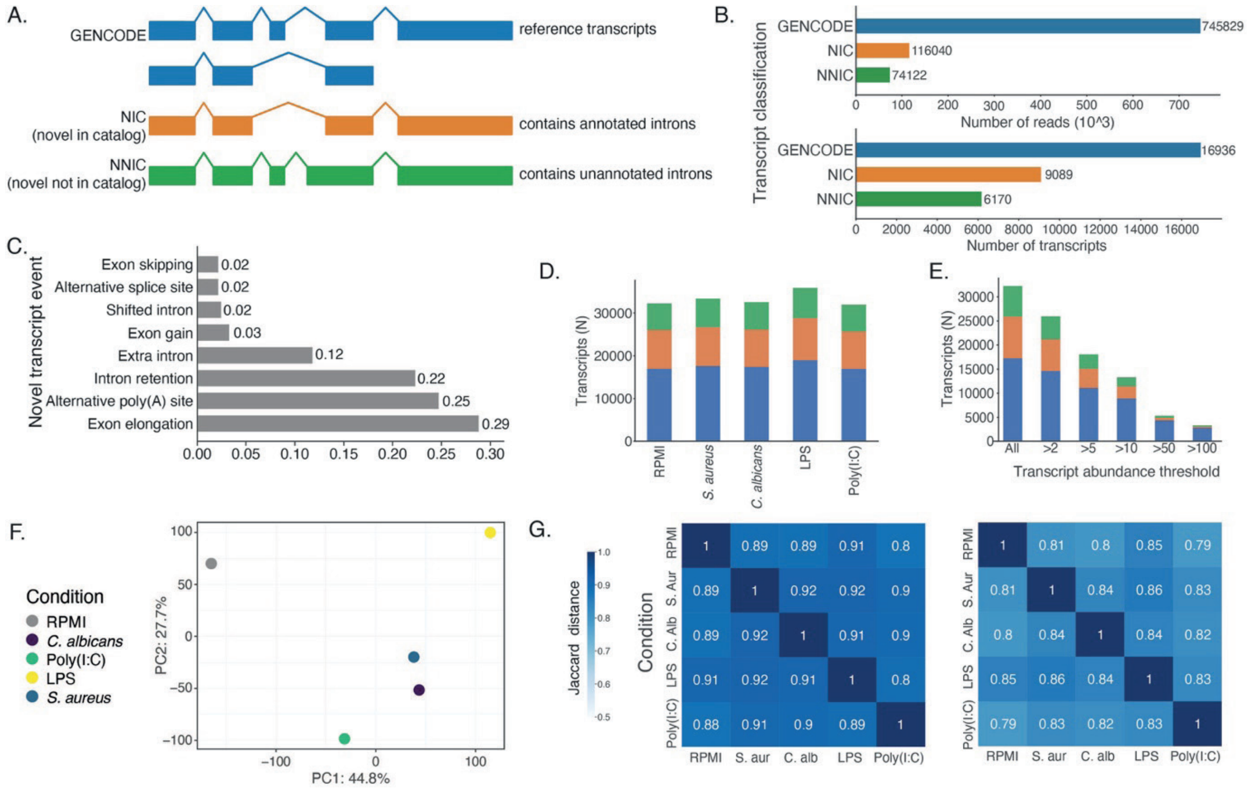
Figure 1: Experimental setup. A) Human peripheral blood mononuclear cells (PBMCs) isolated from five donors were exposed to four different pathogenic stimuli and analyzed after 24 hours. B) PacBio long read RNA-sequencing was performed on samples from one of the five donors. Long-reads were used to estimate differential transcript expression and isoform switching. C) The supernatant from all samples (all donors) was collected and peptides were detected to quantify protein levels in the secretome. D) Short read RNA sequencing (QuantSeq) was performed on all samples (all donors) and differential expression estimates were compared to those measured in long-read sequencing.

Long read transcriptomes of both control and pathogen-stimulated conditions show novelty

Sequences detected using long-read sequencing were categorized in terms of novelty according to their intron chains. Transcripts are divided into three categories that encompass reference transcripts (GENCODE), novel in catalog (transcripts that contain annotated introns) and transcripts that are novel not in catalog (containing unannotated introns) (Figure 2A). We identified a total of 37,312 unique transcript sequences from 11,872 genes across all samples. The majority of transcripts were in protein coding genes (Supplemental figure 2A) including ~10% immune-related genes (Supplemental figure 2B). We found 47.4% of detected transcripts to be novel, while these accounted for only 20.3% of the total reads (Figure 2B). The distribution of reads per novel transcript was similar to that of known transcripts with a slight skew towards lower abundance (Supplemental figure 3A). Exon elongations were the most observed feature distinguishing novel from known transcripts, occurring in nearly a third of the novel transcripts found in RPMI. This was similar for the stimulated conditions (Figure 2C, Supplemental figure 3B). The percentage of novel transcripts and transcript deviations were similar for all conditions (Figure 2D). To corroborate the existence of novel transcripts, we analyzed FANTOM5 CAGE peaks in the vicinity of the transcription start sites for novel transcripts with novel 5' ends. We found 8,233 (51.3%) novel 5' end transcripts across all conditions to be supported by a CAGE peaks from unstimulated human monocytes (within 150 nucleotides)[242].

Principal component analysis of the expression levels for each transcript indicated that stimulated conditions were more similar to each other than to RPMI. *S. aureus* and *C. albicans* were most similar to each other (Figure 2F). Genes and transcripts expressed were similar in the stimulated conditions with average Jaccard similarity indices of 0.9 and 0.82 for genes and transcripts, respectively (Figure 2G). Novel transcripts had similar Jaccard indices to each other than for known transcripts (not shown). Differential expression analysis yielded an average of 949 DEGs and 2,076 differentially expressed transcripts per condition (Supplemental figure 4, Supplemental table 3-4).

We validated the DEGs through 3' transcript counting (QuantSeq)[243]. We gathered a set of representative transcripts based on sequence differences at the 3' end of transcripts (29,760 transcripts, 79.8% of total) and investigated the correlation of differential expression in the long-read sequencing data with the separately generated short read dataset of the same donor. The DEGs that overlapped between both datasets correlate well (R^2 0.62-0.81). Best matching pairs of stimulated conditions between the short- and long-read confirmed the concordance of both sequencing approaches (Supplemental figure 5, Supplemental table 5).



	RPMI	S. aur	C. alb	LPS	Poly(I:C)
RPMI	1	0.89	0.89	0.91	0.8
S. Aur	0.89	1	0.92	0.92	0.9
C. Alb	0.89	0.92	1	0.91	0.9
LPS	0.91	0.92	0.91	1	0.8
Poly(I:C)	0.88	0.91	0.9	0.89	1

	RPMI	S. aur	C. alb	LPS	Poly(I:C)
RPMI	1	0.81	0.8	0.85	0.79
S. Aur	0.81	1	0.84	0.86	0.83
C. Alb	0.8	0.84	1	0.84	0.82
LPS	0.85	0.86	0.84	1	0.83
Poly(I:C)	0.79	0.83	0.82	0.83	1

Figure 2: Transcriptome novelty in the control condition and comparison between stimuli transcriptomes in the five long-read samples. A) Transcript novelty categories. GENCODE (blue) is the set of all known reference transcripts. Novel in catalog (orange) contains a novel combination of annotated introns. Novel not in catalog (green) contains one or more unannotated introns. B) Reads (top) and unique transcripts (bottom) of events in each pre-defined transcript novelty category in RPMI. C) Novelty-inducing events occurring in the RPMI transcriptome. D) Unique transcripts by novelty category for each of the stimulus conditions. E) Unique transcripts by novelty category that remain at various transcript abundance thresholds in the C. *albicans* condition. F) Transcriptomes of the samples plotted on the first two principal components of PCA. G) Jaccard distances of genes (left) and known transcripts (right) of the transcriptomes, not considering transcript abundance.

Pathogen stimuli display upregulation of different pathways

Differential gene expression analysis using the long-read sequencing data resulted in a total of 1,733 genes that were differentially expressed in stimulated conditions compared to control. We performed pathway analysis for each condition using gProfiler[244] (Supplemental table 6). By overlapping the gene sets enriched in each of the four conditions, we discerned biological processes/functions specific to certain pathogen-stimulated conditions. There are a lot of constants in host response regardless of the pathogen, and indeed the largest set of pathways was in the overlap between all stimulated conditions (211 pathways, Figure 3A). This set has an enrichment of genes involved in type II interferon (IFN-γ) responses. Genes involved in tertiary and specific granules, which play a role in the defense against pathogens were found to be enriched among upregulated genes in all conditions. Surprisingly, we also find these and related gene sets to be enriched among downregulated genes as a result of *S. aureus* and Poly(I:C) stimulation, potentially a result of the regulation of the inflammatory response. Further gene sets included the response to molecules of bacterial origin (including LPS), innate immune response signaling such as PRR signaling, antigen processing and presentation and IL-1 production (Figure 3B).

Some pathogen-stimulated conditions had more enriched pathways in common than others. There was a notable overlap of 131 gene sets enriched in *C. albicans-, S. aureus*- and Poly(I:C)-stimulated conditions. Some of these were common to the set overlapping between all conditions, such as interferon responses. The LPS-excluding set showed particular enrichment related to viral processes such as the defense against viruses, regulation of the viral lifecycle, likely due to interferon-stimulated gene expression, such as *STAT1, OAS1/3, OASL and IFIH1*. Also, transcription factor binding matches (TRANSFAC) such as *IRF-2, 5, 8* and *9* were enriched, reflecting downstream signaling through various signaling pathways leading to the regulation of the production of interferons and immune cell development (Figure 3C)[245].

LPS and Poly(I:C) were the 2 stimuli with the most enriched pathways unique to a single stimulus. For 55 gene sets unique to LPS, there was a downregulation of T cell receptor signaling, in part due to the downregulation of *CD4* expression, which has previously been described as a result of endogenous production of TNF-α and IL-1β as a result of LPS stimulation[246]. We further found an upregulation of gene sets involved in metabolic processes such as oxidoreductase complexes and cellular responses to oxygen, possibly reflecting metabolic changes previously described to occur in immune cells such as monocytes upon LPS stimulation[247]. Furthermore, there was an upregulation of genes involved in humoral immune responses (Figure 3D).

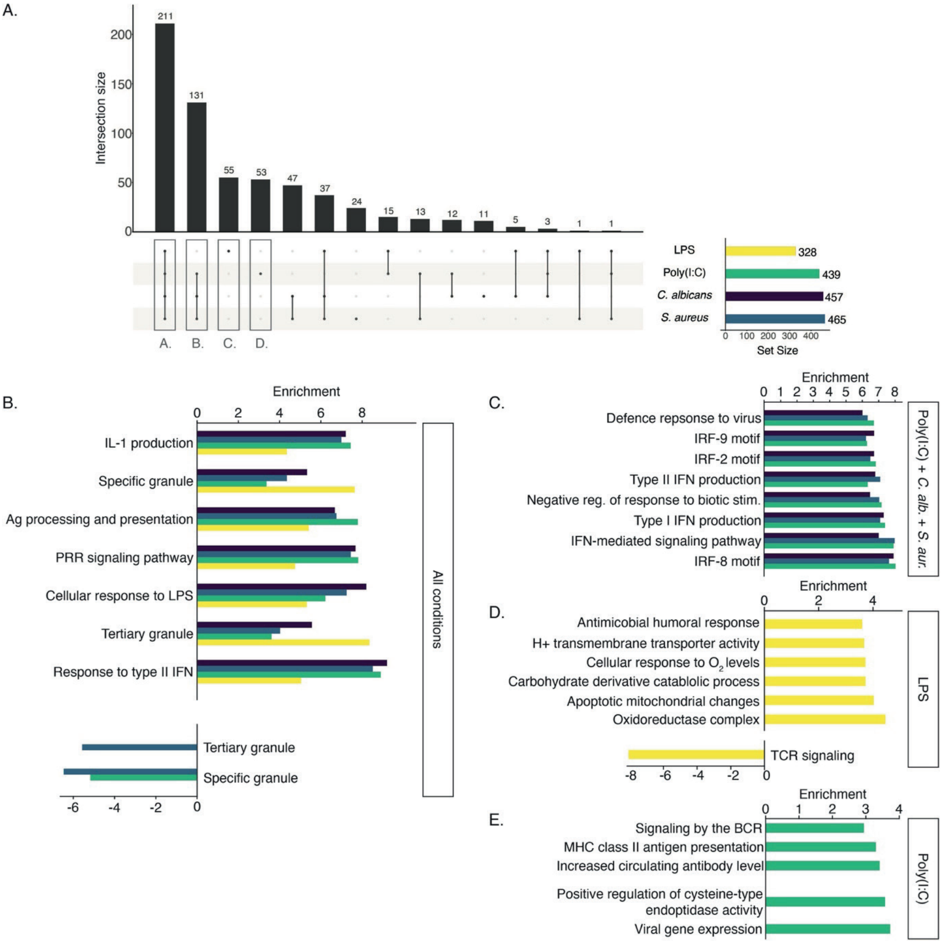


Figure 3: Differential pathway analysis originating from differentially expressed genes on the RNA level. A) Overlap between enriched pathways generated from the differentially expressed genes from the four conditions. B) Selected pathways found to be enriched for all conditions, C) three of the four conditions (Poly(I:C), *C. albicans* and *S. aureus*), D) specifically for LPS and E) specifically for Poly(I:C).

For 53 gene sets enriched uniquely in Poly(I:C), we found functions including viral gene expression, apoptosis related signaling (regulation of cysteine-type endopeptidase activity) and B-cell related gene sets such as increased antibody levels and BCR signaling. Finally, there was an enrichment of MHC class II antigen presentation (Figure 3E).

Isoform switches highlight transcriptome differences between conditions and control

Isoform switching (IS) genes are defined by a change (increase/decrease) of expression of a particular transcript isoform as measured by percent of total reads for a gene. In different samples/conditions, a particular transcript isoform may comprise a different isoform fraction (dIF) value for a given gene. Here, a change of at least 10% (0.10 dIF) in control and the opposite change (decrease/increase) of expression of a different transcript isoform in the same gene of at least 10% in the pathogen-stimulated condition is considered an IS.

A total of 999 IS were detected in 398 genes. Nearly half (N=192, 48.2%) of these IS genes were unique to their respective stimulus conditions, while 10.3% were found in all conditions (N=41) (Figure 4A, Supplemental table 7-8). The majority of genes demonstrating IS were not differentially expressed in their respective conditions (327 genes; 77%). Most genes that were found to undergo IS displayed only one IS instance (Supplemental figure 6A). Pathway analysis of genes undergoing IS were enriched for gene sets involved in metabolic processes, mRNA splicing, protein transport and catabolism. Furthermore, immune and stress-related pathways such as MHC type I antigen processing and transport through vesicles, inflammasomes, oxidative stress and apoptosis were found to be represented in genes undergoing IS (Figure 4B, Supplemental tables 9-13).

We sought to understand the molecular consequences of IS upon pathogen stimulation by categorizing the differing features of the isoform pairs involved in the switch. Each of the IS was annotated with one or more of the following predicted protein characteristics: change in ORF length, ORF gain/loss, domain gain/loss, NMD sensitivity, intron retention (IR) gain/loss, coding probability (ORF presence), and signal peptide gain/loss. These consequences are not independent and often multiple consequences could be attributed to one IS (Supplemental Figure 6B). We observed general IS trends on a genome-wide scale (Supplemental figure 6C, Supplemental table 14). Strikingly, we found IR loss to be the most common consequence of IS in this dataset. Isoforms with retained introns comprised a higher isoform fraction for genes in the control condition, while their respective intron-excluding counterparts had a higher isoform fraction for genes in the pathogen-stimulated conditions. Genes displaying loss of IR were enriched for pathways involved in mRNA processing, including spliceosome-related gene sets, antigen processing and IL-1 production (Supplemental figure 7). IR has previously been described as a regulatory mechanism of RNA processing, splicing, vesicle transport and type I interferon production in the development of various immune cell types, including macrophages[248,249], granulocytes[250] and B cells[251,252]. Our findings support previously described associations of IR losses in immune-related processes, and adds new genes regulated by IR loss during immune responses (Supplemental figure 7, Supplemental table 15).

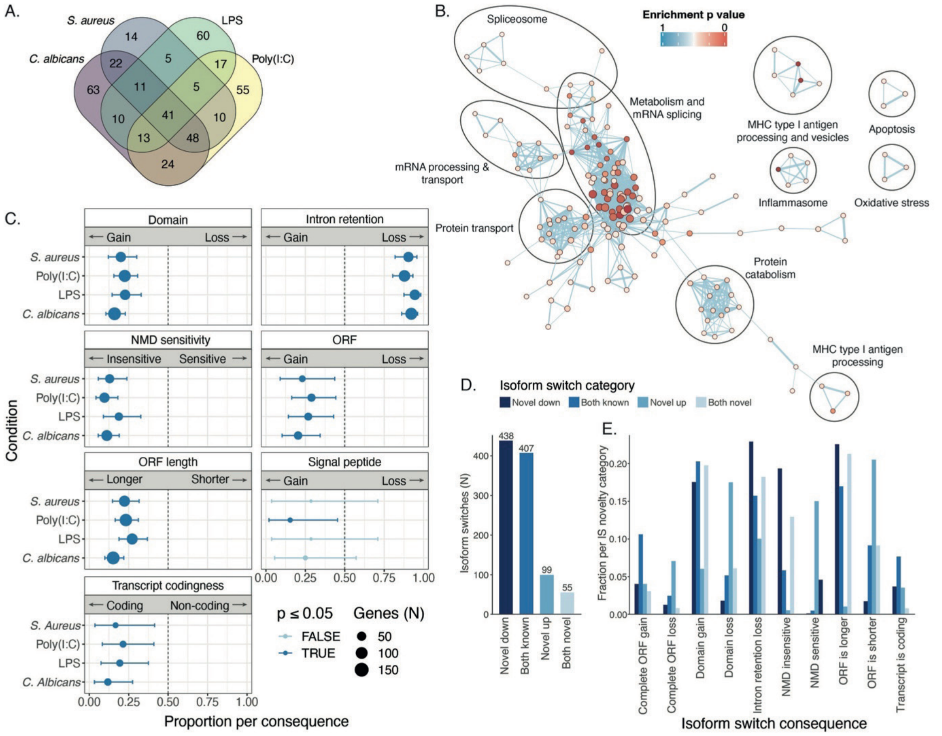


Figure 4: Isoform switching induced by pathogen stimulation. A) Overlap of isoform switching genes between the four stimulus conditions. B) Pathway network analysis derived from genes found to undergo isoform switching (IS) upon pathogen stimulation. Each pathway is colored by p value, where a darker red indicates a lower p value. C) Proportions of total IS events in each stimulated condition per IS consequence. D) Number of IS by category of switch pairs. Categories are defined by involvement of novel transcripts in a given IS. "Novel down" indicates that the isoform switched from a higher proportion of the novel transcript in control to a higher proportion of a known transcript in the stimulus condition. "Both known" indicates that the IS occurs between 2 reference transcripts. E) Fraction of each transcript novelty combination per IS consequence. Normalized by total number of IS events per novelty category.

In addition, we found a higher proportion of transcripts to have domain gains than domain losses. This could indicate that stimulation by a pathogen causes a gene to switch expression to a transcript isoform that codes for a protein with an extra function. Other observed trends included longer ORFs and NMD insensitivity in transcript isoforms induced by pathogen stimulation (Figure 4C).

Since the addition or loss of domains could directly reveal protein function changes, we explored the IS that had this consequence type. We found that genes with domain gain/loss (N=158, Supplemental table 16) were enriched for involvement in various catabolic processes. We also found enrichment of T cell activation genes, an effect previously described as a functional consequence of CD8+ T cell co-stimulation[253]. Other enriched gene sets include leukocyte cell-cell adhesion and activation and general innate immune response genes (Supplemental table 17). When looking more specifically at the molecular functions of the gained domains themselves, we found an enrichment of domains with potassium channel regulator activity, kinase- and transferase activity concerning phosphorus-containing groups and nucleic acid binding. These results potentially indicate functional and cell-type specific effects of domain gains as a result of IS in immune responses (Supplemental figure 8, Supplemental table 18).

Novel transcripts play an important role in IS. Of the 999 IS, more than half (N=592) had at least one novel transcript involved in the IS. In most cases (N=438), the switch was from a novel transcript to a known transcript (Figure 4D, Supplemental table 19). Compared to IS cases where only known transcripts were involved, the IS consequences were more often NMD insensitivity and IR loss (Figure 4E). Conversely, shorter ORFs, domain losses and NMD sensitivity were more common effects when the IS was from a known to a novel transcript isoform. In conclusion, the unstimulated condition is characterized by the presence of many novel transcripts with retained introns, which are difficult to detect with short read sequencing. IR is likely a mechanism to prepare a cell for fast action after an immune stimulus, when splicing of the retained intron could quickly generate a functional transcript with coding potential, which has been for instance been described in CD4+ T cells[254].

A novel read-through transcript including CARD16 and CASP1

As an example of a remarkable finding with possible biological impact once validated, we identified a read-through transcript that includes both *CARD16* and *CASP1* (Figure 5A). Read-through transcripts involve transcription that extends beyond the normal polyadenylation site (PAS), terminating at the PAS of an adjacent gene or other nearby locus[255]. These transcripts have been found to be expressed in specific circumstances, including malignancy and infection[255,256]. This particular novel transcript encompasses the coding region of *CASP1* and has an extended 5’ UTR which spans *CARD16,* and thus contains two ORFs. This IS was annotated as an intron retention loss, as the novel transcript loses an intronic region in its 3’ UTR (Figure 5B). Both the known and novel transcripts in this IS are predicted to be coding (both 100%). *CASP1* was found to be differentially expressed upon Poly(I:C) stimulation (log2FC 1.73, p=0.049; Figure 5C). The isoform expression of the known transcript was found to decrease upon Poly(I:C) stimulation, while the novel transcript was found to increase (Figure 5D). This is further reflected in the isoform fraction, increasing from 8.3% to 24.8%, while the known transcript decreased from 85.5% to 74.2% (Figure 5E).

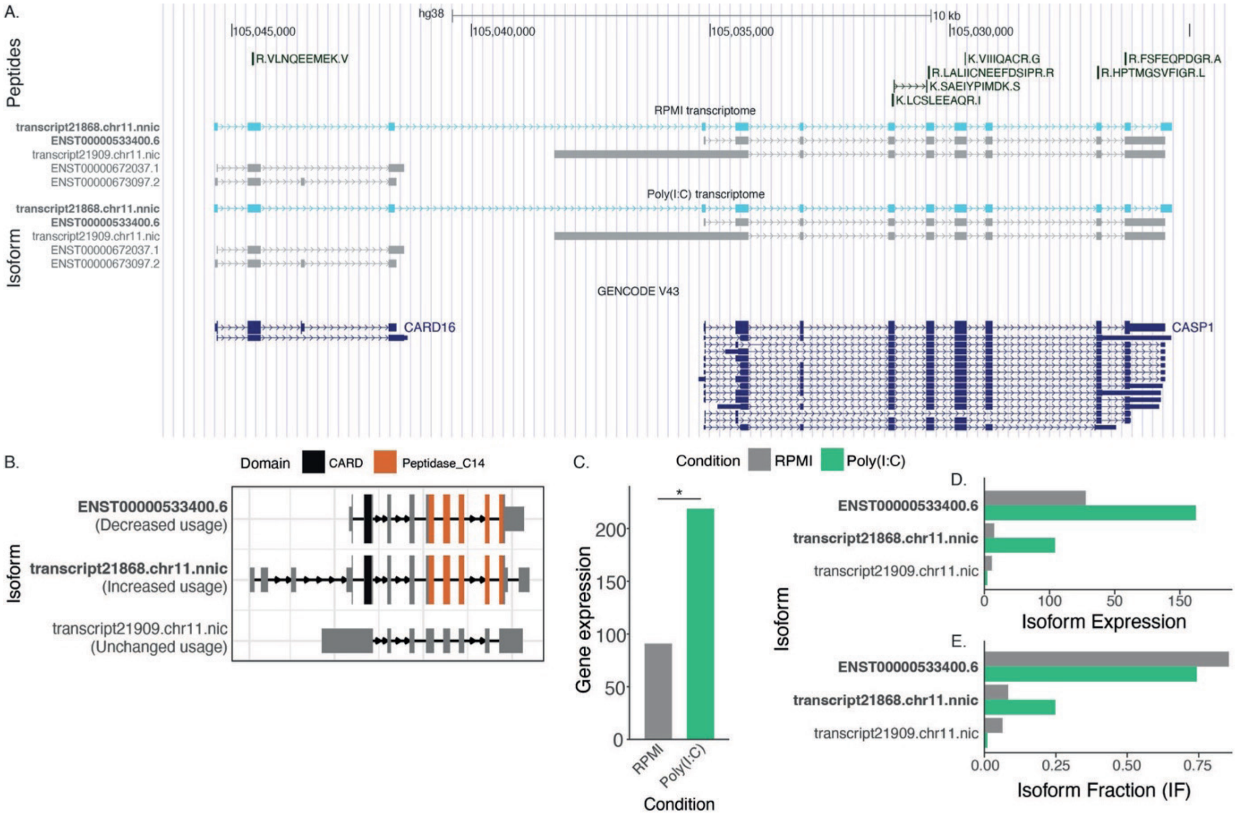


Figure 5: *A novel readthrough transcript of CASP1.* A) USCS genome browser track of the transcripts detected in the control condition (RPMI) and stimulated condition (Poly(I:C)). The novel readthrough transcript containing both CASP1 and CARD16 is presented in light blue. Known transcripts in in GENCODE are presented below. B) Representation of the domains in the novel *CASP1* transcript, indicating that *CARD16* is entirely included in the 5' UTR of the transcript. C) Gene and transcript expression and isoform fraction of the *CASP1* transcripts that were detected.

