Comprehensive Proteogenomics Identification and Validation of Cancer Associated Proteoforms in MCF7 cells

Avinash Yadav

Department of Chemistry
Scuola Normale Superiore
Pisa, Italy

This thesis is submitted for the degree of
Doctor of Philosophy
December 2018

Abbreviations

AmBic	Ammonium Bicarbonate
BED	Browser Extensible Data
CDS	Coding DNA Sequences
cRAP	common Repository of Adventitious Proteins
DNL	Dynamic Noise Level
dORF	Downstream Open Reading Frame
DTT	DL-dithiothreitol
EIC	Extracted Ion Chromatogram
Exome-seq	Exome Sequencing
GFF	General Feature Format
GTF	General Transfer Format
IAA	Iodoacetamide
InDel	Insertion/Deletion
LC-MS/MS	Liquid chromatography coupled to tandem mass spectrometry
LncRNA	Long non-coding RNA
mRNA	Messenger RNA
NGS	Next Generation Sequencing
pre-mRNA	Precursor mRNA
QC	Quality Control
RNA	Ribo Nucleic Acid
RNA-seq	RNA Sequencing
rRNA	Ribosomal RNA
SIS	Stable Isotope labeled Synthetic peptides
SNV	Single Nucleotide Variation
TEC	To be Experimentally Confirmed
uORF	Upstream Open Reading Frame
VCF	Variant Call Format
MAF	Minor Allele Frequency

Summary

Proteomics investigations rely on reference proteomes for the identification of proteins. These reference proteomes reflect the proteins that can be produced by an ideal organism, and so explicitly exclude protein isoforms that may be produced as a result of genetic mutation. In order to identify non-reference, or non-canonical, proteoforms the results of genomics analyses must be incorporated into the protein identification workflow. I developed such a proteogenomics workflow for the comprehensive identification and validation of non-canonical proteins. This development was performed using MCF7 cells, a widely used in-vitro model of breast cancer, because it includes a large number of pathogenic mutations. The comprehensive proteogenomics analysis of MCF7 cells was performed using customized protein sequence database searches. In addition to confirming the protein forms of variants identified by next-generation sequencing, multiple novel proteoforms were identified and validated with synthetic isotopically-labeled standards. Peptides originating from single nucleotide variants, in-frame Insertion/Deletion, upstream open reading frames, transcripts in non-canonical reading frame, long non-coding RNA, transcripts with retained intron, exon extensions, novel exons, non-consensus splicing, variants not detected by next-generation sequencing, and novel isoforms were all identified and validated. Many of the proteins have previously been reported to play a role in tumor development, but many specific proteoforms are reported here for the first time. The results amply demonstrate that the reference proteome databases from UniProt, RefSeq and GENCODE widely underestimate the complexity of the oncoproteome space.

The proteogenomics pipeline reported here was developed to be able to understand how cancer associated mutations affect the proteome, as many mutations do not lead to stable protein product. Furthermore, mutations may act through secondary routes and affect the regulation of which protein isoforms are produced, and so it is insufficient to limit the search to the direct protein analogues of the genetic mutation (i.e. altered peptide sequences produced by single-nucleotide variants and insertion/deletion events).

Table of Contents

1. Introduction 10
1.1. From genes to proteins 10
1.2. Genome sequencing 12
1.3. Genome annotation 13
1.4. Reference and variant sequences 15
1.5. Next-generation sequencing 15
1.6. Exome sequencing 17
1.7. RNA-sequencing 17
1.8. LC-MS/MS based proteomics 18
1.9. Proteogenomics 19
1.9.1. Databases for proteogenomics searches 21
1.9.2. Types of peptides in proteogenomics 21
1.9.3. Proteogenomics mapping 23
1.9.4. Limitations with current proteogenomics mapping tools 24
1.10. Objectives 26
2. Methods 28
2.1. MCF7 cells 28
2.2. Exome-seq 28
2.3. RNA-seq 28
2.4. Genomic variants in MCF7 cells 29
2.5. LC-MS/MS 29
2.5.1. Reagents for LC-MS/MS 29
2.5.2. Sample Preparation for LC-MS/MS 29

2.5.3. LC-MS/MS analysis 31
2.5.4. Reference proteome databases 32
2.6. Data formats 32
2.6.1.1. Genome annotation and sequence file formats 32
2.6.1.2. VCF file format 33
2.6.1.3. BED file format 33
2.7. Algorithm development 34
2.7.1. Extraction of reference sequences 34
2.7.2. Generation of variant sequences 36
2.7.3. Generation of customized protein databases from the reference genome and transcriptome 37
2.7.4. Noise detection in the MS/MS spectra 37
2.8. Comprehensive annotation of matched MS/MS spectra 38
2.8.1. Proteogenomic mapping 40
2.8.1.1. Mapping reference peptides to the reference genome 40
2.8.1.2. Mapping SNV peptides to the reference genome 43
2.8.1.3. Mapping exon-skipped peptides to the reference genome 44
2.8.1.4. Mapping peptides from 6 frame transcript sequences to the reference genome .. 45
2.8.1.5. Mapping InDel peptides to the reference genome 48
2.8.1.6. Mapping peptides from six-frame gene sequences 49
2.8.1.7. Mapping peptides from GNOMON predicted proteins to reference and alternate assemblies 50
2.8.1.8. Mapping peptides from 6 frame full genome searches 50
2.8.2. Generating peptide co-ordinates in BED format 51
2.9. Validation of the non-canonical peptides 52
2.10. Proteogenomics databases 53

2.11. Database searches 54
2.11.1. Discovery searches 54
2.12. Validation searches 57
2.13. Filtering proteogenomics peptides 57
2.14. Proteogenomics mapping 59
2.15. Validation of the non-canonical peptides 60
3. Results 62
3.1. Application of the DNL algorithm 62
3.2. Comprehensive annotation of matched MS/MS spectrum 63
3.3. Peptides from reference proteomes 66
3.4. Sensitivity of database search decreases with increasing database size 67
3.5. Classification of non-canonical peptides identified by proteogenomics searches 69
3.6. Validation of non-canonical peptides identified by proteogenomics search 72
3.7. Peptides from variants identified by next-generation sequencing 74
3.8. Peptides from exon-skipping events 83
3.9. Peptides from non-coding regions of protein coding transcripts 85
3.10. Peptides from alternate frame of protein coding transcripts 89
3.11. Peptides from non-coding transcripts and genes 91
3.12. Peptides from introns and exon boundaries of protein coding genes 97
3.13. Peptides from novel isoforms 100
4. Discussion 108
4.1. Choice of reference proteomes can impact which proteoforms are identified 109
4.2. Validation of peptides with missed cleavages 110
4.3. Variant missed in next-generation sequencing 111
4.4. Ambiguous proteogenomics peptides 117

4.5. Unmapped proteogenomics peptides 120
4.6. Novel proteoforms are expressed in MCF7 cells 121
5. Conclusions 123
6. References 126

1. Introduction

In the last decades an unprecendented rate of improvement has been observed in the molecular analytical technolgies for nucleic acids and proteins. The high-throughput technologies for the analysis of nucleic acids and proteins have evolved in parallel but independently. Proteogenomics is a multi omics research area that integrates the results of mass spectrometry (MS) based proteomics with next-generation sequencing (NGS) based genomics, transcriptomics or translatomics to better characterize and understand cellular systems [1]. In the past these methods were applied and studied independently to one another. Decreasing costs and improved ease-of-use (i.e. accessibility) has made it feasible to interrogate the same sample by multiple approaches. In proteogenomics the results of genome and RNA sequencing, gene expression, and protein expression are simultaenously investigated and integrated. The application of different molecular analysis methods on the same sample cohort is used to validate results and provide new insight that are not possible using any single technique. In proteogenomics customised protein sequence databases are generated using information from genome and RNA sequencing, which are then used to identify novel peptides that are not present in reference proteome databases [1].

1.1. From genes to proteins

In biology a gene is regarded as a unit of heredity that is passed from one generation to another. The genome is the complement of all genes that make up an organism. After the discovery of DNA as the genetic material [2] a gene was defined as a segment of DNA contained in a larger DNA strand that comprised a chromosome. The genome is the set of DNA strands on all the chromosomes in an individual. The genome is the blueprint of the organism, and can be said to contain the instructions in the form of protein coding and non-coding genes. The information flow from the gene begins with transcription, in which the genomic DNA is transcribed into RNA and finally translated into a protein, as depicted in **Figure 1**. The non-coding genes code for functional RNA molecules that are not translated into proteins. In the case of a protein-coding gene the transcribed RNA molecule is known as a premature mRNA (pre-mRNA). The pre-mRNA molecule undergoes a process called splicing whereby some specific segments of the pre-RNA molecule are removed and the remaining are then concatenated together. The removed segments are known as introns and the retained segments are known as exons.

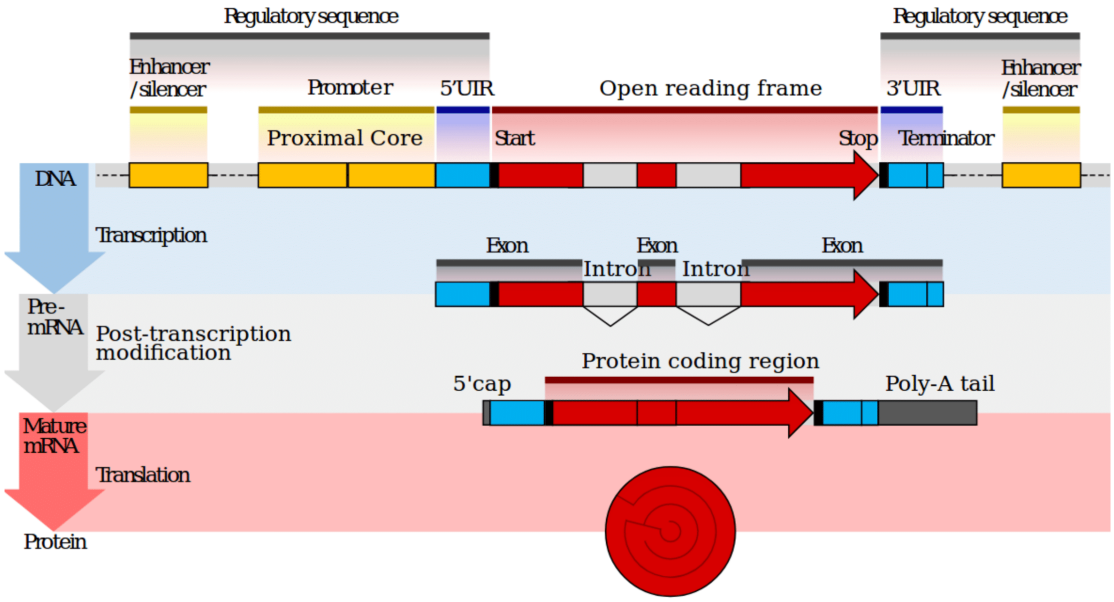

Figure 1. Steps of gene expression.

The spliced RNA sequence is known as messenger RNA (mRNA). The mRNA sequence contains an Open Reading Frame (ORF) flanked by untranslated sequences (UTR). The UTR upstream of the ORF is known as 5'-UTR and the UTR downstream of the ORF is known as 3'-UTR. During protein synthesis by ribosomes the mRNA is read three nucleotides at a time (codon) and, depending on the nucleotide code, a specific amino acid is incorporated into the growing protein chain (Figure 2A). The relationship between the codons and the incorporated amino acid is determined by the genetic code, Figure 2B.

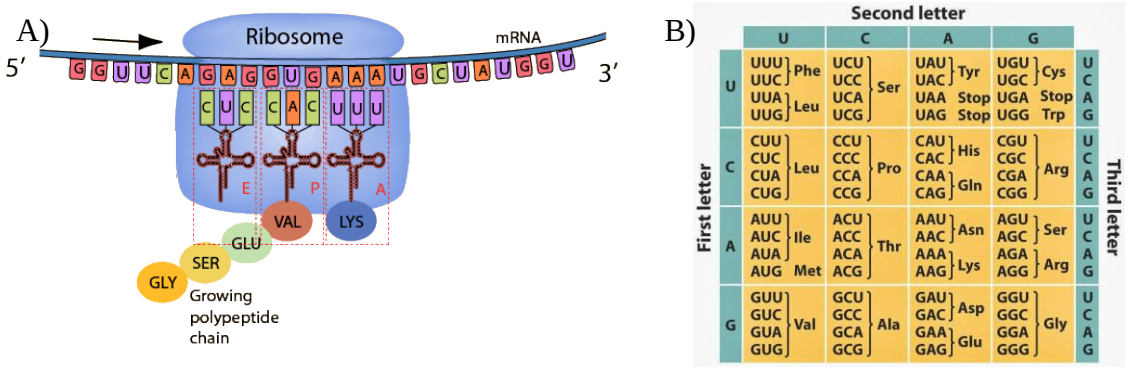

Figure 2. Protein synthesis, the mRNA is read three nucleotides at a time. B) The genetic code translates the 3-nucleotide code into the identity of the amino acid.

The genetic code is the set of rules used by cells to translate the genetic sequence contained in the mRNA into a protein sequence. Each of the 20 amino acids are carried by specific transfer RNA (tRNA) molecules, which recognize a specific codon. Translation is carried out in the ribosomes where the codon on the mRNA is recognized by the anticodon on the tRNA, after which the amino acid carried by the tRNA is incorporated into the growing protein chain. During synthesis the protein is also folded into its conformation, the three-dimensional structure essential for its biological function.

1.2. Genome sequencing

Genome sequencing is the determination of the order of the DNA nucleotides in a genome. The genome of two individuals of a species is different due to the presence or the absence of DNA variants. Thus, for comparisons between the genomes of different individuals a reference genome is required. Individual genomes can then be compared against this reference genome. To provide such a standard reference genome for our own species the Human Genome Project (HGP) was launched in 1990 [3]. In 2004, HGP published a final version of the euchromatin region of the human genome [3–5]. At the time, this was the highest quality vertebrate genome ever published. The source DNA for this project was sampled from several donors and analyzed by hierarchical shotgun assembly [4]. In this approach, a set of large insert clones of 100-200 kb each, covering the genome are generated [4]. Shotgun sequencing is then performed on selected clones [4]. The shotgun process generates fragments from random positions in the target molecule. The fragments from sequenced clones are then assembled into a linear sequence up to the total chromosome length [4]. Following the success of the HGP the Genome Reference Consortium (GRC) was established to continuously maintain and improve the human genome [6]. Besides the reference human genome the GRC also maintains and updates the mouse and zebrafish genomes [7]. The data model of the reference genome released by GRC in the public domain is called an assembly [8]. HGP and GRC provided the first comprehensive information on human gene structure which led to the growth of follow-up technologies such as genome wide association studies and genome wide gene expression profiling using microarrays [9]. Later, the introduction of next generation sequencing (NGS) allowed sequencing of the whole genome of single individuals at ever decreasing cost (currently available for approximately 600 Euro at BGI-Europe) [10].

1.3. Genome annotation

Genome annotation confers the structural and functional significance of every nucleotide in a genome. Since the sequencing of the human genome more than a decade ago, the process of annotation of the human genome is still ongoing. The result of genome sequencing is a DNA sequence containing a long string of the four nucleotides Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Genome annotation adds layers of information, including the precise location of genes, exons, introns, coding DNA sequence (CDS), and UTR, onto the DNA sequence [11]. This is known as structural genome annotation. Functional annotation concerns the biological function, regulation and expression analysis of its structural elements. The use of the word genome annotation in this thesis specifically relates to its structural annotation.

The reference human genome assembly released by GRC is annotated independently by the National Center of Biotechnology Information (NCBI) [12] and GENCODE [13]. The NCBI Eukaryotic Genome Annotation Pipeline is an automated pipeline that produces structural annotations of coding and non-coding genes, transcripts and proteins on the finished and unfinished public genome assemblies [8]. The annotation pipeline outputs the set of genes and their placement on the genome sequence. It provides content for various NCBI resources including Nucleotide, Protein, BLAST, Gene, and the Genome Data Viewer (GDV) [8]. Core components of the pipeline are the alignment programs Splign [14], ProSplign [11], and Gnomon [11], a gene prediction program combining information from experimental evidence and from ab initio models [11]. The GENCODE Consortium aims to identify all gene features in the human genome using a combination of computational analysis, manual annotation and experimental validation [13]. It provides refined annotations by integrating Ensembl automated predictions and the Human and Vertebrate Genome Analysis and Annotation (HAVANA) manual annotations [13]. The annotated gene models are divided into categories on the basis of their functional potential and the type of available supporting evidence [13]. The genes are categorized into protein-coding gene, long noncoding RNA (lncRNA) gene and pseudogenes [13]. At the transcript level additional biotypes reflect functionality, for example, protein coding or subject to nonsense mediated decay (NMD) [13]. A status is assigned at both the gene and transcript level: known (represented in the HUGO Gene Nomenclature Committee (HGNC) database and RefSeq); novel (not currently represented in HGNC or RefSeq databases but supported by

transcript evidence or evidence from a paralogous or orthologous locus); or putative (supported by transcript evidence of lower confidence) [13]. An example of a segment of the reference human genome (GRCh38) annotated by NCBI and GENCODE (Ensembl) is shown in Figure 3. The annotation process is an ongoing effort thus each new release may update previous annotations. These updates can affect the number and structures of some genes, their corresponding transcripts and proteins.

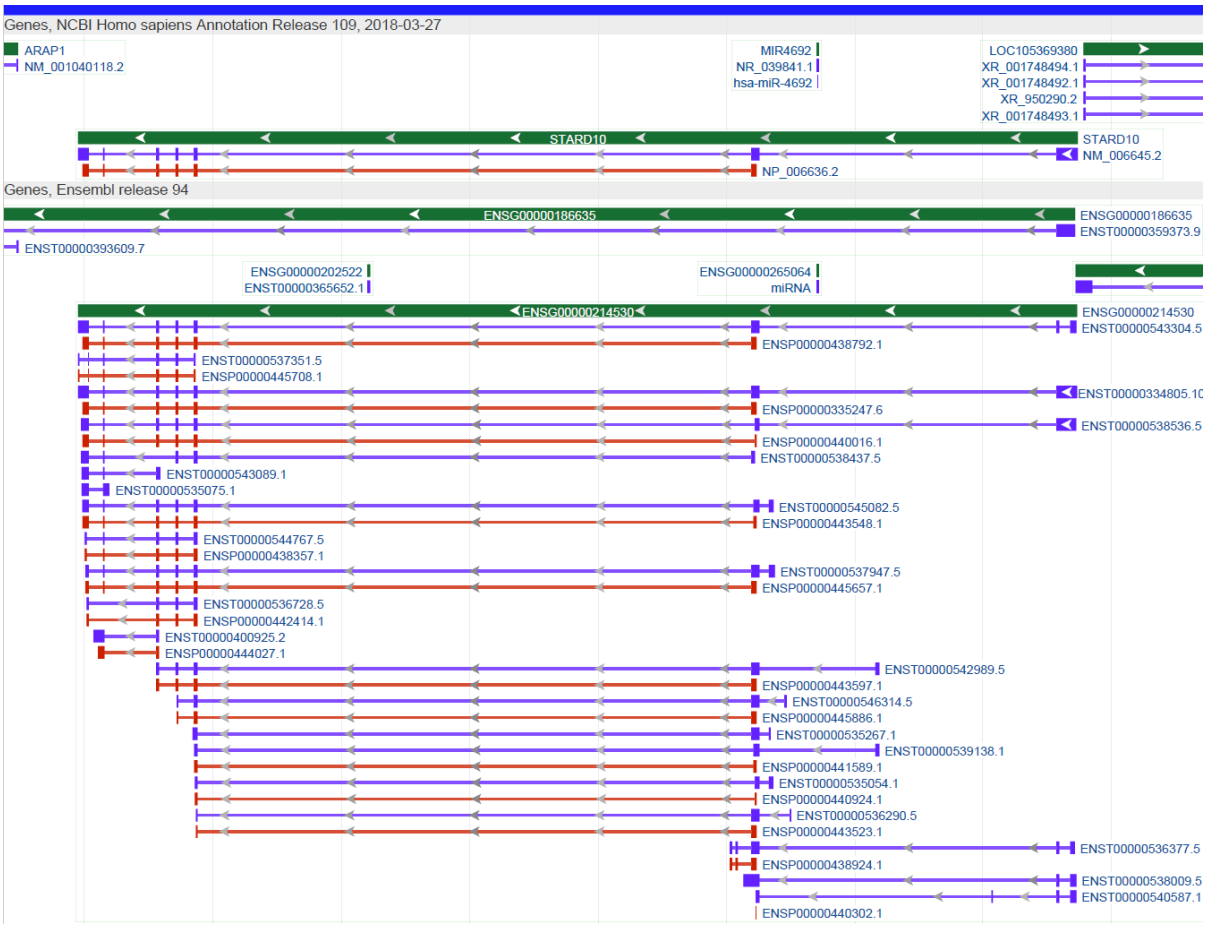

Figure 3. Genome annotations of a section of the human genome as seen in the Genome Data Viewer (GDV) from NCBI. The displayed region is from chr11: 72751835-72798640. The DNA sequence is shown at the top as a blue block. Annotations from NCBI release 109 and Ensembl (GENCODE) release 94 are shown in separate tracks. In each track, genes, transcripts and proteins are colored green, purple and red respectively. Exons are shown as blocks and introns as straight lines connecting exons. The annotation from the NCBI shows a single gene labeled STARD10, a single transcript labeled NM_006645.2 (purple line) and a single protein NP_006636.2 (red line). The Ensembl annotation contains two protein coding genes, Ensembl ids: ENSG00000186635 (ARAP1) and ENSG00000214530 (STARD10), the latter of which has multiple transcripts (purple lines) and multiple protein isoforms (red lines). Note: the start of ARAP1 is extended upstream in Ensembl annotation compared to NCBI annotation.

1.4. Reference and variant sequences

The assembled human genome provided by GRC represent a haploid assembly sampled from many individuals [6]. The human genome (GRCh38), is a composite representation (a consensus) of the human genome [6]. It is not a person's genome. Variations in some regions of the human genome are so extreme that it is not possible to represent them in a single consensus sequence [7]. These regions are represented as alternate loci assembly units in the genome, for example highly variable histocompatibility gene segments on chromosome 6 [8]. Genome annotation pipelines utilized by NCBI and GENCODE provide the set of genes, transcript and proteins and their precise locations in the genome. This set of genes, transcripts and proteins are known as reference sequences. The reference sequences are the same for all individuals of a species. The genome of every individual is different due to the presence of DNA variants such as SNPs, block substitutions, homozygous indels, heterozygous indels, inversions, segment duplications and copy number variations [15]. Furthermore the real genome of a person is diploid and is different than the genome of another person. The differences in the genome manifest at the transcript and the protein level, giving rise to molecules that will be different than the reference sequences. Thus, two individuals may produce a protein that is different in its primary sequence, or two individuals may produce a protein that are identical in their primary sequence but the encoding mRNA is different. The different sequences in different individuals are known as variant sequences. dbSNP is an NCBI resource of short variants (SNVs and InDels) [16]. As of dbSNP build 151, more than 660 million small variants have been reported in humans, out of which more than 381 million variants are localized in various genes of the human genome [17]. This massive explosion in the numbers of variants identified in humans is largely due to the advent of NGS technologies.

1.5. Next-generation sequencing

First generation genetic sequencing technologies, such as capillary based Sanger sequencing and shotgun sequencing, were utilized for the generation of the sequence data of the HGP [3]. The shotgun sequencing was performed on the insert clones and not on the whole genome, in order to eliminate the issue of long range misassembly and reduce the risk of short range misassembly [4]. These assembly problems were perceived to be profound for the human genome due to the presence of almost 50% repeat sequences [4]. In 2007, the introduction of

NGS technologies substantially lowered the cost of genetic sequencing. The application of NGS methods require the target DNA to be fragmented into smaller segments. These fragments are then fixed onto a medium and amplified by polymerase chain reaction (PCR) into colonies. The nucleotide sequences of the fragments in a colony are then determined using modified nucleotides. The modified nucleotides emit a light signal when integrated in to the growing chain, thus revealing their identity and the genome sequence. An image capture device is used to record the light signals, Figure 4.

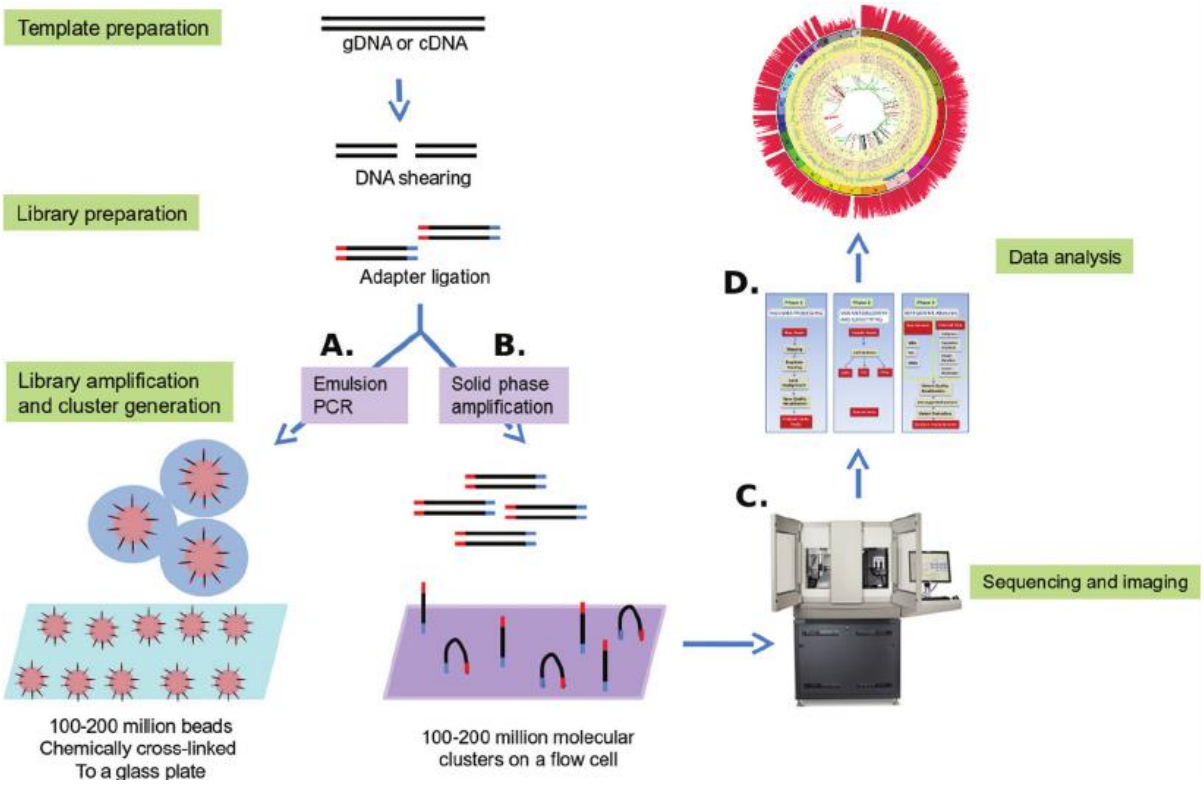

Figure 4. NGS technologies: template preparation, sequencing and data analysis. DNA is sheared by sonication or nebulisation to form fragments of 300–500 bp. Library amplification by either emulsion PCR or solid-phase amplification, followed by sequencing and data analysis.

Different NGS platforms differ in the way the colonies are formed, amplified and how the nucleotide sequence is determined [18]. High throughput is essential and is achieved by sequencing millions of colonies in parallel. The sequence data generated by the NGS platforms are referred to as reads, which is typically between 75 to 500 base pairs depending on the

platform. The ability to sequence massive amounts of DNA has enabled the investigation of genome sequences. Wang *et al.* utilized it to sequence the diploid genome of a person [10]. Yi *et al*. analyzed the exonic regions from 50 individuals for discovering adaptations to higher altitude [19]. By converting mRNA into DNA one can evaluate gene expression at the genome scale and discover novel transcripts and splice isoforms [20].

1.6. Exome sequencing

The sequences corresponding to the exons in a genome is known as the exome. In humans, the exome comprises approximately 1% of the full genome sequence[3]. In whole exome sequencing (WES) the targets are the exons of all protein coding genes. Targeting only a subset of the genome in WES lowers the sequencing cost and simplifies the data analysis (compared to WGS). WES can be used for the identification of genetic variants that affect heritable phenotypes, which includes both pathological and natural variants [21–24]. To selectively capture the exon sequences two types of technologies exist, solution-based and array-based exome capture. In solution-based exome capture the genome is fragmented and biotinylated probes are used to selectively capture and hybridize the exon sequences [25]. The hybridized targets are captured using magnetic streptavidin beads and the untargeted sequences are washed away. In array-based capture the probes are bound on high-density microarrays [26]. The probes selectively capture the exon sequences and the untargeted fragments are washed away. The captured targets are then amplified by PCR and sequenced by NGS [25]. The reads generated by the NGS platforms are then aligned to the reference genome followed by variant calling to detect small mutations (SNVs and InDels) in the sample [27].

1.7. RNA-sequencing

RNA-sequencing (RNA-seq) with NGS technology is used to identify and quantify RNA molecules. It is widely utilized for cataloging transcript species, such as mRNAs and lncRNAs, to determine the transcriptional structure of genes, and differential gene expression analysis, for example between normal vs disease state [20]. Unlike Exome-seq where precise knowledge of target sequences (exons) is required RNA-seq can be performed without any prior knowledge of the target sequences (transcripts) [20]. It can be applied for de-novo construction of the transcriptome for those species whose reference genome is not yet sequenced [28]. It can

also be applied for the detection of DNA variants [29]. In RNA-seq the target RNAs (for ex. poly-A containing mRNAs) are extracted using poly-T oligo attached magnetic beads. The purified mRNAs are then fragmented into small pieces. The cleaved RNA fragments are then copied into cDNA sequences using reverse transcriptase. The single stranded cDNA fragments are then copied into double stranded cDNA sequences. Adapters are then attached to the fragments. Each fragment with or without amplification is then sequenced by NGS. After sequencing, the resulting reads are either aligned to a reference genome or reference transcriptome, or assembled de-novo to construct the transcriptome map if the assembled genome is unavailable [18].