CARD16 and *CASP1* both have a function in proinflammatory IL-1β signaling, where *CARD16* has been shown to play a role in *CASP1* assembly, although there remains discussion on the exact regulatory effect of *CARD16* on this process [257,258]. We have identified an IS specifically for Poly(I:C) stimulation, where a novel transcript of *CASP1* was found to harbor *CARD16* in its 5' UTR was upregulated upon stimulation. This finding could suggest a novel molecular mechanism in IL-1β signaling, potentially through the regulation of *CASP1* by its regulator *CARD16*.

A novel coding transcript of NFKB1

We identified a novel *NFKB1* transcript that demonstrated IS in all four conditions. This novel transcript was shorter than the canonical transcripts (Figure 6A). Further analysis revealed that the novel transcript start site was supported by multiple nearby CAGE peaks (Figure 6B). Strikingly, this novel transcript lacks a part of its Rel homology domain, a conserved domain responsible for functions such as dimerization and DNA binding (Figure 6C)[259]. *NFKB1* was not found to be significantly differentially expressed, although gene expression was found to be higher in pathogen-stimulated condition compared to unstimulated condition (only *C. albicans* shown, Figure 6D). The expression of the novel transcript was found to increase upon pathogen stimulation (Figure 6E). This is reflected in the isoform fraction, which increases from 23.5% to 50.7%, while the known transcript decreases from 39.0% to 21.2% (Figure 5F).

NFKB1 plays a central role in immune responses, regulating the response to infections through transcriptional activation[57]. Furthermore, the Rel homology domain region is known to harbor disease-causing variants responsible for common viable immunodeficiency (CVID)[260], highlighting the importance of this domain in normal B cell function. This finding could suggest a novel regulatory mechanism of *NFKB1*.

Isoform switching in CLEC7A and OAS1 in a stimulus-specific manner

We sought to identify genes with stimulus-specific IS patterns. We identified an IS in *CLEC7A,* which codes for Dectin-1, a receptor that recognizes fungal glucans, triggering the immune response[261]. While this gene was not differentially expressed upon pathogen stimulation, we did identify an IS specific to *C. albicans* stimulation in this gene, involving a decrease in expression of an NMD sensitive transcript, with the increase in expression of the canonical coding transcript and a non-coding transcript (Supplemental figure 9A). In contrast, IR loss in *CLEC7A* was previously identified as a result of stimulation with multiple pathogen stimuli in monocytes. Additionally, no difference in gene expression levels was found between IAV-stimulated and resting cells, where the change in splicing was most pronounced[262]. While we find IR loss in *CLEC7A* specifically upon *C. albicans* stimulation, this could therefore also indicate a shared transcriptional response to pathogen stimuli.

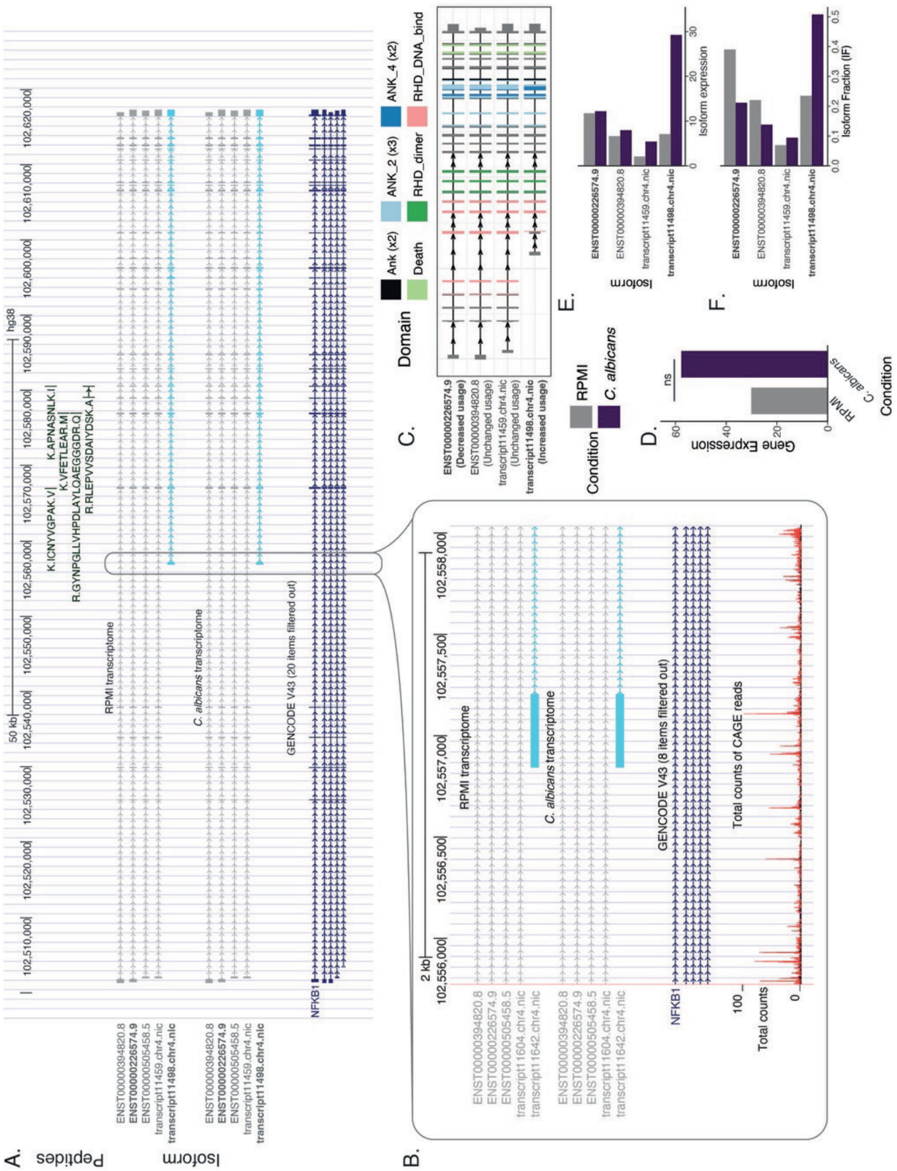


Figure 6: A novel transcript of *NFKB1*. A) UCSC genome browser track of the transcripts detected in the control condition (RPMI) and stimulated conditions. The novel transcript is presented in light blue. Known transcripts in GENCODE are presented below. B) Zoomed view of the transcription start site of the novel transcript with CAGE peaks (monocyte) in this region. C) Representation of the domains in the known and novel *NFKB1* transcripts that were detected. D) Gene and transcript expression and isoform fraction of the *NFKB1* transcripts that were detected.

Additionally, we identified an IS involving *OAS1*, which is involved in antiviral immunity. This gene is differentially expressed in response to *C. albicans, S. aureus* and Poly(I:C) (Supplemental figure 9B). We identified an IS in this gene resulting in an intron loss and a domain gain. This IS was only observed for Poly(I:C) stimulation, potentially indicating this transcript is necessary for antiviral immune responses (Supplemental figure 9C). Previous work has identified common *OAS1* haplotypes responsible for a decrease in protein abundance through the expression of NMD sensitive transcript p42, which contributes to COVID-19 severity[263]. We find this transcript to be downregulated upon Poly(I:C) stimulation. However, the transcript we find to be upregulated lacks the prenylation site needed for antiviral function, as shown for p46[264].

Detecting secreted peptides

We sought to obtain evidence of the protein-coding potential of novel transcripts found through long-read RNA sequencing. Mass spectrometry was performed for 30 secretome samples from five donors' stimulated PBMCs, which includes the samples from the individual for which long read RNA sequencing was performed (see methods). These include 2 control samples and 1 of each 24-hour pathogen stimulation condition for each individual.

We designed a search database comprising all proteins that we suspected could be in the sample. This includes the GENCODE human proteome, the proteomes of the pathogens used, as well as ORFs derived from novel transcripts found using long-read RNA sequencing. Novel transcripts do not always correspond to novel ORFs; 32% of the novel transcripts had an ORF that was present in the GENCODE reference database (Supplemental Figure 10). In the collection of 30 samples, a total of 38,703 peptides from 15,964 proteins were identified. We found 404 (7.37%) of identified proteins were known to be secreted according to the human protein atlas, which constitutes a significant enrichment (OR=2.12, $p=3.88x10^{-21}$, Fisher's exact test). We did not detect microbial proteins in the samples. Many of the novel ORFs predicted from the transcriptome have high similarity to GENCODE ORFs, resulting in a small number of novel peptides that could uniquely identify these. After rigorous filtering, we were unable to confidently identify peptides that mapped uniquely to the predicted novel ORFs.

Wider deviations in expression in the secretome

To assess whether differences in transcript expression resulted in differences in the amounts of secreted proteins, we performed a label-free quantification of the proteins in the cells' supernatants. Using PCA, we found that a large portion of variation in the proteome was explained by inter-individual differences and that these differences were larger than the differences induced by the immune stimuli (Supplemental figure 11).

We found a total of 418 differentially expressed proteins (DEPs) between the stimuli and control when controlling for individual variation. Differential protein expression was not equally distributed between stimuli with over a third (N=131) of the DEGs unique to Poly(I:C) stimulation (Supplemental figure 12A). With the exception of the *S. aureus* condition, more proteins were significantly downregulated than significantly upregulated in the secretome (Supplemental figure 12, Supplemental table 20). We found few overlapping proteins per condition, which could indicate either a high specificity in response to different pathogens or a lack of protein secretion in a subset of samples.

To determine which explanation is more likely, we visualized the specific (groups of) proteins associated with response stemming from the stimuli. We clustered protein expression values normalized by individual and stimulus (Figure 7A, Supplemental table 21). The clustering revealed a separation between poly(I:C) samples and the rest of the stimuli. *C. albicans* showed a large overlap with poly(I:C) in the protein expression profiles. Some *C. albicans* samples were grouped with poly(I:C) samples, which confirms the results from the differential protein expression analysis (34 common DEPs, Supplemental figure 12A). Other stimulus conditions could not reliably be separated from RPMI.

We identified a cluster of proteins that are highly expressed in Poly(I:C) and *C. albicans* (cluster 4, Figure 7A). This group of proteins is enriched for genes with functions in leukocyte migration and chemotaxis, exemplified by neutrophil migration. We identified further enrichments of gene sets involved in the response to IL-1, humoral antimicrobial response, and cellular responses to LPS and type II interferons. Analysis of the molecular functions of these genes indicated an enrichment of cytokine activity and receptor binding, GPCR receptor binding and various catalytic functions, likely due to immune cell differentiation and immune responses involving the degradation of extracellular matrix proteins during immune cell migration[265] (Figure 7B, Supplemental table 22). We further assessed the proteins in cluster 4 through a gene network analysis (Figure 7C, Supplemental table 23-24). Of the 84 proteins in this network, 61 were differentially expressed on the protein level (72.6%, any condition). Of these DEPs, 18 are involved in cytokine signaling (29.5%), of which 13 genes are chemokines (71.2%). A high proportion of proteins are found in the extracellular region (n=47, 77.0%), for instance through secretion in granules. The biological functions of the DEPs in cluster 4 reflect those found for the complete set of proteins in cluster 4, mainly corresponding to pathways associated with functions in neutrophil migration and chemotaxis (Supplemental table 25). As these pathways are not necessarily specific to these two stimuli, this may indicate Poly(I:C) and *C. albicans* may be more effective at eliciting differential protein secretion or have less delay in secretion compared to the other stimuli.

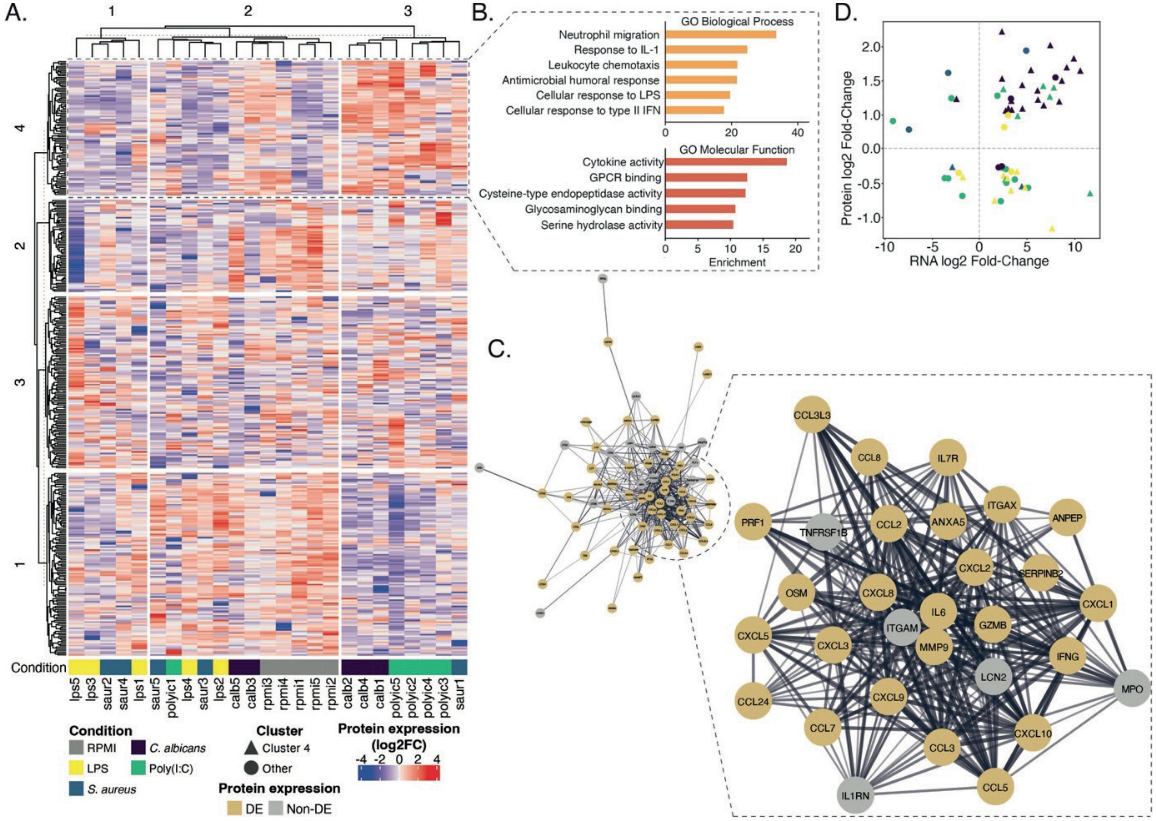


Figure 7: Protein expression in the secretome. A) Normalized protein expression detected from five donors in the five conditions. Donors are denoted with numbers 1 through 5. Clusters originating from kmeans clustering are shown in the heatmap. B) Gene ontology Biological process (top) and molecular function (bottom) pathways of proteins found in cluster 4. C) Clustering of genes found in cluster 4. Genes are found to be differentially expressed on the protein level are in color, others greyed. D) Fold change of gene differential expression versus fold change of protein DE for genes that were differentially expressed on both levels, colored by stimulus. Triangle-shaped points correspond with cluster 4 genes from A. Genes with concordant protein and RNA expression are in the upper right and lower left quadrants.

Comparison with RNA expression

As established earlier, a multi-omics approach is currently the best way to understand the human immune response. Correlation between the RNA and protein levels, or lack thereof, can provide important clues about the host response to pathogens. To assess the correlation of differential gene and protein expression levels, we assessed the concordance of differential expression on the RNA and protein level. This metric corresponds to the percentage of genes for which differential expression on both levels matched in directionality (out of all genes where DE was observed on both levels) (Figure 7D, Supplemental table 26).

We observed an overall poor concordance of directionality and fold change of expression levels at the RNA and protein levels in the different stimulus conditions, with the exception of *C. albicans* with 73% overall concordance. We overlaid the genes in group 4 from our clustering analysis with the genes found to be DE on both RNA and protein levels. There was an overrepresentation of the genes in this cluster in the total group of dual-level DE genes (OR=6.99, p=4.813e-16). Further analysis of concordant differential expression matches arising from proteins in cluster 4 (triangles in Figure 7D), we observed high concordance in the genes induced by *C. albicans* and/or Poly(I:C). Directionality concordance for Poly(I:C) and *C. albicans* for genes in in cluster 4 was significantly higher than overall directionality concordance (p=0.0313 Poly(I:C), p=0.0003 *C. albicans*, Fisher's test one-tailed). The cluster 4 proteins in the LPS and *S. aureus* conditions are in the lower right quadrant, indicating that the increase of RNA translated into a decrease of secreted proteins for these genes (Figure 7D).

We hypothesized in the IS analysis that a major regulatory mechanism in the host response to pathogens was the loss of intron retention for rapid protein generation. We cross-referenced the secreted proteins to support this conjecture. By overlapping upregulated isoforms from intron retention loss events, we found 20 cases from 7 genes (Supplemental table 27). Of these genes, 2 were upregulated on the protein level, supporting our hypothesis. The genes were *GZMB* and *B2M*, which are important immune-regulatory genes that are both secreted[266,267]. Considering the remaining 5 genes that were *downregulated* on the protein level, however, this is not convincing evidence that intron retention loss in general provides a rapid increase of protein production.

Discussion

The identification of novel transcripts and subsequent production of additional protein isoforms could help identify molecular mechanisms that play a role in various biological processes, including immune responses. Various immune system processes have previously been found to be regulated by alternative splicing [268,269] [270] [271]. Immune responses display significant inter-individual differences. Donor-specific effects such as sex and ancestry have been shown to significantly influence the transcriptome. Previous studies have further shown the impact of QTLs in the heritability of cytokine production capacity[202,272–274]. However, the effect of these processes on host defense mechanisms against pathogens, together with the large inter-individual differences in transcription and protein expression, remain to be elucidated.

We have generated a long-read transcriptome of pathogen-challenged primary immune cells (PBMCs) together with the secreted proteome to investigate mechanisms underlying immune responses during infection. We described the accurate identification of known and novel transcripts in both control and pathogen-challenged conditions. Of these transcripts, we identified a subset that is differentially expressed as a result of pathogen stimulation, which we validated by short read RNA sequencing data (including 4 additional individuals) and publicly available CAGE data from neutrophils.

We examined isoform switching that occurred as a result of pathogen stimulation, insight into transcripts that may play a role in pathogen responses. On a genome-wide level, widespread intron retention losses were observed. Retained introns that rendered the transcript unusable in the control condition were spliced out as a result of microbial stimulation; a trend we observed in all conditions regardless of microbe. We postulate that these are examples of unproductive splicing in unstimulated cells switching to productive splicing after stimulation enabling fast production of proteins relevant for the immune response. Genes that undergo intron retention loss mainly have functions in mRNA splicing and processing and in immunity. Tissue- or cell-type specific unproductive splicing has been widely observed as an autoregulatory process for mRNA splicing factors[275], which is supported by our data in immune cells. We were however not able to confirm changes in protein expression of genes that underwent IR losses using our secretome proteomics data, likely because these proteins are not generally secreted. A couple of pertinent examples have been illustrated in greater detail. We identified an IS specific to the viral stimulus that involves a novel read-through transcript of *CASP1* and *CARD16*. We found an instance of IS to a novel *NFKB1* transcript with a shortened DNA binding domain that was found in all four conditions. Additionally, we describe IS in *CLEC7A* and *OAS1* for *C. albicans* and Poly(I:C), which highlight stimulus-specific alternative splicing. Taken together, these results highlight the potential for long-read sequencing to accurately resolve novel transcripts with potential relevance in immune responses, including intron retention loss events that are generally difficult to detect using short-read sequencing.

The extent to which conclusions can be drawn about immune response mechanisms is limited by the low sample size for long-read sequencing. In this explorative study meant to provide insights into the novel technical possibilities utilizing latest sequencing approaches, we generated long-read sequencing data for only a single individual because of the expensive nature of this technology in combination with the required sequencing depth and the number of conditions studied. This design did not allow us to investigate the inter-individual differences in the transcriptome. Novel transcripts that were detected could thus be specific to this individual. Future follow-up including the sequencing of more individuals using accurate long-read sequencing methods and functional studies could provide additional insight into the more general relevance of these transcripts in immune responses. This study focused on the appraisal of the transcriptome and proteome in PBMCs, which consist of multiple cell types. Use of freshly isolated PBMCs accurately represents the complete immune cell population in the peripheral blood and allows for communication between cell types during pathogen stimulation, thereby potentially giving an accurate representation of this cell population *in vivo*. However, no information on cell type specificity of transcripts is available. This could be resolved by recent developments in single cell long-read sequencing[276].

The proteome, in contrast, was generated for all samples from all 5 individuals and highlighted significant differences between the secretome of individual donors, before and after response to immune stimuli. Concordance between the transcriptome and proteome levels was high in Poly(I:C) and *C. albicans*, and lower in LPS and *S. aureus*. We found that genes with high correlation on the RNA- and protein levels form a cluster of protein expression, separating the former two stimuli from the latter. These proteins are enriched for secreted immune-related proteins, indicating that pathogen stimulation successfully led to secretion of relevant proteins. This would indicate that cells have responded faster to the Poly(I:C) and *C. albicans* stimuli than to the LPS and *S. aureus* stimuli, because RNA and protein were isolated simultaneously from our samples. Delay in protein production after expression of an mRNA may partially explain the lack of correlation of differential expression on RNA and protein level. This delay is presumably even longer in the secretome as proteins need to be first produced and subsequently secreted[277].

We focused our study on the secretome to reduce the complexity of the protein mixture analyzed, and to obtain better peptide coverage of the secreted proteins that play an important role in immune signaling. However, this limited our view on the complete proteome affected by immune stimuli. Also, there is the added complication that only a small number of peptides exist that could discriminate between proteoforms. To detect the proteoforms derived from our long-read sequencing data, much deeper shotgun proteomics must be performed[278]. These limitations are reasons why no evidence of novel transcripts could be validated with the proteome.

Multi-omics approaches are a promising method to further our understanding of immune responses. Our study scratches the surface of biological insight to be reaped from a combination of multi-omics and long-read sequencing data and was hindered only by the aforementioned limitations in the samples themselves. Removing these limitations will undoubtedly result in deeper mechanistic understanding and will translate into better outcomes for patients. Insights gained from this methodology can be used immediately in rare disease diagnostics applications, such as the reannotation of variants using more accurate reference transcriptomes for specific tissues[279], contributing to the development of more personalized medicine.

Acknowledgements

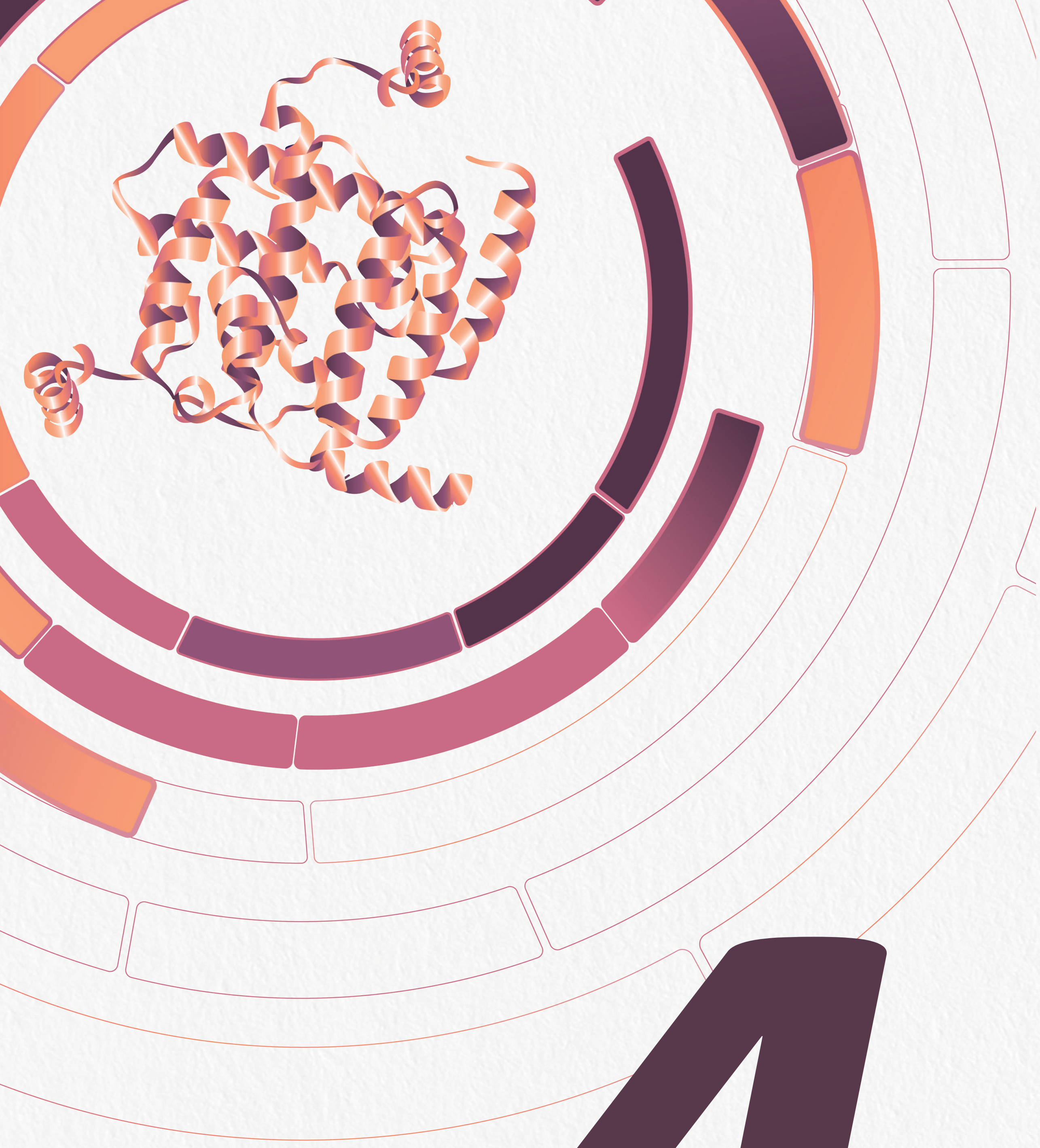
MGN was supported by an ERC Advanced Grant (#833247) and a Spinoza Grant of the Netherlands Organization for Scientific Research. AH was supported by the Solve-RD project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 779257.

Author contributions

Conceptualization: RS, PJV, AH, PACH; data curation: RS; formal analysis: RS; investigation: RS, EEV, CIM, TVR, TH; resources: SK, MS; software: RS, TVR; supervision: MM, MGN, PJV, AH, PACH; visualization: RS, EEV; writing—original draft: RS, EEV; writing—review and editing: All authors. All authors read and approved the final manuscript.

Supporting information

https://www.biorxiv.org/content/10.1101/2023.09.13.557514v2.supplementary-material



4

Chapter 4

SUsPECT: A pipeline for variant effect prediction based on custom long-read transcriptomes for improved clinical variant annotation

Renee Salz, Nuno Saraiva-Agostinho, Emil Vorsteveld, Caspar I van der Made,
Simone Kersten, Merel Stemerdink, Jamie Allen, Pieter-Jan Volders, Sarah E Hunt,
Alexander Hoischen, Peter AC ’t Hoen

Published in BMC Genomics (2023)

Abstract

Our incomplete knowledge of the human transcriptome impairs the detection of disease-causing variants, in particular if they affect transcripts only expressed under certain conditions. These transcripts are often lacking from reference transcript sets, such as Ensembl/GENCODE and RefSeq, and could be relevant for establishing genetic diagnoses. We present SUsPECT (Solving Unsolved Patient Exomes/gEnomes using Custom Transcriptomes), a pipeline based on the Ensembl Variant Effect Predictor (VEP) to predict variant impact on custom transcript sets, such as those generated by long-read RNA-sequencing, for downstream prioritization. Our pipeline predicts the functional consequence and likely deleteriousness scores for missense variants in the context of novel open reading frames predicted from any transcriptome. We demonstrate the utility of SUsPECT by uncovering potential mutational mechanisms of pathogenic variants in ClinVar that are not predicted to be pathogenic using the reference transcript annotation. In further support of SUsPECT's utility, we identified an enrichment of immune-related variants predicted to have a more severe molecular consequence when annotating with a newly generated transcriptome from stimulated immune cells instead of the reference transcriptome. Our pipeline outputs crucial information for further prioritization of potentially disease-causing variants for any disease and will become increasingly useful as more long-read RNA sequencing datasets become available.

Background

The advent of next-generation sequencing (NGS) and the exponential increase in human genomes sequenced has caused a similarly strong increase in the number of genetic variants detected. The identification of novel genetic variants has outpaced the understanding of their functional impact. Since only a small fraction of all observed variants can be characterized clinically or by functional tests, there is a heavy reliance on computational methodology for prioritization. Several computational methods predict the effect of genetic variant effects on function such as PolyPhen-2 [39], SIFT [280], and MutPred2 [281]. Variant annotators such as the Ensembl Variant Effect Predictor (VEP) [282] and ANNOVAR [283] predict molecular consequences and integrate reference data and pathogenicity scores from different resources including dbNSFP [284].

Short-read RNA sequencing has provided us with the majority of knowledge we currently have about the transcriptome, but has some intrinsic limitations when it comes to discovery of alternative transcripts [56,285]. Short read RNA sequencing is done on transcript fragments and the assembly into full-length transcripts is far from perfect, which has resulted in an incomplete reference transcriptome [286]. Long-read sequencing allows for the accurate elucidation of alternative transcripts [287] and long-read RNA sequencing datasets are proving that the human transcriptome has much more diversity than previously thought [51,288,289]. In addition, both short and long-read sequencing have shown that gene expression is highly variable in a context dependent manner, with divergent expression of transcripts expressed under different conditions (infection, stress, disease) or in different tissues or cell-types [213,290–292].

Some newly discovered transcripts result in open reading frames (ORFs) coding for novel proteoforms [236,293,294]. Knowledge on novel ORFs is key to predicting functional consequences of variants within them. There are several computational methods available to predict ORFs of these novel transcripts either based on sequence features [114,295,296] or homology to existing protein coding transcripts [111,297,298]. The prediction of ORFs on novel sequences is an essential first step for the detection of new proteoforms, as mainstream proteogenomics technologies for the discovery of proteoforms rely on databases with peptide sequences present in the predicted ORFs. Transcripts derived from long-read sequencing can provide better predictions of (novel) proteoforms (Figure 1).

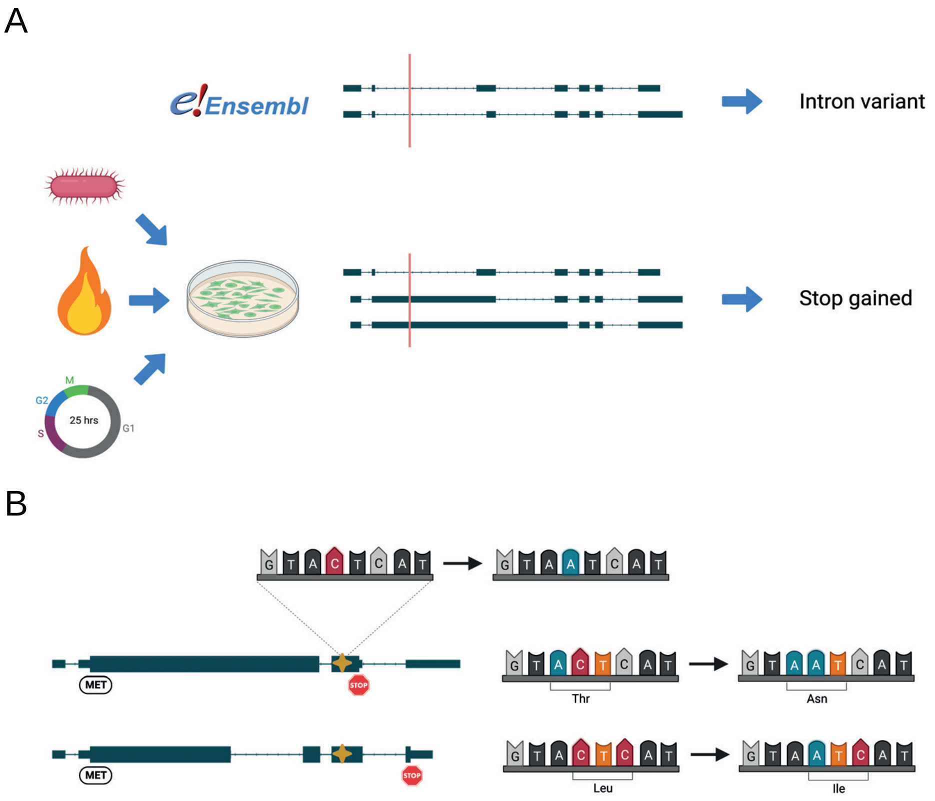


Figure 1: Premise for the creation of SUsPECT. A) Some pathogenic variants may be missed without actual information about all alternative transcripts expressed in a relevant sample. A variant in a particular genomic position may be incorrectly predicted to be non-deleterious. B) A variant at the same genomic position may cause a different missense variant in different transcript structures due to varying open reading frames per transcript.

Current variant annotation tools do not take full advantage of the knowledge of novel transcripts because they work with precalculated pathogenicity scores calculated with respect to a fixed set of reference transcripts. This necessitates manual evaluation of the functional effects of variants on alternative proteoforms, since disruption of their function may have implications for clinical diagnosis and treatment. The pipeline presented here, SUsPECT (Solving Unsolved Patient Exomes/gEnomes using Custom Transcriptomes), is designed to leverage cell/tissue-specific alternative splicing patterns to reannotate variants and provide missense variant functional effect scores necessary for downstream variant prioritization. This pipeline was designed to be generalizable to any type of rare disease variant set paired with a relevant (long-read) transcriptome.

For example, a researcher interested in annotating variants in a patient with a rare intellectual disability could consider using this tool along with a brain transcriptome dataset. We demonstrate the usefulness of this tool by reannotating ClinVar variants with a newly generated immune-related long-read RNA sequencing dataset.

Results

Analysis pipeline overview

We developed SUsPECT to reannotate variants using custom transcriptomes (Figure 2). This pipeline takes a custom transcriptome (GTF file) and a VCF file as input and returns a VCF file with alternative variant annotations for downstream evaluation and prioritization. SUsPECT predicts the ORFs in the alternative transcripts, calculates the molecular effects of the input variants with respect to these transcripts and predicts the pathogenicity of missense variants in the alternative proteoforms. SUsPECT displays subsets of variants predicted to have more severe effects when based on the custom transcriptome instead of the reference transcriptome. The predicted molecular consequences can be one of five severity levels, ranging from “modifier” to “high” (Figure 2A). A schematic overview of the pipeline is presented in Figure 2B. The main steps in the pipeline are:

- Validate pipeline input, including 1) an assembled (long-read) transcriptome in GTF format with novel transcripts. A long-read transcriptome assembly tool such as TALON will output a suitable file. 2) A VCF containing patient(s) variants.
- ORF prediction is performed on the transcripts that are not present in the human reference transcriptome.
- Ensembl VEP predicts molecular consequence annotations based on the user-provided set of transcripts/ORFs. Variants considered as missense in the user-provided transcriptome are reformatted and submitted to Polyphen-2 and SIFT.
- Polyphen-2 and SIFT calculate functional effect scores. These are reformatted and incorporated into the final VCF annotation file.
- A sub-list of variants that have a more severe molecular consequence in the custom transcriptome are provided in tabular format.

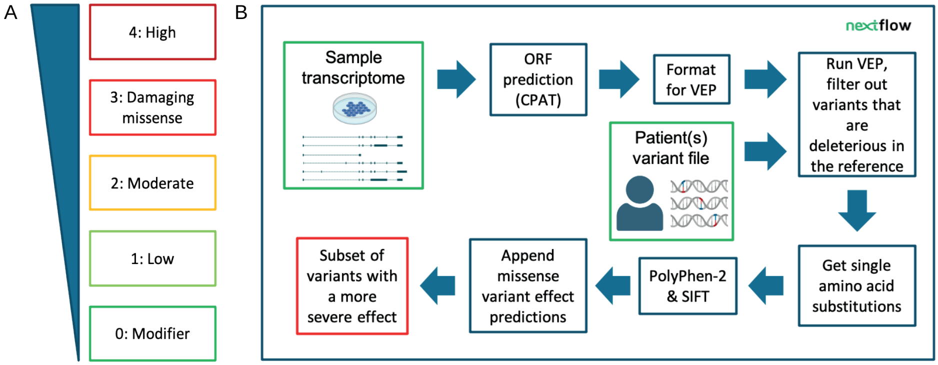

A
4: High
3: Damaging missense
2: Moderate
1: Low
0: Modifier
B
nextflow
Sample transcriptome
ORF prediction (CPAT)
Format for VEP
Run VEP, filter out variants that are deleterious in the reference
Patient(s) variant file
Get single amino acid substitutions
PolyPhen-2 & SIFT
Append missense variant effect predictions
Subset of variants with a more severe effect

Figure 2: Reannotation with SUsPECT. A) Defining "more severe". The five categories of severity are modifier, low, moderate, damaging missense and high. We consider levels 3 and 4 to be deleterious, and thus potentially pathogenic. B) The schematic of the pipeline.

A long-read sequencing transcriptome of stimulated peripheral blood mononuclear cells

We have generated long-read sequencing data on atypical, *i.e. in vitro* stimulated samples - provoking a strong expression response, to illustrate the use of the pipeline. We chose this dataset to exemplify less-studied tissues/conditions because novel transcripts are more numerous in these samples and SUsPECT is most likely to yield interesting results when the input transcriptome has many novel transcripts. Our custom transcriptome is based on long-read transcript sequences related to host-pathogen interactions and is derived from human peripheral blood mononuclear cells (PBMCs) exposed to four different classes of pathogens. We combined the transcript structures of all four immune stimuli and control samples for the reannotation. We identified a total of 80,297 unique transcripts, 37,434 of which were not present in the Ensembl/ GENCODE or RefSeq reference transcriptomes. Relative abundances of novel transcripts were lower than of reference transcripts (Suppl. Figure 1). The custom transcriptomes resulted in prediction of 34,565 unique novel ORFs passing CPAT's coding capacity threshold. The majority of transcripts had at least one ORF predicted (Suppl. Figure 2).

Reannotation of ClinVar variants

Variants may be predicted to have a more severe molecular consequence in novel (non-reference) transcripts, but the functional and ultimately clinical implications remain unclear. To demonstrate that SUsPECT can suggest new candidate pathogenic variants associated with clinical outcomes, we reannotated ClinVar variants. ClinVar contains variants with clinical significance asserted by different sources. We hypothesized that

ClinVar variants that were annotated as pathogenic and not predicted to be deleterious with the reference transcript annotation, but predicted deleterious with a (relevant) sample transcriptome, would support the utility of this pipeline.

We tested SUsPECT on a recent ClinVar [299] release (April 2022), excluding all variants that were annotated in ClinVar to be (probably) benign. We compared the predicted severity of the 776,866 variants using our custom transcript annotation versus the reference. After applying filters as described in the Methods section, 1,867 candidate variants remained. Of these variants, 145 were associated with monogenic immune-related disorders (Suppl. Table 1), which is significantly more than expected by chance (odds ratio=5.46, p=1.51x10^{-55}, Fisher's exact test). This could indicate that annotation with an immune-relevant transcriptome is better suited for the identification of variants with an impact on immune function than annotating with a reference transcriptome. The strongest argument for the utility of this pipeline can be made with variants that are curated in ClinVar to be pathogenic rather than those of uncertain significance. After excluding variants of unknown significance (VUS) from the full candidates list, there are 90 variants remaining, of which 5 immune-related. These 90 variants had an enrichment of severity level 4 events (Suppl. Figure 3). An overview of the number of variants remaining after the different filter steps is given in Suppl. Figure 4.

Five immune-related variants curated in ClinVar to be pathogenic were reannotated from a low severity molecular consequence in the Ensembl/GENCODE and Refseq transcript set to a moderate or high severity in our transcriptome (Table 1). Two were missense variants in the custom annotation and three were start-loss/stop-gain. We visualized the variants in the context of the transcript structures/ORFs on the UCSC genome browser. Two examples can be seen in Figure 3. The variant in *IFNGR1* (dbSNP identifier rs1236009877) is associated with IFNGR1 deficiency. It is curated by a single submitter in ClinVar as 'likely pathogenic' using clinical testing. Annotation of the variant with reference transcripts results in a low severity (intronic variant) result, but results in a stop-gain variant (high severity) when annotating with our transcriptome. Our custom transcriptome contained multiple novel transcripts with a retained intron at the site of the variant, but only 1 of these transcripts had a predicted ORF in this intron. The particular transcript affected by this stop gained variant was found in all samples sequenced with minimum 3 and up to 10 supporting reads, indicating that it is unlikely an artifact. The predicted ORF extended 30 base pairs into the retained intron in the region of this variant. It was the most probable ORF for that transcript with a coding probability by CPAT of 0.934.

Table 1. Five ClinVar pathogenic immune-related variants annotated as low severity in the reference transcript set but high severity in the custom transcriptome.

Variant	Location GRCh38	Allele	Gene	Consequence reference	Consequence custom	ClinVar condition	ClinVar evidence
rs80358236	1:172665641	C	*FASLG*	In-frame deletion	Start lost & in-frame deletion	Autoimmune lymphoproliferative syndrome	No assertion criteria provided. Citation; PMID: 8787672. No functional evidence.
rs1573262398	2:97724319	T	*ZAP70*	Benign missense	Missense (unknown)	Combined T and B cell immunodeficiency	Criteria provided, single submitter. No functional evidence, no citation
rs113994173	2:97733464	A	*ZAP70*	Intron	Missense (unknown)	Combined immunodeficiency due to ZAP70 deficiency	No assertion criteria provided. Citation; PMID: 20301777. No functional evidence.
rs387906763	2:190999647	G	*STAT1*	Benign missense	Start lost	Immunodeficiency 31C	Criteria provided, single submitter. Citation; PMID: 21727188. No functional evidence.
rs1236009877	6:137203727	A	*IFNGR1*	Intron	Stop gained	Immunodeficiency 27A	Criteria provided, single submitter. No functional evidence, no citation.

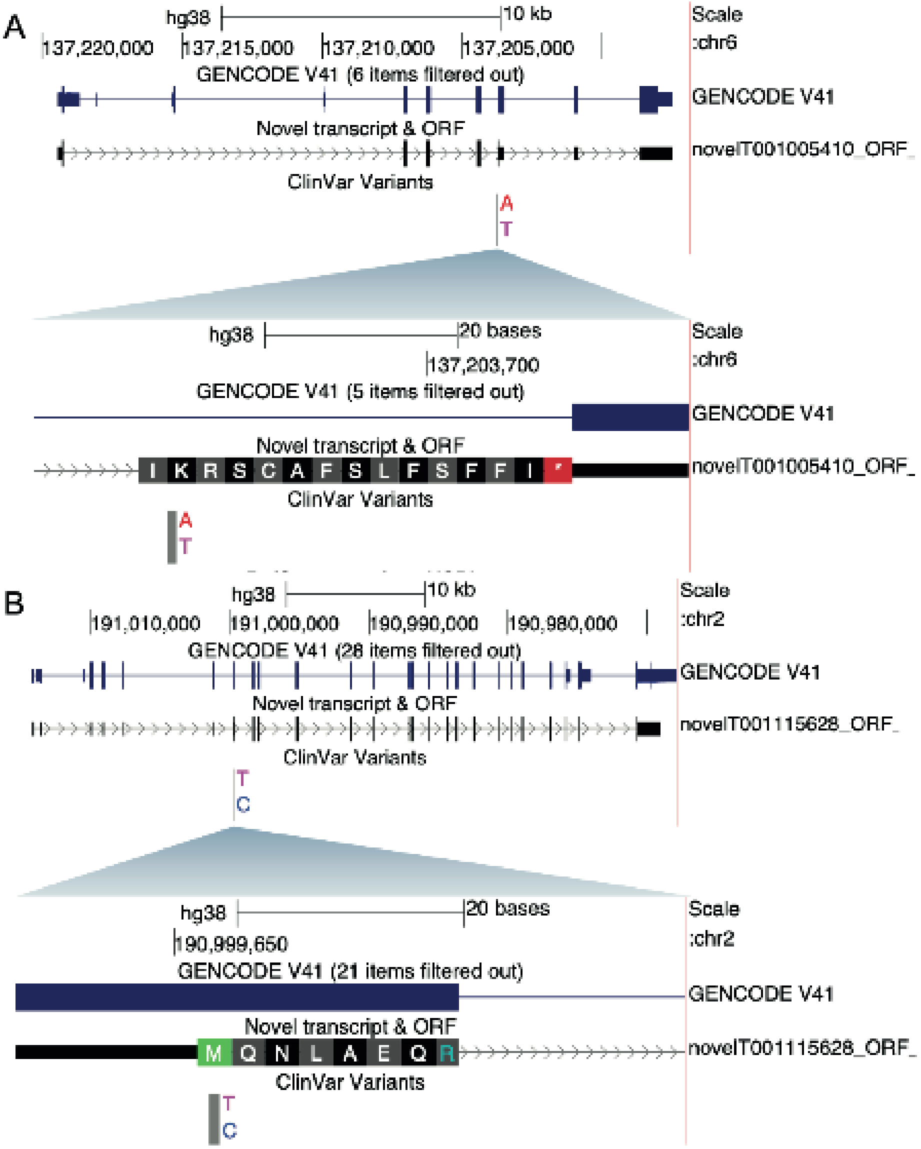


Figure 3: Two examples of ClinVar pathogenic variants being reannotated. Both variants were considered low severity variants when using hg38 reference transcriptome to annotate. A) IFNGR1 whole view and close-up of region around the variant. Variant causes a stop-gain effect (K>*) in the custom transcript novelT001005410. B) STAT1 whole view and close-up of region around variant. Variant causes a start loss (M>T) in the custom transcript novelT001115628.

In addition, the variant in *STAT1* (dbSNP identifier rs387906763) was pathogenic according to the LitVar [300] literature mining tool and a clinical testing submission. It is a missense variant (Tgc/Cgc) in the reference annotation that is predicted by PolyPhen-2 to be benign. However, in one novel transcript it causes an M/T substitution, leading to loss of translation start site. Further inspection revealed that the transcript affected by the start-loss was expressed in *C. albicans*, *S. aureus* and PolyIC stimulated conditions by up to 6 supporting reads, but not in the control condition. STAT1 is previously described to be involved in the immune disease (chronic mucocutaneous candidiasis) linked to this variant by weakened response to *C. albicans* [301], which is a condition where this novel transcript was expressed. The ORF affected was the most probable ORF for that transcript and had a coding probability of almost 1 by CPAT.

Discussion

SUsPECT predicts the functional consequences of genetic variants in the context of novel open reading frames predicted from a user-defined transcriptome. It is important to underline that the pipeline does not return a statement on the pathogenicity of variants. The pipeline simply brings new candidates forward for further interpretation; the user may choose to cross-reference the clinical phenotypes of the patients with the functions of the genes that the patients' variants are found to disrupt. In our use case, ClinVar variants were used as they already have widely accepted annotations. However, 40% of ClinVar variants are of unknown significance, some of which are suspected to have some impact on clinical phenotype. Nearly 2% of these variants changed rating to be predicted as deleterious in our reannotation. As more people generate sample-specific transcriptomes to annotate variant sets, an increasing number of VUS may be classified as benign or deleterious.