1.8. LC-MS/MS based proteomics

In mass spectrometry (MS) based proteomics the goal is to identify and quantify all proteins present in the sample. In the bottom-up proteomics approach the proteins are extracted and digested with a proteolytic enzyme such as trypsin. The resulting mixture of peptides is separated by Liquid Chromatography (LC) and injected into a mass spectrometer, where the peptides are ionized and accurate measurements made of the peptide's mass and isotopic profile [30]. Modern mass spectrometers are also able to isolate peptides (on the basis of the measured molecular mass) and then dissociate the isolated peptides into structurally informative fragments. This process, termed tandem mass spectrometry or MS/MS for short, provides the raw data used for the identification of proteins. The precursor ion mass (i.e. isolated and subject to MS/MS) and the masses of the structurally informative fragments are submitted to proteomics search engines [31]. The search engines statistically match the experimental data (masses of precursor and MS/MS fragments) to theoretical data (masses of precursor and expected fragments) of the peptides predicted from a database of known protein sequences and assign a score for every match. The peptide with the highest scoring match between the experimental and theoretical data is reported as a Peptide Spectrum Match (PSM). Proteins are then inferred from the identified peptides, Figure 5. To evaluate the False Discovery Rate (FDR), a target decoy search strategy is utilized in which the same data is also searched against a database of decoy proteins. The decoy protein sequences are created from the target proteins by reversing or randomizing the amino acid sequences. The proportion of

PSMs in the decoy database to the total PSMs (target PSMs + decoy PSMs) above a score threshold is utilized to estimate the rate of false positives [32].

The proteomics search engines utilized for protein identification require a database of protein sequences, usually this database is the reference proteome of the organism of the sample. A protein can only be identified if it is contained in the database. For many applications, especially cancer, the reference proteome may not contain all of the proteins that may be expressed.

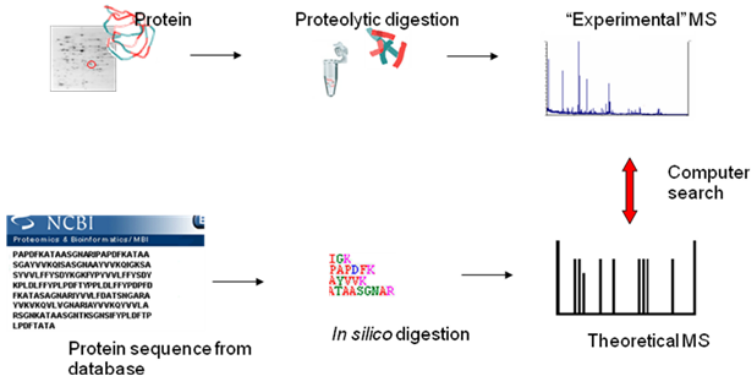

Figure 5. Summary of protein identification by mass spectrometry.

1.9. Proteogenomics

Proteogenomics integrates the data generated from genome and transcriptome sequencing into the proteomics data analysis pipeline. The central idea in proteogenomics is to interrogate proteomics data using customized protein sequence databases that are derived from genome or transcriptome sequencing. Peptides identified from these custom databases but which are not part of the reference proteome can be utilized to discover novel genes, correct existing gene annotations and confirm the expression of variant proteins (e.g. resulting from mutation) [33 34]. Initially proteogenomics was used for the correction of existing gene models [35]. Lately it has emerged as a powerful tool in the study of cancer [36]. Cancer is driven by genomic alterations that result in a series of genomic changes that include mutations, methylations, copy number aberrations and translocations [1]. To understand the molecular changes associated with cancer deep genome sequencing has been performed, for example the International Cancer Genome

Consortium and The Cancer Genome Atlas (TCGA) projects [37]. It was later understood that the definition of the cancer proteome was also vital to link cancer genotypes to phenotypes. To accelerate the knowledge of the molecular basis of cancer through the application of quantitative, proteomic technologies the Clinical Proteomics Tumor Analysis Consortium (CPTAC) was launched under the auspices of the National Cancer Institute (NCI) [38]. It carries out large scale proteome characterization of matched tumor samples which had undergone genome and transcriptome sequencing in the TCGA projects [38] . The complexity and high-throughput nature of each omics technology is not amenable to manual interpretation. Thus, bioinformatics plays a vital role in proteogenomics for data integration and its interpretation [39].

Genetic mutations accumulate during cancer progression and change the proteome landscape by translation of variant proteins [40], aberrant proteins [41], alternative splice isoforms [42], upstream open reading frames (uORFs) [43,44], long non-coding RNAs (LncRNAs) [45] and novel protein coding sequences (CDS) [46]. uORFs are protein translations from the 5'-UTR of the mRNA and always precede the natural start site in the mRNA [47,48]. uORFs have been found to regulate the expression of the main ORF [49,50]. Likewise, dORFs are translations from the 3'-UTR of the mRNA and always end after the natural stop site in the mRNA [48]. LncRNAs are transcript classes that do not code for proteins because they do not contain a long open reading frame [51], nevertheless evidence for their active translation has been reported [48,52–54].

The compact reference proteome databases from UniProt [55], RefSeq [56], GENCODE [13] used to identify proteins in LC-MS/MS experiments only contain curated reference protein sequences, and so cannot be used to identify peptides from genomic variants or novel proteoforms. LC-MS/MS analysis of a cancer proteome using such a reference protein database limits the analysis to proteins that are expressed by normal human cells. As such it under utilizes the sensitivity of MS methods to identify mutant or aberrant peptide signatures present in the sample that cannot be explained by a reference human proteome [57].

Many specific germline mutations are strongly associated with disease and have been essential to our understanding of the molecular basis of many cancers [58–61]. For example the identification of germline mutations in succinate dehydrogenase genes in patients with head and neck paragangliomas [62] and pheochromocytomas [63], inactivating mutations of fumarate hydratase in hereditary leiomyomas and type 2 papillary renal cell carcinoma syndrome [64], and

mutations in the isocitrate dehydrogenases genes in patients with Ollier disease [65] and low grade/secondary gliomas [66].

In personalized proteogenomics sample specific mutation data is used to generate a patient-and-sample-specific protein sequence database [57,67]. Proteogenomics pipelines have been reported for the mutations and isoforms identified by NGS experiments [68]. It should be noted that all of these types of genomic alterations may occur simultaneously, and protein sequences may also originate from supposedly non-coding transcripts and non-coding regions of protein coding transcripts [69,70]. Furthermore there are many common protein modifications that may occur of biological origin (acetylation, deamidation, methylation, etc.), or that occur during the preparation of the samples (e.g. oxidation) or during the mass spectrometry analysis itself (loss of water, ammonia). When proteogenomics is used to identify novel proteoforms resulting from genomic mutation it is essential to first consider these common protein modifications, and to consider all types of mutation. The evidence level needed for the confident identification of novel proteoforms, especially those related to disease, necessarily exceeds that used for the routine identification of normal proteins.

1.9.1. Databases for proteogenomics searches

In proteogenomics the first step is the construction of customized protein sequence databases. The types of database utilized depend upon the goals of the proteogenomics experiment [71]. If the goal is to discover protein variants, a database of variant proteins predicted from the genomic variants must be constructed [71]. Likewise, if the goal is to identify novel protein splice isoforms, a database of novel proteoforms predicted from the reference genome or measured transcriptome can be generated. A database of ORFs from transcript sequences can be generated to discover uORFs, dORFs, alternate frame translations and translations from supposedly non-coding transcripts. A database of ORFs from gene sequences or full genome can be generated to discover intronic and novel CDS translations.

1.9.2. Types of peptides in proteogenomics

Peptides identified from the customized databases used in proteogenomics searches must be mapped onto the reference genome. By mapping these peptides onto the same genomic co-ordinate system that is utilized to view NGS data the results can be placed in their genomic

context. All peptides in a proteogenomics search can be divided into two major groups, intragenic and intergenic peptides [71]. Intragenic peptides map onto the annotated gene segments of the genome whereas intergenic peptides map onto the unannotated segments of the genome, Figure 6. The intragenic peptides can then be classified as protein-coding or non-coding based on the biotypes of the genes onto which they map. The non-coding peptides map onto non-coding genes such as pseudogenes and lncRNA genes. Most peptides from an MS experiments map onto exons or between two or more exons (exon-spanning) of the protein coding genes. A small fraction can have a different origin, such as non-coding regions of protein coding genes (Intron, 5'-UTR, 3'-UTR). Others may map onto the protein coding transcripts in an alternate frame (Alt-frame). Intragenic peptides can also span the boundary of exons and introns (Exon-extension), or map onto unannotated alternative splice junctions (Exon-skipping). Peptides discovered from a personalized proteome (SNV and InDel) may map onto CDS regions of the gene. Finally, some peptides may map onto novel exons predicted by the RNA-seq data (novel isoforms).

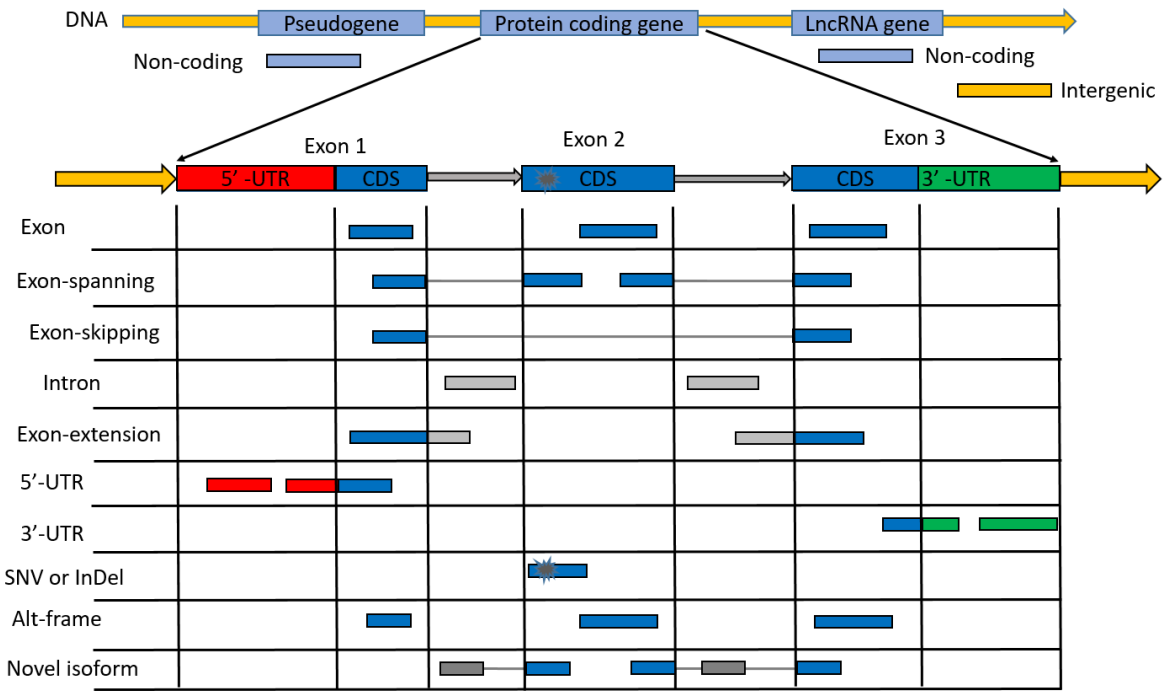

Figure 6. Types of peptides that can be identified in a proteogenomics experiment.

1.9.3. Proteogenomics mapping

Search engines utilized in the identification of proteins from MS/MS spectra report identified proteins by their names or identifiers. The protein names or identifiers are extracted from the fasta headers of the sequences contained in the protein database supplied to the search engines. If gene level identification is desired a link between the identified protein and its encoding gene is established by matching the protein identifiers to the encoding genes. Peptides in proteogenomics searches may originate from reference proteins, SNVs, InDel, Exon-skipping, uORFs, altCDS, dORFs, Exon-extension, noncoding RNAs, novel isoforms, pseudogenes and novel CDS. Although all peptides identified in an LC-MS/MS experiment have a genomic origin: Genome → Gene → transcript → protein → peptide, search engines do not report peptides by their genomic co-ordinates. By mapping them onto a coordinate system that also enables the visualization of the corresponding genomic features one can learn about their genomic context. The coordinate system of choice is the reference genome of the organism. This coordinate system is extensively utilized for the display of NGS data, such as sequence reads, variants, and novel isoforms. Genome browsers such as the Integrative Genomics Viewer (IGV) and Genome Data Viewer (GDV) enable the visualization of genomic features alongside the NGS data. Thus, by mapping peptides onto the reference genome one can visualize them with their associated genomic features. The mapping procedure should report the result in a format that can be easily understood by the genome browsers. This is known as (proteo)genomics mapping. Currently, several tools exist that can be utilized for this purpose, such as PGx [72], Peppy [73], Proteogenomics Mapping Tool [74], Pepline [75], ProteoAnnotator [76], MSProGene [77], GalaxyP [78] and PoGo [79]. Most available tools can be broadly divided into two groups: pipeline dependent and pipeline independent. The pipeline dependent tools such as Peppy and GalaxyP limit the users to a specific analysis pipeline, and thus a specific method for FDR, search engine etc... Although pipeline dependent tools can be user friendly in the sense they are designed with considerations to proteogenomics mapping, they limit the user's freedom with regard to the bioinformatics options for peptide identification. Pipeline independent tools such as PoGo require the set of peptides as input, but generate the peptide co-ordinates in Browser Extensible Data (BED) file format, which is easily understood by genome browsers such as IGV and GDV.

1.9.4. Limitations with current proteogenomics mapping tools

Although, many tools are available for mapping peptides to their genomic origin they have limitations in sensitivity and specificity [79] . BLAST is a powerful tool for alignment of query nucleic acid or protein sequences to a database of nucleic acids or proteins [80] . TBLASTN is one of the tools in the BLAST tool suite that can map a query protein sequence to a database of nucleic acids such as a reference transcriptome or reference genome. It can also be utilized to map peptides by customizing the TBLASTN search parameters for short sequences. TBLASTN suffers from sensitivity if the peptide sequence is too small and/or contains low sequence diversity. For example, the SNV peptide "L**L**LEEEQKEEEER" produced due to a mutation at chr10 position 3200020, cannot be mapped by TBLASTN (version 2.8.1) onto human RefSeq transcripts, genes or the full human genome. Furthermore, the TBLASTN output cannot be directly utilized in genome browsers and has to be converted to BED format or General Transfer Format (GTF).

PoGo is another tool that can map reference and SNV peptides to their genomic co-ordinates. To perform mapping PoGo requires the peptides and associated PSMs as a text file, a reference annotation file (GTF), and a reference protein sequence file (FASTA). One of the output formats of PoGo is BED file format, which can be directly utilized in genome browsers for visualization. PoGo is a fast peptide mapper and can be successfully applied to map thousands of peptides directly onto the reference genome, and is also able to accommodate up to 2 SNVs on the peptides [79]. Although PoGo is adept at mapping reference and SNV peptides onto the reference genome, it cannot map peptides produced due to Exon-skipping, InDel mutations, non-coding regions of protein coding transcripts, or non-coding transcripts. PoGo maps all peptides onto the reference proteome then transforms the peptide locations onto the reference genome without incorporating any knowledge of the detected SNVs. Even if the peptide has been generated from a VCF file of detected variants (NGS guided), PoGo does not utilizes this information and can map the SNV peptide to all possible genomic co-ordinates or reference proteins. For example, the SNV peptide "TN**T**FPLLEDEDDLFTDQKVK" of the gene WASHC2A and produced due to a mutation at chr10 position 50129923, was mapped by PoGo onto WASHC2A and WASHC2C. Both WASH genes are highly homologous with their corresponding UniProt proteins "Q641Q2" and "Q9Y4E1" sharing 97% sequence identity. The NGS data demonstrated that the mutation was present in WASHC2A and not in

WASHC2C. Specific RNA-seq reads were aligned onto the WASHC2A gene but not on the WASHC2C gene. Mapping the peptide with PoGo unnecessarily introduced ambiguity about its origin because it did not utilize all of the available information. In other words, despite NGS evidence that the peptide has a single genomic origin, event level classification of the peptide using PoGo would result in it being classified as ambiguous due to multiple possible genomic co-ordinates.

1.10. Objectives

In this work computational methods for proteogenomics have been developed to identify and validate non-canonical proteins. The data for this project was generated from Exome-seq, RNA-seq and LC-MS/MS analysis of MCF7 cells.

It is known that cancer is driven by mutations in the genome. Mutations accumulated during tumor development can alter the primary sequence and expression level of the mutated gene, and through dysregulation can affect the expression or other proteins. However the reference proteome databases used to identify proteins are curated to only include those proteins produced by normal cells, and thus cannot be utilized to identify those due to mutation. Proteogenomics analysis enables the analysis of such non-canonical proteins by creating customized protein sequence databases that include the mutant proteins, as well as novel protein isoforms that may result from dysregulation of protein expression.

In the first part of my work I developed Python scripts to generate customized protein sequence databases from the reference genome, transcriptome and from variants identified in NGS experiments. These *proteogenomics* databases were then subsequently used to identify peptides from non-canonical proteins (i.e. those not included in the curated reference protein databases).

In the second part of my work I focused on quality control of the non-canonical peptides. For example it is not uncommon that peptides will be confidently identified based on MS/MS spectra with a poor signal-to-noise-ratio. Such identifications have an increased risk of being a false positive because noise peaks may contribute to the identification. To guard against such false positives a dynamic noise level (DNL) algorithm was incorporated to remove peptides identified from spectra with poor signal quality.

Peptides are frequently identified using a fraction of the peaks contained in the MS/MS spectrum, because the statistical match between experimental data and predicted MS/MS spectra is greater if the database search utilizes only the most common fragmentation channels for the MS/MS method. Although it is not necessary for the identification to use all high intensity fragment ions, the presence of non-annotated peaks increases the risk of false positive identifications. To further guard against the non-canonical peptides being false positive identifications an MS/MS spectrum annotation tool was developed to comprehensively

annotate the matched spectrum. The tool was utilized post database search to ensure that the other possible fragmentation channels could explain the non-annotated peaks.

Non-canonical peptides identified through the proteogenomics analysis pipeline are best contextualized by mapping them on to the genome. In the third part of my work, I developed an algorithm to map these peptides onto the reference genome. The genomic mapping was utilized to assign context to the identified peptides and in their genomic classification.

In the final part of my work I developed methods for the validation of the non-canonical peptides. Synthetic isotopically-labeled standards (SIS) of the non-canonical peptides were purchased; the SIS peptides were then added to the cell extracts and targeted proteomics experiments performed on the non-canonical peptides (endogenous and SIS). I wrote Python scripts to quantitatively compare the MS/MS spectra and elution profiles of the endogenous and SIS peptides. A two tier automated validation scheme was implemented in which the cosine similarity was utilized to compare the fragmentation patterns and elution profiles.

The result of the proteogenomics searches, data quality control, genomic mapping, and validation, is a list of confident non-canonical peptides that can be classified in terms of the underlying genome. The development has been performed using the MCF7 breast cancer cell line, because it is a widely used model system and is known to include a large number of genomic mutations.

2. Methods

2.1. MCF7 cells

In this project proteogenomics analysis of MCF7 cells line has been performed. Data from Exome-seq and shotgun proteomics of MCF7 cells were produced in-house whereas and the RNA-seq data on MCF7 cells was obtained from a public resource (Ion Community website [81]).

2.2. Exome-seq

The library was sequenced to mean 33x coverage using a Life Technologies Ion Proton sequencing platform (Ion Torrent, Life Technologies, Grand Island, NY) for NGS. Genomic DNA was extracted from MCF7 cells and the DNA concentration was measured using a Qubit system (Life Technologies). Each fragment library was constructed from 1 μg of DNA using the IonTargetSeq exome enrichment kit. The quality and the quantity of the amplified library was checked using an Agilent 2100 Bioanalyzer instrument with the Agilent High Sensitivity DNA Kit. A total of 500 ng of amplified, size-selected library DNA was used for exome enrichment: a probe hybridization reaction followed by recovery of the probe-hybridized DNA and amplification of the exome-enriched library. The exome-enriched library was analyzed on an Agilent 2100 Bioanalyzer instrument with the Agilent High Sensitivity DNA Kit and the dilution required for template preparation was determined. For sequencing an Ion PI Sequencing 200 kit (Ion Torrent, Life Technologies, Grand Island, NY) was used. After the Ion Proton run the data was analyzed in the Ion Torrent server (Ion Torrent, Life Technologies, Grand Island, NY), set for alignment to the reference human genome build 38. The variants were called using the Variant Caller Plugin included in the Torrent suite. Variant analysis was restricted to variants occurring in exome regions.

2.3. RNA-seq

RNA-seq data of MCF7 cells was obtained from the Ion community website [81]. Two data analysis pipelines, one using TopHat2 [82] and the other using BWA MEM [83], were used. For both pipelines the reference human genome GRCh38 was used and variant calling was performed identically. In TopHat2 analysis the unmapped reads, generated from the first step,

were re-aligned using Bowtie2 [84] . The reads mapped with Tophat2 and Bowtie2 were then merged using the Picard [85] command SamMerge. We applied the samtools [86] command rmdup on the merged file to remove PCR duplicates. Variant calling was performed by samtools command "mpileup". A further filter with a Perl script (Vcf_filter, 2010, Ion Torrent System, modified by Nils Homer) was also performed.

2.4. Genomic variants in MCF7 cells

Variants identified in Exome-seq, RNA-seq and previously reported in the Catalogue Of Somatic Mutations In Cancer (COSMIC) [87] were obtained in Variant Call File (VCF) format and merged together into a single VCF file with VCFtools [88]. The merged file contained the union of all variants identified in our NGS experiments and publicly available through COSMIC.

2.5. LC-MS/MS

2.5.1. Reagents for LC-MS/MS

Urea, Ammonium Bicarbonate (AmBic), Iodoacetamide (IAA), DL-dithiothreitol (DTT), bovine trypsin, water, acetonitrile (ACN), formic acid, ammonium hydroxide were purchased from Sigma-Aldrich (St. Louis, MO, USA). Lys-C was produced by Wako (Neuss, Germany). Complete mini EDTA-free Cocktail and PhosSTOP phosphatase inhibitor Cocktail in tablets were purchased from Roche (Basel, Switzerland). C18 and Reversed Phase, S (RPS) cartridges were purchased from Agilent Technologies (Santa Clara, CA, USA). 326 crude synthetic heavy stable isotope standard (SIS) peptides containing one C-terminal heavy Lysine (composition C12[-6]N14[-2]C13[6]N15[2]: +8 Da) or heavy Arginine (composition C12[-6]N14[-4]C13[6]N15[4]: +10 Da) were purchased from JPT technologies (Berlin, Germany).

2.5.2. Sample Preparation for LC-MS/MS

MCF7 cells were lysed by sonication with an ultrasonic processor (Q125A, QSonica - 5 times 50% power, 2 sec pulses, 40 J) in lysis buffer (8M urea, Complete mini EDTA-free Cocktail, PhosSTOP phosphatase inhibitor Cocktail and 50mM AmBic in MilliQ Water) and spinned down for 40 min at 21000 g. Supernatant was recovered in a clean eppendorf tube and the

protein concentration determined with a microBCA protein assay (Thermo Fisher Scientific). Reduction was performed using DTT at a final concentration of 4 mM during 25 min incubation at 56 °C. Alkylation was performed with 8 mM IAA, incubating at room temperature in the dark for 30 mins. Digestion was performed in two steps. Lys-C was added (1:75 enzyme/protein) and incubated for 4 hours at 37 °C. Trypsin was added at 1:100 enzyme/protein after dilution to 2 M urea with 50 mM AmBic and incubated overnight at 37 °C. In the morning the solution was acidified with 10 % formic acid in H_2O until pH < 2.

The resulting MCF7 proteolytic peptides were desalted and fractionated using an AssayMAP Bravo (Agilent Technologies) equipped with C18 and RPS cartridges, respectively. Peptide desalting was performed using the peptide cleanup V2 protocol. Briefly, C18 cartridges were primed with ACN, equilibrated with 50 µL of 0.1% Formic Acid, 100 µL of diluted samples were loaded at 5 µL/min. Two cup washes and a cartridge wash were performed with 50 µL of water 0.1% formic acid at 10 µL/min, followed by a stringent syringe wash with ACN and then peptides were eluted with 30 µL of 80% ACN and 0.1% formic acid at 5 µL/min.

After desalting the proteolytic peptides were dried in vacuum, and resuspended in 10 mM NH_4OH pH 10 solution for high pH fractionation. Two sets of fractions were generated for the proteogenomics discovery runs. For each run 150 µg of peptides were divided in seven fractions using different percentages of ACN in 10 mM aqueous NH_4OH pH 10 for peptide elution. Specifically the elution solutions were Set A: 0%, 12.5%, 20%, 27.5%, 35%, 42.5%, 70%; Set B: 0%, 12.5%, 17.5%, 22.5%, 27.5%, 32.5%, 70%. High pH fractionation was performed using the fractionation protocol V1.0. RPS cartridges were primed with 100 µL of ACN and equilibrated with 50 µL of 10 mM aqueous NH_4OH pH 10. The peptide sample was loaded in 100 µL 10 mM NH_4OH pH 10 at 5µL/min and the flow through was collected. A cartridge and cup wash were performed with 100 and 50 µL of the same high pH buffer, and fractions eluted with 35 µL plugs at 5 µL/min. All the fractions, including the flow through, were dried in vacuum and resuspended in 10% Formic Acid in water before individual LC-MS/MS analysis.

For the validation experiments 326 stable isotopically-labeled standard (SIS) peptides were individually resuspended in ACN/H_2O/formic acid, ratio 50/49/1, and mixed in equimolar quantities using the “Reagent transfer” utility of the AssayMAP Bravo. Five sets of seven fractions were generated for the validation runs. In the first run, the MCF7 sample was

reanalyzed without addition of the SIS peptides. In the second, third and fourth runs the MCF7 sample was spiked with the SIS peptides at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). In the fifth and final run, the SIS equimolar peptide mix was analyzed separately. For each validation run high pH fractionation was performed on the AssayMAP Bravo using 0%, 12%, 18%, 24%, 30%, 36%, 80% of ACN in 10 mM aqueous NH_4OH pH 10.

2.5.3. LC-MS/MS analysis

LC-MS/MS experiments were performed using an Easy-nLC 1000 (Thermo Scientific) coupled to an Orbitrap Fusion mass spectrometer (Thermo Scientific). Proteolytic peptides were separated using an EASY-spray C18 column (2 µm particle size, 75 µm x 50 cm, Thermo Fisher Scientific) equipped with a trap column (2 µm particle size, 75 µm x 2 cm Thermo Fisher Scientific). The eluents had the following composition, eluent A: H_2O 0.1 % Formic Acid, and eluent B: ACN 0.1 % Formic Acid. Each peptide sample was loaded on to the column at 800 bar with 100 % A and peptides were eluted at 300 nl/min using a segmented gradient: 5 % B for 5 min, 5 - 22 % B in 104 min, 22 - 32 % B in 15 min, 32 – 90 % B in 10 min, 90 % B for 15 min. The electrospray voltage was set at 2.2 kV and the ion transfer tube temperature at 275 °C.

Data dependent tandem mass spectrometry was performed using top speed mode (3 sec max cycle time). The Full MS scan was acquired in the Orbitrap, *m/z* 350 to 1500, at 120,000 resolution. The Automatic Gain Control (AGC, which controls the number of ions to ensure consistent high mass accuracy) target was set at 4e5 with 100 ms maximum injection time. Monoisotopic precursor selection and a dynamic exclusion of 60 s were adopted. Peptide ions with a charge state from +2 to +7 and an intensity greater than 5e3 counts were selected for high energy collision dissociation (HCD) using a normalized collision energy (NCE) of 35 % and a 1.6 *m/z* isolation window. MS/MS spectra were acquired in the ion trap mass analyzer with a rapid scan rate, 1e4 AGC target and 35 ms of maximum injection time. Data acquisition was performed in profile mode for the MS scans and in the centroid mode for MS/MS.

2.5.4. Reference proteome databases

Reference human protein databases, in fasta file format, were obtained from UniProt (release September 2016), GENCODE (release 25), and RefSeq (release 78). Mapping files containing mapped protein sequences between GENCODE and RefSeq as well as between GENCODE and UniProt were obtained from GENCODE release 25. A reference protein sequence database was created by taking 82,636 non-redundant protein sequences from GENCODE. This protein set was extended with any UniProt protein sequence that had not been mapped onto GENCODE proteins. This included 1329 manually annotated "SwissProt" proteins and 610 computationally predicted "TrEMBL" [89] protein sequences. Finally, we added 2643 manually annotated RefSeq protein sequences, which had not been mapped onto GENCODE. Computationally predicted RefSeq protein sequences were not added to the reference proteins to keep the number of reference sequences fairly small and complete. Finally, our reference protein database contained a total of 87218 non-redundant protein sequences.

2.6. Data formats

In this section the file formats utilized in this work are briefly described. Genome annotations were obtained in GTF and GFF3 format. The reference sequences (genes, transcripts and proteins) were obtained in FASTA format and the genomic co-ordinates of peptides were produced in BED file format. The genomic variants in MCF7 cells were obtained in VCF format.

2.6.1.1. Genome annotation and sequence file formats

The reference human genome is released as a fasta file that contains the DNA sequences of all chromosomes, haplotypes, patches and scaffolds on the forward strand. Sequences are composed of 5 letters: Adenine (A), (Guanine) G, Cytosine (C), (Thymine) T and Unknown (N). The DNA sequence of the reverse strand is not included in the fasta file but can be inferred, since bases in the forward strand are complementary to the bases on the reverse strand (e.g. A pairs with T and G pairs with C).

The annotations on the sequences in the fasta file are released in GTF or GFF3 format. The GTF or GFF3 format is a structured tab separated hierarchical feature file. This file format is extensively used to record genomic features such as gene, transcript, exon, CDS, start codon,

end codon and UTRs. A GTF file consists of 9 columns. The columns in the file are: seqname, source, feature, start, end, score, strand, frame and attribute. A line in the annotation file contains the chromosome (seqname), start position (start), end positon (end), and strand information for any feature localized in the genome. The chromosome can be one of autosomes (1 to 22), sex chromosomes (X and Y) and mitochondrial (M). The feature could be one of gene, transcript, exon, CDS, start codon, stop codon and UTR. The start and end values are positive integers which defines the inclusive range of nucleic acids that belong to a feature. The strand is either positive (+) or negative (-). The annotation file is hierarchically structured. A gene may have multiple transcripts and any transcript may have one or more exons. An exon may have a CDS if the transcript is protein coding.