Alternative splicing is known to increase the proteomic diversity, but it is less well understood how the novel transcripts contribute to the diversity of proteoforms and their function, and how these are impacted by genetic variants [302–305]. One of the most commonly used variant annotators, Ensembl VEP, predicts molecular consequences for variants in custom transcripts in standard formats, but lacks functional effect predictions for missense variants in those transcripts. Considering the well-established importance of missense variants on a variety of diseases [306–308], this presents a hurdle in the reannotation of variants with a custom transcriptome data.

We observed that many missense variants were predicted to have more severe effects when annotated based on custom transcriptomes. This may be due to the numerous new ORFs. Multiple ORFs passing CPAT's 'human threshold' were often predicted per novel sequence; for our 37,434 novel transcript sequences we predicted 34,565 novel

ORFs. Some proteogenomics tools choose the 'best' ORF per sequence, but we have decided to keep all that passed the probability threshold. We do not filter out non-coding genes when predicting ORFs, because some of them may still have protein coding capacity. Missense results implicitly depend on the confidence of the ORF predictions that are produced by CPAT. New deleterious missense variants will not be relevant if the predicted protein is not produced in the cell. Coding ability of novel transcripts is an area of active research [309–311] and new techniques to identify credible ORFs may be added to the pipeline as they become available. In the meantime, it may be prudent to validate interesting candidates using targeted proteomics techniques before establishing a genetic diagnosis.

SUsPECT is flexible; it takes transcriptomes from either short-read or long-read sequencing, PacBio or Oxford Nanopore, cDNA or direct RNA, as long as novel transcripts exist in the dataset. SUsPECT may produce the most comprehensive results if the transcriptome dataset comes from patient cells or tissues that are affected by the condition under study. However, it is also possible to use existing or newly generated long-read transcriptomes from relevant cells or tissues of healthy individuals, like we have demonstrated in the current work. The modularity of the tool means its components are also adaptable. The module that reads input can be updated as new (long-read) transcript analysis tools become available, which is useful considering new tools are actively being developed [223]. Its modularity facilitates incorporation of other functional effect prediction tools [312–315] than the currently implemented PolyPhen-2 and SIFT software. The current implementation and future extensions of SUsPECT may thus contribute to increase the diagnostic yield for disorders that are associated with transcripts expressed in specific tissues or under specific conditions.

Conclusions

The full complexity of the human transcriptome is not represented in the current reference annotation. Analysing variants using alternative transcripts may aid in explaining missed genetic diagnoses, especially when disease or tissue-specific transcripts are used. SUsPECT puts genetic variants in the context of alternative transcript expression and can contribute to an increase in diagnostic yield. We used missense variants with ClinVar assertions of pathogenicity to demonstrate the potential of this methodology and have demonstrated a higher yield of missense variants are predicted to be deleterious. The enrichment of immune-related variants after reannotation suggests there is biological significance to these findings. Thus, long-read transcriptome data relevant to the disease of interest may not only improve our understanding of the ever-growing number of genetic variants that are identified in human disease context, but also aid in diagnoses for rare and/or unsolved disease [316,317].

Methods

Severity classification

SUsPECT classifies variants according to their expected impact and their molecular consequence. Impact scores used by SUsPECT are based on the predicted molecular consequence groupings in Ensembl VEP (Figure 2A) with higher numbers corresponding to more severe consequences: zero being equivalent to “modifier”, one to “low” severity, two to “moderate” severity, and four to “high” severity. SUsPECT uses Polyphen-2 predictions to distinguish between (likely) benign (score: 2) and (likely) deleterious (score: 3) missense variants.

Additional filters for output variant list

SUSPeCT initial output is a list of variants with higher severity scores based on the custom transcriptome annotation compared to the reference annotation (homo_sapiens_merged cache version 104 which includes both Refseq and Ensembl/GENCODE transcripts). The variants that remain in the final list of “increasing severity” are filtered to retain only variants that are potentially interesting for establishing a disease diagnosis. Thus, the pipeline removes variants that are already considered deleterious based on the reference annotation, *i.e.* variants that already have scores of 3 or 4. An additional criterion was applied for missense variants. Missense variants for which the same amino acid substitution found in the custom and reference annotation are also removed. To reduce computational time further, missense variant alleles in novel sequences that are common (AF > 0.01) are removed. These filters are integrated in SUsPECT. For the use case described in this manuscript, missense variants present in the custom annotation that are predicted by PolyPhen-2 to be “benign” in both custom and reference annotation are removed. In our ClinVar example, we define “immune-related” variants as those variants that contain the string “immun” somewhere in the clinical description.

Software details

A pipeline was built to streamline the process of variant prioritization using custom transcript annotation. The pipeline is written in Nextflow [318], using Ensembl VEP as the variant annotator. Each step of the pipeline runs Singularity/Docker containers pulled automatically from Docker Hub. The input of the pipeline is the sample-specific/non-reference long-read transcriptome in GTF format, variants in a VCF file, and a FASTA file of the genome sequence. It is designed for use with output from TALON [319].

First, the GTF file is converted to BED format with AGAT v0.9.0 [320]. ORFs for any novel sequences are predicted based on the BED annotation and FASTA genome reference using CPAT v3.0.4. CPAT output is converted to BED format with the biopj python package and filtered for a coding probability of at least 0.364, which is the cutoff for human ORFs recommended by the authors of CPAT [114]. Conversion from CPAT CDS to protein FASTA is performed with EMBOSS transeq v6.5.7. This ORF BED file is combined with the BED file of transcripts to make a complete BED12 file with ORF/transcript information. Then, we convert this BED12 file to GTF with UCSC's bedToGenePred and genePredToGtf. The resulting GTF file is used for a preliminary annotation of the variants with Ensembl VEP to fetch variants predicted as missense in the custom transcript sequences. Next, variant filtering was performed as outlined in the previous section with the filter_vep utility distributed with Ensembl VEP as well as bedtools v2.30.0. The functional effect predictions from Polyphen-2 and SIFT are reformatted and one final run of Ensembl VEP (with the custom plugin enabled) integrates these predictions to the VCF. The output is the annotated VCF, as well as a VCF with the subset of variants predicted to have higher severity.

Ex vivo PBMC experiments

Venous blood was drawn from a healthy control [321] and collected in 10mL EDTA tubes. Isolation of peripheral blood mononuclear cells (PBMCs) was conducted as described elsewhere [222]. In brief, PBMCs were obtained from blood by differential density centrifugation over Ficoll gradient (Cytiva, Ficoll-Paque Plus, Sigma-Aldrich) after 1:1 dilution in PBS. Cells were washed twice in saline and re-suspended in cell culture medium (Roswell Park Memorial Institute (RPMI) 1640, Gibco) supplemented with gentamicin, 50 mg/mL; L-glutamine, 2 mM; and pyruvate, 1 mM. Cells were counted using a particle counter (Beckmann Coulter, Woerden, The Netherlands) after which, the concentration was adjusted to 5 × 10^6/mL. *Ex vivo* PBMC stimulations were performed with 5×10^5 cells/well in round-bottom 96-well plates (Greiner Bio-One, Kremsmünster, Austria) for 24 hours at 37°C and 5% carbon dioxide. Cells were treated with lipopolysaccharide (*E. Coli* LPS, 10 ng/mL), *Staphylococcus aureus* (ATCC25923 heat-killed, 1×10^6/mL), TLR3 ligand Poly I:C (10 µg/mL), *Candida albicans* yeast (UC820 heat-killed, 1×10^6/mL), or left untreated in regular RPMI medium as normal control. After the incubation period of 24h and centrifugation, supernatants were collected and stored in 350uL RNeasy Lysis Buffer (Qiagen, RNeasy Mini Kit, Cat nr. 74104) at −80°C until further processing.

RNA isolation and library preparation

RNA was isolated from the samples using the RNeasy RNA isolation kit (Qiagen) according to the protocol supplied by the manufacturer. The RNA integrity of the isolated RNA was examined using the TapeStation HS D1000 (Agilent), and was found to be ≥7.5 for all samples. Accurate determination of the RNA concentration was performed using the Qubit (ThermoFisher). Libraries were generated using the Iso-Seq-Express-Template-Preparation protocol according to the manufacturer's recommendations (PacBio, Menlo Parc, CA, USA). We followed the recommendation for 2-2.5kb libraries, using the 2.0 binding kit, on-plate loading concentrations of final IsoSeq libraries was 90pM (*C. albicans*, *S. aureus*, PolyIC, RPMI) and 100pM (LPS) respectively. We used a 30h movie time for sequencing. The five samples were analyzed using the isoseq3 v3.4.0 pipeline. Each sample underwent the same analysis procedure. First CCS1 v6.3.0 was run with min accuracy set to 0.9. Isoseq lima v2.5.0 was run in isoseq mode as recommended. Isoseq refine was run with '--require-polya'. The output of isoseq refine was used as input for TranscriptClean v2.0.3. TranscriptClean was run with '--primaryOnly' and '--canonOnly' to only map unique reads and remove artifactual non-canonical junctions of each of the samples. The full TALON pipeline was then run with all five samples together using GRCh38 (https://www.encodeproject.org/files/GRCh38_no_alt_analysis_set_GCA_000001405.15/@@download/GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta.gz). Assignment of reads to transcripts was only allowed with at least 95% coverage and accuracy. A minimum of 5 reads was required to keep alternative transcripts in the final transcript set (default of talon_filter_transcripts). GENCODE annotation (v39) was used by TALON to determine novelty of transcripts in the sample.

Authors' contributions

RS wrote the manuscript. RS and NSA developed the code for the pipeline. JA oversaw the development and organized collaboration. EV and CIvdM contributed to the interpretation of the results. MS and SK performed the library preparation for our samples. PV contributed his python package as one of the components of the pipeline. SEH, AH and PACtH supervised the study and software development and revised the manuscript draft.

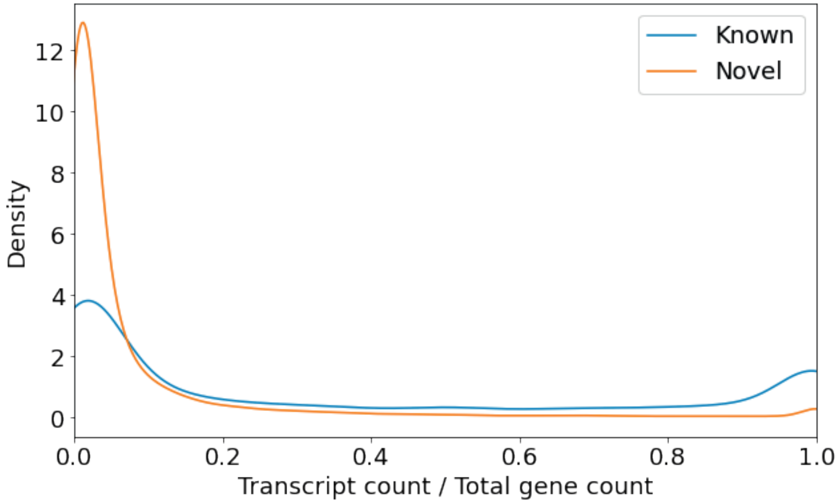
Acknowledgements

This work received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825575 (EJP RD), and the

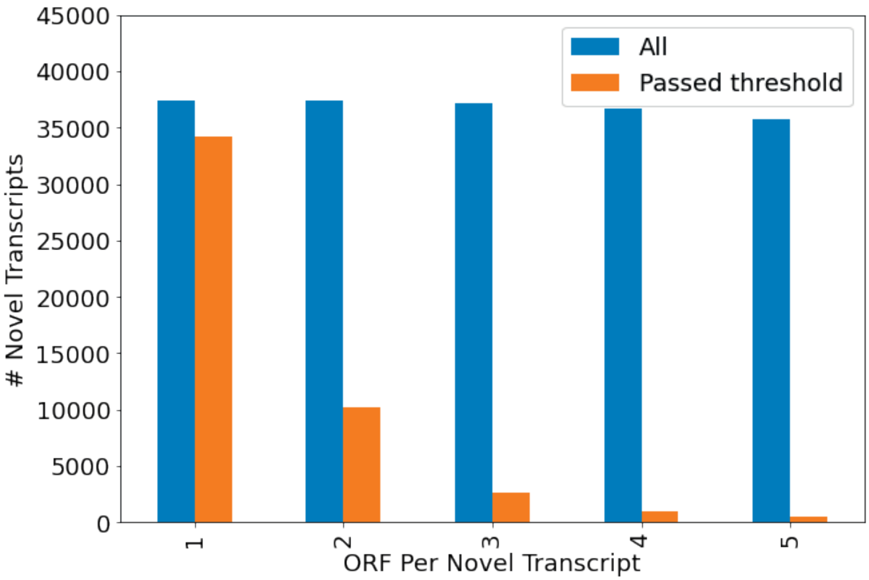
Netherlands X-omics initiative, which is funded by the Dutch Research Council under as project no. 184.034.019. The work was also supported by the European Molecular Biology Laboratory.

We would like to thank Simon V. van Reijmersdal for his contribution to the library preparation.

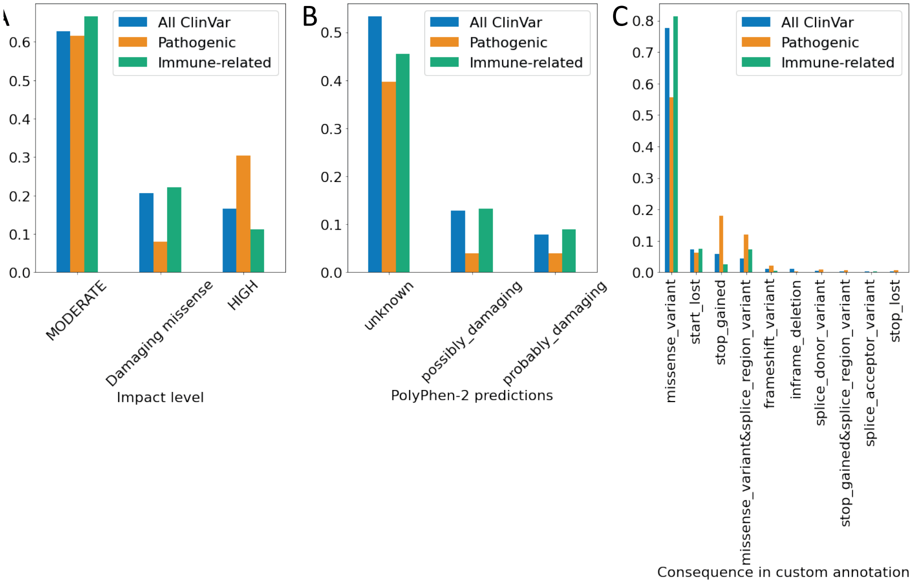
Supporting information



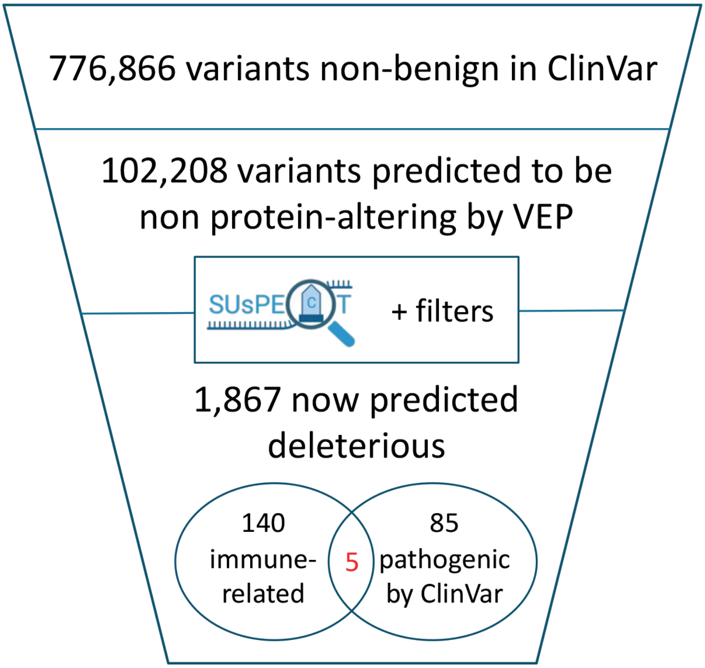
Supplementary figure 1: Relative abundance of novel transcripts relative to known transcripts. Transcript counts were summed for all 5 conditions per transcript.



Supplementary figure 2: Number of predicted ORFs passing CPAT's human coding threshold. ORFs were predicted per novel transcript and sorted by most to least likely to be coding (1 being most likely). All predicted ORFs in blue, those that passed the human coding threshold in orange.



Supplementary figure 3: Comparing subsets of variants that changed annotation from benign to deleterious. "All ClinVar" corresponds to all reannotated clinvar variants including VUS and all clinical phenotypes that were reannotated from benign to deleterious with our transcriptome (N=1867). "Pathogenic" is a subset of all reannotated variants that excludes VUS variants (N=90). "Immune-related" is a subset of all reannotated variants that includes only immune-related clinical phenotypes (N=145). A) Impact level of variants after reannotation. The impact level shown is associated with annotation in the custom transcriptome. B) PolyPhen-2 predictions of variants after reannotation. C) Specific molecular effects of variants after reannotation.


776,866 variants non-benign in ClinVar
102,208 variants predicted to be non protein-altering by VEP
SUsPECT
+ filters
1,867 now predicted deleterious
140 immune-related
5
85 pathogenic by ClinVar

Supplementary figure 4: Overview of variant prioritzation on our SUsPECT test case.

https://link.springer.com/article/10.1186/s12864-023-09391-5#Sec5

5

Chapter 5

General Discussion and Future Perspectives

Applications

Proteogenomics can shed light on important biological processes, including mechanisms underlying post-transcriptional regulation. Since many diseases disrupt or intercept these processes, the methods presented in this thesis potentially uncover those disease mechanisms too. Understanding post-transcriptional regulation has real-world, clinical relevance. In this section, some common applications of proteogenomics methodology in disease diagnostics and biomarker discovery will be outlined.

Diagnostics of rare disease

Rare diseases tend to have wide phenotypic and genetic variability, which makes them particularly challenging to diagnose. The majority (80%) have a genetic cause; NGS has helped greatly in these cases by identifying variants in an unbiased and high-throughput manner[322]. Decisions about what variants proceed through diagnostic evaluation are largely done on a case-by-case basis according to what information was available at the time. Generally, patient variants are narrowed down to a small handful that are not common in the general population and/or in genes that are involved in processes relevant to the disease, but the information about population frequencies and gene panels are being constantly updated. As a result, many rare disease patients remain undiagnosed. Definitively unraveling mechanisms of variant pathogenicity can be accomplished with functional studies[323]. In these studies, the effects of variant on proteins and biological systems can be tested in model organisms[324,325] or in patients' own cells[326,327]. The process of functional validation is time consuming, however. Recently, the introduction of protein information has shown promise in the diagnosis of these cases by revealing biochemical consequences of variants, providing a valuable bridge between early- and late-stage diagnostic processes[328–330]. SUsPECT, developed in Chapter 4, aims to replicate that bridge *in silico* to reduce the number of variants that need to undergo functional validation.

Biomarker discovery in cancer

Proteogenomics has emerged as a powerful approach in biomarker discovery, providing a holistic understanding of the molecular landscape of diseases. A biomarker is a measurable and quantifiable indicator of a biological process, condition or response to a therapeutic intervention[331]. Insights into the relationship between genetic alterations and protein expression enables the identification of novel biomarkers associated with various diseases. Variant peptides, the focus of this thesis, have great biomarker potential, as they can be used to distinguish diseased versus healthy phenotypes (diagnostics) and provide potential therapeutic targets (treatment).

The most striking example of success in proteogenomic biomarker discovery is in cancer. Full proteogenomic characterization (like in Chapter 4) has been performed on countless cancer types[332]. Considering the low correlation (0.3-0.45 range) of RNA-protein correlation reported in cancer and elsewhere[333–335], the addition of proteomics data has been a real benefit for patient stratification. Recent studies in, among others, prostate cancer[336], pediatric brain cancer[337], medulloblastoma[338], clear cell renal cell carcinoma[339] identified outcome-correlated subtypes of their respective cancers. In these studies, addition of proteomics data in addition to existing RNAseq data either contributed to or was solely responsible for the subtype definition. Post-translationally modified proteins were also included in most of the aforementioned cases; in colorectal cancer, phosphorylation separated primary tumors with metastasis from those without[340]. The defined subtypes characterized by proteogenomics will inform treatment and predict treatment response in the clinic.

Proteogenomics does not only aid in the definition of the cancer (sub)types, it also facilitates in development of the treatments. Neoantigens may arise from variant peptides arising from single nucleotide variants, intron retention and cryptic splicing. They are critical targets for immunotherapy since they are derived from tumor-specific mutations and thus presented only on cancerous tissue. Proteogenomic analyses are key for the identification and validation of neoantigens, enhancing our understanding of the tumor immunopeptidome and guiding the development of personalized cancer vaccines[341–344]. T-cell responses against tumor-specific antigens were successfully mounted using immunotherapy[345,346], and the therapy has been shown to improve long-term survival[347]. The use of proteogenomics for neoantigen detection is so pervasive that multiple software pipelines have been built for this purpose[348–352]. Cancer is one illustrative example of where proteogenomics has proved valuable for both disease subtyping and therapeutics. Proteogenomics requires large volumes of omics data. The breadth of proteogenomics research can thus be attributed to the resources allocated to cancer research; many of the studies cited here originate from cancer-related consortiums with many members involved. The consortiums are well organized and make large multi-omics datasets available to the scientific community, encouraging re-use[353].

The applications of proteogenomics in biomarker discovery extend beyond cancer; many complex and heterogenous diseases benefit from the multi-faceted picture that proteogenomics provides. In neurodegeneration, disease stratification using proteogenomics was successful with Parkinson’s disease[354] and alternatively spliced proteoforms were found differentially expressed in Alzheimer’s brains[355]. Non-canonical proteoforms are also prevalent in the human heart[356], with age-related differential expression that could be used as biomarkers in age-related heart disease[357]. Host-defense peptides, which selectively alter innate immune pathways in immune cells in response to pathogen infection, have been indicated as suitable biomarkers for both infectious and non-infectious diseases regardless of whether constituently expressed or only as response to pathogenic infection[358,359]. Proteogenomics may aid in uncovering the latter, as indicated in Chapter 3. Proteogenomics has demonstrated great diagnostic and therapeutic potential, even with methodology challenges in sequencing (short-reads) and in proteomics. Methods like those developed in this thesis (Chapter 4) are well-positioned to discover candidate biomarkers in a high-throughput manner. We are only beginning to scratch the surface of clinically-useful proteogenomics findings; fully addressing limitations in the manner described in the remainder of this chapter will increase both sensitivity and specificity of biomarkers candidates in any disease model.

Challenges

Proteogenomics, like any less well-established method, faces a variety of challenges. There is much novel biology to be found, but the success of proteogenomics is very dependent on the content of the search database. Utilizing long-read sequencing technology results in a database with fewer likely artifacts, but being a newer technology itself, faces its own set of limitations. The improvement in search database enables more novel proteoform discovery in general. However, accurate reporting of these novel protein products is important. Published results can be incorrect without established standards and may compromise our understanding of biological processes under study.

Defining transcript novelty

There are clear discrepancies in transcriptomes produced by different tools for transcriptome assembly/annotation of long-read sequencing data. Estimates for the number of unique transcripts in a dataset can vary substantially depending on the tool used, as seen in the transcriptome comparison in Chapter 2. This variation begins at sequence correction, an essential step in processing long-read sequencing data. Mismatches, indels and splice junctions are corrected by using external information such as reference annotations or high-accuracy short reads, or self-correction in a *de novo* fashion. There are numerous algorithms to perform the correction that result in different sets of corrected reads; fortunately, these algorithms have recently been evaluated and benchmarked [360]. Algorithms for assembly of transcripts based on the corrected reads have lagged behind. Several algorithms became available for transcriptome assembly of long reads during the course of this thesis, each incrementally better than the last, but with varying definitions of transcript novelty[64,65,223,319]. A proper benchmark using synthetic spike-in RNAs was published near the completion of this thesis[361]. For this benchmark, an *in silico* mixture strategy was used to generate a ground truth dataset that allowed the performance evaluation of several transcript analysis tools. Some tools performed better than others in isoform detection, but all suffered from frequent identification of artifactual isoforms. These methods will need to improve to enable more accurate observations of differential isoform expression and usage in various biological settings, such as those of the pathogen stimulated versus unstimulated case in Chapter 3[362].

"Good" ORF prediction is not good enough

Improved isoform characterization will also impact potential open reading frames (ORFs). ORF length is known to be the most important feature of coding potential prediction, and basic ORF predictors using the longest ORF per transcript perform reasonably well. Developments like CPAT, a popular ORF prediction tool, show that the addition of just a few sequence-based features lead to even better predictions[114]. Some ORF predictors were developed to specifically predict ORFs for long read transcripts (SQANTI and ANGEL used in Chapter 2), but these did not out-perform CPAT and were ultimately depreciated. In practice however, CPAT tends to predict several high confidence ORFs per transcript. While not implausible that multiple proteins are produced from the same transcript in some cases (discussed in detail below), the overabundance of ORFs predicted by CPAT has a material effect on the work in Chapters 3 and 4; it increases peptide search database size leading to less sensitive detection, and causes potentially inaccurate variant reannotations with SUsPECT. A recent long-read proteogenomics pipeline includes in-house scripts to further filter CPAT output predictions[363]. Considering the potential biological impact of novel ORFs, "good" predictions are not good enough[364,365]. Luckily, promising improvements are being made in eukaryotic ORF detection, for instance by using more contextual information[366]. The use of context relevant to protein production regulation mechanisms will continue to yield improvements in ORF prediction accuracy.

Detecting other protein products

The reduction in database size from eventual improved ORF prediction opens the possibility to shed light on the "dark proteome" (largely unexplored alternative protein products) by including it in the search database. Proteins/peptides can be produced from alternative ORFs[255]. One example are peptides originating from short ORFs, or sORF-encoded proteins (SEPs). The longest ORF is considered to be the most likely to produce protein, so SEPs are not generally included in search databases even though they have been found to be a common class of protein products that are likely to be functional[367–369]. Likewise, peptides from upstream ORFs in the 5' untranslated regions have long been known to be important translational regulators[370–372]. Proteins can also be produced from unexpected start codons, as in the case of alternative translation initiation. Ribosomes may skip AUG start codons (leaky scanning) or start translation at non-AUG start codons, depending on transcript sequence or cellular conditions[373,374]. Proteins can even be produced from transcripts that were not expected to be protein coding at all, such as lncRNA and pseudogenes[375,376]. All these alternative protein products would need to be specifically added to the database to find them, and searched for in a single step to avoid statistical irregularities. This is, however, not an attractive option due to the database size considerations. Ribosome profiling has been key technology in discovery of alternative protein products[377], and remains the best method to corroborate findings amidst low identification power from large search database size. However, it is currently too expensive and labor-intensive to be practical for use in already resource intensive genome-wide proteogenomics studies.

Quality assessment of proteogenomics findings

Assessing the quality of findings in proteogenomics is much more challenging than in classic proteomics. Novel peptide identifications from proteogenomics strategies suffer from high false negative rates when using target-decoy FDR control, but abundant false positives when reducing the cutoff[378]. Methods to circumvent the issue (detailed in Chapter 1) do not have a solid statistical basis. Without standards or validation requirements prior to publication, false positive and false negative variant peptide identifications are abundant in current literature[379,380]. Standards have been painstakingly created by the Human Proteome Organization (HUPO) for reporting peptide findings to the proteomics community[381]. However, these are not widely adopted for proteogenomics as they are considered too conservative for variant peptide detection[145]. Instead, reports of variant peptides are frequently validated in the form of manual inspection of PSMs, as done in Chapter 2. Several software tools to aid manual visual assessment have been developed[382–384]; these largely exclude AI and deep learning efforts as these deep learning algorithms are primarily used for initial spectrum matching or result re-scoring rather than perpendicular validation of individual PSMs. Manual PSM inspection has become common practice because FDR is a global quality metric that is unable to distinguish true vs false positives, and often results in incorrect peptide identifications with big proteogenomics databases[385]. Proteogenomics will eventually require its own set of standards for objective assessment of the quality of findings. A consensus on acceptable database creation/analysis approaches will be a prerequisite of creating such standards, which is itself challenging due to ongoing improvements in the field. This consensus will need to include an alternative method for FDR estimation in proteogenomics.

Biology versus detectability

With all the aforementioned challenges, underlying biological truth in regards to proteomic variation remains elusive. Most studies, including Chapter 2, detect much fewer variant peptides than could be expected with known genetic information. We observe that methodology falls short; proteogenomics remains too biased to the composition of the search database and novel computational proteomics (as defined in the introduction) suffers from too many false positives to reliably detect the variant peptides in a sample. An alternative explanation for poor detection could be their absence[386]. However, the reality may be more nuanced. A recent, comprehensive study hinted at the existence of widespread proteomic variation[278]. This study makes clear that proteomic data depth plays a crucial role in detectability, implying a lower relative abundance of protein-level variants and begging the question of their biological importance. Lower abundance does not equate to biological irrelevance[387]. The impact of low abundance proteins and proteoforms is an important topic to address (perhaps on a case-by-case basis) considering the re-annotated disease variants found in the SUsPECT test case were more often found in non-dominant transcript isoforms. Relativity aids interpretation; low abundance proteoforms may be higher abundant in certain cellular contexts or time points. While variant peptide detection challenges persist, the best policy would be to always validate any significant findings using synthetic peptides before application to a diagnostic setting or going into functional validation studies.

Improvements to come

The plethora of challenges in proteogenomics are bound to be addressed with clever solutions in the coming years. The interdisciplinary nature of proteogenomics has the advantage that improvement is possible in multiple areas. These improvements can be made in the experimental or bioinformatics methodologies. Since sequencing methods have already become quite well-established in comparison to proteomics, proteomics has some of the most exciting experimental method developments to come. The new developments will in turn lead to more data and method generation, which will need the appropriate infrastructure to be shared and accessible for the benefit of the whole scientific community.

Experimental method development in proteomics

Data independent acquisition

In terms of new technologies, proteomics will see the most innovation. There is already significant improvement in the domain of peptide spectrum acquisition. Data independent acquisition (DIA) is a method to analyze complex samples that has several significant improvements over the standard DDA[388]. DIA comprehensively targets all precursor ions in a defined mass range per run instead of DDA's stochastic selection, increasing reproducibility and reducing bias. DIA is able to capture both highly and lowly abundant peptides with its increased dynamic range as compared to DDA. However, this increase of precursor ions results in composite fragment-ion spectra which are so complex that their analysis is non-trivial. Deconvoluting the multiplexed output spectra produced by DIA is in fact the biggest challenge of the method. The use of spectral libraries became somewhat of a necessity to extract data used to identify peptides in DIA[389–393]. Typically, DDA scans from the same or similar samples are used to create these spectral libraries(Guan et al., 2020; Lam et al., 2007; F. Zhang et al., 2020), but large publicly available libraries exist for some species. Of course, this comes with the caveat that peptides that are not in the spectral library cannot be analyzed. While most publications with DIA use spectrum-library based approaches, newer library-free methods are most interesting for the identification of non-canonical peptides sought in proteogenomics[397,398].

DIA applied in a DM1 biomarker study

We attempted to use DIA data in a proteogenomics search for biomarker peptides associated with myotonic dystrophy type 1 (DM1). DM1 is an inherited neuromuscular disorder caused by a CTG repeat expansion, thereby causing abnormal RNA splicing[399–401]. Aberrant splicing as a result of the disorder could potentially yield distinct proteoforms that can be used as a biomarker for the disease. We had whole blood samples originating from 248 DM1 patients including DDA, DIA, and both short- and long-read RNA seq. Since both DDA and DIA data were available, DIA data was analyzed using two methods for comparison; once with a DDA data-based search database and one spectral-library prediction-based database. We expected novel transcript isoforms to be present in patient samples due to the nature of the disease, and there were: 7,683 novel isoforms were detected in a set of 14,053 total by long-read RNA sequencing. Despite the abundance of novel transcripts detected, no viable peptide biomarker candidates were detected. Analysis of DDA data yielded two potential novel exon/splice peptide candidates. However, one of them was in an exon that was not supported by short-read data, and the other was more common in controls than disease samples. Neither of the candidates could be found in the DIA data, regardless of the post-processing method. The discrepancy between DIA and DDA findings from the same samples demonstrates how improvements still need to be made in DIA analysis. The failure of the biomarker search project can be partly attributed to the samples themselves; whole blood is challenging to analyze relative to other tissues due to a higher dynamic range in protein abundance, and DM1 patients from which the samples were derived were only mildly affected by the disease. A recent study performing a similar DM1 biomarker search in mouse muscle tissue was successful[402], forecasting potential success for a future attempt with a different study design including targeted proteomics.

Top-down proteomics

Much information is lost when digesting a protein to peptides in LC-MS/MS protocol. Sequence similarities between proteins are common, which leads to many cases of a peptide that could originate from multiple different proteins (as seen in Chapter 2). The process of protein inference is riddled with uncertainty and error[403]. Ideally, digestion of a protein into peptides would not be necessary to identify them. Somewhat analogous to long-read nucleotide sequencing versus short reads, top-down proteomics is a detection method to characterize whole proteins. Intact protein undergoes fragmentation instead of its peptides, eliminating the necessity for protein inference. Top-down proteomics unfortunately is much more challenging to execute in practice due to data complexity and technical limitations[404,405]. Updates to instrumentation and protocol (mainly in protein separation) are ongoing[406–409]. We may yet see top-down proteomics become a more viable option to observe proteomic diversity.

Nanopore-based peptide sequencing

Reading out amino acid sequences in the same way NGS reads out nucleotide sequences is currently not possible. However, attempts to repurpose ONT sequencers into peptide sequencers have shown some recent success. Discerning the electrical signals of the 20 distinct amino acids is a considerable challenge; amino acid sequences are heterogeneously charged unlike nucleic acid sequences, and thus do not translocate neatly through the pore. Tackling this challenge required creative engineering. Several strategies were proposed to unfold whole proteins prior to translocation through a nanopore[410–413]. Pores were fitted with blockades to keep amino acids in the pore long enough to detect the differences between them[414]. The newer methods employ DNA-peptide conjugation to encourage movement through the pore[415–417]. The newest development made near the completion of this thesis is a nickel ion-modified nanopore (specifically *Mycobacterium smegmatis* porin A) able to distinguish all 20 amino acids. Peptide sequence reading using nanopores still has a long road ahead, but will slowly become a reality. If it does live up to its promise, it will be a feat of engineering that will render peptide database searching largely unnecessary.

Bioinformatic improvements in proteogenomics

Towards a complete database

The working assumption in proteogenomics is that the reference database is incomplete. Research questions relating to samples that come from e.g. a less well-studied tissue, species or a disease-affected individual may require sequencing/proteomics data generated from that specific sample. Using comparable samples originating from other labs or tissues can be informative in some cases, and save resources. Similar samples will become increasingly abundant over time. Increasing accessibility and accuracy of long-read sequencing will lead to generation of more (publicly available) datasets, more proteogenomics studies and eventually a more complete reference transcriptome and proteome. There are considerable efforts in the scientific community to collect experimentally verified genetic variant and alternative proteoform data in comprehensive databases such as Ensembl[418], RefSeq[419] and UniProt[420]. NextProt[421] and UniProt only include information if these are verified also in the proteome. Classically, these databases and the Human Proteome Project have focused on the set of all canonical proteins in the human proteome[422]. Recently a consortium was created to document all proteoforms, acknowledging the biological information to be gained outside of canonical proteins[423]. There are also databases that specifically document variant peptides identified in proteogenomics studies[147–149]. These knowledgebases are added to and revised continuously. As the knowledge of the human transcriptome deepens and a complete, refined search database becomes possible, additional variant peptide discoveries can be made using existing proteomics data in re-analysis[424,425]. The continuation of these efforts and improved data centralization will lead to a more complete database, facilitating proteogenomics efforts.

Proteogenomics accessibility

Proteogenomics pipelines are highly heterogeneous as the components are all customizable to researchers' questions and available datasets[426]. In addition, proteogenomics can be resource intensive as data must be generated for multiple omics levels. Research design must be carefully thought out to properly allocate available resources, and design decisions determine the appropriate analysis protocol. New technologies demand parallel updates to methodology and accompanying software; analysis of long-read RNA sequencing uses different tools than short-read, DDA different from DIA, etc. Many tools are developed to do only one piece of proteogenomics protocol to accommodate variation in research design. They must be performed in the correct order with attention to formats, pre- and post-processing requirements and computational needs, and useful visualizations should then be produced. Carrying out these tasks requires a skilled bioinformatician. To make proteogenomics methods accessible to biologists who could benefit from it, centralization and pipeline development that accommodate diverse research questions are critical. Galaxy for Proteomics (Galaxy-P) is one such option that is web-based, flexible and accessible. Galaxy-P provides training materials to teach users implement their proteogenomics pipelines, and has been successful in aiding proteogenomics research in a variety of studies. Other more comprehensive one-stop-shop options have been developed and include additional features[363,427–430]. Maintenance is crucial however; increased customizability via containerization as was implemented in SUsPECT, along with user-friendliness and continuous updates are needed to keep proteogenomics accessible.

Improved prediction of effects of variants on protein function

While proteogenomics enables the detection of protein variants, the true utility lies in their interpretation. We would like to know how the functioning of the protein in question changes as a result of the observed variation, which eventually can lead to correlation to a phenotype. As proteins are three-dimensional molecules whose functions are tightly linked to their structures, understanding of protein structure and perturbations thereof are essential to predicting functional change[431]. Proteogenomics involves the prediction of potential new protein sequences, whose structures were until recently quite challenging to predict based on sequence alone[432]. During this PhD, a new method called Alphafold was created to predict structures from sequences with unprecedented accuracy[433]. Prediction of probable protein structures can help assess protein-coding potential of novel ORFs detected in proteogenomics, thus providing a valuable tool to filter search databases. During the construction of SUsPECT, many variants changed predicted effect from benign to missense when annotated with sample-specific novel transcripts. Missense variations are a very heterogeneous class of variant effect with widely varying outcomes on proteins' structure and function[434]. Shortly after the release of Alphafold, AlphaMissense was released to tackle the challenge of structural change in response to missense variation[435]. These tools are invaluable to interpretation of protein variation uncovered by proteogenomics.

Concluding remarks

Proteogenomics provides the lens that brings the whole picture of post-transcriptional regulation into focus, and the picture is becoming sharper with new innovations. While its multi-omics nature means that proteogenomics takes on the challenges in each of the omics fields individually atop the existing challenges of data integration, it also means that the innovations per field directly influence the quality of proteogenomics findings for the better. This is certainly the case for long-read transcriptome sequencing, which is the innovation that this thesis focuses on specifically. The reduced noise in the peptide search database leads to more discovery of variation in the proteome. Despite the advantages this sequencing technology brings, it cannot completely compensate for the considerable limitations in proteomics. The challenges in proteogenomics seem vast, but addressing them will be worth the hassle. A complete understanding of proteome variation gets us a big step closer to a world where every person receives the correct diagnosis and treatments for them, every time.

A

Appendix

Bibliography

Data management plan

Summary

Samenvatting

Curriculum vitae

List of publications

Portfolio

Acknowledgements/Dankwoord

Bibliography

1. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, (2007).
2. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243 (2002).
3. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature Genetics 2002 30:1* **30**, 13–19 (2002).
4. Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* **17**, 100–107 (2001).
5. Martinez, N. M. & Lynch, K. W. Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. *Immunol Rev* **253**, 216–236 (2013).
6. Tian, G. G., Li, J. & Wu, J. Alternative splicing signatures in preimplantation embryo development. *Cell Biosci* **10**, 1–10 (2020).
7. Revil, T., Gaffney, D., Dias, C., Majewski, J. & Jerome-Majewska, L. A. Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* **11**, 1–17 (2010).
8. Yang, X. *et al.* Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805 (2016).
9. Choi, S., Cho, N. & Kim, K. K. The implications of alternative pre-mRNA splicing in cell signal transduction. *Experimental & Molecular Medicine 2023 55:4* **55**, 755–766 (2023).
10. Chabot, B. & Shkreta, L. Defective control of pre–messenger RNA splicing in human disease. *J Cell Biol* **212**, 13 (2016).
11. Poulos, M. G., Batra, R., Charizanis, K. & Swanson, M. S. Developments in RNA splicing and disease. *Cold Spring Harb Perspect Biol* **3**, 1–14 (2011).
12. Ouyang, J. *et al.* The role of alternative splicing in human cancer progression. *Am J Cancer Res* **11**, 4642 (2021).
13. Gebauer, F. & Hentze, M. W. Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology 2004 5:10* **5**, 827–835 (2004).
14. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol* **3**, 1–10 (2002).
15. Duchaine, T. F. & Fabian, M. R. Mechanistic Insights into MicroRNA-Mediated Gene Silencing. *Cold Spring Harb Perspect Biol* **11**, 32771–32772 (2019).
16. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology 2015 16:11* **16**, 665–677 (2015).
17. Vanderperre, B. *et al.* Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS One* **8**, 70698 (2013).
18. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics 2012 13:4* **13**, 227–232 (2012).
19. Ramakrishnan, S. R. *et al.* Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics* **25**, 1397–1403 (2009).
20. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biol* **4**, e309 (2006).
21. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature 2008 455:7209* **455**, 64–71 (2008).
22. Dahan, O., Gingold, H. & Pilpel, Y. Regulatory mechanisms and networks couple the different phases of gene expression. *Trends in Genetics* **27**, 316–322 (2011).
23. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature 2001 409:6822* **409**, 860–921 (2001).

24. Nurk, S. *et al.* The complete sequence of a human genome. *Science (1979)* **376**, 44–53 (2022).
25. Craig Venter, J. *et al.* The sequence of the human genome. *Science (1979)* **291**, 1304–1351 (2001).
26. Abdellah, Z. *et al.* Finishing the euchromatic sequence of the human genome. *Nature 2004 431:7011* **431**, 931–945 (2004).
27. Shastry, B. S. SNP alleles in human disease and evolution. *J Hum Genet* **47**, 561–566 (2002).
28. Shastry, B. S. SNPs: impact on gene function and phenotype. *Methods Mol Biol* **578**, 3–22 (2009).
29. Momozawa, Y. & Mizukami, K. Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics 2020 66:1* **66**, 11–23 (2020).
30. Gorlov, I. P., Gorlova, O. Y., Frazier, M. L., Spitz, M. R. & Amos, C. I. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet* **79**, 199–206 (2011).
31. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science (1979)* **335**, 823–828 (2012).
32. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665–677 (2017).
33. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics 2003 33:3* **33**, 228–237 (2003).
34. Wright, C. F. *et al.* Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. *The American Journal of Human Genetics* **108**, 1083–1094 (2021).
35. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102 (2015).
36. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science (1979)* **373**, 1464–1468 (2021).
37. Boucher, J. I., Bolon, D. N. A. & Tawfik, D. S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Science* **25**, 1219–1226 (2016).
38. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature Biotechnology 2017 35:2* **35**, 128–135 (2017).
39. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* (2013) doi:10.1002/0471142905.hg0720s76.
40. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003).
41. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).
42. Satam, H. *et al.* Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology 2023, Vol. 12, Page 997* **12**, 997 (2023).
43. Mardis, E. R. Next-Generation Sequencing Platforms. *https://doi.org/10.1146/annurev-anchem-062012-092628* **6**, 287–303 (2013).
44. Patel, R. K. & Jain, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS One* **7**, e30619 (2012).
45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**, 10–12 (2011).
46. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* **28**, 3169–3177 (2012).

47. Keel, B. N. & Snelling, W. M. Comparison of Burrows-Wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: Application to illumina data for livestock genomes 1. *Front Genet* **9**, 319058 (2018).
48. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Medicine 2020 12:1* **12**, 1–13 (2020).
49. Karlsson, K. & Linnarsson, S. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics* **18**, 1–11 (2017).
50. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications 2017 8:1* **8**, 1–11 (2017).
51. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**, 1297–1305 (2019).
52. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338 (2018).
53. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology 2019 37:10* **37**, 1155–1162 (2019).
54. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology 2018 36:4* **36**, 338–345 (2018).
55. Esteller-Cucala, P. *et al.* Y chromosome sequence and epigenomic reconstruction across human populations. *Communications Biology 2023 6:1* **6**, 1–11 (2023).
56. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology 2020 21:1* **21**, 1–16 (2020).
57. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009–1014 (2013).
58. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **34**, 666–681 (2018).
59. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).
60. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009–1014 (2013).
61. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods 2013 10:12* **10**, 1185–1191 (2013).
62. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods 2013 10:12* **10**, 1177–1184 (2013).
63. Sahlin, K. & Mäkinen, V. Accurate spliced alignment of long RNA sequencing reads. *Bioinformatics* **37**, 4643–4651 (2021).
64. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 1–13 (2019).
65. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications 2020 11:1* **11**, 1–12 (2020).
66. Tardaguila, M. *et al.* SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**, 396–411 (2018).
67. Kuosmanen, A., Norri, T. & Mäkinen, V. Evaluating approaches to find exon chains based on long reads. *Brief Bioinform* **19**, 404–414 (2018).
68. Križanović, K., Echchiki, A., Roux, J. & Šikić, M. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**, 748 (2018).

69. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput Biol* **18**, e1009730 (2022).
70. Fu, S. *et al.* IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics* **34**, 2168–2176 (2018).
71. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
72. Karpievitch, Y. V., Polpitiya, A. D., Anderson, G. A., Smith, R. D. & Dabney, A. R. Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects. *Ann Appl Stat* **4**, 1797 (2010).
73. 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* **10**, 1785–1793 (2011).
74. Mueller, L. N., Brusniak, M. Y., Mani, D. R. & Aebersold, R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res* **7**, 51–61 (2008).
75. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193–4201 (2004).
76. Wang, H., Chang-Wong, T., Tang, H. Y. & Speicher, D. W. Comparison of Extensive Protein Fractionation and Repetitive LC-MS/MS Analyses on Depth of Analysis for Complex Proteomes. *J Proteome Res* **9**, 1032 (2010).
77. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res* **9**, 1323–1329 (2010).
78. Rauniyar, N. & Keck, W. M. Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry. *International Journal of Molecular Sciences 2015, Vol. 16, Pages 28566-28581* **16**, 28566–28581 (2015).
79. Domon, B. & Aebersold, R. Challenges and Opportunities in Proteomics Data Analysis. *Molecular & Cellular Proteomics* **5**, 1921–1926 (2006).
80. Baldwin, M. A. Protein identification by mass spectrometry: issues to be considered. *Mol Cell Proteomics* **3**, 1–9 (2004).
81. Carr, S. *et al.* The Need for Guidelines in Publication of Peptide and Protein Identification Data: Working Group On Publication Guidelines For Peptide And Protein Identification Data. *Molecular & Cellular Proteomics* **3**, 531–533 (2004).
82. Muth, T. & Renard, B. Y. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform* **19**, 954–970 (2018).
83. Nesvizhskii, A. I. *et al.* Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides. *Molecular & Cellular Proteomics* **5**, 652–670 (2006).
84. Maillet, N. Rapid Peptides Generator: fast and efficient in silico protein digestion. *NAR Genom Bioinform* **2**, (2020).
85. Barton, S. J. & Whittaker, J. C. Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrom Rev* **28**, (2009).
86. Lam, H. & Aebersold, R. Spectral library searching for peptide identification via tandem MS. *Methods Mol Biol* **604**, 95–103 (2010).
87. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nature Biotechnology 2019 38:3* **38**, 365–373 (2019).
88. Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33**, 743–749 (2015).

89. Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534–1536 (2004).
90. Bugyi, F. *et al.* Influence of Post-Translational Modifications on Protein Identification in Database Searches. *ACS Omega* **6**, 7469–7477 (2021).
91. Lysiak, A., Fertin, G., Jean, G. & Tessier, D. Evaluation of open search methods based on theoretical mass spectra comparison. *BMC Bioinformatics* **22**, 1–17 (2021).
92. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature Biotechnology 2018 36:11* **36**, 1059–1061 (2018).
93. Na, S., Kim, J. & Paek, E. MODplus: Robust and Unrestrictive Identification of Post-Translational Modifications Using Mass Spectrometry. *Anal Chem* **91**, 11324–11333 (2019).
94. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods 2021 18:11* **18**, 1363–1369 (2021).
95. Moruz, L. & Käll, L. Peptide retention time prediction. *Mass Spectrom Rev* **36**, 615–623 (2017).
96. Degroeve, S., Martens, L. & Jurisica, I. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199–3203 (2013).
97. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nature Biotechnology 2022 41:1* **41**, 33–43 (2022).
98. Nesvizhskii, A. I. Proteogenomics: Concepts, applications and computational strategies. *Nature Methods* vol. 11 1114–1125 Preprint at https://doi.org/10.1038/NMETH.3144 (2014).
99. Blakeley, P., Overton, I. M. & Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* **11**, 5221–5234 (2012).
100. Krug, K. *et al.* Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol Cell Proteomics* **12**, 3420–3430 (2013).
101. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol* **16**, 1–26 (2015).
102. Kim, D. *et al.* FragGeneScan-plus for scalable high-throughput short-read open reading frame prediction. *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2015* (2015) doi:10.1109/CIBCB.2015.7300341.
103. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods 2013 10:12* **10**, 1177–1184 (2013).
104. Kannan, S., Hui, J., Mazooji, K., Pachter, L. & Tse, D. N. Shannon: An Information-Optimal de Novo RNA-Seq Assembler. *bioRxiv* 039230 (2016) doi:10.1101/039230.
105. Lau, E. *et al.* Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. *Cell Rep* **29**, 3751-3765.e5 (2019).
106. Sheynkman, G. M., Shortreed, M. R., Frey, B. L. & Smith, L. M. Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq* ⍰ S. *Molecular & Cellular Proteomics* **12**, 2341–2353 (2013).
107. Fermin, D. *et al.* Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* **7**, (2006).
108. Khatun, J. *et al.* Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genomics* **14**, (2013).
109. Baerenfaller, K. *et al.* Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **320**, 938–941 (2008).
110. Woo, S. *et al.* Proteogenomic database construction driven from large scale RNA-seq data. *J Proteome Res* **13**, 21–28 (2014).

111. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols 2013 8:8* **8**, 1494–1512 (2013).
112. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, (2011).
113. Washietl, S. *et al.* RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17**, 578–594 (2011).
114. Wang, L. *et al.* CPAT: coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucl. Acids Res.* **41**, e74 (2013).
115. Pohl, M., Theien, G. & Schuster, S. GC content dependency of open reading frame prediction via stop codon frequencies. *Gene* **511**, 441–446 (2012).
116. Tong, X. & Liu, S. CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res* **47**, e43–e43 (2019).
117. Schandorff, S. *et al.* A mass spectrometry–friendly database for cSNP identification. *Nature Methods 2007 4:6* **4**, 465–466 (2007).
118. Bunger, M. K. *et al.* Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *J Proteome Res* **6**, 2331–2340 (2007).
119. Li, J. *et al.* A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol Cell Proteomics* **10**, (2011).
120. Chang, K. Y., Ryan Georgianna, D., Heber, S., Payne, G. A. & Muddiman, D. C. Detection of alternative splice variants at the proteome level in aspergillus flavus. *J Proteome Res* **9**, 1209–1217 (2010).
121. Edwards, N. J. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol Syst Biol* **3**, (2007).
122. Fermin, D. *et al.* Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* **7**, 1–13 (2006).
123. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
124. Elias, J. E., Haas, W., Faherty, B. K. & Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods 2005 2:9* **2**, 667–675 (2005).
125. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods 2007 4:3* **4**, 207–214 (2007).
126. Blakeley, P., Overton, I. M. & Hubbard, S. J. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *J Proteome Res* **11**, 5221 (2012).
127. Li, H. *et al.* Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics* **17**, (2016).
128. Yi, X. *et al.* Quality control of single amino acid variations detected by tandem mass spectrometry. *J Proteomics* **187**, 144–151 (2018).
129. Fu, Y. Bayesian false discovery rates for post-translational modification proteomics. *Stat Interface* **5**, 47–59 (2012).
130. Fu, Y. & Qian, X. Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry. *Molecular and Cellular Proteomics* **13**, 1359–1368 (2014).
131. Tharakan, R., Edwards, N. & Graham, D. R. M. Data maximization by multipass analysis of protein mass spectra. *Proteomics* **10**, 1160–1171 (2010).
132. Jagtap, P. *et al.* A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**, 1352–1357 (2013).

133. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **73**, 2092–2123 (2010).

134. Nagaraj, N. & Mann, M. Quantitative analysis of the intra-and inter-individual variability of the normal urinary proteome. *J Proteome Res* **10**, 637–645 (2011).

135. Kushner, I. K. *et al.* Individual Variability of Protein Expression in Human Tissues. *J Proteome Res* **17**, 3914–3922 (2018).

136. Li, J. *et al.* A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Molecular and Cellular Proteomics* **10**, (2011).

137. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).

138. Subbannayya, Y., Pinto, S. M., Gowda, H. & Prasad, T. S. K. Proteogenomics for understanding oncology: Recent advances and future prospects. *Expert Review of Proteomics* vol. 13 297–308 Preprint at https://doi.org/10.1586/14789450.2016.1136217 (2016).

139. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).

140. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755–765 (2016).

141. Tan, Z. *et al.* Comprehensive Detection of Single Amino Acid Variants and Evaluation of Their Deleterious Potential in a PANC-1 Cell Line. *J Proteome Res* **19**, 1635–1646 (2020).

142. Ruppen-Cañás, I. *et al.* An improved quantitative mass spectrometry analysis of tumor specific mutant proteins at high sensitivity. *Proteomics* **12**, 1319–1327 (2012).

143. Su, Z. D. *et al.* Quantitative detection of single amino acid polymorphisms by targeted proteomics. *J Mol Cell Biol* **3**, 309–315 (2011).

144. Dimitrakopoulos, L. *et al.* Variant peptide detection utilizing mass spectrometry: Laying the foundations for proteogenomic identification and validation. *Clin Chem Lab Med* **55**, 1291–1304 (2017).

145. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol* **15**, (2019).

146. Guillot, L. *et al.* Peptimapper: Proteogenomics workflow for the expert annotation of eukaryotic genomes. *BMC Genomics* **20**, (2019).

147. Wen, B., Wang, X. & Zhang, B. PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res* **29**, 485–493 (2019).

148. Flores, M. A. & Lazar, I. M. XMAn v2—a database of Homo sapiens mutated peptides. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz693.

149. Cao, R. *et al.* dbSAP: single amino-acid polymorphism database for protein variation detection. *Nucleic Acids Res* **45**, D827–D832 (2017).

150. Lichti, C. F. *et al.* Systematic identification of single amino acid variants in glioma stem-cell-derived chromosome 19 proteins. *J Proteome Res* **14**, 778–786 (2015).

151. Nie, S. *et al.* Quantitative analysis of single amino acid variant peptides associated with pancreatic cancer in serum by an isobaric labeling quantitative method. *J Proteome Res* **13**, 6058–6066 (2014).

152. Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C. & Yates, J. R. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal Chem* (2000) doi:10.1021/ac991025n.

153. Roth, M. J. *et al.* Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Molecular and Cellular Proteomics* (2005) doi:10.1074/mcp.M500064-MCP200.

154. Noble, W. S. Mass spectrometrists should search only for peptides they care about. *Nature Methods* vol. 12 605–608 Preprint at https://doi.org/10.1038/nmeth.3450 (2015).

155. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* vol. 73 2092–2123 Preprint at https://doi.org/10.1016/j.jprot.2010.08.009 (2010).
156. Li, J. *et al.* A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Molecular and Cellular Proteomics* **10**, (2011).
157. Song, C. *et al.* Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. *J Proteome Res* **13**, 241–248 (2014).
158. Wang, X. *et al.* Protein identification using customized protein sequence databases derived from RNA-seq data. *J Proteome Res* **11**, 1009–1017 (2012).
159. Wen, B. *et al.* PGA: An R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. *BMC Bioinformatics* **17**, 244 (2016).
160. Li, Y. *et al.* JUMPg: An integrative proteogenomics pipeline identifying unannotated proteins in human brain and cancer cells. *J Proteome Res* **15**, 2309–2320 (2016).
161. Wang, X. & Zhang, B. Data and text mining customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. **29**, 3235–3237 (2013).
162. Zickmann, F. & Renard, B. Y. MSProGene: Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. in *Bioinformatics* vol. 31 i106–i115 (2015).
163. Cesnik, A. J. *et al.* Spritz: A proteogenomic database engine. *bioRxiv* Preprint at https://doi.org/10.1101/2020.06.08.140681 (2020).
164. Sticker, A., Martens, L. & Clement, L. Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nat Methods* **14**, 643–644 (2017).
165. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* (2008) doi:10.1038/nature07509.
166. Blakeley, P., Overton, I. M. & Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* **11**, 5221–5234 (2012).
167. Tanner, S. *et al.* InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology* **422**, 4626–4639 (2003).
168. Tabb, D. L., Ze-Qiang, M., Martin, D. B., Ham, A. J. L. & Chambers, M. C. DirecTag: Accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* (2008) doi:10.1021/pr800154p.
169. Dasari, S. *et al.* TagRecon: High-Throughput Mutation Identification through Sequence Tagging NIH Public Access. *J Proteome Res* **9**, 1716–1726 (2010).
170. Abraham, P., Adams, R. M., Tuskan, G. A. & Hettich, R. L. Moving away from the reference genome: Evaluating a peptide sequencing tagging approach for single amino acid polymorphism identifications in the genus populus. *J Proteome Res* **12**, 3642–3651 (2013).
171. Han, Y., Ma, B. & Zhang, K. Spider: Software for protein identification from sequence tags with de novo sequencing error. *J Bioinform Comput Biol* **3**, 697–716 (2005).
172. Devabhaktuni, A. *et al.* TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat Biotechnol* **37**, 469–479 (2019).
173. Gabriels, R., Martens, L. & Degroeve, S. Updated MS^{2}PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res* (2019) doi:10.1093/nar/gkz299.
174. Silva, A. S. C., Bouwmeester, R., Martens, L. & Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz383.
175. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* (2019) doi:10.1038/s41592-019-0617-2.

176. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499**, 79 (2013).
177. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–1372 (2008).
178. Neph, S. *et al.* BEDOPS: High-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
179. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
180. Adusumilli, R. & Mallick, P. Data conversion with proteoWizard msConvert. in *Methods in Molecular Biology* (2017). doi:10.1007/978-1-4939-6747-6_23.
181. Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V. & Gorshkov, M. V. Pyteomics - A python framework for exploratory data analysis and rapid software prototyping in proteomics. *J Am Soc Mass Spectrom* **24**, 301–304 (2013).
182. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *bioRxiv* 2020.03.28.013003 (2020) doi:10.1101/2020.03.28.013003.
183. Hirsch, C. & Schildknecht, S. In vitro research reproducibility: Keeping up high standards. *Frontiers in Pharmacology* vol. 10 Preprint at https://doi.org/10.3389/fphar.2019.01484 (2019).
184. Shao, W. *et al.* Comparative analysis of mRNA and protein degradation in prostate tissues indicates high stability of proteins. *Nat Commun* **10**, 1–8 (2019).
185. Mamie Lih, T. S., Choong, W. K., Chen, Y. J. & Sung, T. Y. Evaluating the Possibility of Detecting Variants in Shotgun Proteomics via LeTE-Fusion Analysis Pipeline. *J Proteome Res* **17**, 2937–2952 (2018).
186. Hwang, H. *et al.* Next Generation Proteomic Pipeline for Chromosome-Based Proteomic Research Using NeXtProt and GENCODE Databases. *J Proteome Res* **16**, 4425–4434 (2017).
187. Ma, S., Menon, R., Poulos, R. C. & Wong, J. W. H. Proteogenomic analysis prioritises functional single nucleotide variants in cancer samples. *Oncotarget* **8**, 95841–95852 (2017).
188. Ruggles, K. V. *et al.* An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Molecular and Cellular Proteomics* **15**, 1060–1071 (2016).
189. Dimitrakopoulos, L. *et al.* Proteome-wide onco-proteogenomic somatic variant identification in ER-positive breast cancer. *Clin Biochem* **66**, 63–75 (2019).
190. Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J Proteome Res* **17**, 3463–3474 (2018).
191. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat Biotechnol* **36**, 1059–1066 (2018).
192. Yu, F. *et al.* Identification of modified peptides using localization-aware open search. *Nat Commun* **11**, 1–9 (2020).
193. Chang, H. Y. *et al.* Crystal-C: A Computational Tool for Refinement of Open Search Results. *J Proteome Res* **19**, 2511–2515 (2020).
194. Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L. & Degroeve, S. The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* 1900351 (2020) doi:10.1002/pmic.201900351.
195. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* **114**, 8247–8252 (2017).
196. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* **16**, 63–66 (2019).

197. Shi, J. *et al.* Determining Allele-Specific Protein Expression (ASPE) Using a Novel Quantitative Concatamer Based Proteomics Method. *J Proteome Res* **17**, 3606–3612 (2018).
198. Medzhitov, R. & Horng, T. Transcriptional control of the inflammatory response. *Nature Reviews Immunology 2009 9:10* **9**, 692–703 (2009).
199. Carpenter, S., Ricci, E. P., Mercier, B. C., Moore, M. J. & Fitzgerald, K. A. Post-transcriptional regulation of gene expression in innate immunity. *Nature Reviews Immunology 2014 14:6* **14**, 361–376 (2014).
200. Wells, C. A. *et al.* Alternate transcription of the Toll-like receptor signaling cascade. *Genome Biol* **7**, 1–17 (2006).
201. Oosting, M. *et al.* Functional and Genomic Architecture of Borrelia burgdorferi-Induced Cytokine Responses in Humans. *Cell Host Microbe* **20**, 822–833 (2016).
202. Li, Y. *et al.* A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell* **167**, 1099-1110.e14 (2016).
203. Lu, Y. C., Yeh, W. C. & Ohashi, P. S. LPS/TLR4 signal transduction pathway. *Cytokine* **42**, 145–151 (2008).
204. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).
205. Van Der Made, C. I. *et al.* Presence of Genetic Variants Among Young Men With Severe COVID-19. *JAMA* **324**, 663–673 (2020).
206. Smeekens, S. P. *et al.* Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans. *Nat Commun* **4**, (2013).
207. Bruno, M. *et al.* Transcriptional and functional insights into the host immune response against the emerging fungal pathogen Candida auris. *Nat Microbiol* **5**, 1516–1531 (2020).
208. Askarian, F., Wagner, T., Johannessen, M. & Nizet, V. Staphylococcus aureus modulation of innate immune responses through Toll-like (TLR), (NOD)-like (NLR) and C-type lectin (CLR) receptors. *FEMS Microbiol Rev* **42**, 656–671 (2018).
209. Netea, M. G. *et al.* Immune sensing of Candida albicans requires cooperative recognition of mannans and glucans by lectin and Toll-like receptors. *J Clin Invest* **116**, 1642–1650 (2006).
210. Heinhuis, B. *et al.* Inflammation-dependent secretion and splicing of IL-32γ in rheumatoid arthritis. *Proc Natl Acad Sci U S A* **108**, 4962–4967 (2011).
211. Oberdoerffer, S. *et al.* Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science* **321**, 686–691 (2008).
212. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009–1014 (2013).
213. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature 2022 608:7922* **608**, 353–359 (2022).
214. Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nature Biotechnology 2023* 1–5 (2023) doi:10.1038/s41587-023-01815-7.
215. Vollmers, A. C., Mekonen, H. E., Campos, S., Carpenter, S. & Vollmers, C. Generation of an isoform-level transcriptome atlas of macrophage activation. *J Biol Chem* **296**, (2021).
216. Cole, C., Byrne, A., Adams, M., Volden, R. & Vollmers, C. Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing. *Genome Res* **30**, 589–601 (2020).
217. Inamo, J. *et al.* Immune Isoform Atlas: Landscape of alternative splicing in human immune cells. *bioRxiv* 2022.09.13.507708 (2022) doi:10.1101/2022.09.13.507708.
218. Kanno, T. *et al.* Characterization of proteogenomic signatures of differentiation of CD4+ T cell subsets. *DNA Research* **30**, 1–11 (2023).
219. Shi, Z. R. *et al.* Integrated proteogenomic characterization reveals an imbalanced hepatocellular carcinoma microenvironment after incomplete radiofrequency ablation. *Journal of Experimental and Clinical Cancer Research* **42**, 1–18 (2023).

220. Proteogenomics, R. *et al.* Proteogenomics Analysis Reveals Novel Micropeptides in Primary Human Immune Cells. *Immuno 2022, Vol. 2, Pages 283-292* **2**, 283–292 (2022).
221. Rivero-Hinojosa, S. *et al.* Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nature Communications 2021 12:1* **12**, 1–15 (2021).
222. Oosting, M. *et al.* Functional and Genomic Architecture of Borrelia burgdorferi-Induced Cytokine Responses in Humans. *Cell Host Microbe* **20**, 822–833 (2016).
223. Prjibelski, A. *et al.* IsoQuant: a tool for accurate novel isoform discovery with long reads. (2022) doi:10.21203/RS.3.RS-1571850/V1.
224. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**, 10–12 (2011).
225. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
226. Tarazona, S. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* **43**, e140 (2015).
227. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics 2000 25:1* **25**, 25–29 (2000).
228. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**, 238–241 (1996).
229. Reese, F. & Mortazavi, A. Swan: a library for the analysis and visualization of long-read transcriptomes. *Bioinformatics* **37**, 1322–1323 (2021).
230. Vitting-Seerup, K., Sandelin, A. & Berger, B. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics* **35**, 4469–4471 (2019).
231. Eddy, S. R. HMMER User's Guide Biological sequence analysis using profile hidden Markov models. (2020).
232. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology 2019 37:4* **37**, 420–423 (2019).
233. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
234. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One* **5**, e13984 (2010).
235. Fang, H. dcGOR: an R package for analysing ontologies and protein domain annotations. *PLoS Comput Biol* **10**, (2014).
236. Miller, R. M. *et al.* Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol* **23**, (2022).
237. Miller, R. M., Millikin, R. J., Rolfs, Z., Shortreed, M. R. & Smith, L. M. Enhanced Proteomic Data Analysis with MetaMorpheus. *Methods Mol Biol* **2426**, 35–66 (2023).
238. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J Proteome Res* **18**, 4108–4116 (2019).
239. Millikin, R. J., Shortreed, M. R., Scalf, M. & Smith, L. M. Fast, Free, and Flexible Peptide and Protein Quantification with FlashLFQ. *Methods Mol Biol* **2426**, 303–313 (2023).
240. Digre, A. & Lindskog, C. The Human Protein Atlas-Spatial localization of the human proteome in health and disease. *Protein Sci* **30**, 218–233 (2021).
241. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
242. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci Data* **4**, (2017).
243. Moll, P., Ante, M., Seitz, A. & Reda, T. QuantSeq 3⍰ mRNA sequencing for RNA quantification. *Nature Methods 2014 11:12* **11**, i–iii (2014).

244. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191–W198 (2019).
245. Paun, A. & Pitha, P. M. The IRF family, revisited. *Biochimie* **89**, 744–753 (2007).
246. Herbein, G., Doyle, A. G., Montaner, L. J. & Gordon, S. Lipopolysaccharide (LPS) down-regulates CD4 expression in primary human macrophages through induction of endogenous tumour necrosis factor (TNF) and IL-1 beta. *Clin Exp Immunol* **102**, 430–437 (1995).
247. Lachmandas, E. *et al.* Microbial stimulation of different Toll-like receptor signalling pathways induces diverse metabolic programmes in human monocytes. *Nat Microbiol* **2**, (2016).
248. Green, I. D. *et al.* Macrophage development and activation involve coordinated intron retention in key inflammatory regulators. *Nucleic Acids Res* **48**, 6513–6529 (2020).
249. Song, R. *et al.* Dynamic intron retention modulates gene expression in the monocytic differentiation pathway. *Immunology* **165**, 274–286 (2022).
250. Wong, J. J. L. *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, (2013).
251. Ullrich, S. & Guigó, R. Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. *Nucleic Acids Res* **48**, 1327–1340 (2020).
252. O'Grady, T. M. *et al.* Reversal of splicing infidelity is a pre-activation step in B cell differentiation. *Front Immunol* **13**, (2022).
253. Karginov, T. A., Ménoret, A. & Vella, A. T. Optimal CD8+ T cell effector function requires costimulation-induced RNA-binding proteins that reprogram the transcript isoform landscape. *Nat Commun* **13**, (2022).
254. Ni, T. *et al.* Global intron retention mediated gene regulation during CD4+ T cell activation. *Nucleic Acids Res* **44**, 6817–6829 (2016).
255. Brunet, M. A., Levesque, S. A., Hunting, D. J., Cohen, A. A. & Roucou, X. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res* **28**, 609–624 (2018).
256. Heinz, S. *et al.* Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**, 1522-1536.e22 (2018).
257. Karasawa, T. *et al.* Oligomerized CARD16 promotes caspase-1 assembly and IL-1β processing. *FEBS Open Bio* **5**, 348–356 (2015).
258. Devi, S. *et al.* CARD-only proteins regulate in vivo inflammasome responses and ameliorate gout. *Cell Rep* **42**, (2023).
259. Oeckinghaus, A. & Ghosh, S. The NF-kappaB family of transcription factors and its regulation. *Cold Spring Harb Perspect Biol* **1**, (2009).
260. Fliegauf, M. *et al.* Detrimental NFKB1 missense variants affecting the Rel-homology domain of p105/p50. *Front Immunol* **13**, (2022).
261. Mata-Martínez, P., Bergón-Gutiérrez, M. & del Fresno, C. Dectin-1 Signaling Update: New Perspectives for Trained Immunity. *Front Immunol* **13**, 812148 (2022).
262. Rotival, M., Quach, H. & Quintana-Murci, L. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nat Commun* **10**, 1–15 (2019).
263. Banday, A. R. *et al.* Genetic regulation of OAS1 nonsense-mediated decay underlies association with COVID-19 hospitalization in patients of European and African ancestries. *Nature Genetics 2022 54:8* **54**, 1103–1116 (2022).
264. Wickenhagen, A. *et al.* A prenylated dsRNA sensor protects against severe COVID-19. *Science (1979)* **374**, (2021).
265. Perišić Nanut, M., Pečar Fonović, U., Jakoš, T. & Kos, J. The Role of Cysteine Peptidases in Hematopoietic Stem Cell Differentiation and Modulation of Immune System Function. *Front Immunol* **12**, (2021).

266. Thiery, J. *et al.* Perforin activates clathrin- and dynamin-dependent endocytosis, which is required for plasma membrane repair and delivery of granzyme B for granzyme-mediated apoptosis. *Blood* **115**, 1582–1593 (2010).
267. Momoi, T. *et al.* Amino acid sequence of a modified β2-microglobulin in renal failure patient urine and long-term dialysis patient blood. *Clinica Chimica Acta* **236**, 135–144 (1995).
268. Fukuhara, K. *et al.* A study on CD45 isoform expression during T-cell development and selection events in the human thymus. *Hum Immunol* **63**, 394–404 (2002).
269. Orta-Mascaró, M. *et al.* CD6 modulates thymocyte selection and peripheral T cell homeostasis. *J Exp Med* **213**, 1387–1397 (2016).
270. De Arras, L. & Alper, S. Limiting of the innate immune response by SF3A-dependent control of MyD88 alternative mRNA splicing. *PLoS Genet* **9**, (2013).
271. Pozzi, B. *et al.* Dengue virus targets RBM10 deregulating host cell splicing and innate immune response. *Nucleic Acids Res* **48**, 6824–6838 (2020).
272. Stein, M. M. *et al.* Sex-specific differences in peripheral blood leukocyte transcriptional response to LPS are enriched for HLA region and X chromosome genes. *Sci Rep* **11**, (2021).
273. Piasecka, B. *et al.* Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc Natl Acad Sci U S A* **115**, E488–E497 (2018).
274. Li, Y. *et al.* Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nature Medicine 2016 22:8* **22**, 952–960 (2016).
275. Mironov, A. *et al.* Tissue-specific regulation of gene expression via unproductive splicing. *Nucleic Acids Res* **51**, 3055–3066 (2023).
276. Hardwick, S. A. *et al.* Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nat Biotechnol* **40**, 1082–1092 (2022).
277. Meissner, F., Scheltema, R. A., Mollenkopf, H. J. & Mann, M. Direct proteomic quantification of the secretome of activated immune cells. *Science (1979)* **340**, 475–478 (2013).
278. Sinitcyn, P. *et al.* Global detection of human variants and isoforms by deep proteome sequencing. *Nature Biotechnology 2023* 1–11 (2023) doi:10.1038/s41587-023-01714-x.
279. Salz, R. *et al.* SUsPECT: a pipeline for variant effect prediction based on custom long-read transcriptomes for improved clinical variant annotation. *BMC Genomics* **24**, 1–10 (2023).
280. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* (2009) doi:10.1038/nprot.2009.86.
281. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications 2020 11:1* **11**, 1–13 (2020).
282. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 1–14 (2016).
283. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
284. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* **12**, 1–8 (2020).
285. Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: their purpose and place. *Hum Mol Genet* **27**, R234–R241 (2018).
286. Morillon, A. & Gautheret, D. Bridging the gap between reference and real transcriptomes. *Genome Biol* **20**, (2019).
287. Dong, X. *et al.* Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *bioRxiv* 2022.07.22.501076 (2022) doi:10.1101/2022.07.22.501076.
288. Sun, Y. H. *et al.* Single-molecule long-read sequencing reveals a conserved intact long RNA profile in sperm. *Nature Communications 2021 12:1* **12**, 1–12 (2021).
289. de Paoli-Iseppi, R., Gleeson, J. & Clark, M. B. Isoform Age - Splice Isoform Profiling Using Long-Read Technologies. *Front Mol Biosci* **8**, (2021).

290. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics 2013 45:6* **45**, 580–585 (2013).
291. Gibson, G. The environmental contribution to gene expression profiles. *Nature Reviews Genetics 2008 9:8* **9**, 575–581 (2008).
292. Wright, D. J. *et al.* Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes. *BMC Genomics* **23**, (2022).
293. Tay, A. P., Hamey, J. J., Martyn, G. E., Wilson, L. O. W. & Wilkins, M. R. Identification of Protein Isoforms Using Reference Databases Built from Long and Short Read RNA-Sequencing. *J Proteome Res* **21**, 1628–1639 (2022).
294. Mehlferber, M. M. *et al.* Characterization of protein isoform diversity in human umbilical vein endothelial cells via long-read proteogenomics. *RNA Biol* **19**, 1228–1243 (2022).
295. Li, A., Zhang, J. & Zhou, Z. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* **15**, 1–10 (2014).
296. Tong, X. & Liu, S. CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res* **47**, e43–e43 (2019).
297. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**, W345–W349 (2007).
298. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
299. Landrum, M. J. *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* (2014) doi:10.1093/nar/gkt1113.
300. Liu, L. *et al.* Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. *J Exp Med* **208**, 1635–1648 (2011).
301. van de Veerdonk, F. L. *et al.* STAT1 Mutations in Autosomal Dominant Chronic Mucocutaneous Candidiasis . *New England Journal of Medicine* **365**, 54–61 (2011).
302. Rodriguez, J. M., Pozo, F., di Domenico, T., Vazquez, J. & Tress, M. L. An analysis of tissue-specific alternative splicing at the protein level. *PLoS Comput Biol* **16**, e1008287 (2020).
303. Pozo, F. *et al.* Assessing the functional relevance of splice isoforms. *NAR Genom Bioinform* **3**, 1–16 (2021).
304. Rodriguez, J. M. *et al.* APPRIS: selecting functionally important isoforms. *Nucleic Acids Res* **50**, D54–D59 (2022).
305. Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics 2022* 1–14 (2022) doi:10.1038/s41576-022-00514-4.
306. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular Mechanisms of Disease-Causing Missense Mutations. *J Mol Biol* **425**, 3919–3936 (2013).
307. Capriotti, E. & Altman, R. B. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics* **98**, 310–317 (2011).
308. Kryukov, G. v., Pennacchio, L. A. & Sunyaev, S. R. Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *The American Journal of Human Genetics* **80**, 727–739 (2007).
309. Sieber, P., Platzer, M. & Schuster, S. The Definition of Open Reading Frame Revisited. *Trends Genet* **34**, 167–170 (2018).
310. Martinez, T. F. *et al.* Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol* **16**, 458–468 (2020).
311. Prensner, J. R. *et al.* Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nature Biotechnology 2021 39:6* **39**, 697–704 (2021).
312. Jagadeesh, K. A. *et al.* S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nature Genetics 2019 51:4* **51**, 755–763 (2019).
313. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16**, 1–12 (2015).

314. Steinhaus, R. *et al.* MutationTaster2021. *Nucleic Acids Res* **49**, W446–W451 (2021).
315. López-Ferrando, V., Gazzo, A., de La Cruz, X., Orozco, M. & Gelpí, J. L. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res* **45**, W222 (2017).
316. Swamy, V. S., Fufa, T. D., Hufnagel, R. B. & McGaughey, D. M. A long read optimized de novo transcriptome pipeline reveals novel ocular developmentally regulated gene isoforms and disease targets. *bioRxiv* 2020.08.21.261644 (2020) doi:10.1101/2020.08.21.261644.
317. Miller, D. E. *et al.* Targeted long-read sequencing identifies missing disease-causing variation. *The American Journal of Human Genetics* **108**, 1436–1449 (2021).
318. di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology 2017 35:4* **35**, 316–319 (2017).
319. Wyman, D. *et al.* A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. (2019) doi:10.1101/672931.
320. Dainat, J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format.
321. Li, Y. *et al.* A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell* **167**, 1099-1110.e14 (2016).
322. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nature Reviews Genetics 2018 19:5* **19**, 253–268 (2018).
323. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Medicine 2022 14:1* **14**, 1–22 (2022).
324. Wangler, M. F. *et al.* Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research. *Genetics* **207**, 9–27 (2017).
325. Hmeljak, J. & Justice, M. J. From gene to treatment: Supporting rare disease translational research through model systems. *DMM Disease Models and Mechanisms* **12**, (2019).
326. Anderson, R. H. & Francis, K. R. Modeling rare diseases with induced pluripotent stem cell technology. *Mol Cell Probes* **40**, 52–59 (2018).
327. Li, Y. *et al.* Establishment and Maintenance of Primary Fibroblast Repositories for Rare Diseases—Friedreich’s Ataxia Example. *https://home.liebertpub.com/bio* **14**, 324–329 (2016).
328. Crowther, L. M., Poms, M. & Plecko, B. Multiomics tools for the diagnosis and treatment of rare neurological disease. *J Inherit Metab Dis* **41**, 425–434 (2018).
329. Kopajtich, R. *et al.* Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. *medRxiv* 2021.03.09.21253187 (2021) doi:10.11 01/2021.03.09.21253187.
330. Roos, A., Thompson, R., Horvath, R., Lochmüller, H. & Sickmann, A. Intersection of Proteomics and Genomics to “Solve the Unsolved” in Rare Disorders such as Neurodegenerative and Neuromuscular Diseases. *Proteomics Clin Appl* **12**, 1700073 (2018).
331. Aronson, J. K. & Ferner, R. E. Biomarkers—A General Review. *Curr Protoc Pharmacol* **76**, 9.23.1-9.23.17 (2017).
332. Mani, D. R. *et al.* Cancer proteogenomics: current impact and future prospects. *Nature Reviews Cancer 2022 22:5* **22**, 298–313 (2022).
333. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature 2014 513:7518* **513**, 382–387 (2014).
334. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755–765 (2016).
335. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature 2016 534:7605* **534**, 55–62 (2016).
336. Sinha, A. *et al.* The Proteogenomic Landscape of Curable Prostate Cancer. *Cancer Cell* **35**, 414 (2019).
337. Petralia, F. *et al.* Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Cell* **183**, 1962-1985.e31 (2020).

338. Archer, T. C. *et al.* Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell* **34**, 396-410.e8 (2018).
339. Clark, D. J. *et al.* Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **179**, 964-983.e31 (2019).
340. Li, C. *et al.* Integrated Omics of Metastatic Colorectal Cancer. *Cancer Cell* **38**, 734-747.e9 (2020).
341. Creech, A. L. *et al.* The Role of Mass Spectrometry and Proteogenomics in the Advancement of HLA Epitope Prediction. *Proteomics* **18**, (2018).
342. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun* **7**, (2016).
343. Zhang, X., Qi, Y., Zhang, Q. & Liu, W. Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomed Pharmacother* **120**, (2019).
344. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature 2017 547:7662* **547**, 222–226 (2017).
345. Robbins, P. F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nature Medicine 2013 19:6* **19**, 747–752 (2013).
346. Rivero-Hinojosa, S. *et al.* Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nature Communications 2021 12:1* **12**, 1–15 (2021).
347. Lu, Y.-C. *et al.* Mutated PPP1R3B Is Recognized by T Cells Used To Treat a Melanoma Patient Who Experienced a Durable Complete Tumor Regression. *The Journal of Immunology* **190**, 6034–6042 (2013).
348. Tan, X. *et al.* PGNneo: A Proteogenomics-Based Neoantigen Prediction Pipeline in Noncoding Regions. *Cells* **12**, 782 (2023).
349. Li, Y. *et al.* ProGeo-neo: A customized proteogenomic workflow for neoantigen prediction and selection. *BMC Med Genomics* **13**, 1–11 (2020).
350. Hundal, J. *et al.* pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol Res* **8**, 409–420 (2020).
351. Kim, S. *et al.* Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology* **29**, 1030–1036 (2018).
352. Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A. & Anderson, A. R. A. NeoPredPipe: High-throughput neoantigen prediction and recognition potential pipeline. *BMC Bioinformatics* **20**, 1–6 (2019).
353. Li, Y. *et al.* Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell* **41**, 1397–1406 (2023).
354. Kaiser, S. *et al.* A proteogenomic view of Parkinson’s disease causality and heterogeneity. *npj Parkinson’s Disease 2023 9:1* **9**, 1–13 (2023).
355. da Silva, E. M. G. *et al.* Proteogenomics reveals orthologous alternatively spliced proteoforms in the same human and mouse brain regions with differential abundance in an alzheimer’s disease mouse model. *Cells* **10**, (2021).
356. Han, Y. *et al.* Computation-assisted targeted proteomics of alternative splicing protein isoforms in the human heart. *J Mol Cell Cardiol* **154**, 92–96 (2021).
357. Han, Y. *et al.* Proteogenomics reveals sex-biased aging genes and coordinated splicing in cardiac aging. *Am J Physiol Heart Circ Physiol* **323**, H538–H558 (2022).
358. de la Fuente-Núñez, C., Silva, O. N., Lu, T. K. & Franco, O. L. Antimicrobial peptides: Role in human disease and potential as immunotherapies. *Pharmacol Ther* **178**, 132–140 (2017).

359. Silva, O. N., Porto, W. F., Ribeiro, S. M., Batista, I. & Franco, O. L. Host-defense peptides and their potential use as biomarkers in human diseases. *Drug Discov Today* **23**, 1666–1671 (2018).
360. Zhang, H., Jain, C. & Aluru, S. A comprehensive evaluation of long read error correction methods. *BMC Genomics* **21**, 1–15 (2020).
361. Dong, X. *et al.* Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nature Methods 2023 20:11* **20**, 1810–1821 (2023).
362. Soneson, C., Matthes, K. L., Nowicka, M., Law, C. W. & Robinson, M. D. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol* **17**, 1–15 (2016).
363. Miller, R. M. *et al.* Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol* **23**, 1–28 (2022).
364. Matthew, D. C. N. *et al.* A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. *Genome Res* **31**, 327–336 (2021).
365. Erady, C. *et al.* Pan-cancer analysis of transcripts encoding novel open-reading frames (nORFs) and their potential biological functions. *npj Genomic Medicine 2021 6:1* **6**, 1–17 (2021).
366. Wei, C., Ye, Z., Zhang, J. & Li, A. CPPVec: an accurate coding potential predictor based on a distributed representation of protein sequence. *BMC Genomics* **24**, 1–9 (2023).
367. Ma, J. *et al.* Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* **13**, 1757–1765 (2014).
368. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science (1979)* **367**, 140–146 (2020).
369. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology 2012 9:1* **9**, 59–64 (2012).
370. Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol* **20**, 8635–8642 (2000).
371. Somers, J., Pöyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *Int J Biochem Cell Biol* **45**, 1690–1700 (2013).
372. Ito, K. & Chiba, S. Arrest peptides: cis-acting modulators of translation. *Annu Rev Biochem* **82**, 171–202 (2013).
373. James, C. C. & Smyth, J. W. Alternative mechanisms of translation initiation: an emerging dynamic regulator of the proteome in health and disease. *Life Sci* **212**, 138 (2018).
374. Andreev, D. E. *et al.* Non-AUG translation initiation in mammals. *Genome Biology 2022 23:1* **23**, 1–17 (2022).
375. Prabakaran, S. *et al.* Quantitative profiling of peptides from RNAs classified as noncoding. *Nature Communications 2014 5:1* **5**, 1–10 (2014).
376. Brosch, M. *et al.* Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res* **21**, 756–767 (2011).
377. Brar, G. A. & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology 2015 16:11* **16**, 651–664 (2015).
378. Aggarwal, S., Raj, A., Kumar, D., Dash, D. & Yadav, A. K. False discovery rate: the Achilles' heel of proteogenomics. *Brief Bioinform* **23**, (2022).
379. Raj, A., Aggarwal, S., Singh, P., Yadav, A. K. & Dash, D. PgxSAVy: A tool for comprehensive evaluation of variant peptide quality in proteogenomics – catching the (un)usual suspects. *Comput Struct Biotechnol J* **23**, 711–722 (2024).
380. Levitsky, L. I. *et al.* Massive Proteogenomic Reanalysis of Publicly Available Proteomic Datasets of Human Tissues in Search for Protein Recoding via Adenosine-to-Inosine RNA Editing. *J Proteome Res* **22**, 1695–1711 (2023).

381. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J Proteome Res* **18**, 4108–4116 (2019).
382. Martín-Campos, T. *et al.* MsViz: A Graphical Software Tool for In-Depth Manual Validation and Quantitation of Post-translational Modifications. *J Proteome Res* **16**, 3092–3101 (2017).
383. Voytik, E. *et al.* AlphaViz: Visualization and validation of critical proteomics data directly at the raw data level. *bioRxiv* 2022.07.12.499676 (2022) doi:10.1101/2022.07.12.499676.
384. Curran, T. G., Bryson, B. D., Reigelhaupt, M., Johnson, H. & White, F. M. Computer Aided Manual Validation of Mass Spectrometry-based Proteomic Data. *Methods* **61**, 219 (2013).
385. Chen, Y., Kwon, S. W., Kim, S. C. & Zhao, Y. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J Proteome Res* **4**, 998–1005 (2005).
386. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem Sci* **42**, 98–110 (2017).
387. Righetti, P. G. & Boschetti, E. Introducing Low-Abundance Species in Proteome Analysis. *Low-Abundance Proteome Discovery* 1–11 (2013) doi:10.1016/B978-0-12-401734-4.00001-4.
388. Guo, T. *et al.* Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. *iScience* **21**, 664 (2019).
389. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology 2014 32:3* **32**, 219–223 (2014).
390. Ludwig, C. *et al.* Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial . *Mol Syst Biol* **14**, 8126 (2018).
391. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications 2018 9:1* **9**, 1–12 (2018).
392. Egertson, J. D., MacLean, B., Johnson, R., Xuan, Y. & MacCoss, M. J. Multiplexed Peptide Analysis using Data Independent Acquisition and Skyline. *Nat Protoc* **10**, 887 (2015).
393. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods 2019 17:1* **17**, 41–44 (2019).
394. Guan, S., Taylor, P. P., Han, Z., Moran, M. F. & Ma, B. Data Dependent-Independent Acquisition (DDIA) Proteomics. *J Proteome Res* **19**, 3230–3237 (2020).
395. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).
396. Zhang, F., Ge, W., Ruan, G., Cai, X. & Guo, T. Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *Proteomics* **20**, 1900276 (2020).
397. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* **16**, 63–66 (2019).
398. Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat Methods* **14**, 903–908 (2017).
399. Nakamori, M. *et al.* Splicing biomarkers of disease severity in myotonic dystrophy. *Ann Neurol* **74**, 862–872 (2013).
400. Wang, E. T. *et al.* Transcriptome alterations in myotonic dystrophy skeletal muscle and heart. *Hum Mol Genet* **28**, 1312–1321 (2019).
401. Otero, B. A. *et al.* Transcriptome alterations in myotonic dystrophy frontal cortex. *Cell Rep* **34**, (2021).
402. Solovyeva, E. M. *et al.* Integrative Proteogenomics for Differential Expression and Splicing Variation in a DM1 Mouse Model. *Molecular & Cellular Proteomics* **23**, 100683 (2024).
403. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**, 1419–1440 (2005).

404. Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nature Methods 2013 10:3* **10**, 186–187 (2013).

405. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem (Palo Alto Calif)* **9**, 499 (2016).

406. Michalski, A. *et al.* Ultra high resolution linear ion trap orbitrap mass spectrometer (orbitrap elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Molecular and Cellular Proteomics* **11**, 1–11 (2012).

407. Ahlf, D. R. *et al.* Evaluation of the compact high-field orbitrap for top-down proteomics of human cells. *J Proteome Res* **11**, 4308–4314 (2012).

408. Compton, P. D., Zamdborg, L., Thomas, P. M. & Kelleher, N. L. On the scalability and requirements of whole protein mass spectrometry. *Anal Chem* **83**, 6868–6874 (2011).

409. Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochem Biophys Res Commun* **445**, 683 (2014).

410. Oukhaled, G. *et al.* Unfolding of proteins and long transient conformations detected by single nanopore recording. *Phys Rev Lett* **98**, (2007).

411. Payet, L. *et al.* Thermal unfolding of proteins probed at the single molecule level using nanopores. *Anal Chem* **84**, 4071–4076 (2012).

412. Rodriguez-Larrea, D. & Bayley, H. Multistep protein unfolding during nanopore translocation. *Nat Nanotechnol* **8**, 288–295 (2013).

413. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat Biotechnol* **31**, 247–250 (2013).

414. Ouldali, H. *et al.* Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nature Biotechnology 2019 38:2* **38**, 176–181 (2019).

415. Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science* **374**, 1509–1513 (2021).

416. Yan, S. *et al.* Single Molecule Ratcheting Motion of Peptides in a Mycobacterium smegmatis Porin A (MspA) Nanopore. *Nano Lett* **21**, 6703–6710 (2021).

417. Chen, Z. *et al.* Controlled movement of ssDNA conjugated peptide through Mycobacterium smegmatis porin A (MspA) nanopore by a helicase motor for peptide sequencing application. *Chem Sci* **12**, 15750–15756 (2021).

418. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682–D688 (2020).

419. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).

420. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* (2018) doi:10.1093/nar/gky092.

421. Lane, L. *et al.* neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res* **40**, D76 (2012).

422. Humphery-Smith, I. A human proteome project with a beginning and an end. *Proteomics* **4**, 2519–2521 (2004).

423. Smith, L. M. *et al.* The Human Proteoform Project: Defining the human proteome. *Sci Adv* **7**, 734 (2021).

424. Cao, X., Sun, S. & Xing, J. A massive proteogenomic screen identifies thousands of novel peptides from the human "dark" proteome. *Molecular & Cellular Proteomics* 100719 (2024) doi:10.1016/J.MCPRO.2024.100719.

425. Brunet, M. A. *et al.* OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res* **49**, D380–D388 (2021).

426. Raj, A. *et al.* Proteogenomics 101: a primer on database search strategies. *Journal of Proteins and Proteomics 2023 14:4* **14**, 287–301 (2023).

427. Krasnov, G. S. *et al.* PPLine: An automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics. *J Proteome Res* **14**, 3729–3737 (2015).

428. Nagaraj, S. H. *et al.* PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization. *J Proteome Res* **14**, 2255–2266 (2015).
429. Jagtap, P. D. *et al.* Flexible and accessible workflows for improved proteogenomic analysis using the galaxy framework. *J Proteome Res* **13**, 5898–5908 (2014).
430. Guilloy, N. *et al.* OpenCustomDB: Integration of Unannotated Open Reading Frames and Genetic Variants to Generate More Comprehensive Customized Protein Databases. *J Proteome Res* **22**, 1492–1500 (2023).
431. Orengo, C. A., Todd, A. E. & Thornton, J. M. From protein structure to function. *Curr Opin Struct Biol* **9**, 374–382 (1999).
432. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
433. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).
434. Chasman, D. & Adams, R. M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* **307**, 683–706 (2001).
435. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science (1979)* **381**, (2023).

Data management plan

Data sharing

All studies in this thesis were/will be published open access. The RNA/proteomics data that were generated for this thesis were deposited in their suitable respective databases. Raw RNA sequencing data was deposited to EGA under the accessions EGAS00001006779 and EGAS50000000007 (Chapter 3) and EGA50000000188 (DM1 project). Proteomics data generated for Chapter 3 was deposited to PRIDE with accession PXD045237. All other data was publicly available as outlined in the respective chapters. The code used in the analysis of the data or the development of tools are stored in the following public Github repositories:

Chapter 2 – https://github.com/cmbi/NA12878-saav-detection

Chapter 3 – https://github.com/cmbi/hpi_isoseq_paper (MIT license)

Chapter 4 – https://github.com/cmbi/SUsPECT (Apache 2.0 license)

Ethics and privacy

PBMCs from Chapter 3 and 4 were retrieved form healthy, anonymized donors, as part of the human functional genomics project (HFGP). The HFGP study was approved by the Ethical Committee of Radboud University Nijmegen, the Netherlands (no. 42561.091.12). Experiments were conducted according to the principles expressed in the Declaration of Helsinki. Samples of venous blood were drawn after informed consent was obtained.

Summary

Large-scale DNA sequencing efforts in the past decade have led to a staggering volume of discoveries in human genetic variation. However, our understanding of genetic variant *effects* is lagging behind. The effect of variants can only be accurately predicted when the expression of DNA throughout the entire biological system is fully understood, including transcripts and proteins. At present, both the reference human transcriptome and proteome are incomplete. They are missing many of the different proteins, or slightly altered versions of the same proteins (called proteoforms), that are only present in a specific tissue, time, person, or cellular condition. These protein variations can reveal much about regulation that occurs in the cell and its malfunctioning in disease.

The missing knowledge is largely attributed to technological limitations. Short-read RNA sequencing, the current standard, does not have the resolution needed to observe the rich diversity of human transcript isoforms. Full transcript sequences must be inferred based on small sequence fragments, and they are often wrong. Long-read RNA sequencing is a relatively recent solution that captures the entire transcript sequence. This new technology is rapidly expanding our understanding of human genetic expression; novel transcripts are being discovered in droves. Understanding how these findings affect the proteome is important, as proteins play crucial roles in the structure and function of cells in living organisms.

Unfortunately, the technological limitations are even worse on the protein level than the RNA level. In a typical proteomics experiment, proteins from a sample are broken down into peptides, which are then measured with a mass spectrometer. A spectrum is produced for each peptide. The spectra cannot be accurately read out as a sequence directly; to identify them, they must be compared to every peptide in a database containing all peptides that are expected to be in the sample. The problem is that if a database contains only previously-known peptides, the discovery of variant peptides is impossible.

One solution is the use of proteogenomics. Proteogenomics is a relatively recent method that leverages nucleotide data to enable identification of variant peptides. The growing abundance of genetic data has made proteogenomics an increasingly powerful tool in this regard. Predicted proteins inferred from sequencing data, including all genetic and transcriptomic variation, are added to the database used to search spectra. The composition of the search database is a central aspect in proteogenomics, and also its greatest challenge: ironically, the larger the database, the less likely spectra are to be identified.

Short-read transcriptomes resulted in larger databases. Several potential coding regions known as open reading frames (ORFs) needed to be added per transcript to observe variant peptides, since ORF could not be resolved from sequence fragments. Using long-read transcriptomes reduces the number of additions to the search database by providing one confident ORF per transcript. In **Chapter 2**, we assess the state-of-the-art in variant peptide identification. We compare the ability of new long-read proteogenomics methods to that of the latest computational proteomics methods to detect genetic variants in peptides. Based on a well-characterized cell line NA12878, we successfully showed that using long-read proteogenomics indeed results in more accurate variant peptide identification.

This thesis also aims to fill the aforementioned knowledge gap in the reference annotation by studying cells under specific conditions. To this end, we performed a multi-omics characterization of pathogen-stimulated human immune cells in **Chapter 3**. Four different pathogens were used, including bacteria, fungal and viral types. Long-read RNA sequencing revealed the presence of many novel transcript isoforms (around 40% of unique transcripts) in both control and all pathogen-stimulated cells. Thanks to the accurate elucidation of transcript isoforms, we were able to study isoform switching (IS) in addition to general gene/transcript differential expression. IS is a lesser-studied phenomenon where the relative isoform expression changes *within* a gene in response to a condition, regardless of overall gene expression. We found 398 genes taking part in IS, the majority of which were not differentially expressed on the gene level. The IS events occurred in a wide variety of genes involved in metabolic processes, mRNA splicing, protein transport and catabolism. About half of all IS cases involved a novel transcript. Protein evidence of the alternative splicing events could not be confirmed in the secreted proteome; we suggest using whole cells for proteomics analysis in future studies to fully uncover the rich transcriptomic and proteomic diversity resulting from pathogen stimulation.