2.6.1.2. VCF file format

A VCF file is used to record genomic variants. It is a structured tab separated file and contains 8 columns. The columns in the VCF file are #CHROM, POS, ID, REF, ALT, QUAL, FILTER and INFO. Any variant detected by the variant calling programs is represented in a single line with its chromosome (#CHROM), position (POS), reference allele in the genome (REF) and the alternate allele (ALT) detected in the sample. The variants in the VCF file are always recorded in the forward strand.

2.6.1.3. BED file format

The BED format is a structured tab separated file that is used to record genomic features. It can easily handle features that are distributed over multiple exons and can be loaded into any genome browser for visualization. A BED file consists of a single feature per line separated by tabs. Each line contains 3 to 12 columns. The 12 columns of the BED file are: chrom, chromStart, chromEnd, name, score, strand, thickStart, thickEnd, itemRGB, blockCount, clockSizes, blockStarts.

1. chrom – name of the chromosome.
2. chromStart – Start position of feature. The system is 0 based.
3. chromEnd – End position of the feature. End position is not included in view.
4. name – Display name of the feature.
5. score –Score of the feature (between 0 to 1000).

6. strand – Strand defined as + or -.
7. thickStart – Co-ordinate to start drawing co-ordinate as solid rectangle.
8. thickEnd – Co-ordinate at which to stop drawing feature as solid line.
9. itemRGB – An RGB color value for the feature.
10. blockCount – The number of sub elements of the feature.
11. blockSizes – The size of these sub-elements.
12. blockStarts – The start coordinate of each sub-element.

2.7. Algorithm development

Python scripts were written for the generation of customized databases, application of a dynamic noise level (DNL) algorithm on MS/MS spectra, comprehensive annotation of MS/MS spectra, proteogenomics mapping of all classes of peptides, chromatogram extraction of endogenous and the synthetic stable isotope-labeled standard (SIS) peptides as well as comparisons of their fragmentation pattern and elution profiles. Besides the modules available in the Python standard library, the following packages were extensively utilized: BioPython [90], Pandas [91] , NumPy [92], SciPy [93], Pyteomics [94] and Matplotlib [95]. In the following sections, the algorithms utilized for performing these tasks are briefly described.

2.7.1. Extraction of reference sequences

With the full genome and the annotation file it is possible to generate the full set of reference genes, transcripts and protein sequences. This might seem to be an unnecessary task since the annotation sources also provide these reference sequences. The reference genes, transcripts and protein sequences are the outputs of the genome annotation process. The utility of this task can be fully exploited when one seeks to perform sequence operations such as; producing a personalized genome, transcriptome and proteome. Sequences from the genome file can be extracted guided by their co-ordinates in the annotation file and customized databases for proteogenomics searches can be constructed. Programming languages such as Python have an extensive collection of modules for dealing with biological sequences. BioPython is one such package that can be used for the extraction, manipulation and translation of DNA sequences. Guided by the co-ordinates in the annotation file an in-silico transcription and translation can

be performed. To generalize this, let P be a protein that is produced from a transcript T of Gene G. Let's assume transcript T has 3 exons, each containing a CDS. To obtain the full sequence of the protein coding gene G, the genomic co-ordinates of gene G in the annotation file can be utilized to extract the DNA sequence from the genome file (FASTA). If G is situated on the reverse strand of the DNA then the extracted gene sequence is reverse complemented. Automating the gene extraction process for all genes in the annotation file will result in a set of all gene sequences. If one wants to construct the spliced mRNA sequence of transcript T, then the DNA sequence of all three exons of transcript T are extracted from the genome file guided by their co-ordinates in the annotation file. If the transcript is located on the reverse strand then the extracted exon sequences are reverse complemented. The three exon sequences are then concatenated together in proper exon order to get the full length spliced mRNA sequence of transcript T. Automating the transcript extraction process for all transcripts in the annotation file will result in a set of all transcript sequences (reference transcriptome). If one wants to construct the full protein sequence P, the DNA sequence of each of the three CDSs on each of the three exons of transcript T are extracted guided by their co-ordinates in the annotation file. If the transcript is situated on the reverse strand, each of the extracted CDS sequences are reverse complemented. The CDS sequences from the exons are then concatenated together in proper CDS order to produce the full length CDS. The full length CDS is then translated into a full length protein P utilizing the proper genetic code. Repeating the protein extraction process for all protein coding transcripts in the annotation file will generate the full set of proteins (reference proteome). The described method for gene, transcript and protein extraction can be performed for any species whose genome has been sequenced and the corresponding annotation is available along with the genetic code for translation of nucleic acid sequences. The set of gene, transcript and protein sequences that can be extracted from the genome guided by their co-ordinates in the annotation file are referred to as reference sequences of the organism. The reference sequences are the same for all individuals of a species. The use of the terms reference genes, reference transcripts and reference proteins in the following sections would relate to these sequences that are accessible directly from the assembled genome (FASTA) and its corresponding annotation file (GTF or GFF3).

2.7.2. Generation of variant sequences

In-house python scripts were utilized for the generation of variant proteins in FASTA format. To create a personalized proteome of a sample incorporating the variants identified by NGS, the VCF file can be utilized along with the annotation file (GTF or GFF3) and the genome file (FASTA). To generalize this, the example of generating a reference protein sequence P in the section above can be extended with the variant information in the VCF file. Let, V be a variant (SNP or InDel) located in the VCF file which causes amino acid changes in the encoded reference protein P. The variant V is stored in the VCF file with its chromosome (#CHROM), position on chromosome (POS), reference bases in the reference genome (REF) and alternate bases detected in the sample (ALT). Start and end co-ordinates of all three CDSs of protein P are scanned for the presence of variants in the VCF file. Reference CDS sequences of protein P that satisfy the condition: CDS start <= POS <= CDS end, are extracted. The CDS is then modified by the substitution of reference bases (REF) at position (POS) with alternate bases (ALT), and a variant CDS is generated. The variant CDS can now be utilized instead of the reference CDS to generate the full length CDS sequence, which upon translation will generate the full length variant protein. Repeating the variant protein generation process for all reference proteins in the annotation file, and utilizing all detected variants in the VCF file, will generate a set of variant proteins (personalized proteome).

If more than one variant (mixed variants) were identified at a genomic location all possible variant CDSs were generated. If a protein contained more than one SNV they all appear on the same variant protein at different amino acid positions. If the variant amino acid in a protein is Arginine (R) or Lysine (K) it creates new tryptic sites. If more than one SNV occur in those proteins, the set of tryptic peptides generated when all SNVs are applied together will be different than the set of tryptic peptides when those SNVs are applied individually. For such cases, we also generated variant proteins in which all SNVs were individually applied. For example, if a protein contains two SNVs, one of them being R or K, a variant protein was generated with both SNVs appearing together on the same variant protein. Two other variant proteins were also generated in which the SNVs were individually applied. This approach provided all the tryptic peptides that could be obtained as a result of the two SNVs on the protein. Insertion or deletion mutations in the VCF file were applied one at a time, so for each insertion or deletion mutation separate variant protein sequences were generated. Thus, if a

protein had two insertion and two deletion mutations, four protein sequences were generated, one for each insertion and deletion mutation.

2.7.3. Generation of customized protein databases from the reference genome and transcriptome

Python scripts were written to generate a database of novel exon-skipped proteins from the reference transcript structures in FASTA format. An exon in the transcript was skipped if its frame of translation was the same as the subsequent exon, Figure 7. The exon-skipped transcript was then translated into an exon-skipped protein.

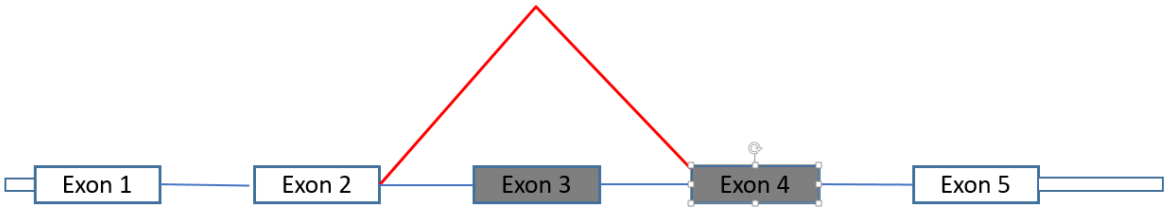

Figure 7. An example of exon-skipping. The figure shows a transcript structure of a gene with 5 exons (1 to 5) and 4 introns as straight blue lines connecting the exons. Exons that have the same frame of translation are shown in grey (exon 3 and exon 4). A novel exon-skipped isoform transcript sequence was generated by skipping exon 3, and splicing exon 2 to exon 4.

Three transcript biotype specific databases were generated. GENCODE transcripts with biotypes protein coding or NMD, retained intron and long non-coding RNA were translated into ORFs in 3 frames (1, 2 and 3). Translation in frame 1 was conducted by taking all nucleic acids in the transcript sequence. Translation in frame 2 and 3 was done by removing one and two nucleic acids from the beginning of the transcript sequence, respectively. GENCODE gene and CDS sequences with 100 base pairs flanking sequences were extracted guided by their co-ordinates in the GTF file. The extracted sequences were translated into ORFs in three frames.

2.7.4. Noise detection in the MS/MS spectra

After the database search it is not uncommon that some peptides, even confidently identified peptides, are reported from spectra with poor signal-to-noise-ratio (SNR). When seeking to identify novel proteoforms it is incumbent to demand the highest data quality, in order to

minimize false positive identifications. In principal, such spectra can be filtered out through visual inspection but it cannot match the very high throughput of proteomics experiments. As part of an automated quality control of identified peptides a dynamic noise level algorithm (DNL) [96] was implemented in Python to filter out peptide spectrum matches that have low SNR. The DNL algorithm makes two assumptions about the MS/MS spectra: 1) in a good quality spectrum the signal peaks are of greater intensity than the noise peaks, and 2) there is at least one noise peak in all spectra [96]. The steps of the DNL algorithm are:

i) All peaks in the spectrum are sorted in order of increasing abundance I_i (i = 1, 2,… N).

ii) The first peak in the sorted spectrum is assumed to be noise.

iii) The SNR of the other peaks are calculated as the ratio of their intensity to the predicted peak intensity of a noise peak, which is predicted by scaling the abundance of the noise peak by a scaling factor (1 + α). Here the default setting of α was 0.5.

iv) If the SNR of the second peak is greater than the minimum SNR threshold (SNR_{min}) the second peak is considered to be signal and the predicted noise level is set as the noise level for the entire MS/MS spectrum. If the SNR of the second peak is not greater than the SNR_{min} the algorithm scans from peaks 3 to N. For any peak k the previous k-1 peaks are considered noise. The noise level for peak k is calculated by fitting a linear regression to the intensities of the previous k-1 peaks. If the SNR for peak k exceeds SNR_{min} it is considered signal and that noise level is set as the noise level for the entire MS/MS spectrum.

In this work the DNL algorithm was modified to not consider the second peak as signal. If the second peak was determined to be signal then the scaling factor α was increased by 0.1 until the second peak was determined to be noise.

2.8. Comprehensive annotation of matched MS/MS spectra

An additional quality control step consisted of assessing how many of the detected fragment ions could be explained by the identified peptide. The database search methods used for protein identification only use a fraction of the possible peptide fragmentation pathways, which is determined by the fragmentation method used, e.g. HCD generates primarily *a*, *b*, and *y* type

fragments, whereas electron transfer dissociation generates primarily *c*, and *z* type fragments, Figure 8.

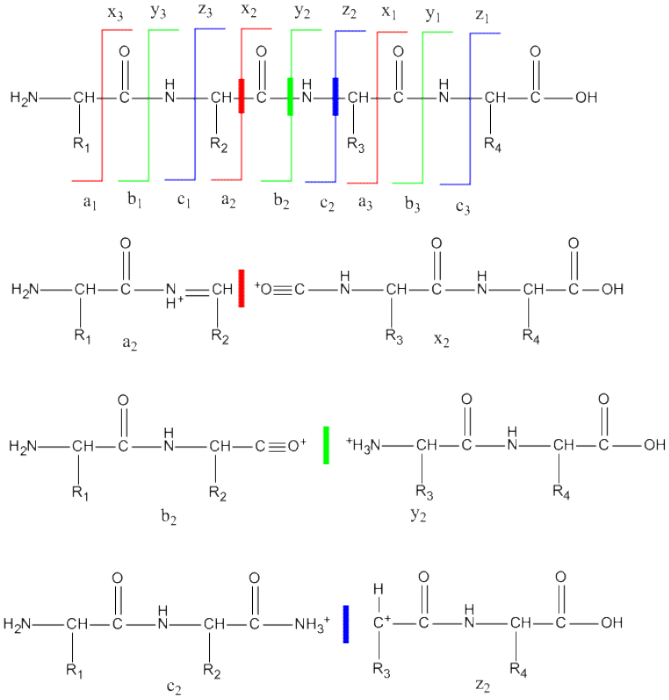

Figure 8. Peptide ion fragmentation channels in MS/MS. Double backbone cleavage can give rise to internal fragments, and cleavage of amino acid side chains can lead to *d*, *v*, and *w* fragment ions.

The specification of the fragment ion series (determined by the MS/MS technique) during database search is performed to maintain statistical power: the inclusion of all possible fragmentation channels would increase the search space but decrease the specificity (the score distribution for the random matches is increased while the score for the correct match is unchanged). Accordingly, the MS/MS spectra from confidently identified peptides can contain high intensity but non-annotated fragment ion peaks that were not used for the identification. Although it is not necessary to annotate all high intensity peaks, annotating them increases the confidence in the identified peptide because more fragment ions can be explained. If the precursor ion is isolated with little or no interference from other co-eluting ions it is important that all high intensity fragment peaks can be rationalized. The MS/MS spectra used to identify peptides may contain fragment peaks that were not included in the database search, for example

from immonium ions, internal fragments, and neutral losses of the peptide precursor. A python based spectrum annotation tool was developed to comprehensively annotate the matched spectrum. The tool was utilized post database search to investigate if unannotated peaks in MS/MS spectra could be explained by other fragment ion types, charge states and neutral losses. These quality control steps were developed to ensure all novel proteoforms were characterized by good signal-to-noise ratio MS/MS spectra, and in which the identification can be used to annotate all good signal-to-noise ratio peaks.

2.8.1. Proteogenomic mapping

If the genomic co-ordinates of the peptides are tracked during the search (as in Peppy), peptide mapping is not required. If peptides are discovered from reference or customized databases, they need to be explained by mapping them onto the reference genome. Genome annotation systems at NCBI and GENCODE provide the co-ordinates of genes, transcripts, exon and CDS in GFF3 or GTF file format, respectively. These files do not contain the genomic co-ordinates of reference proteins, but they can be deduced from the genomic co-ordinates of its CDS. In this way, the peptide mapping problem can be traversed backwards where the peptide position on the protein is transformed into CDS co-ordinates. In the following sections algorithms implemented in python to map different types of peptides are shown. In all cases the algorithms use as inputs: genome annotations (in GTF or GFF3 format), a database of proteins (FASTA format) and input peptide sequences to be mapped (CSV format).

2.8.1.1. Mapping reference peptides to the reference genome

Genomic co-ordinates of peptides from reference proteins can be deduced by transforming the peptide position in the protein to its position on the full CDS. The peptide position on the full CDS can then transformed into its genomic co-ordinates. An algorithm to map peptides of the reference proteome to their genomic co-ordinates is described below. See Figure 9 and Figure 10 for a graphical representation of peptide mapping in forward and reverse strand.

Step 1: Let, *PStart* be the peptide start position on a protein and *len* be the length of the peptide. The peptide end position, *PEnd*, on the protein is given by: *PStart* + *len* -1.

Step 2: The peptide position onto the protein is transformed into its position on the full reference CDS. Let *cdsStart* be the start position of the peptide on the full CDS. Then, *cdsStart* can be computed with the formula:

cdsStart = (*PStart*-1)*3 + frame of the first CDS of the protein + 1.

The frame value (0, 1 or 2) for every CDS is encoded in the GTF. Frame value 0 indicates that the first base of the CDS is the first base of a codon, 1 indicates that the second base of the CDS is the first base of a codon, and 2 indicates that the third base of the CDS is the first base of a codon. Since each amino acid in the protein corresponds to a triplet of nucleic acids the end position of the peptide on the CDS, *cdsEnd*, can be computed as:

cdsEnd = *cdsStart* + (3 * *len*) - 1.

Step 3. The CDS feature which contain the *cdsStart* and *cdsEnd* positions are extracted. This is done by extracting all CDS's of the transcript and calculating their cumulative length until each CDS. These values are contained in the vector *cumlen*. The CDS on which the condition, *cdsStart* <= *cumlen* is satisfied for the first time is located. The peptide begins on this CDS feature which we refer to as *cdsStartFeature*. The subsequent CDS on which the condition *cdsEnd*<=*cumlen* is satisfied for the first time is located. The peptide ends on this CDS which we refer to as *cdsEndFeature*.

Step 4: If the c*dsStartFeature* and *cdsEndFeature* are the same CDS then the peptide is located within a single CDS and is not exon-spanning. If *cdsEndFeature* is consecutive to *cdsStartFeature* the peptide spans CDSs of two exons: starting within the exon of *cdsStartFeature* and ending within the exon of *cdsEndFeature*. If any other CDS's exist between *cdsStartFeature* and *cdsEndFeature* the peptide spans multiple exons. Once, the *cdsStartFeature* and *cdsEndFeature* are located the genomic start and end co-ordinate of the peptide can be calculated as follows:

Genomic start = Start of *cdsStartFeature* + *cdsStart* – cumlen before *cdsStartFeature* - 1

Genomic end = Start of *cdsEndFeature* + *cdsEnd* – cumlen before *cdsEndFeature* -1

If the protein is on the reverse strand,

Genomic start = End of *cdsStartFeature* – (*cdsStart* – cumlen before *cdsStartFeature*) +1

Genomic end = End of *cdsEndFeature* – (*cdsEnd* – cumlen before *cdsEndFeature*) + 1.

The deduced genomic start and end positions can be utilized to generate the peptide co-ordinate in BED file format.

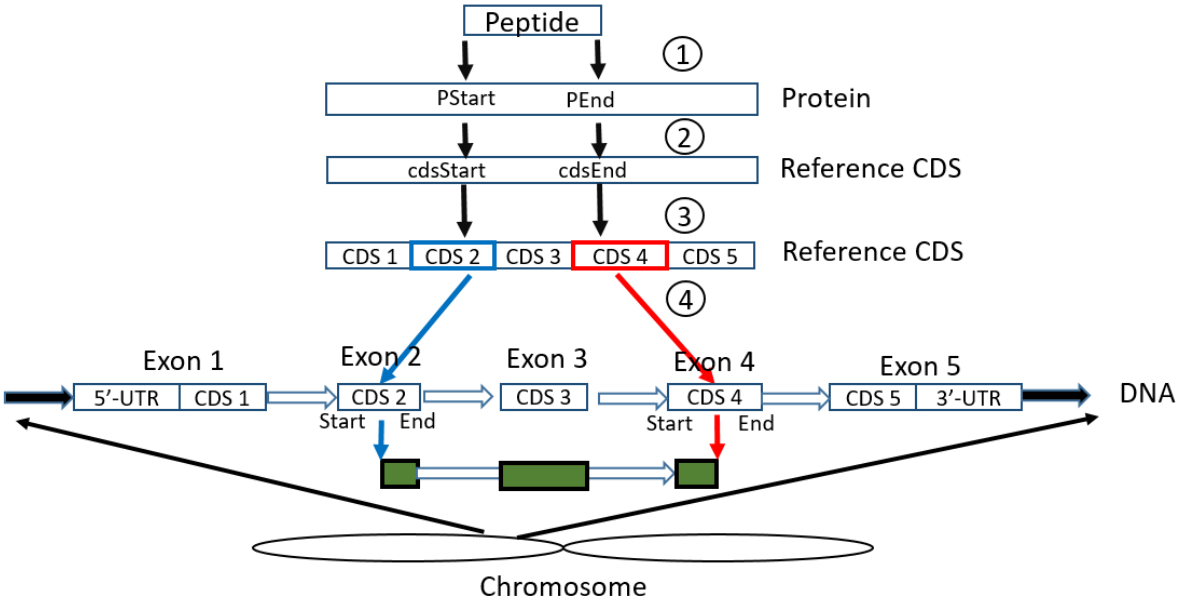

Figure 9. Mapping reference peptides to the reference genome on the forward strand. The cds*StartFeature* is shown as a blue box and the cds*EndFeature* is shown as red box. The mapped genomic co-ordinate of the peptide is shown as a green block. The circled numbers correspond to the steps in the algorithm described in the text.

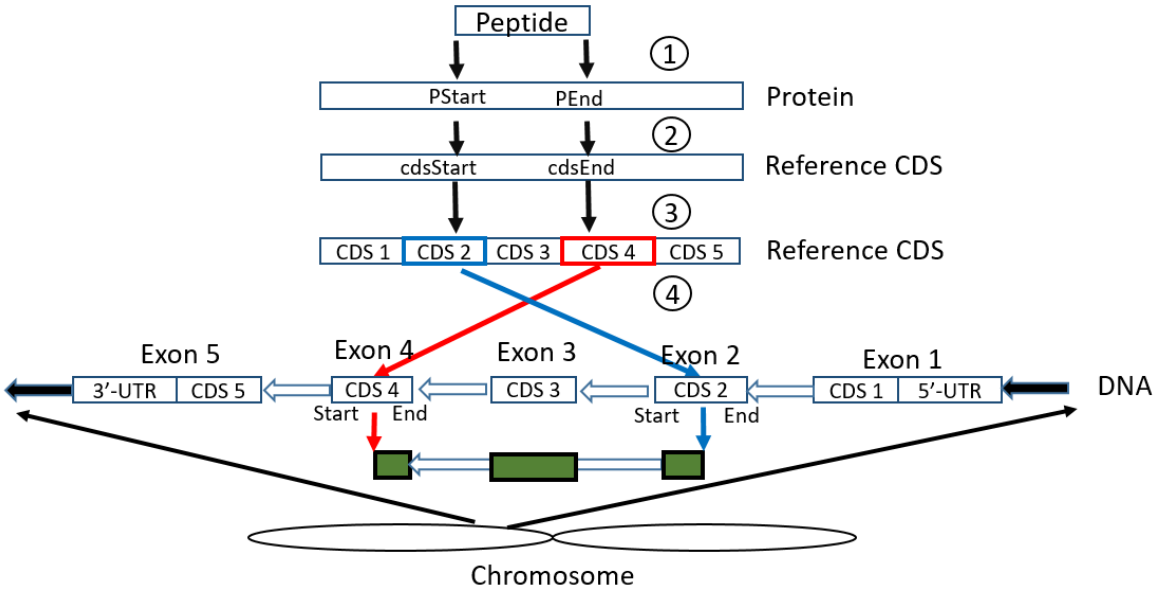

Figure 10. Mapping reference peptides to the reference genome on the reverse strand. The cds*StartFeature* is shown as a blue box and the cds*EndFeature* is shown as a red box. The mapped genomic co-ordinate of the peptide is shown as a green block.

2.8.1.2. Mapping SNV peptides to the reference genome

The algorithm discussed in section 2.8.1.1 can also be utilized to obtain genomic co-ordinates of SNV peptides if their reference proteins can be tracked. Instead of providing reference proteins (FASTA) a database of SNV proteins (FASTA) can be provided. An example fasta header of a SNV protein is shown below.

>ENST00000629481.1_snp_4 chr19 Gene=ENSG00000239998.5 GN=LILRA2 Strand=+ 54574349_T/C_49_ATC/ACC_I/T_snp

The header of the SNV protein links it to reference Ensembl transcript ENST00000629481 of Gene LILRA2. Once the peptide location on the SNV protein is determined the genomic co-ordinate of the reference transcript ENST00000629481 can be utilized to map the peptide, Figure 11. SNVs do not cause any change in the reference transcript length. Thus the mapping of peptides from SNVs can be treated in the same manner as the mapping of reference peptides to the reference genome.

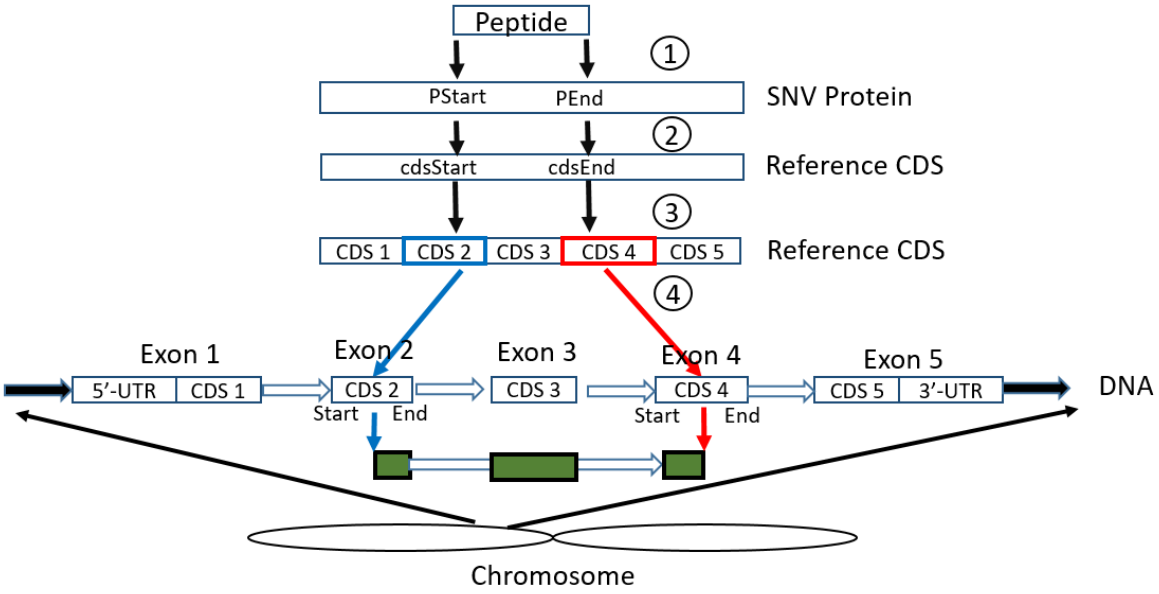

Figure 11. Mapping SNV peptides to the reference genome on the forward strand. The cds*StartFeature* is shown as a blue box and the cds*EndFeature* is shown as a red box. The mapped genomic co-ordinate of the peptide is shown as a green block.

2.8.1.3. Mapping exon-skipped peptides to the reference genome

Exon-skipped proteins are generated from the reference transcript structures by skipping single or multiple exons. If the information regarding which exons were skipped from the reference transcripts is retrievable the mapping of the peptides is readily achieved. The algorithm described in section 2.8.1.1 can also be used for this task by modifying the inputs. Instead of reference annotation (GTF) and reference proteins (FASTA), an exon-skipped annotation (GTF) and a database of exon-skipped proteins (FASTA) are provided as inputs to the algorithm. An example fasta header of an exon-skipped proteins is shown below.

>ENST00000368216.8_NovIso_3 chr1 Gene=ENSG00000143303.11 GN=RRNAD1 Strand=+ Skipping exon 3

The header shows that the proteoform was generated by skipping exon 3 from the reference transcript ENST00000368216 of the RRNAD1 gene. In this case the GTF file needs to conform to the database of exon-skipped proteins. GTF file of exon-skipped isoform transcripts can be generated from the reference GTF file. For every exon-skipped protein in the fasta file a novel transcript feature can be generated by removing the exon and CDS features of the skipped exon from the reference transcript feature. For the proteoform example shown above, exon 3 and its CDS are removed from the reference transcript ENST00000368216. The modified GTF file, the fasta file of exon-skipped proteins and the peptides to map can then be submitted to the mapping algorithm described in section 2.8.1.1 to obtain genomic co-ordinates of the exon-skipped peptides, Figure 12.

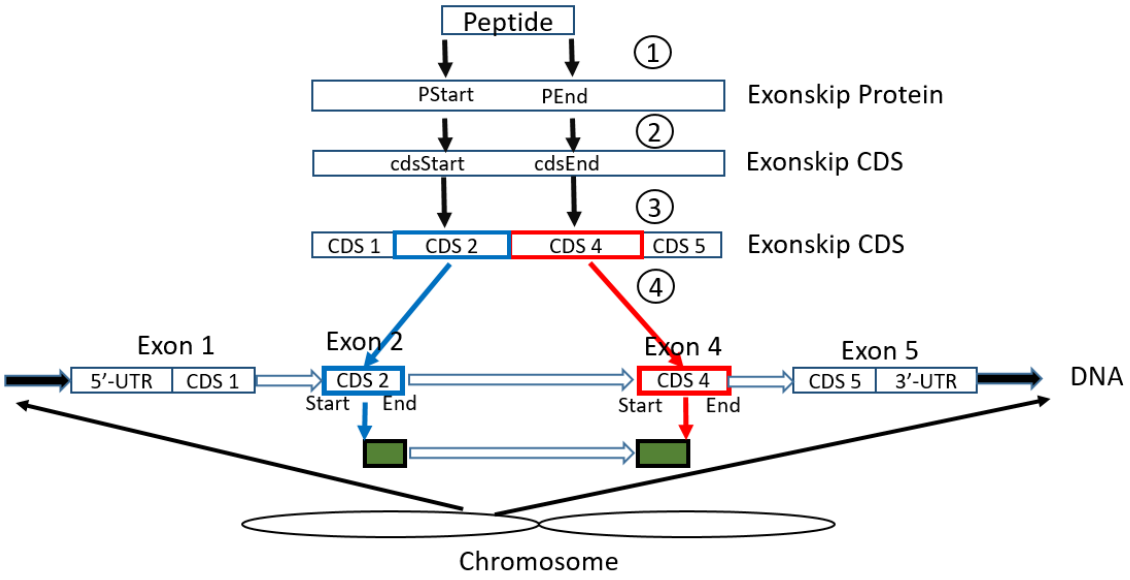

Figure 12. Mapping exon-skipped peptides to the reference genome. The *cdsStartFeature* is colored blue and the *cdsEndFeature* is colored red. The mapped genomic co-ordinate of the peptide is shown as a green block.