The novel transcripts found in long-read transcriptome studies like Chapter 3 can be directly leveraged to benefit patients of rare, undiagnosed diseases. We developed a software pipeline called SUsPECT in in **Chapter 4** that uses sample-specific transcripts to re-analyze genetic variants of patients with rare disease. The pipeline uses a variety of tools to predict ORFs of novel transcripts, predict variant effects on new transcripts, compare these to the old effects, and provide missense effect predictions where applicable. The end result is a list of variants with a more severe predicted effect in the provided sample than the reference. In practice, this could mean that certain heart disease-causing variants in some patients could have been marked as benign when in

reality, they have a severe molecular consequence in heart-specific transcripts/proteins. We used the transcripts from Chapter 3 to show that SUsPECT uncovers candidate variants for disease causality. The pipeline is publicly available on GitHub.

There is a wide knowledge gap in the effects of genetic variation, transcript expression and protein expression on one another. Filling this gap will require the generation of much more data, further innovation in experimental methodology, and continuous development of bioinformatic tools. This thesis explores the current state-of-the-art in both experimental and bioinformatics methodologies for use in capturing proteomic variation. It demonstrates ways in which long-read proteogenomics methods can be used for detecting new biology and provides a tool to directly leverage these discoveries for disease diagnosis.

Samenvatting

Grootschalige inspanningen op het gebied van DNA-sequencing hebben in het afgelopen decennium de variatie in het menselijk genoom in kaart gebracht. Desondanks loopt ons begrip van de effecten van deze genetische varianten achter. Het effect van varianten kan alleen nauwkeurig worden voorspeld wanneer de genexpressie in het hele biologische systeem volledig wordt begrepen, inclusief transcripten en eiwitten. Op dit moment zijn zowel het referentie menselijke transcriptoom als proteoom onvolledig. Er ontbreken nog veel eiwitten, of licht gewijzigde versies van eiwitten (genaamd proteovormen), die alleen aanwezig zijn in een specifiek weefsel, op een bepaald moment, bij één individu of onder een bepaalde cellulaire conditie. Deze eiwitvariaties kunnen veel onthullen over de regulatie die in de cel plaatsvindt en het dysfunctioneren ervan bij ziekte.

De ontbrekende kennis wordt grotendeels toegeschreven aan technologische beperkingen. Short-read RNA-sequencing, de huidige standaard om transcriptomen te karakteriseren, heeft niet de resolutie die nodig is om de rijke diversiteit van menselijke transcriptisovormen waar te nemen. Volledige transcriptsequenties moeten worden afgeleid op basis van kleine sequentiefragmenten, en zijn bijgevolg vaak onjuist. Sequencing van lange RNA-sequenties is een relatief recente oplossing die de volledige transcriptsequentie vastlegt. Deze nieuwe technologie breidt snel ons begrip van menselijke genexpressie uit; er worden in groten getale nieuwe transcripten ontdekt. Het is belangrijk te begrijpen hoe deze het proteoom beïnvloeden, aangezien eiwitten cruciale rollen spelen in de structuur en functie van cellen in levende organismen.

Helaas zijn de technologische beperkingen op eiwitniveau misschien nog wel groter dan op RNA niveau. In een typisch proteomics experiment worden eiwitten uit een monster afgebroken tot peptiden, die vervolgens worden gemeten met een massaspectrometer. Dit toestel bepaalt een massaspectrum voor elk peptide. Deze spectra kunnen niet direct worden gelezen als een sequentie; om ze te identificeren, moeten ze worden vergeleken met elk peptide in een databank die alle peptiden bevat die worden verwacht in het monster. Als een databank alleen eerder bekende peptiden bevat is de ontdekking van variantpeptiden en nieuwe eiwit(vorm)en bijgevolg onmogelijk.

Een oplossing is het gebruik van proteogenomics. Proteogenomics is een relatief recente methode die nucleotidegegevens benut om identificatie van variantpeptiden mogelijk te maken. De groeiende overvloed aan genetische gegevens heeft proteogenomics tot een steeds krachtiger instrument in dit opzicht gemaakt. Voorspelde eiwitten afgeleid uit sequentiegegevens, inclusief alle genetische en transcriptomische variatie, worden toegevoegd aan de databank die wordt gebruikt om spectra te doorzoeken. De samenstelling van de zoekdatabank is een centraal aspect in proteogenomics, en ook de grootste uitdaging: ironisch genoeg, hoe groter de databank, hoe minder waarschijnlijk dat spectra worden geïdentificeerd.

Short-read transcriptomen resulteerden in grotere databanken. Verschillende potentiële coderingsgebieden die bekend staan als open leesframes (ORF's) moesten per transcript worden toegevoegd om variant-peptiden te observeren, aangezien ORF's niet kon worden opgelost uit korte sequencing-fragmenten. Het sequencen van lange RNA sequenties vermindert het aantal toevoegingen aan de zoekdatabank omdat het leesraam per transcript met grotere zekerheid kan worden vastgesteld. In **Hoofdstuk 2** beoordelen we de technologische stand van zaken voor identificatie van variant-peptiden. We vergelijken het vermogen van nieuwe proteogenomische methoden met lange leeslengte met dat van de nieuwste computationele proteomica-methoden om genetische varianten in peptiden op te sporen. Gebaseerd op een goed gekarakteriseerde cellijn NA12878, hebben we succesvol aangetoond dat het gebruik van lange RNA sequenties inderdaad resulteert in meer accurate identificatie van variantpeptiden.

Dit proefschrift heeft ook tot doel eerdergenoemde gaten in de referentieannotatie op te vullen door cellen te bestuderen onder specifieke omstandigheden. Hiertoe voerden we in **Hoofdstuk 3** een multi-omics karakterisering uit van door pathogenen gestimuleerde menselijke immuuncellen. Vier verschillende pathogenen werden gebruikt, waaronder pathogenen van bacteriële, schimmel en virale oorsprong. Lange RNA-sequenties onthulden de aanwezigheid van veel nieuwe transcriptisovormen (ongeveer 40% van de aanwezige transcripten) in zowel controle- als alle door pathogenen gestimuleerde cellen. Dankzij de nauwkeurige opheldering van transcriptisovormen konden we isovorm-switching (IS) bestuderen naast algemene gen-/transcript-differentiële expressie. IS is een weinig bestudeerd fenomeen waarbij de relatieve isovormexpressie verandert binnen een gen als reactie op een conditie, ongeacht het totale expressieniveau. We vonden 398 genen die IS ondergingen, waarvan de meerderheid niet differentieel tot expressie kwam op het gen-niveau. De IS-evenementen vonden plaats in diverse genen die betrokken zijn bij metabole processen, mRNA-splicing, eiwittransport en katabolisme. Ongeveer de helft van alle IS-gevallen betrof een nieuw transcript. Bewijs van deze alternatieve splicing-evenementen kon niet worden gevonden in het uitgescheiden proteoom; we stellen daarom voor om hele cellen te gebruiken voor proteomics-analyse in toekomstige studies om de rijke transcriptoom en proteoom diversiteit als gevolg van pathogenenstimulatie volledig bloot te leggen.

De nieuwe transcripten die zijn gevonden in lange RNA sequenties zoals gegenereerd in Hoofdstuk 3 kunnen direct worden benut om patiënten met zeldzame, ongediagnostiseerde ziekten te helpen. We hebben in **Hoofdstuk 4** een software-pijplijn ontwikkeld genaamd SUsPECT, die gebruikmaakt van monster-specifieke transcripten om genetische varianten van patiënten met zeldzame ziekten opnieuw te analyseren. De pijplijn maakt gebruik van verschillende tools om ORF's van nieuwe transcripten te voorspellen, varianteffecten op nieuwe transcripten te voorspellen, deze te vergelijken met eerder voorspelde effecten, en missense-effectvoorspellingen te geven waar van toepassing. Het eindresultaat is een lijst met varianten met een ernstiger voorspeld effect in het verstrekte monster dan in de referentie. In de praktijk zou dit kunnen betekenen dat bepaalde varianten die een hartziekte veroorzaken als goedaardig zijn aangemerkt hoewel ze in werkelijkheid een ernstig moleculair gevolg hebben in hart-specifieke transcripten/eiwitten. We hebben de transcripten uit Hoofdstuk 3 gebruikt om aan te tonen dat SUsPECT kandidaatvarianten voor ziekte-oorzaak blootlegt. De pijplijn is openbaar beschikbaar op GitHub.

Er is een groot gat in kennis in de effecten van genetische variatie, transcriptexpressie en eiwitexpressie op elkaar. Het vullen van dit gat zal vereisen dat er veel meer data wordt gegenereerd, verdere innovaties in experimentele methodologie plaatsvinden, en voortdurende nieuwe bio-informatica-tools worden ontwikkeld. Dit proefschrift onderzoekt de huidige stand van zaken in zowel experimentele als bio-informatica-methodologieën voor het vastleggen van proteoom variatie. Het demonstreert manieren waarop proteogenomics-methoden op basis van lange RNA sequenties kunnen worden gebruikt om nieuwe biologie op te sporen en voorziet algoritmen om deze ontdekkingen direct te benutten voor ziekte-diagnose.

Curriculum vitae

Renee Salz was born in Boston, Massachusetts, USA on February 7th, 1995. Two years later, she moved to San Diego, California, where she remained until graduation from La Jolla High School. Her interest in biology started early, when Dr. Robert M. Hoffman came to her high school biology class for a guest lecture and brought glow-in-the-dark (GFP-expressing) mice. At just 15 years old, Renee began her internship at AntiCancer Inc. where she used microscopy to investigate cancer cell-killing capabilities of a mutant strain of *Salmonella typhimurium* bacteria. Her research project was carried on by colleagues upon her departure, resulting in a publication.

Renee graduated with a bachelor's degree in Molecular and Cell Biology from University of California, Berkeley, with an emphasis in Biochemistry and Molecular Biology. During her time there, she completed various internships alongside her course work. She interned at the Robert Tjian lab at UC Berkeley, using fluorescence microscopy methods to study OCT4 and SOX2 transcription factor kinetics in pluripotent cells. After that, she completed a summer internship at AbbVie testing antibody drug conjugates for breast and ovarian cancer. Finally, she interned in the Oksenberg group of the Multiple Sclerosis Genetics Research Laboratory of UCSF, where she performed GWAS studies for Multiple Sclerosis and Hodgkin's Lymphoma. Through these experiences and coursework, she developed a keen interest in bioinformatics.

Renee went on to pursue a master's degree in Bioinformatics at KU Leuven in Belgium, and graduated *cum laude*. She completed her master's thesis at KeyGene NV in Wageningen, which involved a thorough investigation of gain-of-function missense variants and the development of text mining software to detect (homologous) protein variants in existing literature.

Shortly thereafter, Renee began her PhD in the group of Prof. Dr. Peter-Bram 't Hoen where she employed novel proteogenomics methods to uncover transcriptomic and proteomic variation as described in this thesis. The pipeline she built to utilize long read-transcriptomes in the diagnosis of rare disease, SUsPECT, garnered much attention in the genetics community. She presented the pipeline at three international genetics conferences and won the Rolduc prize for best presenter at the genetics retreat of BeSHG 2022. In March 2024, she joined ProQR Therapeutics BV as Scientist Bioinformatics.

List of publications

Salz R, Saraiva-Agostinho N, Vorsteveld E, van der Made CI, Kersten S, Stemerdink M, Allen J, Volders PJ, Hunt SE, Hoischen A, ’t Hoen PA. SUsPECT: A pipeline for variant effect prediction based on custom long-read transcriptomes for improved clinical variant annotation. BMC genomics. 2023 Jun 6;24(1):305.

Salz R*, Vorsteveld EE*, van der Made CI, Kersten S, Stemerdink M, Riepe TV, Hsieh TH, Mhlanga M, Netea MG, Volders PJ, Hoischen A. Multi-omic profiling of pathogen-stimulated primary immune cells. bioRxiv. 2023 Sep 15:2023-09.

Salz R, Bouwmeester R, Gabriels R, Degroeve S, Martens L, Volders PJ, ’t Hoen PA. Personalized proteome: Comparing proteogenomics and open variant search approaches for single amino acid variant detection. Journal of Proteome Research. 2021 May 17;20(6):3353-64.

Uchugonova A, Zhang Y, **Salz R**, Liu F, Suetsugu A, Zhang L, Koenig K, Hoffman RM, Zhao M. Imaging the different mechanisms of prostate cancer cell-killing by tumor-targeting Salmonella typhimurium A1-R. Anticancer research. 2015 Oct 1;35(10):5225-9.

Portfolio

Department: **Medical BioSciences**
PhD period: **03/09/2018 – 31/03/2023**
PhD Supervisor(s): **Prof. dr. P.A.C. 't Hoen**
PhD Co-supervisor(s): **Prof. dr. ir. P. Volders**

Training activities	Hours
Courses	
- Radboudumc - Introduction Day (2018)	6.00
- RIMLS - Introduction course "In the lead of my PhD" (2018)	15.00
- Mass spectrometry data processing training (2018)	16.00
- IMM - The Art of Presenting Science (2019)	33.00
- Adobe illustrator workshop (2019)	8.00
- Literature review for PhDs workshop (2019)	28.00
- Radboudumc - Scientific Integrity (2020)	20.00
- RU - Achieving your Goals and performing more successfully in your PhD (2021)	28.00
- Radboudumc - Workshop "The Next Step in my Career" (2021)	21.00
Seminars	
- RTC Bioinformatics event (2018)	6.00
- Deep learning seminar (2018)	3.00
- Theme meetings Bioinformatics (2018-2022)	112.00
- Research integrity round (2018)	3.00
- BioSB soft skills workshop on effective supervision (2018)	8.00
- Storytelling seminar (2019)	4.00
- VIB Valorization and technology transfer (2019)	8.00
- Nanopore Day (2019)	6.00
- Radboud Bioinformatics Christmas event - Oral presentation (2019)	8.00
- SMRT Leiden (2020)	6.00
- FAIR data workshop (2020)	4.00
Conferences	
- EuBIC Proteomics winter school - Poster presentation (2018)	56.00
- BioSB - Poster presentation (2019)	20.00
- X-omics festival Nijmegen (2019)	8.00
- Dagstuhl seminar computational proteomics (2019)	35.00
- EuBIC proteomics developers meeting (2020)	56.00
- BioSB - Poster presentation (2020)	20.00
- Genetics retreat - Oral presentation (2022)	20.00
- BeSHG Annual Meeting - Poster presentation (2022)	20.00
- ESHG - Poster presentation (2022)	35.00
- ASHG - Oral presentation (2022)	42.00
- ESHG - Oral presentation (2023)	35.00
Other	
- Proteogenomics journal club (2019)	56.00
- PhD Organisation Nijmegen board member (2021)	84.00
Teaching activities	
Supervision of internships / other	
- Joint supervision of HAN Bioinformatics student (2019)	28.00
- Supervision of MSc Bioinformatics student (2021)	28.00
- Joint supervision of MSc Medical Biology student (2021)	70.00
- Meet the PhD (2022)	28.00
Total	**1005**

Acknowledgements/Dankwoord

My supervision team was key to the creation of this thesis. Firstly, my supervisor **Peter-Bram ‘t Hoen**. You challenged and encouraged me from the beginning. I appreciated your punctuality, your organized and structured way of working. With your wide network, you facilitated many collaborations including all the ones in this thesis. Even with many commitments as head of the department, you still made time nearly every Wednesday to meet with me. Your excellent feedback and advice on all aspects of this trajectory were greatly appreciated. I feel lucky to have had you in my corner and I am thankful for your guidance and flexibility in these years.

My co-supervisor, **Pieter-Jan Volders**. You joined me after my first year when my original co-supervisor left. Even though we live in different countries, you stepped up to the plate. I’m really glad you did, especially since you had the expertise to guide me on every part of this diverse thesis. Your feedback on my presentations and writing hit the target every time. I quickly learned I could call you for helpful advice in all aspects of the PhD, both scientific and personal. Also, you have an impressive repertoire of great memes used at exactly the right time. Hero!

I would like to thank my mentor **Ronald Roepman** for meeting with me more regularly than was required, listening to me, and kindly offering advice/assistance. Also to my thesis committee consisting of **Peter-Bram ‘t Hoen**, **Pieter-Jan Volders**, **Ronald Roepman**, **Lennart Martens**, **Christian Gillissen**, **Hanka Venselaar** who met yearly to steer me in the right direction, thank you! I would also like to thank the Manuscript Committee, consisting of **Christian Gillissen**, **Jolein Gloerich** and **Victor Guryev** for their careful reading and evaluation of this thesis.

A special thanks goes to my genetics collaborators on the host-pathogen project: **Alex Hoischen**, **Emil Vorsteveld**, **Cas van der Made** and **Simone Kersten**. Neither this thesis nor my first-place presentation prize would have been possible without your input and guidance. Thank you for your positivity, creativity, and hard work. Collaborating with such knowledgeable and kind people was truly a pleasure.

At the end of my first year, I had the privilege of working at CompOmics in Ghent for a two-month research visit. I want to thank you **Lennart**, for welcoming me to your group. You are a constant source of brilliant ideas, and your group is full of incredibly nice people. One of them, **Pieter-Jan**, even became my co-promotor. I would also like to thank **Ralf**, **Robbin**, **Sven**, **Tim**, **Nina**, **Natalia, Surya**, and **all the other members of CompOmics** who welcomed me into the group, brought me to Gentse feesten, exposed me to my first VR gaming experience and taught me a ton about computational proteomics. Thank you for your collaboration and making my research visit so enjoyable.

I would like to thank my paranymphs **Tabea Riepe** and **Dario Marzella**. **Tabea**, you were the only other proteogenomics PhD student in the group. We were clueless together, tackling challenge after challenge of long-read RNAseq analysis together. At least we had each other- and we had fun! Everything from "proteogenomics roundtable" dinners to Los Angeles for ASHG. Thank you for your support and for being such a great listener. **Dario**- my favorite Italian, my officemate, my friend. We could chat about anything, about protein structures, politics, memes, puppies, you name it. Board game nights were always so fun with you, and the Italian delicacies you often brought back to share were beyond heavenly. You're thoughtful and kind, and I feel very lucky to have been able to rely on you time and time again. "Thank you" doesn't really cover it, but here it is anyways. Huge thanks to both of you for several years of good company and friendship.

A big thanks goes to my "god-paranymph", my beer sensei, **Laurens van de Wiel**. I don't even want to imagine how my PhD would have gone without you. You were the more senior PhD student in the department who showed me the ropes and set me on the right path from day 1. Your mentorship, and subsequently your friendship, was and still is invaluable. Thank you for sharing your code, chrome bookmarks, special beer, bluray GoT collection, random household items, and so many other things that wouldn't fit on this page. Thank you for making me laugh so much that the offices next to us complained. Thank you for including me in so many fun activities, whether it be bringing me to crash genetics borrels/parties, Donders events, aesculaaf, pub quizzes, or the like. Thank you for being my memory when I inevitably forget things, and also my speed dial during the most challenging moments of my PhD. Seriously, thank you.

The CMBI was my department and I would like to thank not only my paranymphs but also the rest of my colleagues at CMBI who supported me during my PhD. Shout out to my proteogenomics club members **Tabea**, **Daniel van As**, and **Remco** for the thoughtful discussions proteogenomics-related and otherwise; it was so helpful to have colleagues to check up on the content of my work and to bond over shared experiences. Thank you **Joeri**, for your superb sense of humor that got me through the work week, discussions about code, and being a great officemate/soup buddy/Denmark travel buddy. Thank you **Sander**, for making the department (and honestly any gathering) way livelier with your amazing energy, always adding something interesting/funny to any discussion, and being the dedicated photographer. Also thank you **Josh**, **Casper**, **Tom**, **Martijn**, **Daniel R**, **Junda**, **Bruna**, **Lisette**, **Anouk**, **Joanna**, **Li**, **Hanka**, **Farzaneh**, **Cenna**, **Anna**, **Siqi**,

Prashant, **Dei**, **Coos**, **Barbara**, **Arthur**: from our occasional outings/bbqs to our daily bioinformatics discussions/random lunch or coffee convos, you all had a part to play in making CMBI a great place to work.

Although CMBI was my department, I had the privilege to get to know some members of the genetics department in addition to my aforementioned collaborators. Sometimes this was in the context of my crashing genetics parties, but sometimes they took me in because I was often the only one from CMBI at genetics conferences. I want to thank **Lot**, **Teun**, **Juliette**, **Lex**, **Iris**, **Bart**, **Lisenka**, **Christian**, **Alex**, **Emil** and **Cas** for good times at aesculaaf throughout the years but especially for making my experience at various genetics conferences so positive. I also want to thank my fellow Californian-in-Radboudumc **Brooke Latour**- you are such an amazing, kind, accomplished human who I really look up to. Thanks for showing me great food/wine, sharing your funny stories, empathizing so well with me and helping co-paranymph Laurens. I am an only child, but if I could imagine having a big sister it would look a lot like you.

I met some great friends during my master study in Leuven and thesis in Wageningen. Thank you **Daphne**, **Theo**, **Will**, **David**, and **Christophe** for being the greatest study mates I could have asked for and making my time in Leuven unforgettable! And **Pascal**- thanks for making my time in Wageningen so fun and keeping in touch. Glad we could reconnect in Leiden after all these years!

I had some wonderful friends who were there for me during (different phases of) this period. I want to thank my girls **Aoife** and **Valentina**, who were there from the start of my PhD, and became constant sources of fun and friendship through the years. I cherished all those thought-provoking convos, hilarious gossip over high tea and late nights out dancing. I feel blessed to have such interesting, funny, inspiring women to trade stories and help me make sense of life, especially in such challenging time. **Mariana**- you made life fun with drinks, stories and spontaneity. For example, I would not have tried wakeboarding in Nijmegen if it weren't for you. Your endless positive energy and down-to-earth perspective lifted me up time and time again. **Kate**- I met you later in my PhD and you really helped pull me through the hardest part. You quickly became an important safe space in my life. Thanks for being a true friend I can rely on, go do fun things with, and a little antidote to homesickness! The world can be small, and I was fortunate enough able to reconnect with a couple friends from back home. **Nick**- from high school kids hitting up the beach in SD to go tanning, to PhDs in Europe. You're hilarious and somehow know how to get the best out of any situation. I'm so glad life brought us back together to hang out regularly in Nijmegen or Amsterdam, trying new bars/cafes, going for walks during the day or going to museums. **Falko**- after graduating together from Cal life took us separate ways but we reconnected in the final year of my PhD. Thanks for showing me cool places in Zurich and supporting me through a tough period.

I owe my success to my parents. **Mom** and **Dad**, thank you for whole-heartedly supporting all my passions, interests and choices. Thank you for raising me to be curious, to work hard and to enjoy life. You built a foundation for me to stand on my own two feet and to excel, all the while knowing that you were behind me if/when I ever stumble – regardless of the physical distance between us. Going home (or somewhere together on a family vacation) a couple times a year during my school/PhD really recharged me every time. I'm lucky to remain so well cared for even now, late into my twenties.

Over the past 8 years, I've acquired an extra family. Aan mijn schoonfamilie: **Rene**, **Marion**, **Peter**, **Steven**, **Sophie**, dank jullie wel voor de steun door de jaren heen. Bedankt voor mij eraan helpen te herinneren dat het leven meer dan werk is, voor perspectief in mijn stressvolle tijden maar ook het meevieren wanneer dingen goed gingen. Vooral met mijn eigen familie heel ver weg ben ik enorm dankbaar dat jullie mij van harte hebben verwelkomd. Heel erg bedankt!

Last but definitely not least, my better half, my boyfriend **Oscar**. You were my safe space, my happy place, a well of strength and resilience I pulled from during the past 9 years that we've been together. The years of my PhD were no exception. You cheered me on, you thought along with me whenever I encountered obstacles, provided the best distractions when I needed them, made me laugh every day without fail. Despite my all-consuming academic pursuit and a whole pandemic, we managed to make it some of the best years of our lives (often featuring other people's pets). Every day is an adventure with you- I can't wait to see what the future has in store for us.

Radboudumc
university medical center

IN DEI NOMINE FELICITER

Radboud University