2.8.1.4. Mapping peptides from 6 frame transcript sequences to the reference genome

Proteogenomics searches can be conducted with three or six frame translated transcript sequences. Although most peptides from this type of search will have a reference protein origin some will not. These include peptides originating from (designated) non-coding regions of coding transcripts, peptides from non-coding transcripts and complement sequences of coding or non-coding transcripts. A fasta file of transcript sequences from which the peptides were discovered is required. This can be obtained from annotation resources such as NCBI or Ensembl or can be generated in-silico by utilizing the annotation file (GTF) and the full genome FASTA file. For in-silico generation all exons of a transcript are extracted from the full genome utilizing the co-ordinates of the exons in the annotation file. The exon sequences are concatenated in proper exon order. If the transcript is on the reverse strand the exon sequences are reverse complemented before concatenation. The transcript sequences can then be translated in three forward frames (1, 2 and 3). Translation in frame 1 is conducted by

utilizing all nucleic acids in the transcripts. Translations in frames 2 and 3 are conducted by utilizing all nucleic acids in the transcript except the first and the second base, respectively.

If peptides need to be mapped on the complement sequences of transcripts then the reference annotation file can be appended with annotations for complement sequences. The protein translations from complement sequences need to be generated as well. This is easily achieved by reverse complementing the reference transcript sequences and translating them in frames 1, 2 and 3. To generate annotations for complement sequences a copy of the reference annotation is generated. The strand information of all transcripts and its sub-features (exons and CDS) in the copied GTF file is reversed (+ to – and vice versa). The order of exons in also reversed. For example if a transcript had three exons on the forward strand, the strand is set to negative and the order of exons reversed in the copied GTF file. Exon number 1 is set to exon number 3, and exon number 3 is set to exon number 1. The hierarchical order of exons (exon 1 followed by exon 2 and so on) in the copied GTF file is preserved after these changes. The modified GTF file can then be appended to the reference GTF file. After this step the modified mapping algorithm can be applied for peptide mapping. The inputs to the algorithm are the GTF file with reference and complement annotations, six-frame translated protein database and peptides to map, Figure 13.

Step 1: Let *PStart* be the start position of a peptide on a proteoform and *len* be the length of the peptide. Then the peptide end position on the proteoform is: *PEnd = PStart + len* -1.

Step 2: The peptide position on the proteoform is transformed into its position on the transcript. Let t*Start* be the start position of the peptide on the transcript, computed as:

t*Start* = (*PStart*-1)*3 + frame of the translated proteoform.

The frame value (1, 2 or 3) for every proteoform is recorded while translating the transcript sequences. Since each amino acid in the proteoform corresponds to a triplet of nucleic acids the end position of the peptide on the transcript can be computed:

t*End* = t*Start* + (3 * *len*) -1.

Step 3: The Exon features that contain the *tStart* and *tEnd* positions are extracted. This is done by extracting the cumulative length until each exon of the transcript. The values are stored in a vector referred to as *exon_cumlen*. The exon on which the condition *tStart* <= *exon_cumlen* is satisfied for the first time is located. The peptide starts on this exon, referred to as *exonStartFeature*. The subsequent exon on which the condition *tEnd*<=*exon_cumlen* is

satisfied for the first time is located. The peptide ends on this exon, referred to as *exonEndFeature*.

Step 4: If the *exonStartFeature* and *exonEndFeature* are the same exon then the peptide is located within a single exon and is not exon-spanning. If *exonEndFeature* is consecutive to *exonStartFeature* the peptide spans two exons: starting within *exonStartFeature* and ending within the *exonEndFeature*. If any other exons exist between *exonStartFeature* and *exonEndFeature* then the peptide spans multiple exons. Once *exonStartFeature* and *exonEndFeature* are determined the genomic start and end co-ordinates of the peptide can be calculated as follows:

Genomic start = Start of *exonStartFeature* + *tStart* – cumlen before *exonStartFeature* – 1.

Genomic end = Start of *exonEndFeature* + *tEnd* – cumlen before *exonEndFeature* -1.

If the protein is on the reverse strand,

Genomic start = End of *exonStartFeature* – (*tStart* – cumlen before *exonStartFeature*) +1.

Genomic end = End of *exonEndFeature* – (*tEnd* – cumlen before *exonEndFeature*) + 1.

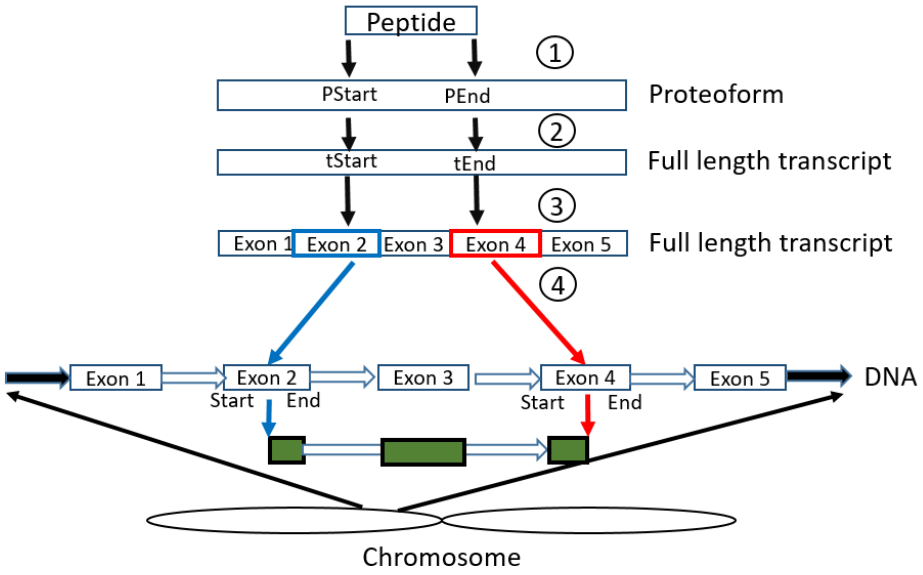

Figure 13. Mapping peptides from 6 frame transcripts to the reference genome. The *exonStartFeature* is colored blue and the *exonEndFeature* is colored red. The mapped genomic co-ordinate of the peptide is shown as a green block.

2.8.1.5. Mapping InDel peptides to the reference genome

InDel mutations change the length of the encoded mRNA and thereby produce proteins whose lengths differ from the reference proteins. If the mutation is in-frame (addition or deletion of nucleotides in multiples of 3), it adds or removes amino acids into the encoded protein. If the mutation is out of frame (addition or deletion of nucleotides not in multiples of three) the protein's primary sequence is modified and the protein's length altered. If the mutation that generated the InDel peptide can be retrieved the genomic co-ordinates of the peptide can be deduced by modifying the mapping algorithm used to generate the *cdsEnd* co-ordinates in section 2.8.1.1 and supplying the reference GTF file, database of InDel proteins along with the peptide sequences to map.

For example, a peptide "***MVSAL-QQQQQQQR***" was identified due to an in-frame deletion in protein TNRC6B. The fasta header of the in-frame deleted proteoform is shown below.

>ENST00000454349.6_del_52_DB_4 chr22 Gene=ENSG00000100354.20 GN=TNRC6B Strand=+
40301172_TGCAGCAGCAGCAGCAGCAGCAGCAG/TGCAGCAGCAGCAGCAGCAG CAG_23_CTG/fs*_L/fs*

The InDel protein is produced from reference transcript ENST00000454349 of TNRC6B. The fasta header shows the chromosome, gene id, gene name, strand and the mutation string. The mutation is in-frame deletion identified by the numbers of bases in the reference allele (REF)"TGCAGCAGCAGCAGCAGCAGCAGCAG" (26) and the alternate allele (ALT) "TGCAGCAGCAGCAGCAGCAG" (23), respectively.

The modified mapping uses:

cdsEnd = *cdsStart* + (3 * *len*) – 1 – (length of (ALT) –length of (REF)).

After this modification the same algorithm described in section 2.8.1.1 can be applied to map InDel peptides to the reference genome, Figure 14.

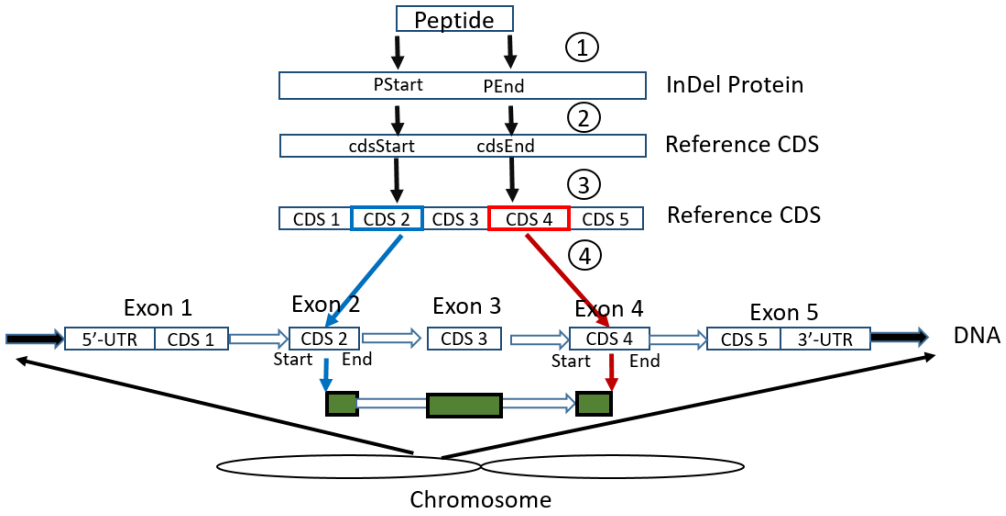

Figure 14. Mapping peptides from InDel mutations to the reference genome. The *cdsStartFeature* is colored blue and the *cdsEndFeature* is colored red. The mapped genomic co-ordinate of the peptide is shown as a green block.

2.8.1.6. Mapping peptides from six-frame gene sequences

The gene sequences can be extracted from the full genome fasta file guided by their co-ordinates in the GTF file. If peptide mapping is desired on the complement strands of genes the annotations for the complement sequences must be generated. This can be achieved by generating a copy of all reference gene annotations and reversing their strand information (+ to – and vice versa). All sequence (original and the copy) are then translated in three frames. Peptide positions are then computed on the translated proteoforms and transformed into genomic co-ordinates, Figure 15. Let *PStart* be the peptide start position in any proteoform and *len* be the length of the peptide. The peptide positions are then transformed into the genomic co-ordinates as follows:

Genomic start = Start of the gene + frame of the proteoform + (*PStart*-1) * 3 - 1

Genomic end = Genomic start + *len**3 -1.

For peptides mapped onto proteoforms in the negative strand,

Genomic start = End of the gene – frame of the proteoform – (*PStart*-1) * 3 + 1

Genomic end = Genomic start – *len**3 +1.

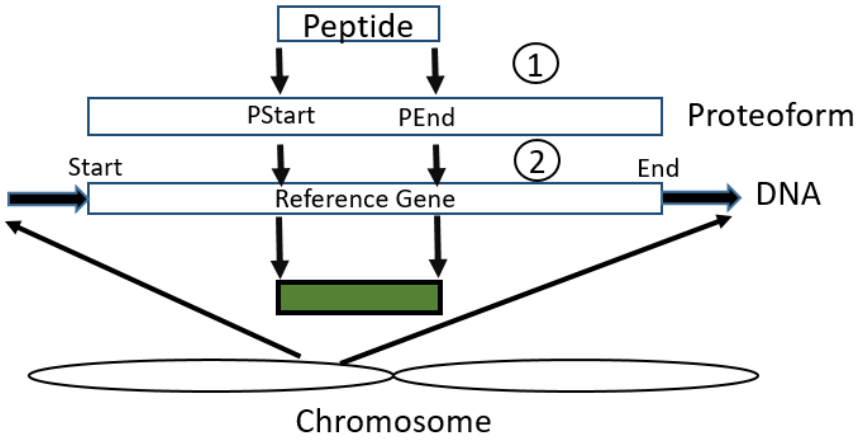

Figure 15. Mapping peptides from six-frame gene sequences to the reference genome. The mapped genomic co-ordinate of the peptide is shown as a green block.

2.8.1.7. Mapping peptides from GNOMON predicted proteins to reference and alternate assemblies

GNOMOM predicted proteins have two different genomic assembly origins. The reference (GRCh38) and the alternate (CHM1_1.1) assembly. Both of these assemblies and the corresponding annotations files generated by GNOMON are available from the NCBI. The algorithm described in section 2.8.1.1 can be used directly for mapping of peptides onto the GNOMON sequences by modifying the inputs. For reference assembly mapping the GNOMOM predicted proteins (FASTA) and the GNOMON predicted annotations (GFF3) from the reference assembly are provided as inputs. For alternate assembly mapping the GNOMOM predicted proteins (FASTA) and the GNOMON predicted annotations (GTF) from the alternate assembly are provided as inputs.

2.8.1.8. Mapping peptides from 6 frame full genome searches

The full genome (FASTA) contains all the chromosomes sequences in forward strand. Nucleic acid sequences of the complement strands can be generated by reverse complementing each chromosome sequence. All sequences can then be translated in three frames (1, 2 and 3). Peptide positions are then computed onto each translated proteoforms and transformed into genomic co-ordinates, Figure 16. Let *PStart* be the start position of a peptide onto a translated

chromosome and let *len* be the length of the peptide. Then the genomic co-ordinate of the peptide can be obtained as follows.

Genomic start = frame of proteoform + (*PStart*-1)*3

Genomic end = Genomic start + *len**3 - 1

If the proteoform was obtained from the reverse strand,

Genomic end = length of chromosome - frame of proteoform – (*PStart*-1)*3 +1

Genomic start = Genomic end – (*len**3) +1.

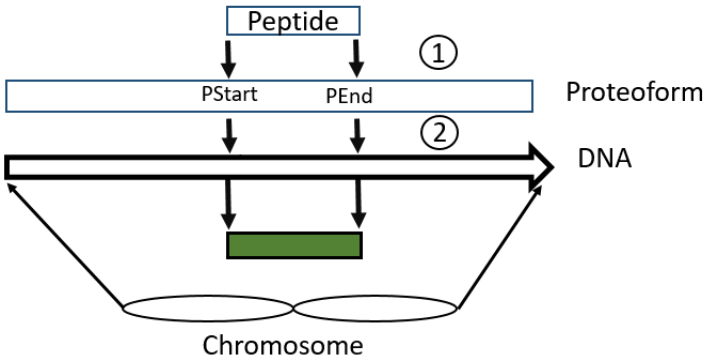

Figure 16. Mapping peptides from six-frame full genome searches.

2.8.2. Generating peptide co-ordinates in BED format

The sections above described the methods developed to obtain the genomic co-ordinates of peptides identified from the proteogenomics searches. All methods produced peptide co-ordinates that were represented as: chromosome, peptide start position on the chromosome, and peptide end position on the chromosome. If the peptide was mapped onto the reverse strand of the DNA the peptide start co-ordinate was greater than the peptide end co-ordinate. Additional information is also available, such as; strand (forward or reverse), starting exon of the peptide (if any) and ending exon of the peptide (if any). These values are sufficient to represent the genomic co-ordinates of the peptides in BED file format, thus enabling the results to be loaded into any genome browser to visualize the location of the mapped peptides. The proteogenomics peptides were color coded into two groups: Ambiguous proteogenomics peptides (Black) and Unambiguous proteogenomics peptides (Red).

2.9. Validation of the non-canonical peptides

Some of the non-canonical peptides identified by proteogenomics searches were selected for validation using synthetic ^{13}C isotopically-labeled standard (SIS) peptides. The physical and chemical properties of the endogenous and SIS peptides are near identical, thus their chromatographic profiles and fragmentation patterns should be near identical. A similarity metric was used to validate the presence of the endogenous non-canonical peptides. A two-tier validation scheme was implemented in Python. Scripts were written for similarity computations between the fragmentation patterns (*i.e.* their MS/MS spectra) and elution profiles of the endogenous and SIS peptides.

Tier1: The cosine similarity was used to quantitatively compare the fragment spectra from endogenous and SIS proteogenomic peptides. In-house python scripts were utilized for the following tasks. We first applied the DNL algorithm to the MS/MS spectra then annotated the signal peaks with fragment ion-types: *a, b, y,* internal ions (*a*-type, *b*-type) of length up to 10 amino acids, and precursors with a maximum loss of 1 water and/or 1 ammonia. The maximum charge state for the fragments was set to 2+ if the precursor was doubly charged, otherwise it was limited to one less than the precursor charge. Fragment ion peaks matching un-fragmented precursors and its isotopes were removed from the spectra. If a peak could be matched to multiple fragment ions we annotated the peaks based on the following priority rule: *N*-terminus or *C*-terminus fragments > neutral losses from *N*-terminus or *C*-terminus fragments > internal fragments > neutral losses from internal fragments. If a peak matched to multiple annotations after priority based selection we selected the annotation that gave the lowest mass deviation. All matched fragment ions present in both the SIS and endogenous MS/MS spectra, above noise level, were utilized for determining MS/MS spectral similarity. The intensities of the fragment ion peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. MS/MS spectral similarity was only computed if the endogenous and SIS MS/MS spectra contained at least 10 common fragment ions. Non-canonical, proteogenomic peptides with a cosine similarity greater than 0.9 were considered validated at tier 1.

Tier 2: The extracted ion chromatogram (EIC) of all tier 1 peptides, endogenous and SIS, were then examined to ensure their retention times were identical. MS raw files from validation runs were converted to .ms1 format with MSConvert (Proteowizard version 3.0.10051). MS1 scans

were centroided during conversion. The resulting peak lists were used for chromatogram extraction using in-house python scripts and a 10 ppm tolerance. Peptide identification time points were extracted from the Mascot [97] search results and the apex of peptide elution peaks calculated. Local intensity minimum time points before and after the elution apex were determined and used as the time range in which the endogenous and SIS peptide elution profiles were compared. If the computed elution profile time range was less than 15 seconds or more than 45 seconds we compared a 30 sec time window spanning the elution apex.

The intensities of the peptides were estimated by summing the intensities of its monoisotopic, $^{1}C_{13}$ and $^{2}C_{13}$ peaks, and the similarity was computed between the endogenous and SIS peptide profiles. Next, a Savitzky Golay filter [98] was applied onto the summed extracted peaks and the similarity computed between the filtered endogenous and SIS peptide profiles. The peptide intensities within the time range were normalized to sum 1 before similarity computation. Peptides with a similarity score greater than 0.9 either in the raw or in the filtered profiles were selected for further processing. Next, we compared the Savitzky Golay filtered peak profiles of each individual isotope (monoisotopic, $^{1}C_{13}$ and $^{2}C_{13}$) of the endogenous and SIS peptides. Peptides with a mean isotope profile similarity greater than 0.9 were selected for further processing. As a final filter the relative intensities of the isotopes were also compared; only those peptides whose isotope composition similarity was greater than 0.9 were considered validated at tier 2.

2.10. Proteogenomics databases

A database of 37,366 SNV proteins and a database of 106905 InDel proteins was created by utilizing all variants in the VCF file (see methods section generation of variant sequences for details). Transcriptome fasta files were obtained from GENCODE release 25 and RefSeq release June 2016. A database of 187,036 novel splice isoforms was created by skipping single exons from the GENCODE mRNAs and "Non-sense mediated decay" (NMD) transcript structures. An exon in the GENCODE transcript was skipped if its frame of translation was the same as the subsequent exon. GENCODE transcripts with biotypes "protein coding" and NMD were translated into ORFs in 3 frames. NMD transcripts contain a premature stop site in its canonical reading frame and are targeted by the NMD pathway for degradation to prevent

production of truncated proteins [99]. ORFs of length less than 10 were discarded and a database of 2,221,980 ORFs from protein coding transcripts was created.

GENCODE gene and CDS sequences with 100 base pairs flanking sequences were translated into three frames and ORFs of length less than 20 were discarded. Two databases containing 20,512,063 ORFs from genes and 1,706,623 ORFs from CDS sequences were created. GENCODE transcripts with biotype "long non-coding RNA" were translated in three frames and ORFs of length less than 20 were discarded. A database of 125590 ORFs from lncRNAs was created. GENCODE transcript sequences with biotype "retained intron" were translated in three frames and ORFs of length less than 50 were discarded. A database of 156169 ORFs from retained intron transcripts was created.

A Fasta file of 316,902 GNOMON predicted human protein sequences was obtained from NCBI Annotation release 107. GNOMON is NCBI's eukaryotic genome annotation pipeline. The fasta file contains model protein sequences based on experimental cDNA sequences and/or ab-initio models from the reference (GRCh38) and the alternate (CHM1_1.1) genome assemblies. Full length protein sequences in the GNOMON fasta file that were not present in the merged reference proteome were extracted and a smaller database of 69,136 GNOMON full length models was created. Henceforth, this database will be referred as GNOMON-small.

Transcriptome fasta files from GENCODE containing 198,093 transcripts and RefSeq containing 176,426 transcripts were utilized for six-frame nucleic acid searches. The human genome fasta file was obtained from GENCODE release 23 and split into chunks with the help of a script (splitter.pl) from Matrix Science and utilized for six-frame nucleic acid search. All sequence databases were uploaded onto Mascot server [97] for database searches.

2.11. Database searches

2.11.1. Discovery searches

The raw LC-MS/MS data from 14 in-depth proteomics investigations of MCF7 cells were converted to Mascot Generic Format (MGF) using Proteome Discoverer version 1.4 (Thermo Scientific). The MGF files were searched using Mascot server version 2.5. A total of 29 discovery searches were performed, the details of which are reported in Table 1.

10 searches were performed with different reference proteomes and search parameters (search numbers 1 to 10 in Table 1). 19 searches were performed with different proteogenomics

databases and search parameters (search numbers 11 to 29 in Table 1). Carbamidomethyl (C) was set as fixed modification and 0.6 Da fragment tolerance was used in all searches. Two searches (search numbers 3 & 4 in Table 1) with the merged reference proteome were conducted with 15 ppm precursor mass tolerance whereas it was set to 10 ppm in all other searches. Trypsin was set as the digestion enzyme in all searches except one search with the merged reference proteome where it was set to semi-trypsin (search number 5 in Table 1). A database of common Repository of Adventitious Proteins (cRAP) was obtained from The Global Proteome Machine (GPM) website and searched alongside all databases. Each of the proteogenomic databases was searched alongside the merged reference proteome database and cRAP database except for the 6 frame nucleic acid searches (search numbers 24,25,26,27 and 29 in Table 1) because Mascot server version 2.5 did not allow amino acid database to be specified alongside nucleic acid databases. A maximum of 2 missed cleavages were allowed in all searches except for two proteogenomics searches where it was set to 0 (search numbers 28 & 29 in Table 1). A total of 7 variable modifications were investigated for the reference proteome searches: Oxidation (M), Acetyl Protein (N-term), Deamidation (NQ), Carbamyl (N-term), Gln->pyro-Glu (N-term Q), Glu->pyro-Glu (N-term E) and Ammonia-loss (N-term C). Because of the much larger search space for proteogenomic databases, these searches were performed using the 3 most common variable modifications, which were determined in a preliminary experiment using the reference proteome. The variable modifications used for the proteogenomic searches were Oxidation (M), Acetyl Protein (N-term), and Deamidation (NQ).

Table 1: Summary of database searches. Variable modifications are abbreviated as follows: Ac: Acetyl (Protein N-term), AL: Ammonia-loss (N-term C), Ca: Carbamyl (N-term), NQ: Deamidated (NQ), Gln: Gln->pyro-Glu (N-term Q), Glu: Glu->pyro-Glu (N-term E), Ox: Oxidation (M). Databases are abbreviated as follows: C: cRAP, MRP: merged reference proteome. Other abbreviations: 3F: 3 frame, 6F: 6 frame, PC: protein coding, MC: missed cleavages, Tol: peptide mass tolerance (ppm).

#	Databases used	# Sequences	# Residues	Cleavage	MC	Variable Modifications	Tol
1	C, MRP	87334	32498123	Trypsin	2	Ac, Ox	10
2	C, MRP	87334	32498123	Trypsin	2	Ac, NQ, Ox	10
3	C, MRP	87334	32498123	Trypsin	2	Ac, Ox	15
4	C, MRP	87334	32498123	Trypsin	2	Ac, AL, Ca, NQ, Gln, Ox	15
5	C, MRP	87334	32498123	semiTrypsin	2	Ac, Ox	10
6	C, GENCODE proteome	94475	35226872	Trypsin	2	Ac, NQ, Ox	10
7	C, UniProt proteome	92633	36837450	Trypsin	2	Ac, NQ, Ox	10
8	C, RefSeq proteome	110502	74023435	Trypsin	2	Ac, Ox	10
9	C, RefSeq proteome	110502	74023435	Trypsin	2	Ac, NQ, Ox	10
10	C, RefSeq proteome	110502	74023435	Trypsin	2	Ac, NQ, Gln, Glu, Ox	10
11	C, SNV proteome, MRP	124700	75219935	Trypsin	2	Ac, Ox	10
12	C, SNV proteome, MRP	124700	75219935	Trypsin	2	Ac, NQ, Ox	10
13	C, InDel proteome, MRP	194239	100112565	Trypsin	2	Ac, Ox	10
14	C, InDel proteome, MRP	194239	100112565	Trypsin	2	Ac, NQ, Ox	10
15	C, Exonskip proteome, MRP	274370	270153356	Trypsin	2	Ac, Ox	10
16	C, Exonskip proteome, MRP	274370	270153356	Trypsin	2	Ac, NQ, Ox	10
17	C, GNOMON small, MRP	156470	87455704	Trypsin	2	Ac, Ox	10
18	C, PC transcripts ORFs 3F, MRP	2309314	140406097	Trypsin	2	Ox	10
19	CDS extensions ORFs 3F, C, MRP	1706623	109281358	Trypsin	2	Ac, NQ, Ox	10
20	C, LncRNA ORFs 3F, MRP	212924	42228922	Trypsin	2	Ac, Ox	10
21	C, Retained intron ORFs 3F, MRP	243503	47876774	Trypsin	2	Ac, Ox	10
22	C, Retained intron ORFs 3F, MRP	243503	47876774	Trypsin	2	Ac, NQ, Ox	10
23	C, GNOMON proteome, MRP	404236	237567077	Trypsin	2	Ac, Ox	10
24	GENCODE transcriptome 6F	1188558	588294950	Trypsin	2	Ox	10
25	GENCODE transcriptome 6F	1188558	588294950	Trypsin	2	NQ, Ox	10
26	RefSeq transcriptome 6F	1058556	1173530098	Trypsin	2	Ox	10
27	RefSeq transcriptome 6F	1058556	1173530098	Trypsin	2	NQ, Ox	10
28	C, GENCODE PC genes 3F, MRP	20599397	870230701	Trypsin	0	Ox	10
29	Genome 6F	1612098	6505830240	Trypsin	0	Ox	10

2.12. Validation searches

35 data files from the validation experiments, namely MCF7 cell extracts spiked with SIS peptides, were converted to MGF using MSConvert (Proteowizard version 3.0.10051). A proteogenomic database was created that contained the non-canonical peptides identified in the discovery proteogenomics searches. The MGF files of the 35 validation experiments were then searched on Mascot server against the merged reference proteome, proteogenomics database and cRAP database. Precursor and fragment mass tolerance were set to 15 ppm and 0.6 Da, respectively. Carbamidomethyl (C) was set as fixed modification and Oxidation (M), Acetyl Protein (n-term), Deamidated (NQ), Carbamy (n-term), Gln->pyro-Glu (N-term Q), Label:13C(6)15N(2) (C-term K) and Label:13C(6)15N(4) (C-term R) were set as variable modifications. A maximum of 2 missed cleavages was allowed.

2.13. Filtering proteogenomics peptides

The results from the proteogenomics database searches were filtered at a significance threshold (p-value) of 0.05 and percolated to a False Discovery Rate (FDR) of 1% utilizing the "show percolator scores" option in Mascot. The FDR estimation was performed using a target decoy strategy inside Mascot using default settings. PSMs with percolated Mascot score below 13 were removed. After utilizing Percolator [100] the Mascot score threshold 13 corresponded to a Mascot expectation score threshold of 0.05. All spectra identified in the discovery proteogenomics searches were collected. Python scripts were then utilized to filter the search results. We applied multiple filters to minimize false positive identifications. Specifically:

i) All peptides identified in the reference database searches were removed. We filtered peptides against UniProt, RefSeq, GENCODE and cRAP databases by string search to ensure no non-canonical peptides could be mapped onto reference proteins or contaminants.

ii) Peptides that differed with reference peptides only by Leucine/Isoleucine were removed.

iii) Non-canonical peptides that contained a deamidation were removed if the peptide, after deamidation, could be mapped onto a reference protein. *i.e.* the non-

canonical peptide and the peptide contained in a reference protein differed only by the deamidation (N→D or Q→E) .

iv) Non-canonical peptides with variable modifications Deamidation (NQ) and/or Oxidation (M) were removed if the corresponding unmodified peptide was not also identified.

v) Non-canonical peptides identified from spectra whose precursor isolation exhibited greater than 70% interference were removed. Interference values for the precursor isolation were exported from Proteome Discoverer (version 1.4).

vi) Only non-canonical peptides from MS/MS spectra that had at least 100 fragment peaks were retained. A dynamic noise level algorithm (DNL) was then applied to the MS/MS spectra [96]. Fragment peaks with intensity above the noise level were regarded as signal peaks. Peptides from MS/MS spectra that had less than 8 signal peaks were also removed.

vii) Only non-canonical peptides with length 10 to 40 amino acids were retained.

viii) Non-canonical peptides originating from reference transcriptome and genome derived databases were also subjected to a local blastp (version 2.6.0+) [101] search against a combined database of the protein sequences from GENCODE, UniProt, RefSeq and cRAP proteomes. Non-canonical peptides originating from transcriptome and genome derived databases were removed if they could be aligned, using a maximum of two amino acid difference, with any protein sequence in the combined proteome. We noted that some of these peptides could not be aligned to reference proteins by blast search due to regions of low amino acid complexity, in this instance a string based search was used to establish if they matched.

Some non-canonical peptides could be identified using multiple proteogenomics databases. We recorded the origin of all filtered peptides for all database searches and assigned a database to the peptide based on the following database priority rule: SNVs > InDel > Exonskip > GNOMON-small > GENCODE protein coding transcripts > retained introns > CDS extensions > GENCODE protein coding genes > lncRNAs > GENCODE transcriptome > RefSeq transcriptome > GNOMON > human genome.

2.14. Proteogenomics mapping

The "*genomic context*" of the non-canonical peptides was provided by mapping them onto the proteome, transcriptome and genome sequences using in-house python scripts. We obtained the genomic coordinates of the peptides in BED file format. Mapping was performed in a step-wise manner in which peptides mapped at each level were classified and filtered out from the next mapping step as follows:

i) Non-canonical peptides identified in the SNVs, InDel and Exonskip database searches were mapped onto their respective proteins. The peptide coordinates on the proteins were then converted to genomic coordinates utilizing the GENCODE release 27 annotation file. Mapped peptides were classified as SNV-pep, InDel-pep and Exonskip-pep, respectively.

ii) Peptides identified from GNOMON databases were mapped onto the GNOMON predicted protein sequences from NCBI annotation release 108. The peptide coordinates were converted to genomic coordinates using the GNOMON predicted annotation files from reference (ref_GRCh38.p7_gnomon_top_level.gff3) and alternate (alt_CHM1_1.1_gnomon_top_level.gff3) assemblies. Mapped peptides whose coordinates overlapped with any known protein-coding gene's coordinates in the main annotation files from reference (ref_GRCh38.p7_top_level.gff3) and alternate (alt_CHM1_1.1_top_level.gff3) assemblies were classified as novel-isoform-pep. Peptides mapping to locations containing non-coding genes in the main annotation files were classified as non-coding-pep, and peptides that mapped to locations that did not contain any known gene in the main annotation files were classified as novel-CDS-pep.

iii) All remaining unclassified peptides were then mapped onto the GENCODE release 27 transcriptome in 3 frames. Mapped peptides were classified as uORF-pep, altCDS-pep and dORF-pep if they mapped onto the 5'-UTR, CDS and 3'-UTR regions of protein-coding or NMD transcripts, respectively. Peptides mapping onto the non-coding RNAs were classified as non-coding-pep.

iv) Peptides were then mapped onto the GENCODE release 27 gene sequences in 3 frames. Peptides mapping onto protein coding genes were classified as intron-pep and exon-extension-pep if they originated from introns and exon-intron boundaries

respectively. Peptides mapping onto the non-coding genes were classified as non-coding-pep.

v) Next, peptides were mapped onto the transcriptome and gene sequences from NCBI annotation release 108 in 3 frames, and classified as mentioned in section iii) and iv) above for GENCODE annotation.

vi) Peptides were then mapped onto the complement sequences of GENCODE transcriptome, GENCODE genes, RefSeq transcriptome and RefSeq genes, respectively. Mapped peptides were classified into novel-CDS-pep.

vii) Finally, peptides were mapped onto the full human genome (GRCh38.p10) in 6 frames. The peptide co-ordinates were converted to genomic coordinates. Peptides were classified into novel-CDS-pep if they mapped outside the annotated gene regions in GENCODE release 27 annotation file or RefSeq annotation from NCBI annotation release 108.

During all mapping steps we accepted only those peptide co-ordinates that obeyed tryptic cleavage rule since all proteogenomics searches were conducted in tryptic mode. Finally, we performed gene based grouping of mapped proteogenomics peptides. Peptides that mapped to more than one gene, to multiple genomic coordinates, or were classified into more than one *genomic context* were reclassified into ambiguous-pep.

2.15. Validation of the non-canonical peptides

326 peptides corresponding to novel proteins identified with the discovery proteogenomics searches were selected for validation using SIS peptides. The peptides were selected on the basis of their length, 10-21 amino acids for easier synthesis, and without cysteine to avoid difficulties associated with Sulphur oxidation. Five LC-MS/MS validation experiments were performed (see LC-MS/MS section for details). After database search (see section database searches for search parameters) the results were again percolated to a target FDR of 1%. We collected the spectra of the SIS and endogenous peptides from the discovery and validation searches. A two tier validation of selected peptides was then performed. In *Tier 1*, the cosine similarity was used to quantitatively compare the fragment spectra from the endogenous and SIS peptides. Peptides with a cosine similarity greater than 0.9 were considered validated at

tier 1. For *Tier 2*, the cosine similarity was used to compare the elution profiles and isotopic composition of SIS and endogenous peptides. Peptides that had a profile similarity and isotopic composition similarity greater than 0.9 were considered validated at tier 2.

3. Results

We performed a comprehensive proteogenomics analysis of MCF7 cells, a popular cancer cell-line routinely utilized in cancer studies. We constructed customized databases utilizing variants from NGS data and COSMIC, GNOMON predicted proteins, reference transcriptome and the human genome. We first conducted extensive searches with reference proteomes from UniProt, GENCODE, RefSeq and a merged reference database. All peptides identified using the reference proteome databases were filtered out from the peptides identified using the proteogenomics searches. In this manner the subsequent data analysis focused exclusively on peptides due to novel (non-canonical) proteins. To guard against false positives we also conducted extensive QC checks of the identified non-canonical peptides. A sub-set of these peptides were then validated using synthetic isotopically-labeled standard (SIS) peptides. This study highlights the presence of proteoforms in MCF7 cells that are missed by proteome profiling experiments that only utilize reference proteomes and thus thereby underestimate the complexity of the oncoproteome.

3.1. Application of the DNL algorithm

MS/MS spectra contain noise. It is known that the database search method used to identify proteins can report confidently identified peptides from MS/MS spectra with poor signal-to-noise. An example of such a PSM (peptide spectral match) with poor signal-to-noise is shown in Figure 17, in which the peaks colored red have been used for the peptide spectral match. When the focus of the study is the identification of novel proteoforms it is important to ensure that the identifications are not based on noise in the MS/MS spectra, as that would increase the chance of false positives. Low signal-to-noise spectra can be identified via visual inspection but given the throughput of modern proteomics experiments visual examination is highly unpractical. Instead we applied a dynamic noise level (DNL) algorithm [96] to the MS/MS spectra to remove PSMs with poor signal quality. After application of the DNL all MS/MS peaks in Figure 17 were found to fall below the noise level.

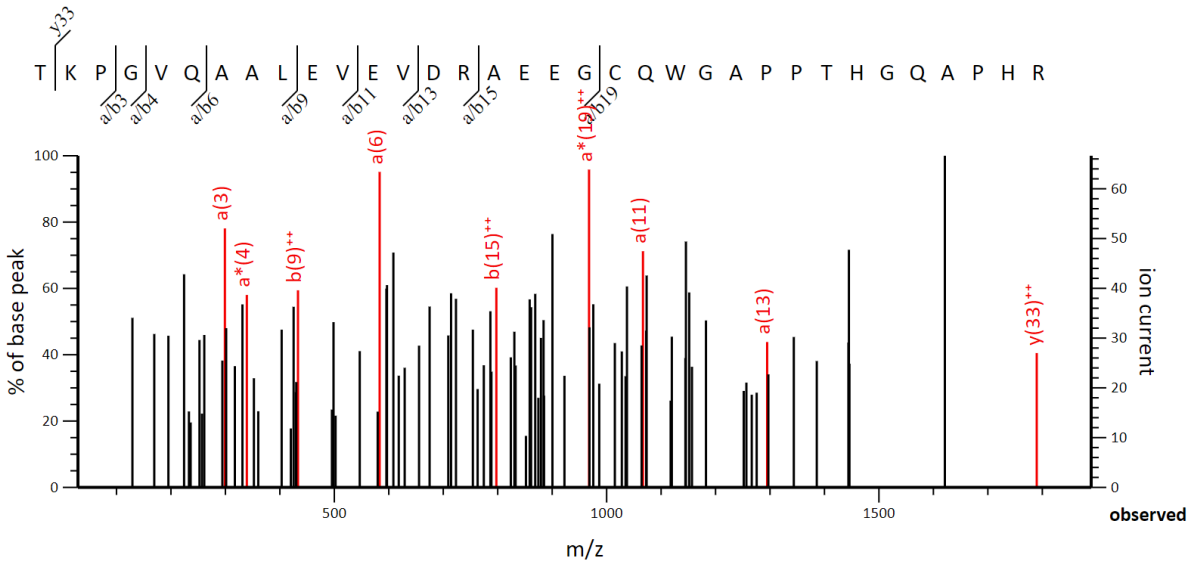

Figure 17. Example of a peptide identification resulting from an MS/MS spectrum with low S/N. The peptide "TKPGVQAALEVEVDRAEEGCQWGAPPTHGQAPHR", charge 3+, *m/z* 1226.993 was confidently identified by the Mascot search engine.

3.2. Comprehensive annotation of matched MS/MS spectrum

The database search of MS/MS spectra uses a set of predefined fragmentation channels. The choice of fragmentation channels depends on the tandem mass spectrometry method utilized to acquire the MS/MS spectra. For the collision induced dissociation method used here Mascot considers *a*, *b* and *y* fragment ions, including neutral loss of a single ammonia molecule. The maximum charge carried by the fragment ions is limited to 2+. These presets define the search space for the peptide-spectral-match.

The MS/MS spectrum of the peptide "GTAAAAAAAAAAAAKVPAK" is shown in Figure 18. The peptide was mapped to the variant protein sequence of 60S ribosomal protein L14 (RPL14). The variant protein contains three extra alanine residues (indicated in red above) when compared with the wild type RPL14 protein. Although the peptide sequence was identified with high confidence (Mascot Score: 43, expectation: 4.50E-05) many high intensity peaks in the spectrum remained unassigned (black peaks in Figure 18). Unassigned peaks in the MS/MS spectra increase the risk of false positive identifications, and so must be avoided when reporting novel proteins. To ensure all identifications of novel proteins could describe the majority of peaks present in the MS/MS spectra I developed a python based "spectrum

annotation tool" to comprehensively annotate the matched MS/MS spectrum. The annotation tool includes additional fragmentation channels not considered during the Mascot search. Note: the additional fragmentation channels were not used for identification, but to ensure the identifications could describe the MS/MS spectrum more completely. Figure 19 and Figure 20 demonstrate that the inclusion of additional fragmentation channels increased the number of MS/MS peaks that could be assigned, thereby increasing the confidence of the non-canonical peptide PSM.

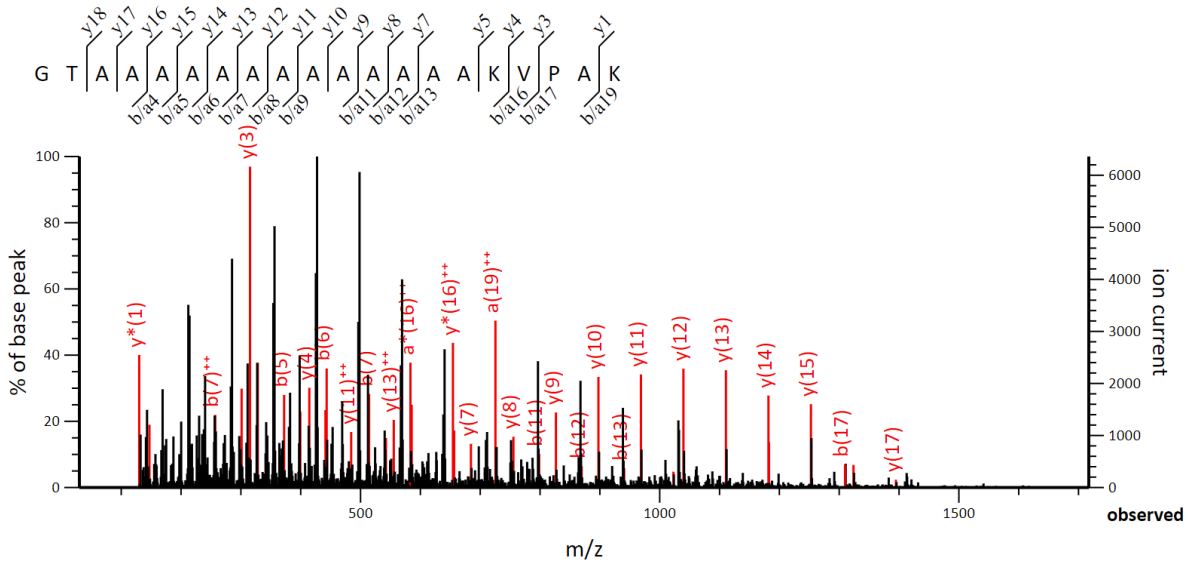

Figure 18. MS/MS spectra of the peptide "GTAAAAAAAAAAAAAKVPAK", charge 2+, *m/z* 812.465, as identified by Mascot (peaks used for identification indicated in red).

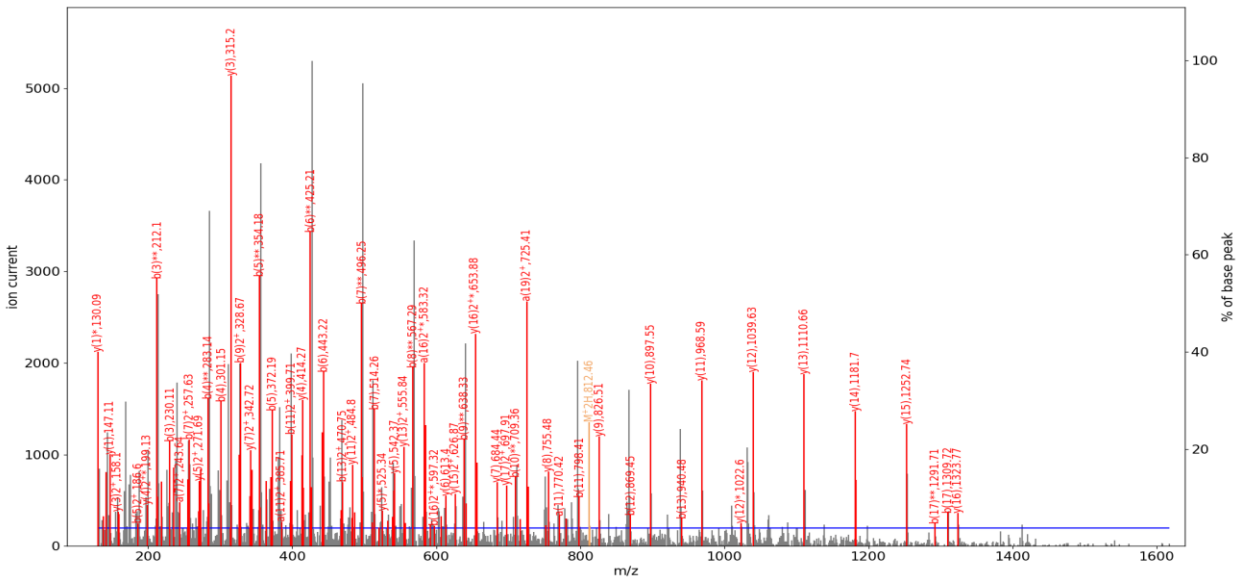

Figure 19. MS/MS spectra of the peptide "GTAAAAAAAAAAAAAKVPAK", charge 2+, *m/z* 812.465. Adding water loss as an additional fragmentation channel (indicated by **) led to additional peak matches, including *b*(3)**, *b*(4)**, *b*(5)**, *b*(6)**, *b*(7)**, *b*(8)**, *b*(9)** and *b*(10)**. Noise level (blue line) was determined by DNL.

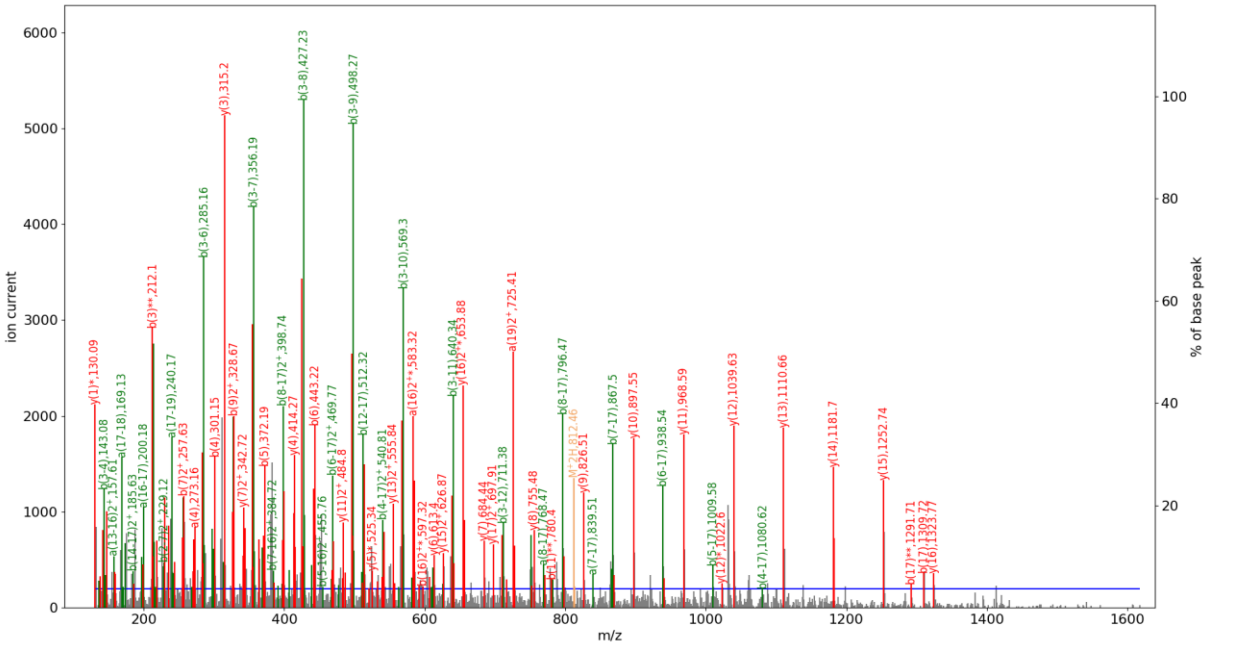

Figure 20. MS/MS spectra of the peptide "GTAAAAAAAAAAAAAKVPAK", charge 2+, *m/z* 812.465. Internal fragment ions (green peaks) were also utilized for annotation. Internal fragments are labeled by the start and end position of the fragment in the precursor peptide. Noise level (blue line) was determined by DNL.

3.3. Peptides from reference proteomes

When we utilized GENCODE, UniProt and RefSeq as reference protein sequence databases (searches 6, 7 and 9 from Table 1) ~95 % of the peptides were identified by all three databases, Figure 21. Not including redundancies (*i.e.* the same peptide identified using different databases) we identified 270,741 spectra grouped into 97,144 peptides from all reference proteome searches (searches 1 to 10 in Table 1). ~25% of the 1,096,963 input spectra were assigned to peptides of the reference proteome. 52,548 spectra (~19 % of identified) were assigned to peptides with PTMs and 20,134 spectra (~7 % of identified) were assigned to peptides with unspecific cleavage.

Using a standard RefSeq search with 2 variable modifications: Acetyl (N-term) and Oxidation (M) we could assign 199,133 MS/MS spectra. To ensure that MS/MS spectra from reference protein sequences were not mistakenly assigned as novel non-canonical proteins, we expanded the search to include different reference proteomes including common PTMs, artifacts and unspecific peptide cleavages. As a result, the number of MS/MS spectra that could be assigned to reference proteins was increased by almost 35%.

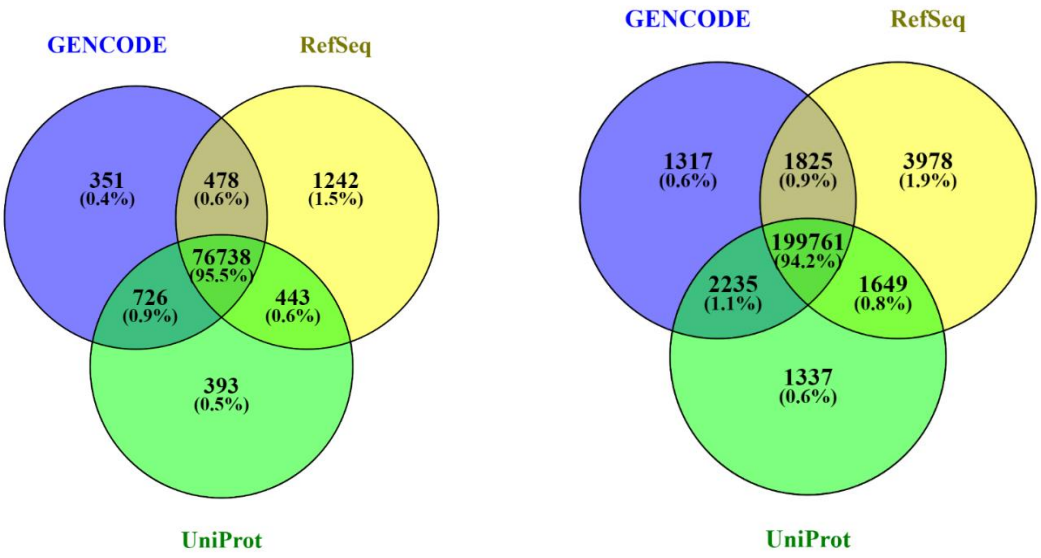

Figure 21. Venn diagram of identified peptides (left) and identified spectra (right) from three common reference proteomes. The data is extracted from searches 6, 7 and 9 from Table 1. Only rank 1 peptide and spectra matches are shown.

3.4. Sensitivity of database search decreases with increasing database size

The inclusion of all possible peptides contained in the transcriptome and genome greatly increase the size of the proteome database and thus decreases the statistical power of matching MS/MS spectra to peptides. To assess the severity of this effect for our analysis we determined how the protein identification rate was affected by database size. Searching the LC-MS/MS data against the merged reference proteome database resulted in 110,498 spectra unambiguously grouped onto 47,300 zero-missed-cleavage unmodified peptides (Figure 22). For the purposes of this database-size investigation we considered these identifications as representing 100% of the true reference spectra and reference peptides. The score distributions of these peptides, identified using different proteogenomics searches, are shown in Figure 22. Note: peptides with missed cleavages were not considered because searches with the ORF database derived from the gene sequence (search number 28 in Table 1) and the full human genome (search number 29 in Table 1) were conducted with 0 missed cleavages. A table summarizing peptides and spectra identified in all discovery searches is provided in online appendix 1. The number of peptides that could be consistently identified differed significantly for the proteome, transcriptome and genome derived databases. When we utilized the SNVs/InDel/Exonskip databases we recovered ~99% of the peptides identified using only the merged reference proteome. ~96% of peptides were recovered when the GNOMON database was used. Approximately, 91% of the peptides were recovered when utilizing the ORF database from CDS extensions and protein coding transcripts. ~74% of the peptides were recovered using an ORF database derived from gene sequences, whereas ~82% were recovered from a six-frame translated full transcriptome database. When the full human genome was searched in 6 frames just ~37 % of the peptides were recovered.

The peptides that are present in the merged reference proteome are also present in the reference transcriptome and genome (except peptides spanning splice sites). Thus, at a constant 1% FDR fewer peptides could be identified when a larger database was utilized (even if all databases contained the peptides). As the peptide search space increases from proteome < transcriptome < genome the probability that an MS/MS spectrum will match to a random peptide sequence within the database increases, thus leading to an increased rate of false discovery. To mitigate the effect of larger database size and hence increased rate of false discovery, we conducted searches with proteogenomics databases targeting specific proteogenomics classes. For

example: to find peptides originating from transcripts with retained introns we utilized the retained intron transcripts from GENCODE and translated it in three frames, which accounts for ~13% of the full GENCODE transcriptome in release 27.

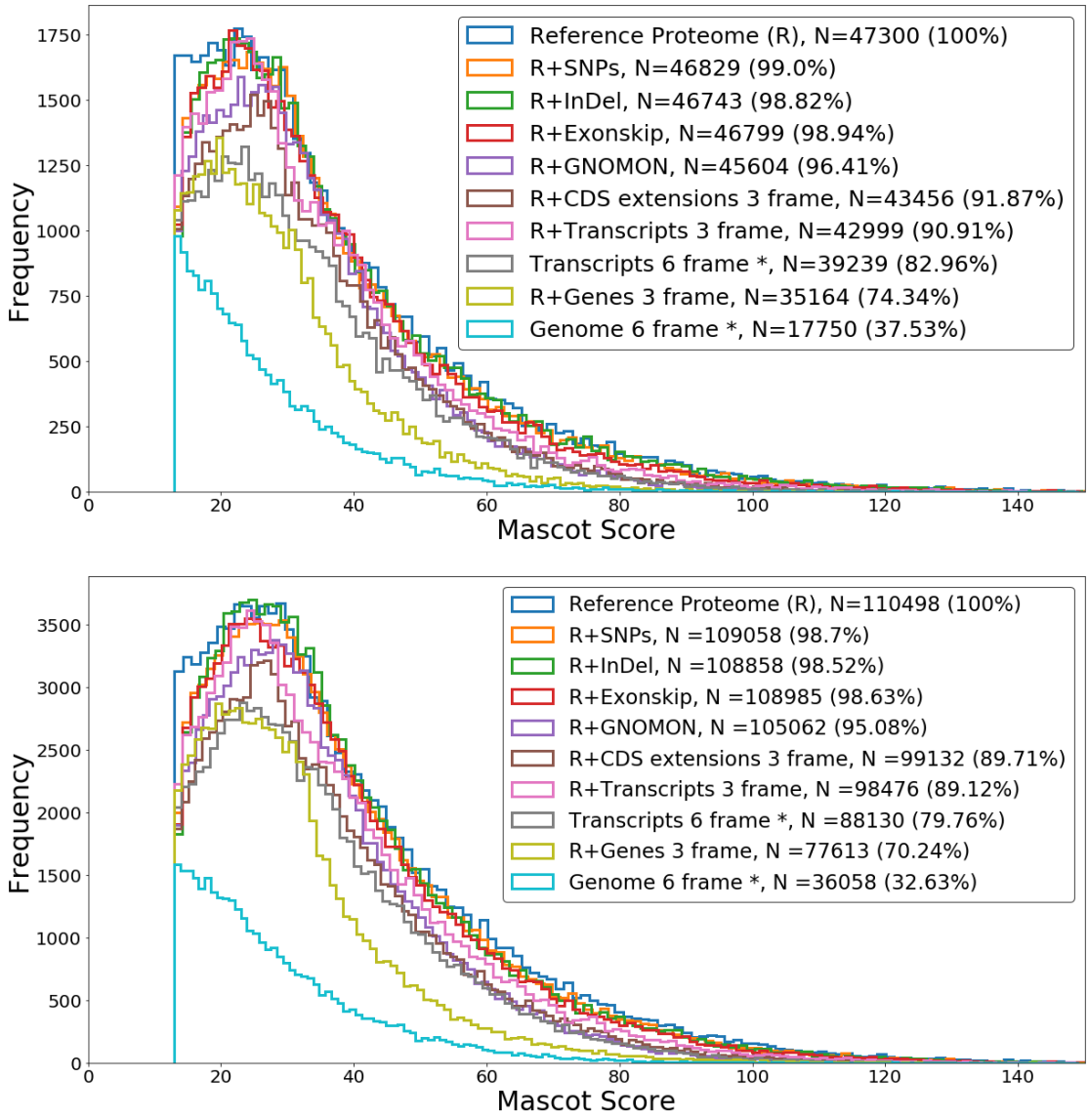

Figure 22. Percolated Mascot score distribution of unambiguous peptides (top) and spectra (bottom) with 0 missed cleavages identified in proteogenomics searches that were previously identified in the merged reference proteome search. The nucleic acid searches are indicated with an *. Search results were percolated to target FDR of 1% with p-value 0.05. Minimum Mascot score is 13.

3.5. Classification of non-canonical peptides identified by proteogenomics searches

We collected all MS/MS spectra identified exclusively with the proteogenomics searches. This included 3760 spectra grouped into 3021 peptides. We subjected these identifications to extensive filtering to ensure all of these non-canonical peptides originated from high quality mass spectra and could not be assigned to reference peptides with common modifications. After this filtration step 1726 spectra remained that could be grouped into 1227 peptides. Annotated spectra of all filtered peptides are included in online appendix 8. We mapped the proteogenomic peptides onto the genome and classified all filtered peptides (Table 2). An example of the mapping of peptides onto the gene StAR related lipid transfer domain containing 10 (STARD10) is shown in Figure 23. The full list of filtered, classified peptides is provided in online appendix 2. Out of 1227 filtered peptides 55 could not be mapped onto the genome (Table 2). The genomic coordinates of all mapped 1172 peptides is provided in online appendix 3 & 4 in BED file format. 59 peptides were classified as ambiguous, which means they could be mapped onto multiple genomic coordinates, multiple genes or were classified into more than one genomic context. The remaining 1113 peptides were unambiguously classified into 11 different classes (Table 2). Of these, 203 peptides were classified as novel-CDS-pep, which means they mapped onto genomic coordinates that are not currently annotated. The remaining 910 peptides were unambiguously grouped into 790 genes (Table 2). A full itemized list of peptides, grouped by genes is provided in online appendix 5.

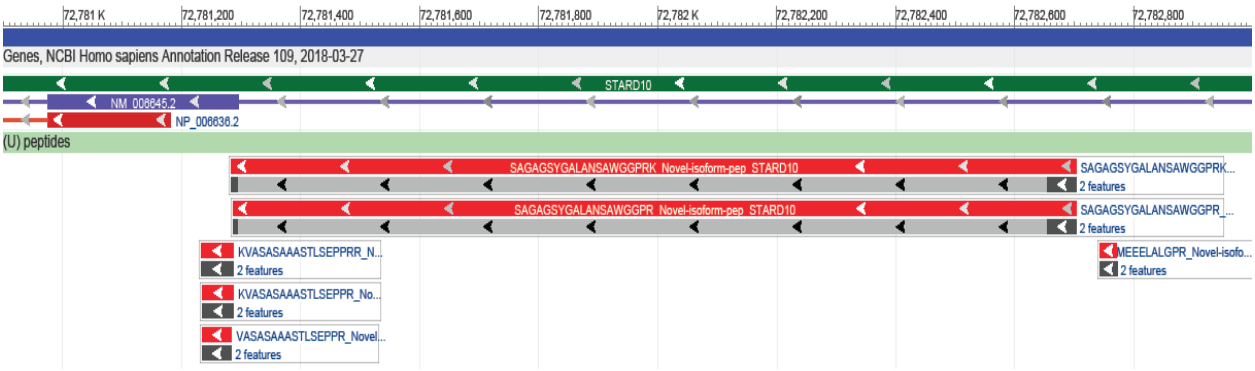

Figure 23. Visualization of the genomic mapping of non-canonical peptides on to the gene STARD10. The peptides (shown in the "peptides" track) were identified from a GNOMON predicted protein model of STARD10 gene. The reference STARD10 gene, transcript (NM_006645.2) and protein (NP_006636.2) are shown in green, purple and red blocks, respectively, in the upper annotation track. The exons are shown as blocks and introns as lines connecting the blocks. The reference protein starts from exon 2 of the STARD10 transcript (note protein is on reverse strand). The GNOMON predicted model protein differs from the reference STARD10 protein, in that its N-terminal region has 77 additional amino acids. Proteogenomic mapping reveals that three of the peptides "R.KVASASAAASTLSEPPR.R", "R.KVASASAAASTLSEPPRR.T", and "K.VASASAAASTLSEPPR.R" map onto the 5'-UTR region of the reference STARD10 mRNA. The results also demonstrate the presence of a novel exon predicted by GNOMON. "R.SAGAGSYGALANSAWGGPR.K" and "R.SAGAGSYGALANSAWGGPRK.V" map onto a region spanning the novel GNOMON exon and exon 2 of the reference STARD10 mRNA, and the peptide "MEEELALGPR.G" maps exclusively onto the novel exon.

Table 2. All filtered non-canonical (proteogenomics) peptides classified by genomic context.

	Discovery			Validation		
Genomic context	**Peptides**	**PSMs**	**Genes**	**Targeted**	**Tier 1**	**Tier 2**
SNV-pep	295	630	219	152	143	108
InDel-pep	36	46	35	6	3	1
Exonskip-pep	8	9	8	0	0	0
Novel-isoform-pep	45	108	36	17	16	12
uORF-pep	87	126	72	27	17	5
altCDS-pep	78	83	78	15	7	3
dORF-pep	61	61	61	7	0	0
Exon-extension-pep	30	35	29	1	1	1
Intron-pep	98	100	97	8	1	0
Non-coding-pep	172	187	169	32	9	2
Novel-CDS-pep	203	213	0	46	4	1
Ambiguous-pep	59	73	0	10	5	4
Unmapped-pep	55	55	0	5	1	0
Total	1227	1726	790	326	207	137

3.6. Validation of non-canonical peptides identified by proteogenomics search

A subset of the non-canonical peptides (length 10-21 amino acids, cysteine free) identified in the discovery searches were selected for validation. Isotopically labeled analogues of 326 proteogenomic peptides were synthesized using heavy lysine or heavy arginine (stable isotopically labeled). Mixtures of the stable isotopically-labeled standard (SIS) peptide analogues were added to the MCF7 proteolytic peptides at 3 different concentrations and the samples analyzed by LC-MS/MS. The presence of endogenous proteogenomic peptides were validated on the basis of the similarity of the tandem mass spectra of the endogenous peptides with their isotopically labeled analogues (tier 1), and on the basis of matched retention times and elution profiles (tier 2). An example of a tier 1 and 2 validated peptide "R.SAGAGSYGALANSAWGGPR.K" from a GNOMON predicted novel isoform of STARD10 protein is shown in Figure 24 and Figure 25. Out of 326 peptides targeted for validation 19 of the isotopically labeled analogues were not detected (online appendix 2). 207 peptides passed tier 1 validation, in which the MS/MS spectrum of the endogenous non-canonical peptide scored at least 0.9 cosine similarity with that of the SIS validation standard (Table 2). A comparison of the MS/MS spectra from the endogenous peptides with their SIS analogues, for all non-canonical peptides that passed tier 1 validation is shown in online appendix 9. Of these, 137 peptides (66%) also passed tier 2 validation (Table 2). A comparison of the extracted ion chromatograms of the endogenous peptides and their SIS analogues, for all tier 2 validated non-canonical peptides, is available in inline appendix 10. In the following sections, we show examples of peptides identified from different genomic events. Peptides validated at tier 1 are shown in *italic* fonts and peptides validated also at tier 2 are shown in ***bold + italic*** fonts.

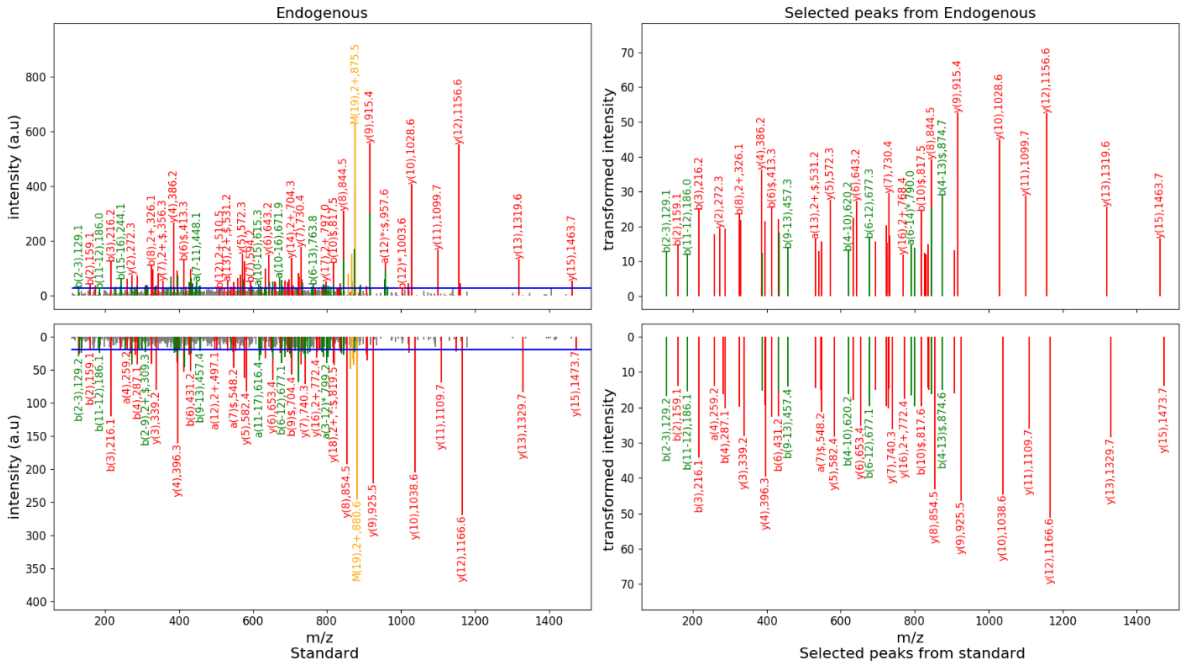

Figure 24. Annotated MS/MS spectra of the endogenous and SIS peptide of a novel isoform of STARD10. Peptide "R.SAGAGSYGALANSAWGGPR.K", charge 2+ (left). Noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments in green and un-fragmented precursors in orange. Ammonia loss is indicated with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity calculation. A total of 46 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.98.

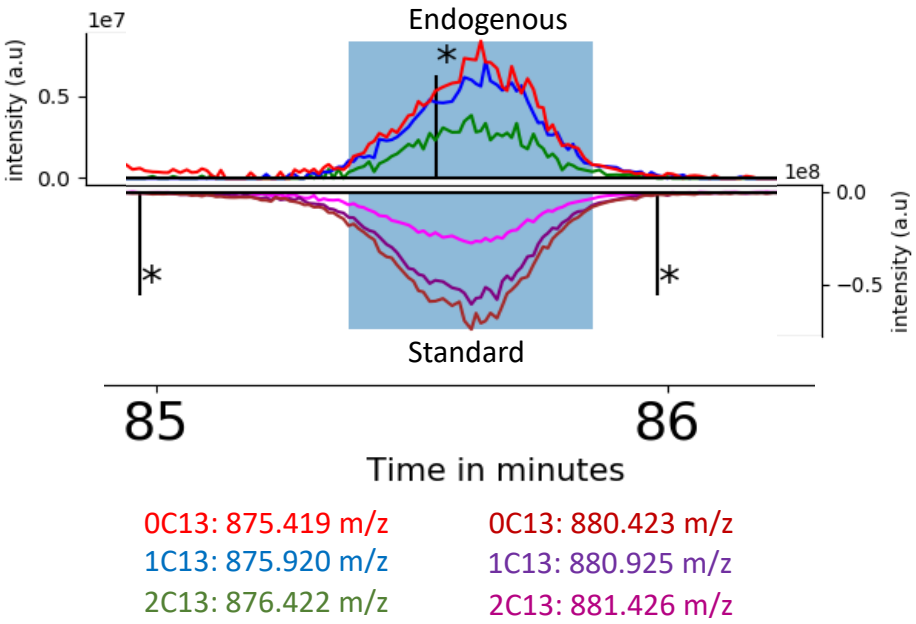

Figure 25. Extracted ion chromatograms of the endogenous peptide and the SIS peptide of a novel isoform of STARD10. Peptide "R.SAGAGSYGALANSAWGGPR.K", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box, and which corresponded to 42 MS scans. The peptide passed tier 2 validation with a profile similarity 0.99. Apex elution times for the SIS and endogenous peptides were both 85.6 minutes. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The mass traces in the extracted ion chromatograms were extracted using an *m/z* tolerance of +/-10 ppm.

3.7. Peptides from variants identified by next-generation sequencing

SNVs were the largest group of non-canonical peptides identified by the proteogenomics searches. 295 SNV peptides were identified with 630 PSMs (Table 2). We performed gene based grouping of the variant peptides and grouped the 295 peptides onto 219 protein coding genes. 152 SNV peptides were selected for validation using SIS peptides, of which 143 (94%) passed tier 1 validation and 108 (71%) also passed tier 2 validation. The largest number of SNV peptides was obtained from Plectin (PLEC) with 10 peptides and 28 spectra (online appendix 5). The SNV peptide "***R.EQLQQEQALLEEIER.H***", variant amino acid shown in grey (Q/R), produced due to the variant "rs11136334" (Highest population MAF: 0.46) [102] was validated at tiers 1 and 2 (Figure 26 and Figure 27).

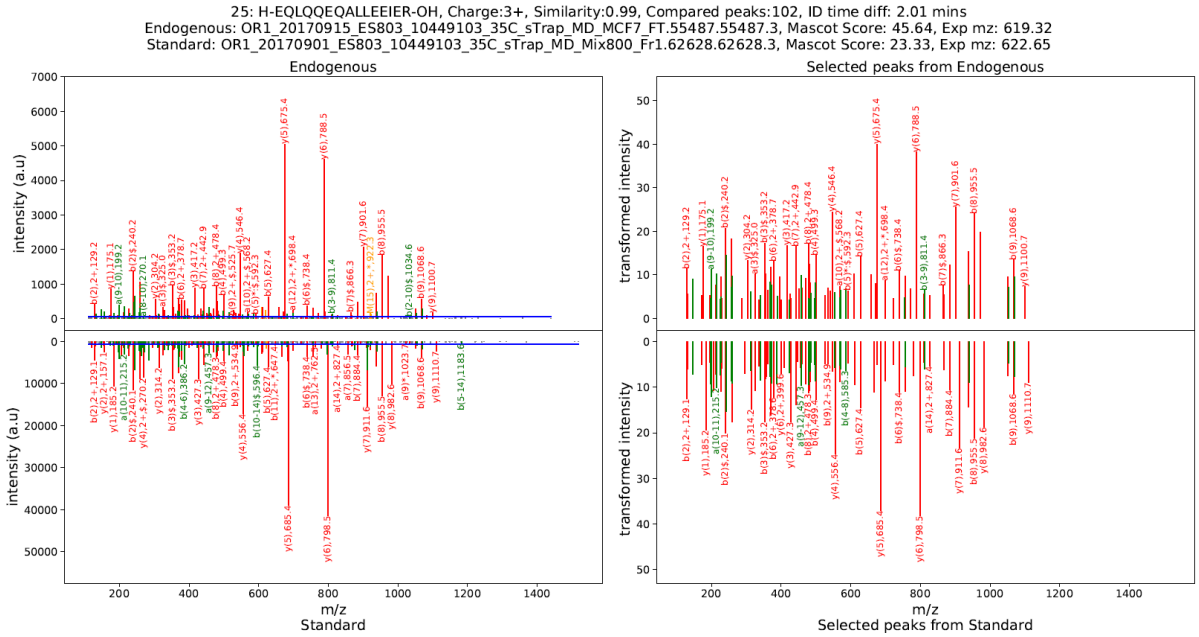

Figure 26. MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from a common variant in PLEC gene. Peptide "***R.EQLQQEQALLEEIER.H***", charge 3+. The noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (a-type, b-type) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensity of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 102 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.99.

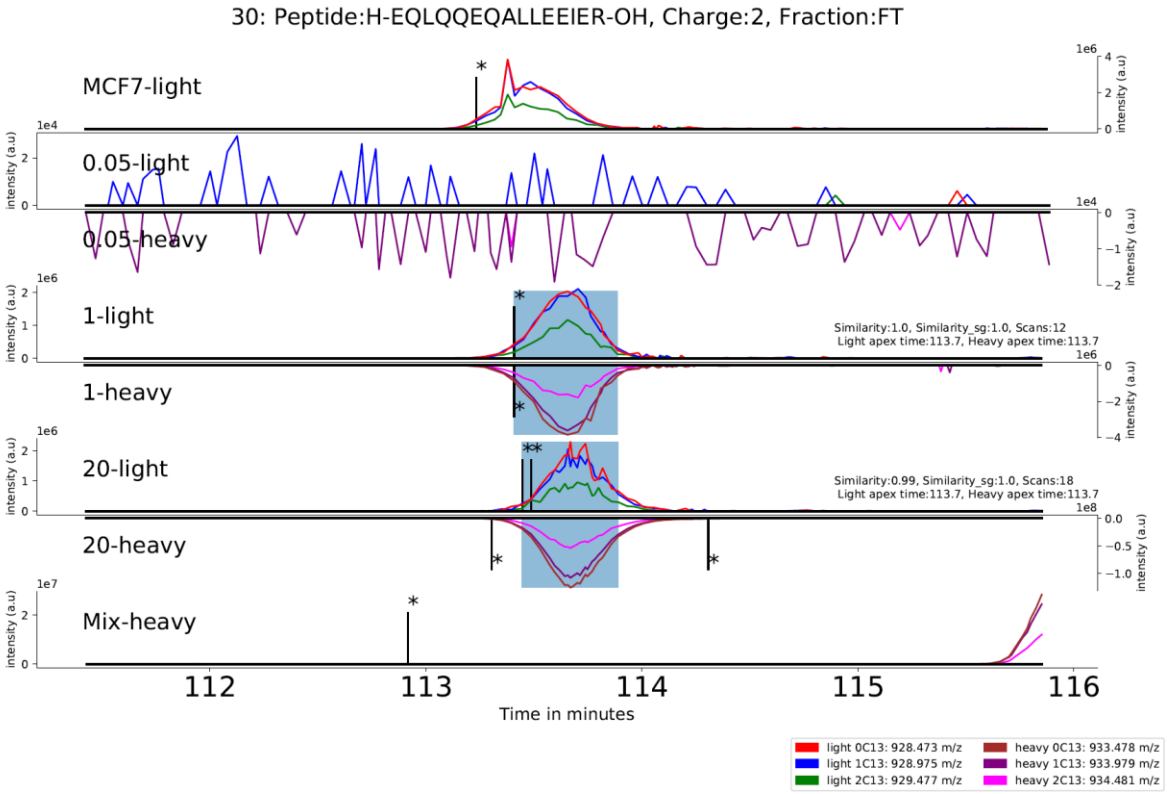

Figure 27. Extracted ion chromatograms (EIC) of peptide "***R.EQLQQEQALLEEIER.H***", charge 2+, in 5 validation runs. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 1.0. Apex elution times for the SIS and endogenous peptides were both 113.7 minutes. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide. The legends show the calculated *m/z* of the peptide isotopes.

A rare variant "rs2669761" (Highest population MAF: <0.01) [103] of WASH complex subunit 2A (WASHC2A) was detected in the NGS experiments. The two peptides "*K.TNTFPLLEDEDDLFTDQKVK.K*", and "***K.TNTFPLLEDEDDLFTDQK.V***", were identified (variant amino acid in grey), the first of which passed tier 1 validation (Figure 28) and the second passed tier 1 and tier 2 validation (Figure 29 and Figure 30).

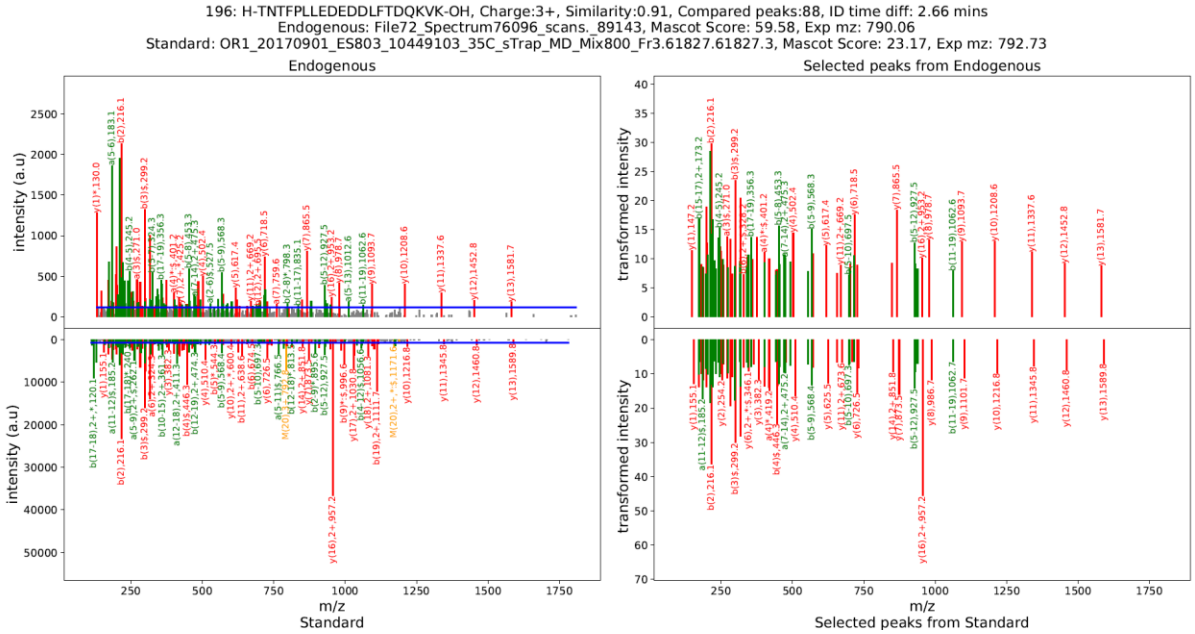

Figure 28. MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from a rare variant of WASHC2A. Peptide "*K.TNTFPLLEDEDDLFTDQKVK.K*", charge 3+. Noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type*, *b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensity of the selected peaks were variance stabilized by square root transform and normalized to sum 100 before similarity computation. The intensities of 88 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.91.

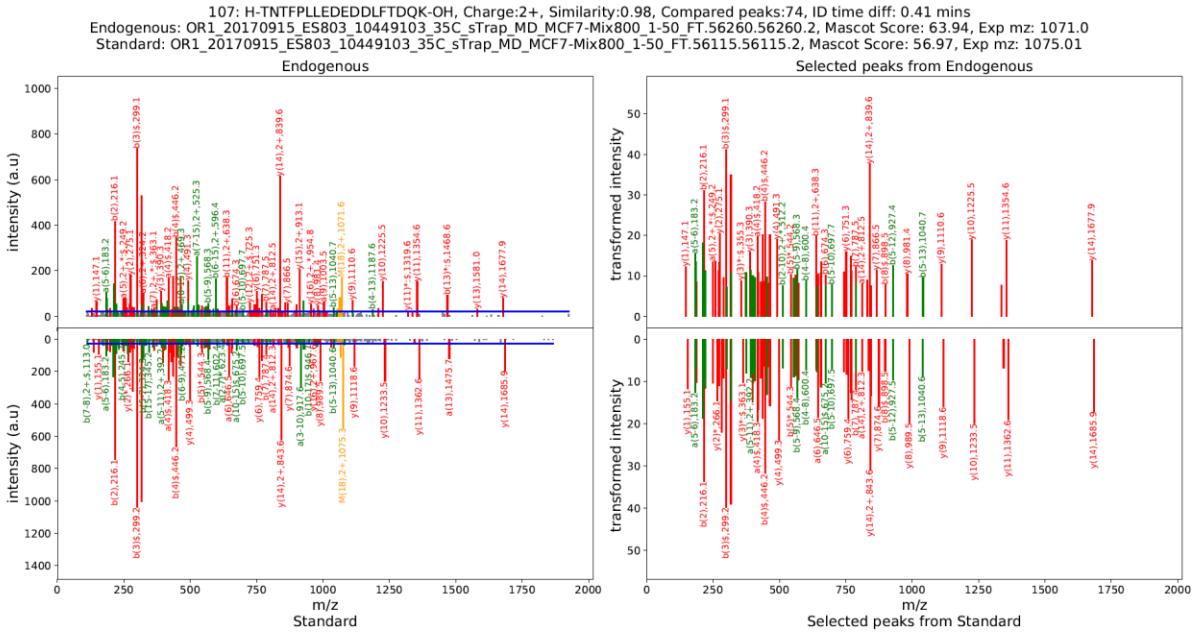

Figure 29. MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from a rare variant of WASHC2A. Peptide "***K.TNTFPLLEDEDDLFTDQK.V***", charge 2+ (left). Noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type, b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensity of the selected peaks were variance stabilized by square root transform and normalized to sum 100 before similarity computation. The intensities of 74 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.98.

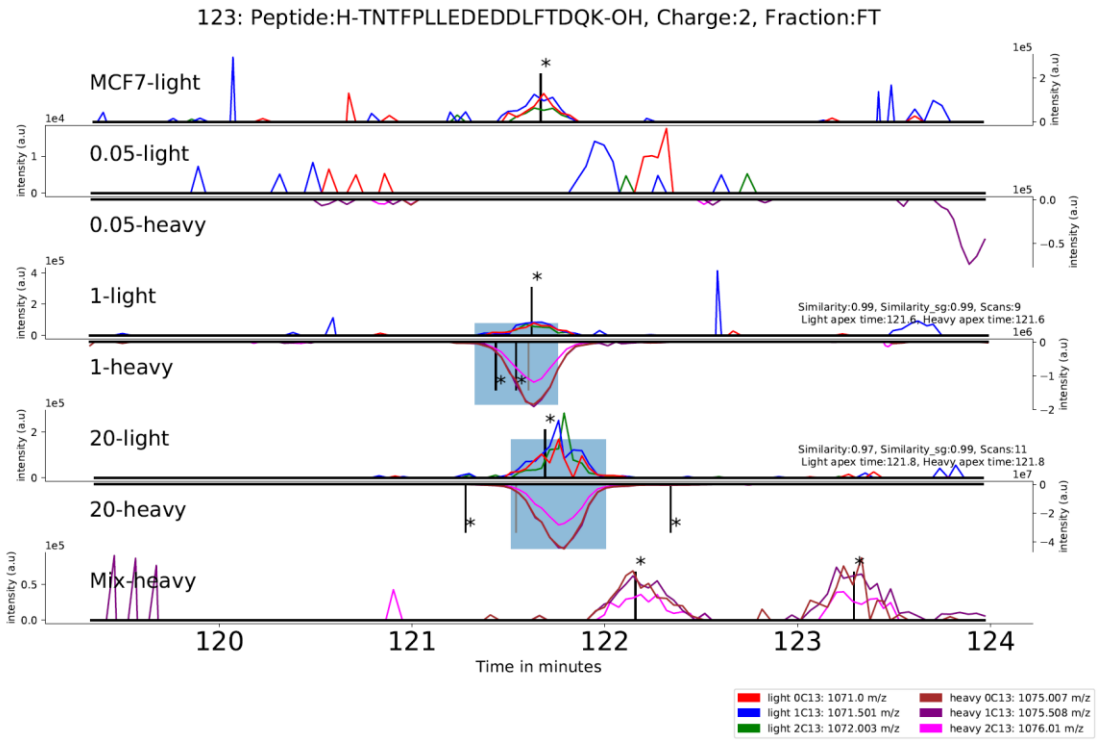

Figure 30. EICs of peptide "***K.TNTFPLLEDEDDLFTDQK.V***", charge 2+, in 5 validation runs. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.99. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison.

A total of 36 peptides from 46 PSMs were identified from insertion/deletion events. 6 InDel peptides were selected for validation, of which 3 passed tier 1 validation and 1 also passed tier 2. An in-frame insertion variant "rs369485042" [104] of a trinucleotide repeat was detected in ribosomal protein L14 (RPL14) in its transcriptome. The corresponding peptides, each with three additional Alanine residues, "*K.GTAAAAAAAAAAAAKVPAK.K*" and

"*K.GTAAAAAAAAAAAAAAKVPAKK.I*" (additional amino acids in grey), were identified by 11 PSMs and validated at tier 1 (Figure 31 and Figure 32).

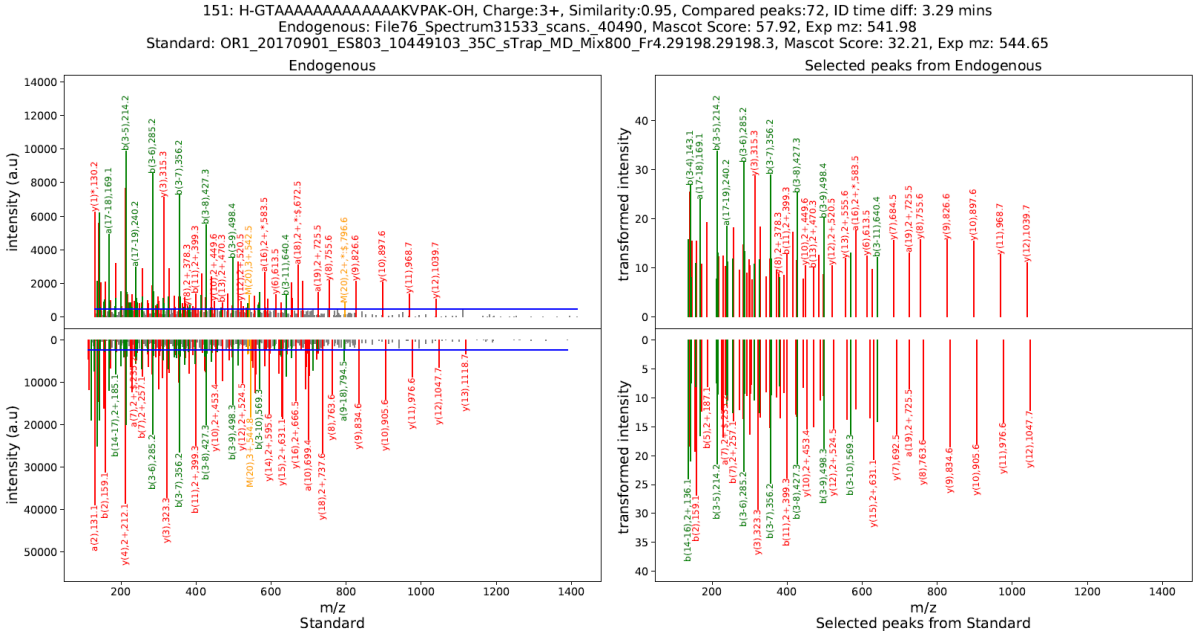

Figure 31. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) of an in-frame insertion variant of ribosomal protein L14. Peptide "*K.GTAAAAAAAAAAAAAKVPAK.K"*, charge 3+. Noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type*, *b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensity of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 74 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.95.

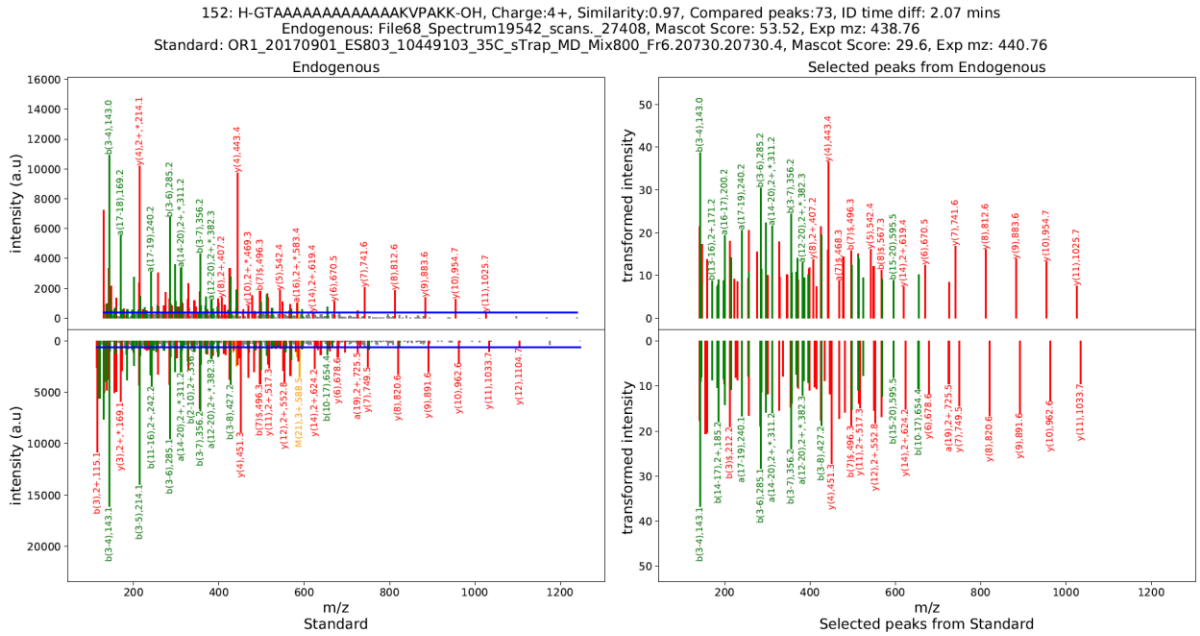

Figure 32. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) of an in-frame insertion variant of ribosomal protein L14. Peptide "*K.GTAAAAAAAAAAAAKVPAKK.I*", charge 4+. Noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type*, *b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensity of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 73 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.97.

An in-frame deletion event was detected in the gene Trinucleotide repeat containing 6B (TNRC6B) in the NGS experiments. The proteogenomic peptide "***MVSAL-QQQQQQQR***" with a deletion of the analogous Glutamine residue (indicated with a hyphen) was identified and validated at the proteome level (Figure 33 and Figure 34).

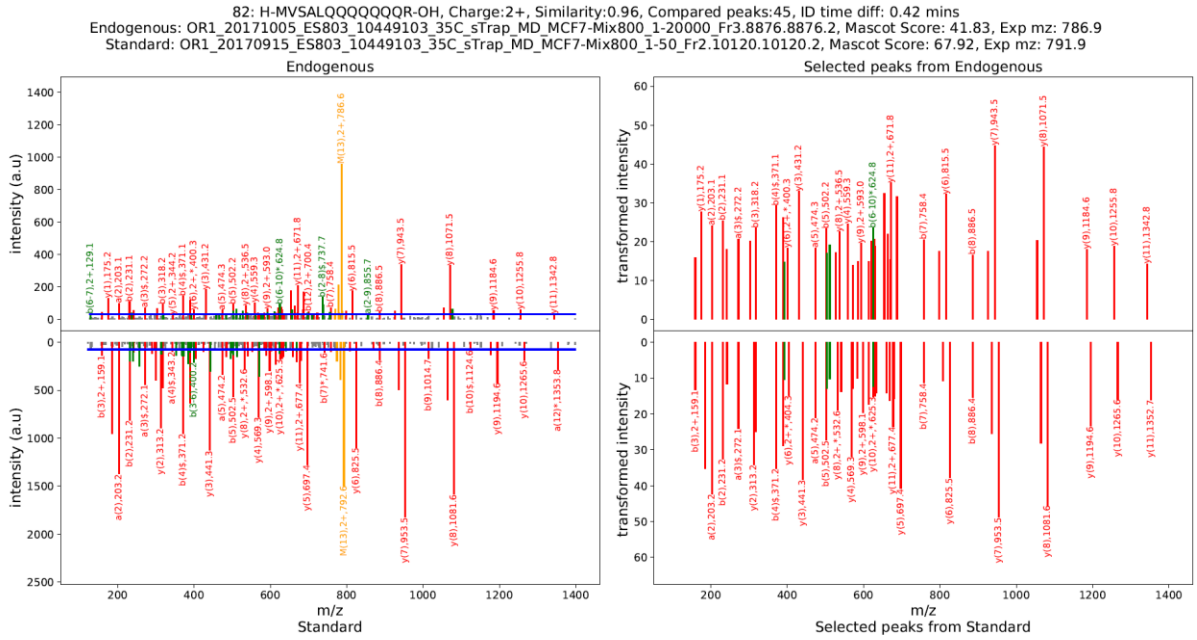

Figure 33. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) of an in-frame deletion variant of TNRC6B. Peptide "***MVSAL-QQQQQQQR***", charge 2+. Noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type, b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensity of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 45 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.96.

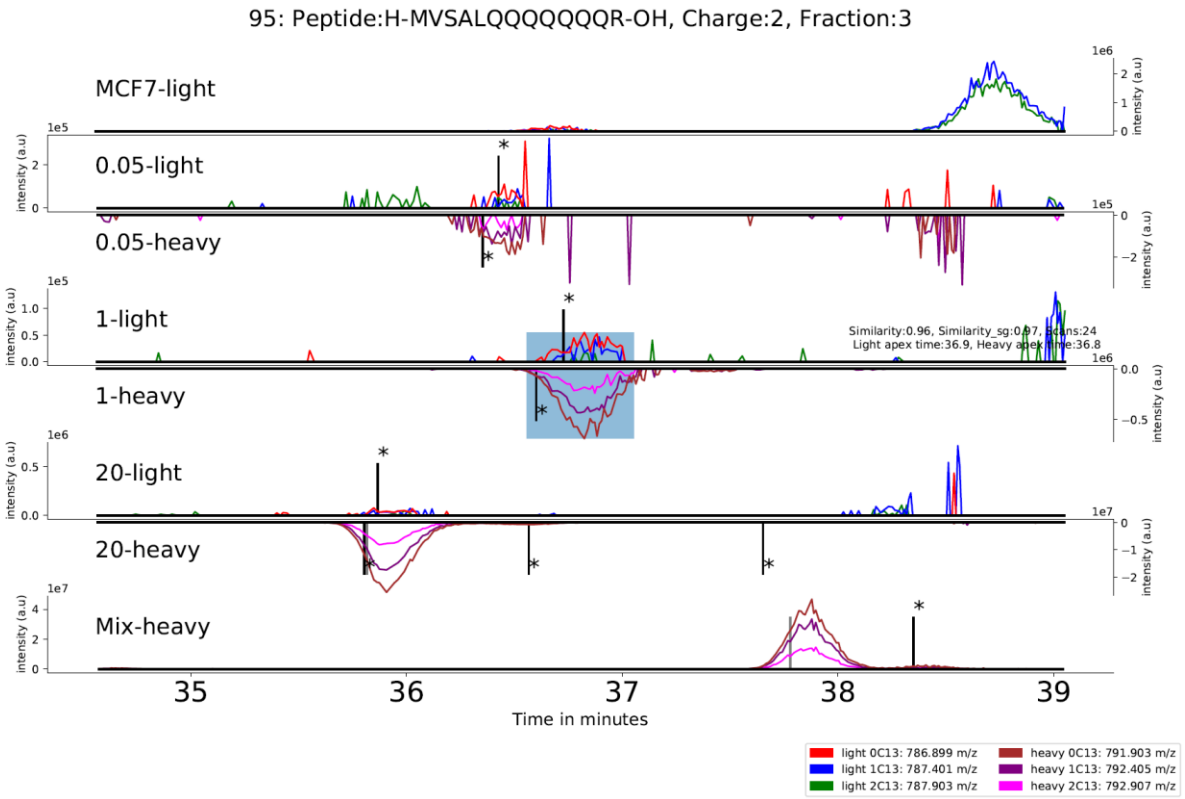

Figure 34. EICs of peptide "***MVSAL-QQQQQQQR***", charge 2+, in 5 validation runs. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.96. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

3.8. Peptides from exon-skipping events

We identified 8 peptides that were produced as a result of exon-skipping events (Table 2). None of the peptides were selected for validation because they did not meet the length requirements (10-21) for heavy peptide synthesis. The gene eukaryotic translation initiation factor 4 gamma 1 (EIF4G1) codes for a protein that is a component of the protein complex EIF4F. The peptide "MNKAPQSTGPPPAPSPGLPQHFYPSR.A" was identified from a novel isoform of EIF4G1, generated by the skipping of exon 3 from the protein coding transcript

"ENST00000342981", Figure 35. The amino acids "MNKAPQSTGPPPAPSPGLPQ" derive from exon 2, and "HFYPSR" derive from exon 4. The peptide was identified in the discovery experiments with a single PSM corresponding to a missed cleavage, but was further supported by the identification of the terminal tryptic peptide "K.APQSTGPPPAPSPGLPQHFYPSR.A" with 3 PSMs in the validation experiments. To the authors' knowledge, no known isoform has been reported that results from the splicing of exons 2 and 4 of transcript "ENST00000342981" (Figure 36). Thus, identification of this peptide indicates that a novel isoform of EIF4G1 is expressed in MCF7 cells, in which exon 3 of transcript "ENST00000342981" is skipped.

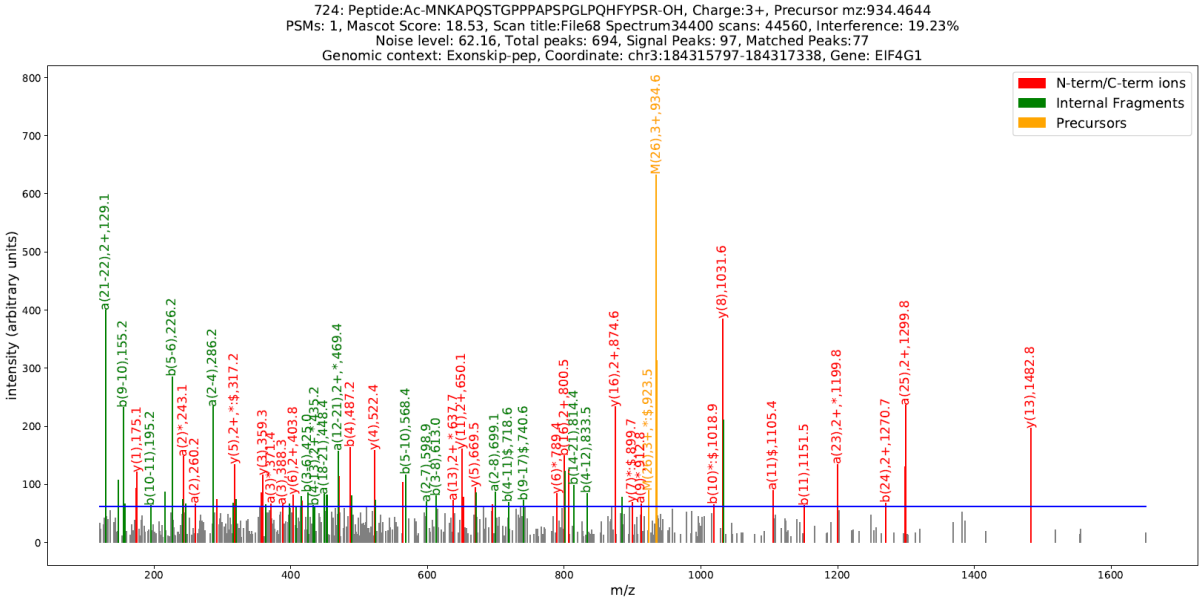

Figure 35. Annotated MS/MS spectra of the endogenous peptide of the exon-skipped variant of EIF4F1 gene. Peptide "MNKAPQSTGPPPAPSPGLPQHFYPSR.A", charge 3+. The noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments are shown in green and un-fragmented precursors in orange. Ammonia loss is shown with a * sign and water loss is shown with a $ sign.

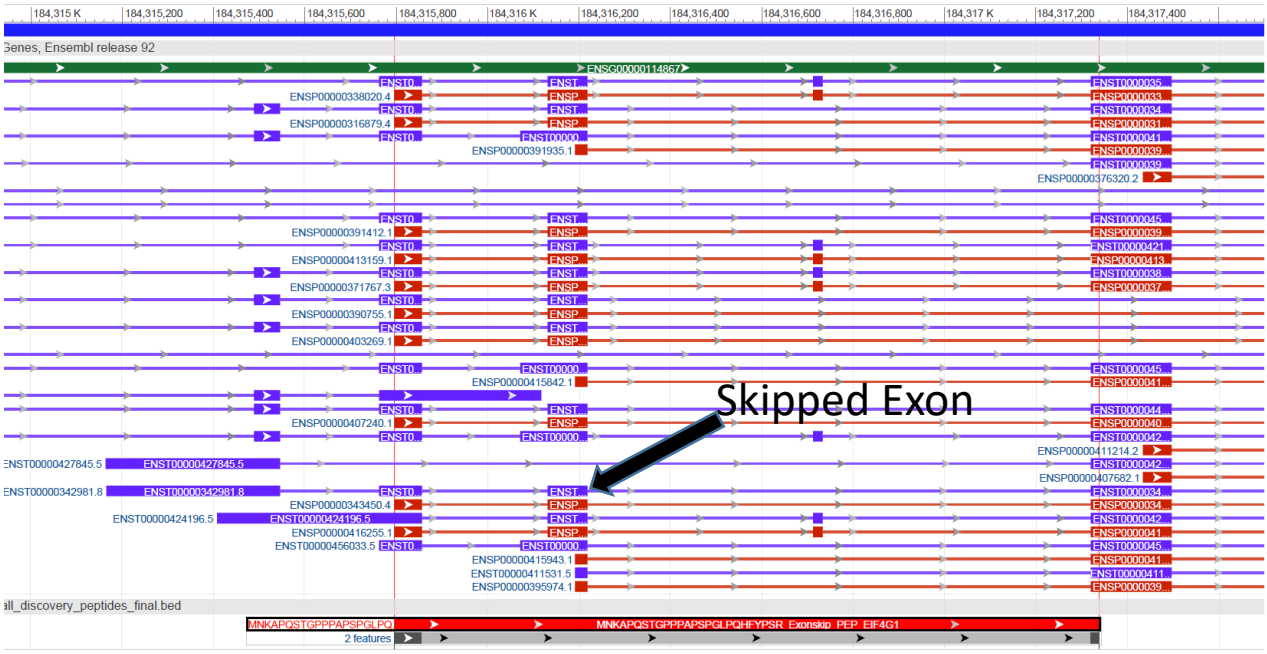

Figure 36. Visualization of exon skipped peptide "MNKAPQSTGPPPAPSPGLPQHFYPSR.A" on the EIF4G1 gene, Ensembl id: ENSG00000114867. The skipped exon, exon 3 of transcript ENST00000342981, is indicated. Green bars denote reference genes. Purple bars denote reference transcripts, and red bars reference proteins. Colored blocks represent exons, and lines introns. The genomic mapping of the exon skipped peptide sequence is shown in the bottom track as dark grey blocks. It can be seen that the peptide sequence span only the flanking exons; the peptide did not contain amino acids from the skipped exon.

3.9. Peptides from non-coding regions of protein coding transcripts

We identified 87 and 61 peptides that mapped onto the 5'-UTR, and 3'-UTR region of protein coding transcripts, respectively (Table 2). For example, the two peptides "R.GGAAPAGGGAEAGPGGGPGGAGGAAAK.A and "***K.AGGAAD*MTDNIPLQPVR.Q***" were identified from the 5'-UTR region of ATPase phospholipid transporting 9A (ATP9A), Figure 37. The amino acids GGAAPAGGGAEAGPGGGPGGAGGAAAK***AGGAAD*** originate from the 5'-UTR region (indicated in grey here for clarity) and the amino acids ***MTDNIPLQPVR*** originate from the main ORF. The peptide

"*K.AGGAAD**MTDNIPLQPVR.Q***" spanned the main ORF and the 5'-UTR region, and was validated at tier 1 and tier 2 (Figure 38 and Figure 39).

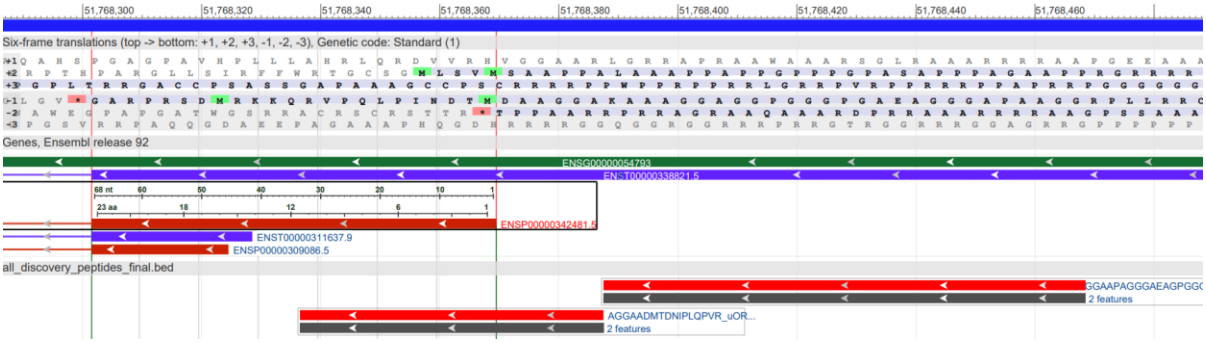

Figure 37. The two peptides "R.GGAAPAGGGAEAGPGGGPGGAGGAAAK.A" and "*K.AGGAAD**MTDNIPLQPVR.Q***" map onto the 5'-UTR region of ATP9A mRNA ENST00000338821. The first peptide maps fully on to the 5'-UTR region of the transcript, whereas the second peptide maps onto the 5'-UTR region and the CDS coding for the first 11 amino acids of ATP9A protein ENSP00000343481. Green bars denote reference genes. Purple bars denote reference transcripts, and red bars reference proteins. Colored blocks represent exons, and lines introns. The genomic co-ordinates of the peptide sequences are shown in the bottom track.

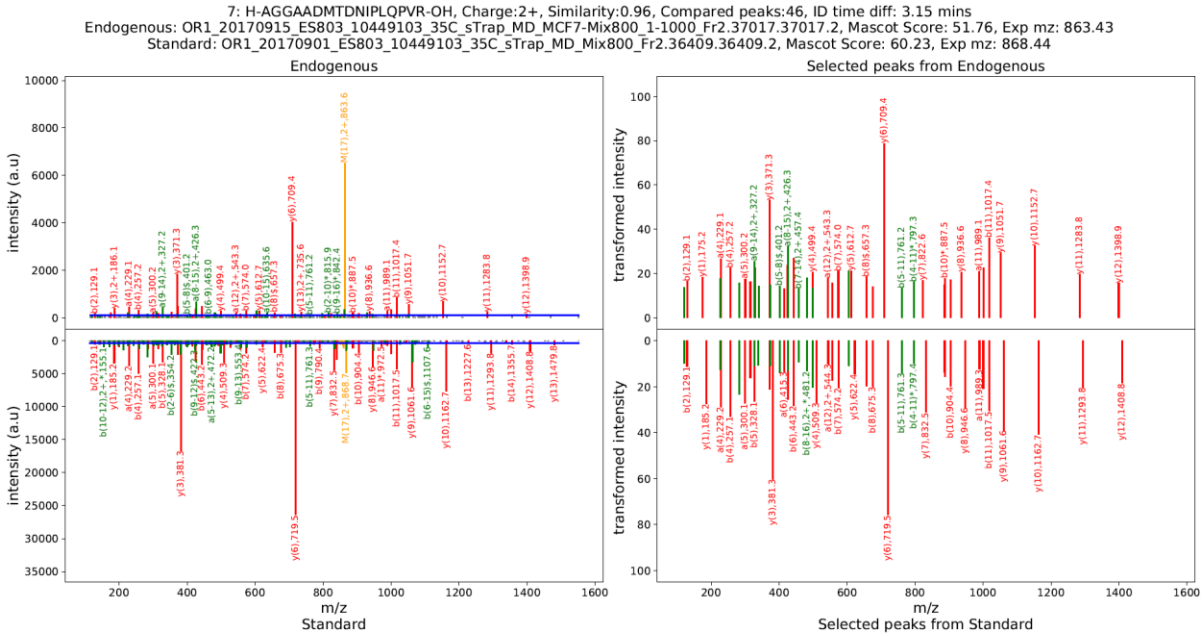

Figure 38. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) that spans the main ORF and the 5'-UTR region of ATP9A. Peptide "***K.AGGAAD*MTDNIPLQPVR.Q***", charge 2+. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type, b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized and normalized to sum 1000 before the similarity calculation. The intensities of 46 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.96.

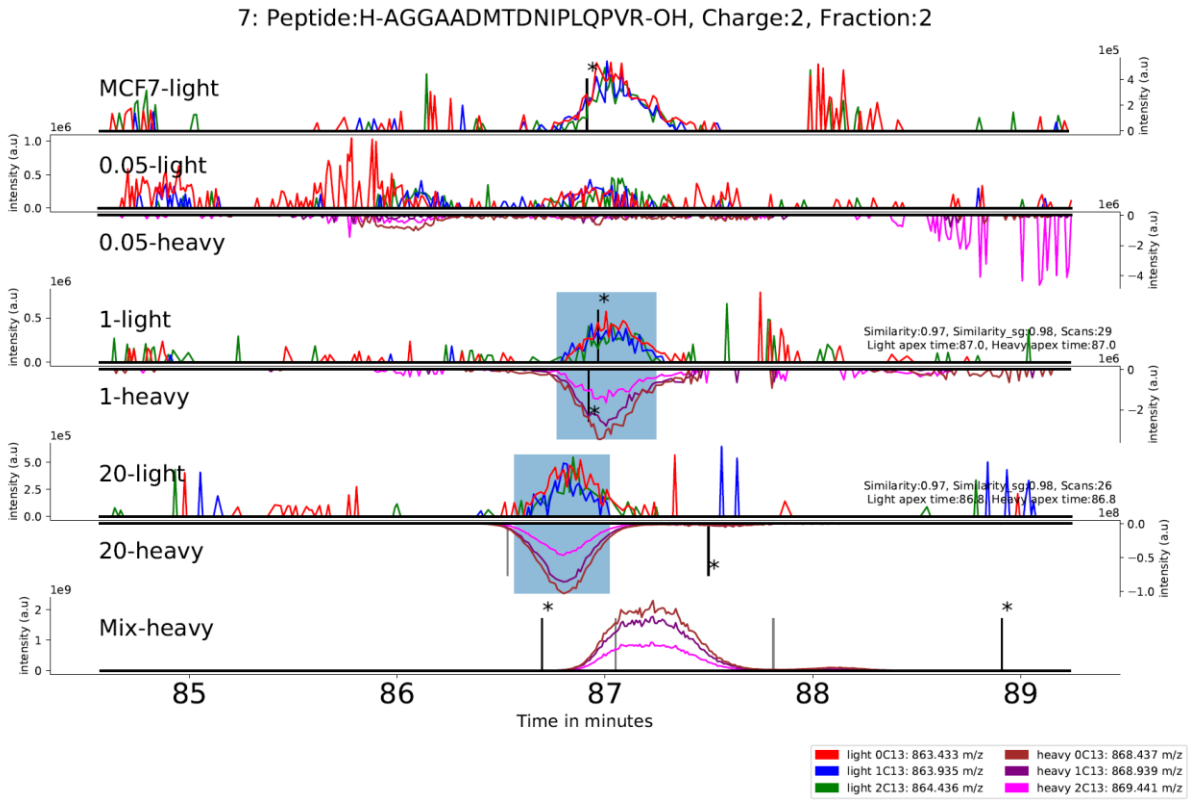

Figure 39. Extract ion chromatogram (EIC) of the endogenous and SIS peptide that spans the main ORF and the 5'-UTR region of ATP9A in 5 validation runs. Peptide "***K.AGGAAD*MTDNIPLQPVR.Q***", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.98. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

3.10. Peptides from alternate frame of protein coding transcripts

We identified 78 peptides that mapped onto the CDS of protein coding transcripts but in a non-canonical frame (Table 2). The gene keratin 8 (KRT8) is situated on the reverse strand of chromosome 12. A peptide "***R.GLQLLQPHQLLQGR.G***" was unambiguously mapped onto KRT8 gene and also validated at tiers 1 and tier 2 (Figure 40 and Figure 41). The reference protein is coded in the +3 frame of the protein-coding transcript "ENST00000552150" of KRT8 gene whereas this peptide exclusively mapped to the +2 frame (Figure 42).

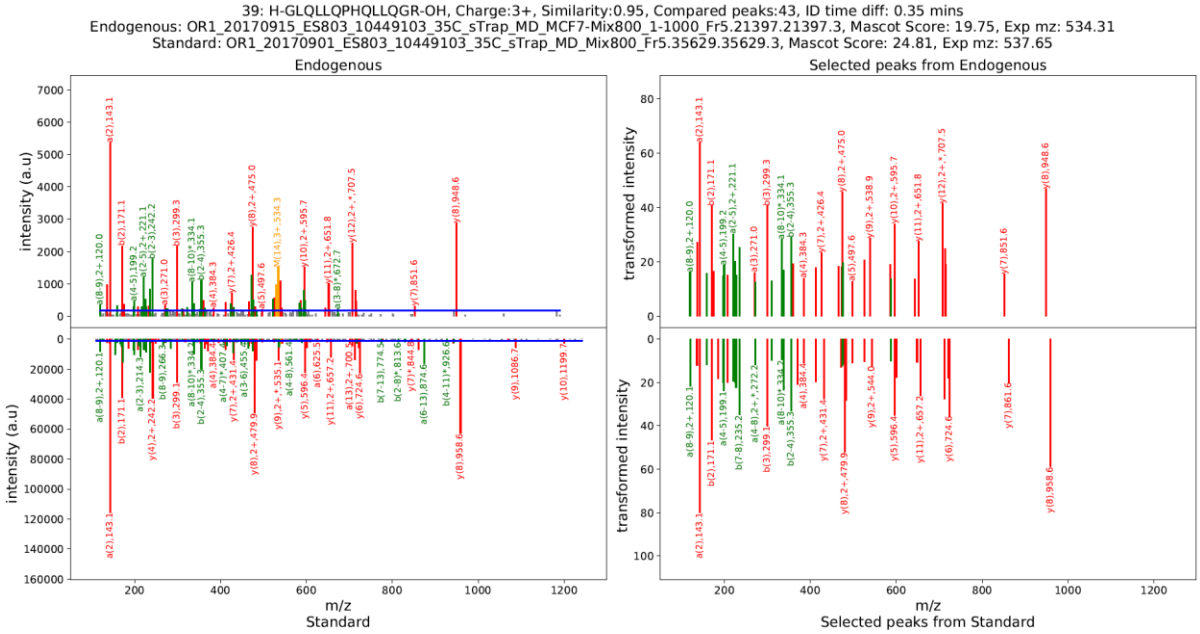

Figure 40. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from the alternate frame translation of KRT8 mRNA. Peptide "***R.GLQLLQPHQLLQGR.G***", charge 3+. Noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type, b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 43 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.95.

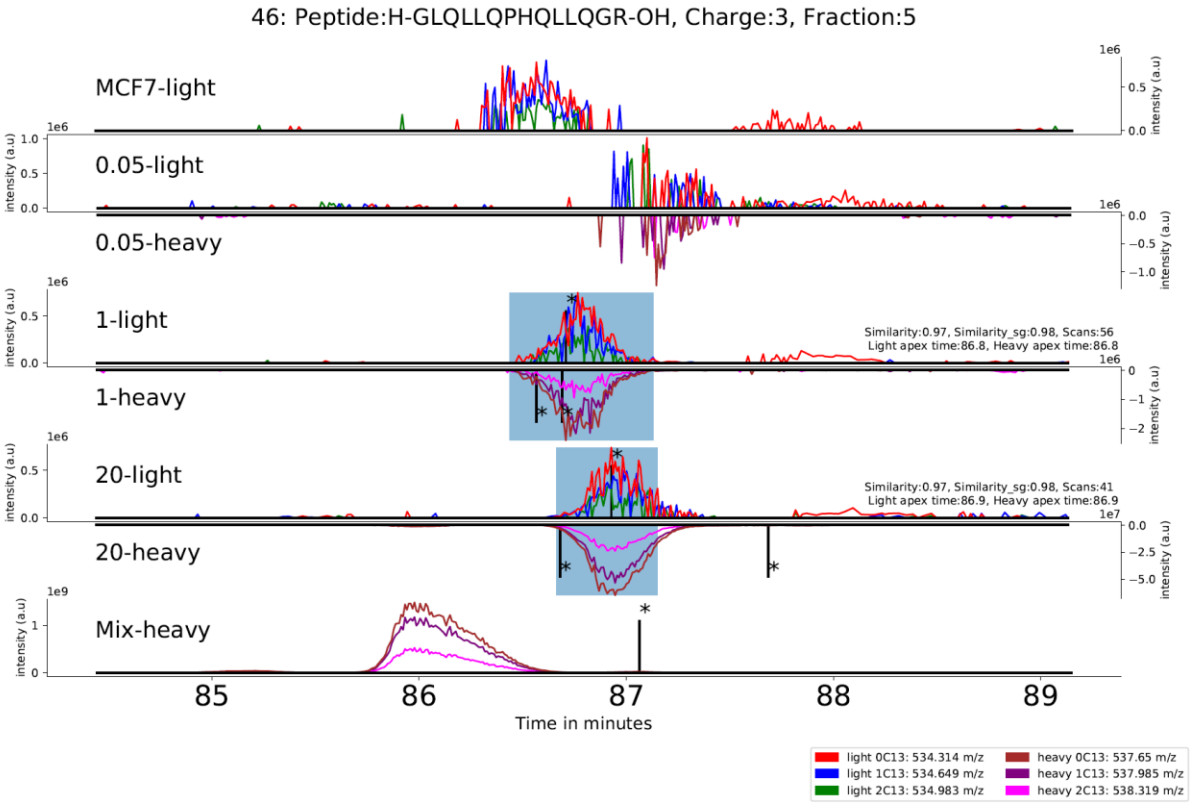

Figure 41. Extracted ion chromatogram (EIC) of the endogenous and the SIS peptide from the alternate frame translation of KRT8 mRNA in five validation runs. Peptide "***R.GLQLLQPHQLLQGR.G***", charge 3+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.97. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9.The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

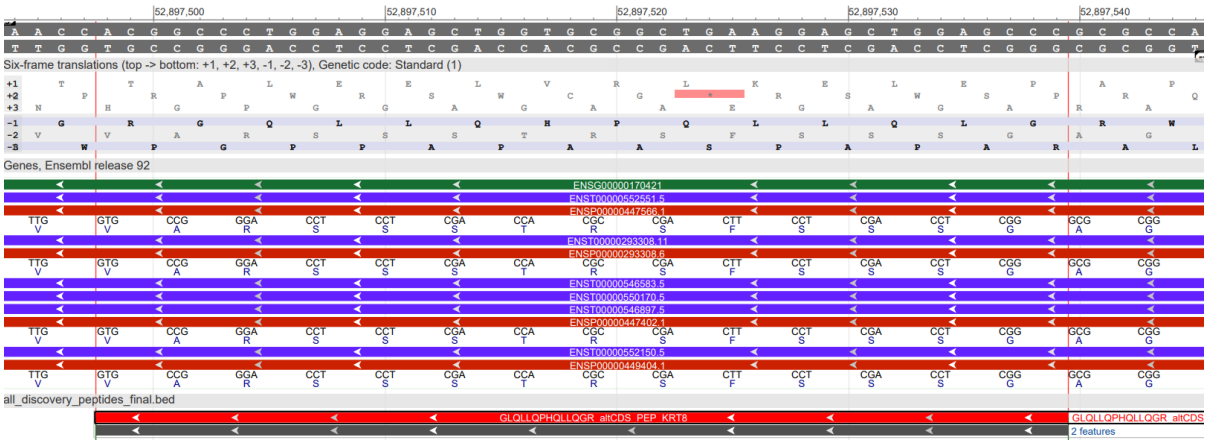

Figure 42. Visualization of the peptide "***R.GLQLLQPHQLLQGR.G*** on the KRT8 gene ENSG00000170421. The peptide maps onto the -1 frame of chromosome 12 whereas all reference KRT8 proteins are coded in the -2 frame (shown in six-frame translation track). Green bars denote reference genes. Purple bars denote reference transcripts, and red bars reference proteins. The locations of the peptide sequence identified by the proteogenomic search are shown in the bottom track (dark grey block).

3.11. Peptides from non-coding transcripts and genes

A total of 172 peptides from 187 spectra were identified from non-coding transcripts and genes (Table 2). This included transcripts with biotypes: retained intron, lncRNA, anti-sense, sense-intronic, rRNA, TEC, and pseudogene transcripts. For example, GATA3 antisense RNA 1 (GATA3-AS1) is a long non-coding RNA gene situated on the reverse strand of chromosome 10. Two peptides "***R.GAEVPGEAAPGAR.A***" and "R.ARQPALPGELR.G" were identified from transcript "ENST00000355358" of GATA3-AS1 (Figure 43), the first of which satisfied tiers 1 and 2 validation (Figure 44 and Figure 45).

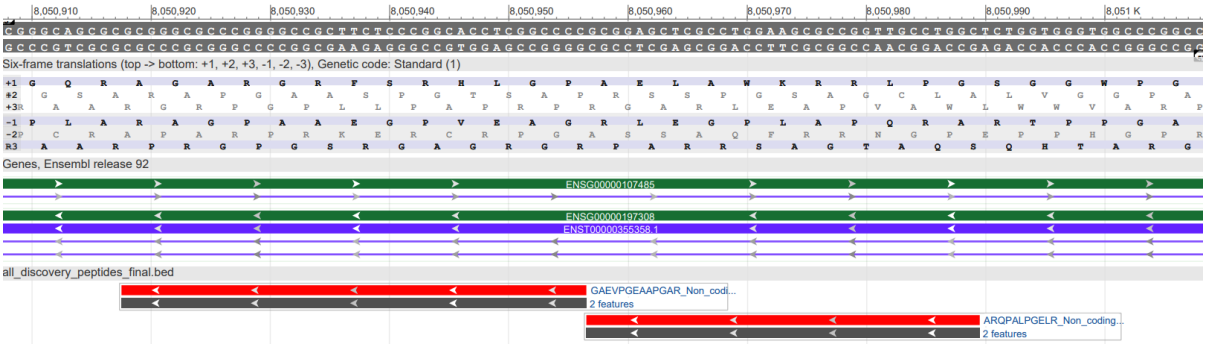

Figure 43. Visualization of the peptides "***R.GAEVPGEAAPGAR.A***" and "R.ARQPALPGELR.G" on GATA3-AS1 long non-coding RNA. The peptides map onto the exon of GATA3-AS1 non-coding transcript ENST00000355358. Green bars denote reference genes. Purple bars denote reference transcripts. Colored blocks represent exons, and lines introns. The locations of the peptide sequences identified by the proteogenomic search are shown in the bottom track.

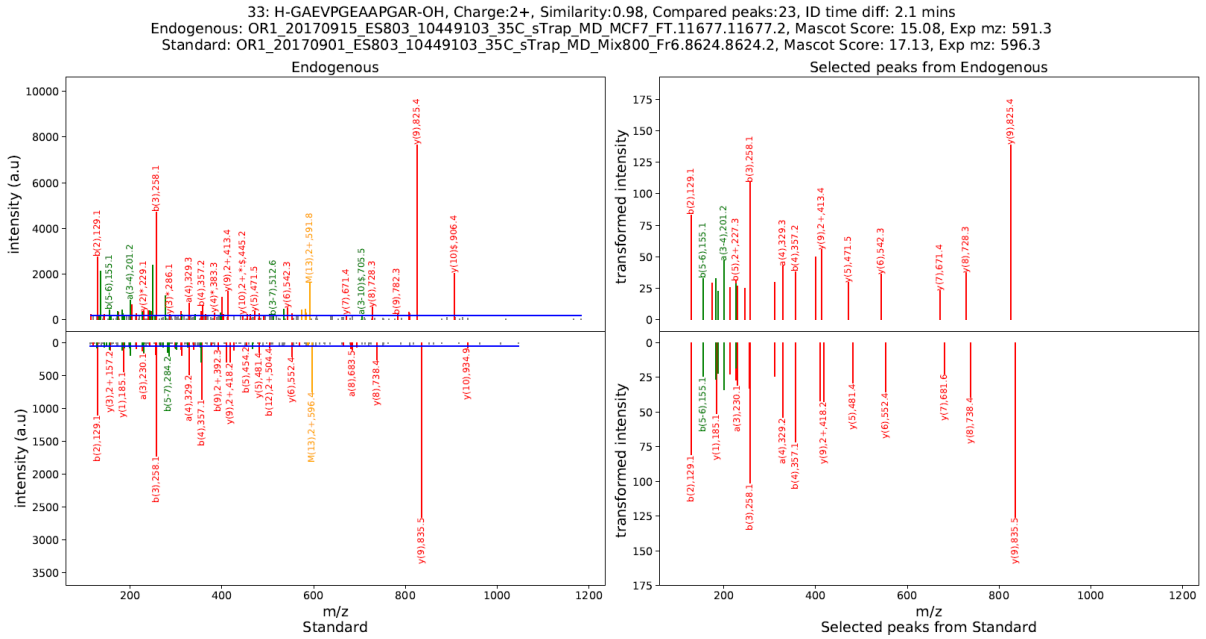

Figure 44. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from long non-coding RNA gene GATA3-AS1. Peptide "***R.GAEVPGEAAPGAR.A***", charge 2+. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type*, *b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 23 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.98.

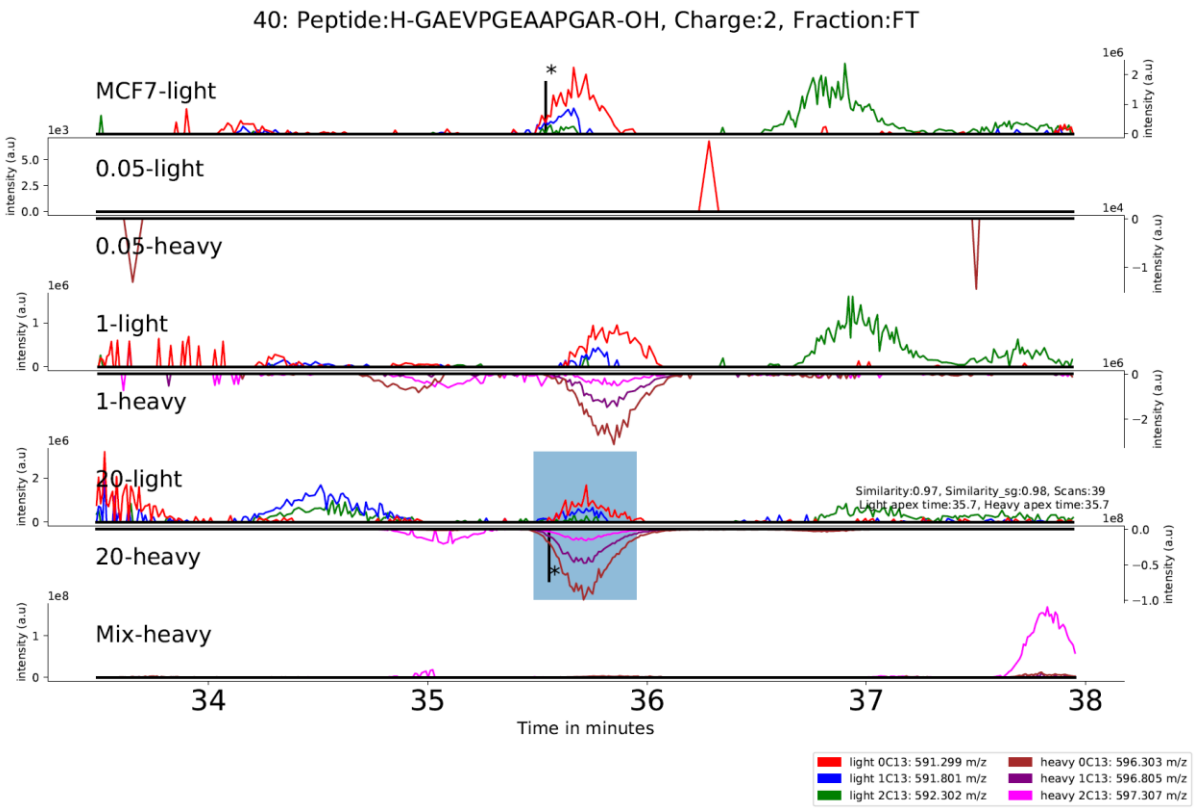

Figure 45. Extracted ion chromatogram (EIC) of the endogenous and SIS peptide from long non-coding RNA gene GATA3-AS1 in five validation runs. Peptide "***R.GAEVPGEAAPGAR.A***", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.97. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

The gene heterogeneous nuclear ribonucleoprotein L (HNRNPL) is situated on the reverse strand of chromosome 19. The missed cleavage peptide "***R.QRQPPLLGDHPAEY****GEGR.G*" was identified with 5 PSMs from the retained intron transcript "ENST00000597731". The peptide spans the boundary of exon 7 (amino acids shown in black) and intron 7-8 (shown in grey) of the protein-coding transcript "ENST00000221419" (Figure 46). Although the peptide was identified from the retained intron transcripts of HNRNPL, it could also be produced due to aberrant splicing of the protein-coding transcripts. For example, the splicing of pre-mRNA transcript "ENST00000221419" that retains intron 7-8 would lead to a larger protein containing the 37 extra amino acids originating from intron 7-8. The peptide was subsequently validated at tier 1 and tier 2 (Figure 47 and Figure 48).

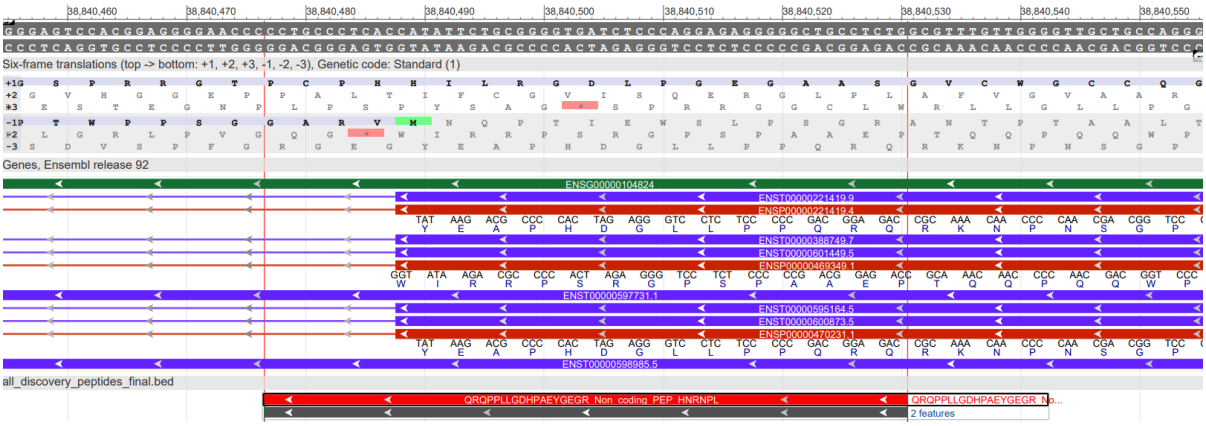

Figure 46. Visualization of peptide "***R.QRQPPLLGDHPAEY****GEGR.G*" on the gene HNRNPL, Ensembl id: ENSG00000104824. The peptide maps onto the exons of retained-intron-transcripts ENST00000597731 and ENST00000598985. The peptide also spans the boundary of exon and intron of protein coding transcripts ENST00000221419, ENST00000601449 and ENST00000600873. Green bars denote reference genes. Purple bars denote reference transcripts, and red bars reference proteins. Colored blocks represent exons, and lines introns. The locations of the peptide sequences identified by the proteogenomic search are shown in the bottom track.

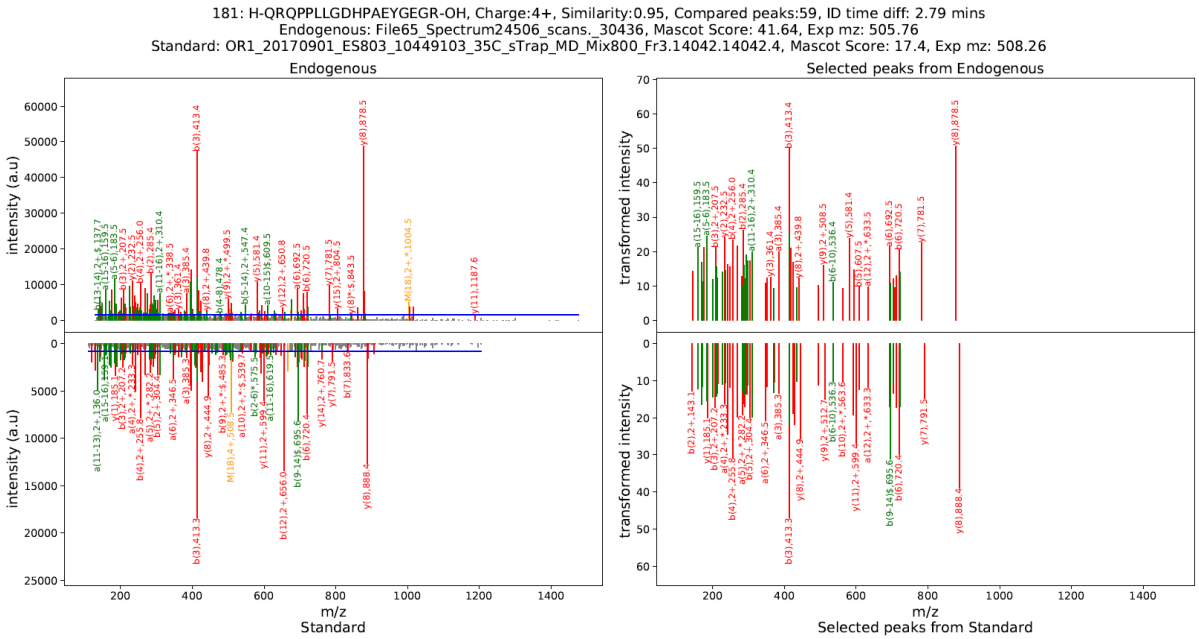

Figure 47. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from the retained-intron transcript of HNRNPL. Peptide "***R.QRQPPLLGDHPAEY****GEGR.G*", charge 4+. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type, b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 59 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.95.

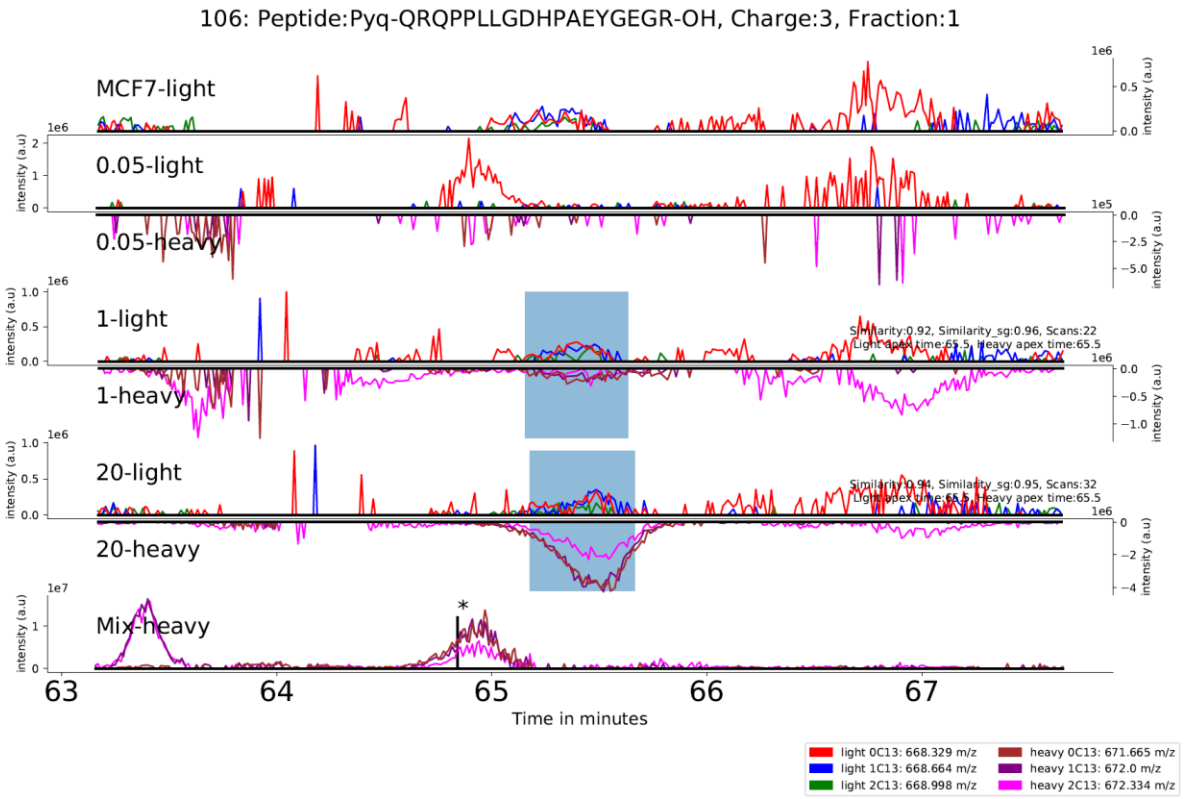

Figure 48. Extracted ion chromatograms (EIC) of the endogenous and SIS peptide from the retained-intron transcript of HNRNPL gene in five validation runs. Peptide "***R.QRQPPLLGDHPAEYGEGR.G***", charge 3+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity score 0.94. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

3.12. Peptides from introns and exon boundaries of protein coding genes

A total of 98 peptides were identified from intronic regions and 30 peptides were identified from the exon-intron boundaries of protein coding genes (Table 2). The peptide "***R.ASAAEGVGEPGASAGR.A***" was identified with 2 PSMs from the gene WD repeat domain 26 (WDR26) in a search of ORFs generated from gene sequences (Figure 49). The peptide mapped on to the 5'-UTR region (amino acids "***EGVGEPGASAGR.A***" shown in black above) and the genomic region upstream of 5'-UTR (amino acids "***R.ASAA***" shown in grey above) of the protein coding transcript "ENST00000414423". This peptide, validated at tier 1 and tier 2, indicates that the first exon of transcript "ENST00000414423" is extended upstream and contains a CDS (Figure 50 and Figure 51).

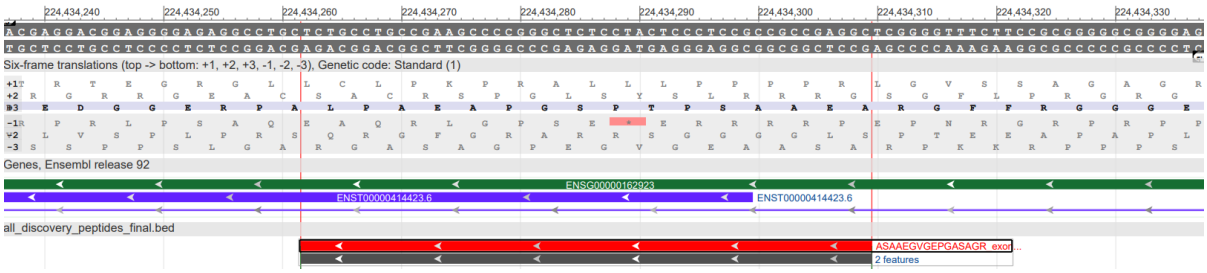

Figure 49. Peptide originating from upstream of exon 1 of WDR26. Visualization of peptide "***R.ASAAEGVGEPGASAGR.A***" on gene WDR26 (ENSG00000162923). The peptide spans the boundary of exon 1 of transcript ENST00000414423 and the genomic region upstream of exon 1. It also maps onto the intron of protein coding transcript ENST00000445239 (purple line) of WDR26. Green bars denote reference genes. Purple bars denote reference transcripts. The locations of the peptide sequences identified by the proteogenomic search are shown in the bottom row (dark grey block extending upstream of the transcript).

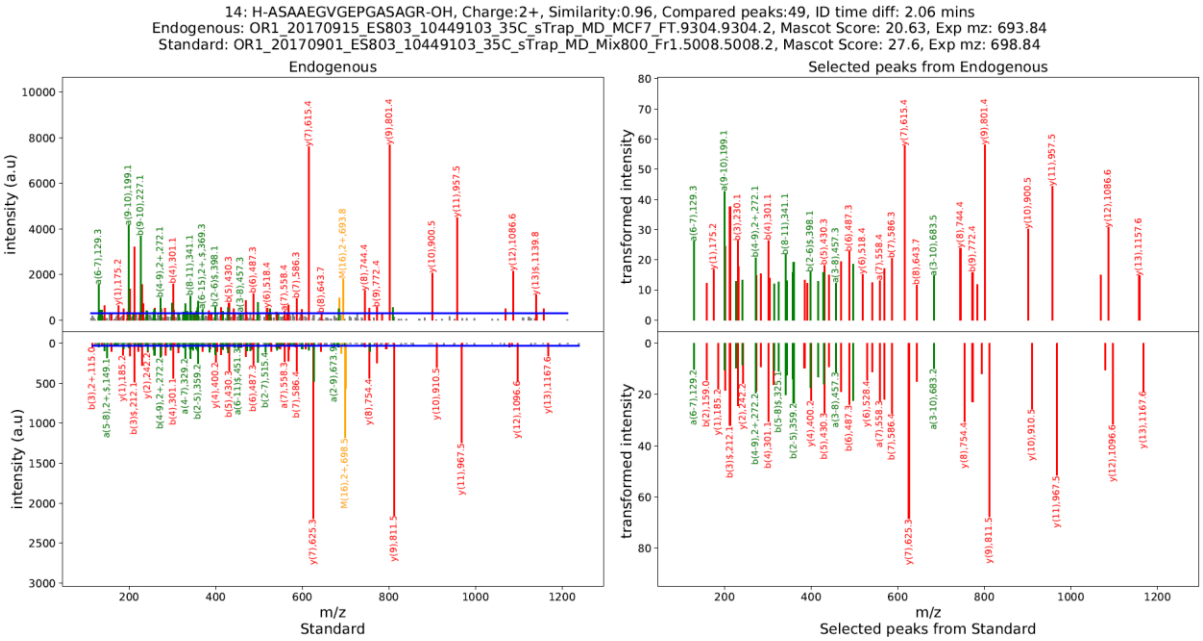

Figure 50. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from the boundary of exon 1 and the genomic region upstream of exon 1 of WDR26. Peptide "*R.ASAA**EGVGEPGASAGR.A***", charge 2+. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type*, *b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 49 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.96.

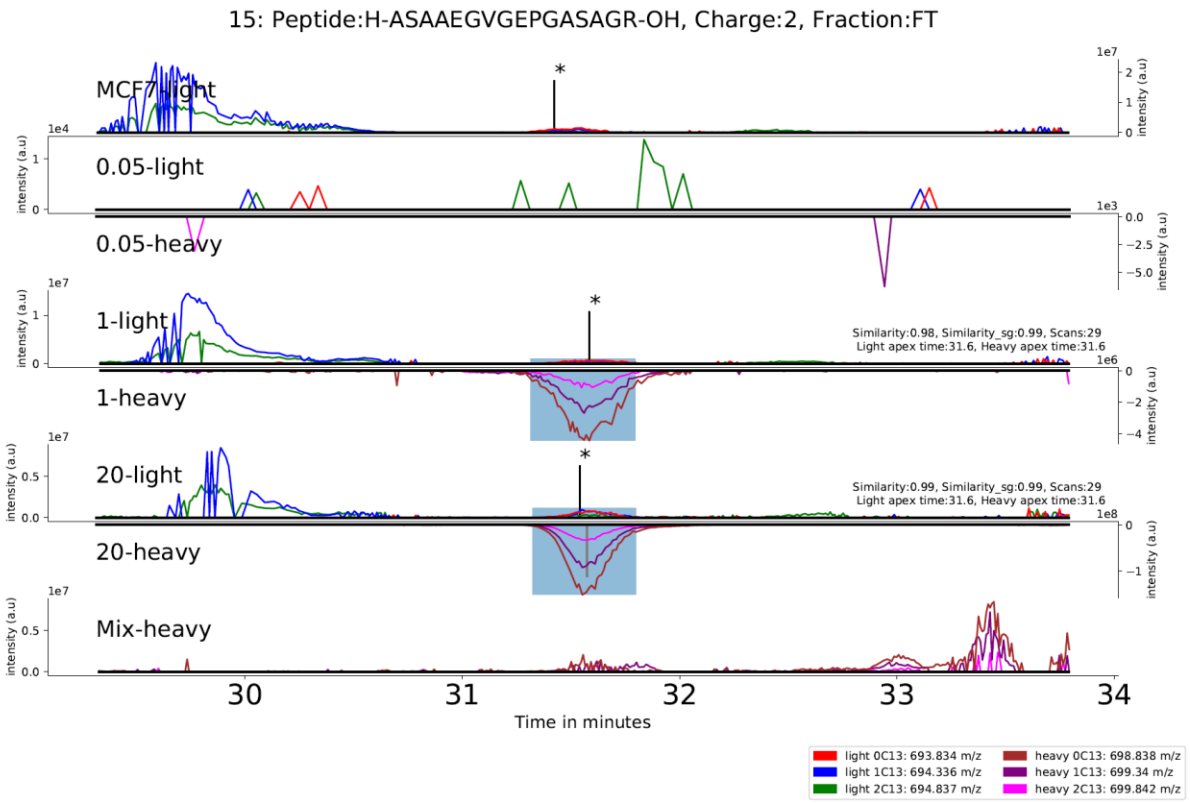

Figure 51. Extracted ion chromatograms of the endogenous and the SIS peptide from five validation runs. Peptide "*R.ASAA**EGVGEPGASAGR.A***", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.99. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest).

3.13. Peptides from novel isoforms

A total of 45 peptides were identified that were classified as novel isoforms of known protein coding genes (Table 2). 6 peptides from 46 PSMs were obtained from a GNOMON predicted model of STARD10 gene (online appendix 5). A blast search with the NCBI non-redundant protein sequences demonstrated that amino acids (78-368) of the protein model showed 100% similarity to reference human STARD10. Thus, the protein could represent an N-terminally extended (1-77) isoform of STARD10. All 6 non-canonical peptides of STARD10 identified by the proteogenomic analysis were mapped to this extended N-terminal region, Figure 52. Four of the peptides passed tier 1 and tier 2 validation, "***R.SAGAGSYGALANSAWGGPR.K***", "***R.SAGAGSYGALANSAWGGPRK.V***", "***R.KVASASAAASTLSEPPR.R***", and "***K.VASASAAASTLSEPPR.R***". Another passed tier 1 *"R.KVASASAAASTLSEPPRR.T*". The final peptide was not included in the peptides selected for validation, "MEEELALGPR.G". The identification and validation of 5 of these peptides indicates that a novel isoform of STARD10 is expressed by MCF7 cells. The tier 1 validation and tier 2 validation of one of the peptides is shown in Figure 53 and Figure 54.

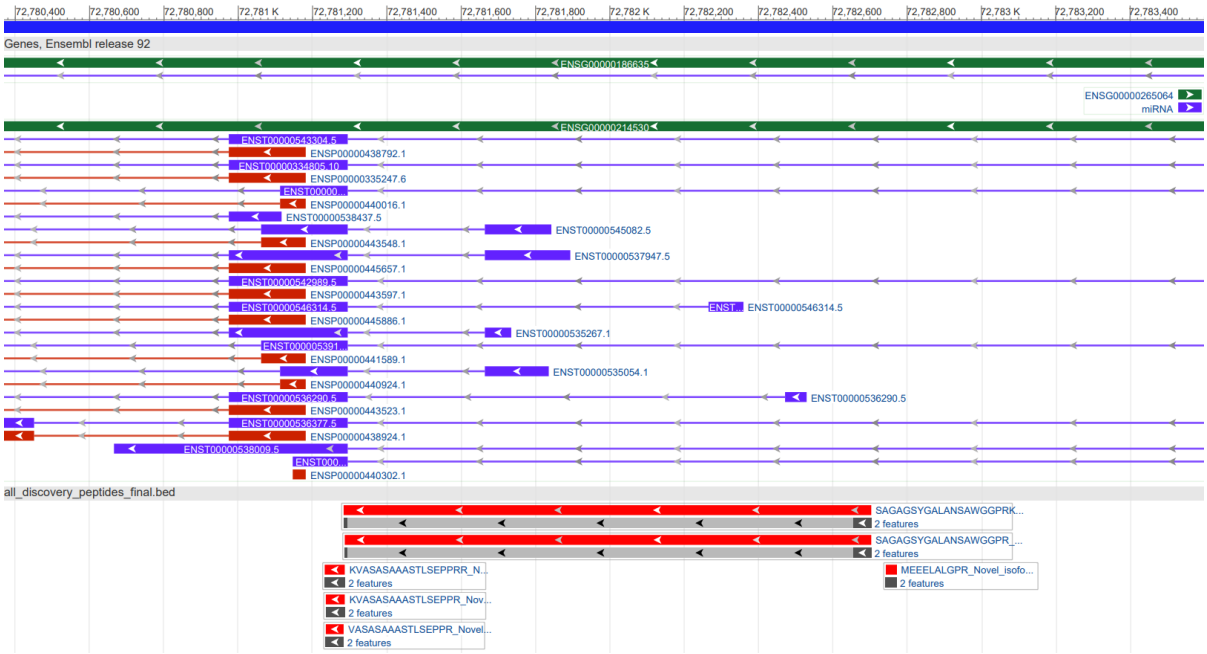

Figure 52. Mapping of the non-canonical peptides identified by the proteogenomic searches demonstrate the presence of an N-terminal extended form of STARD10. The peptides "***R.KVASASAAASTLSEPPR.R***", *"R.KVASASAAASTLSEPPRR.T*" and "***K.VASASAAASTLSEPPR.R***" map onto the 5'-UTR region of protein coding transcripts of STARD10; the other three peptides "***R.SAGAGSYGALANSAWGGPR.K***", "***R.SAGAGSYGALANSAWGGPRK.V***", and "MEEELALGPR.G" do not map onto any reference transcripts. Green bars denote reference genes. Purple bars denote reference transcripts, and red bars proteins. Colored blocks represent exons, and lines introns. The locations of the peptide sequences identified by the proteogenomic search are shown in the bottom track (dark grey blocks (amino acids); light grey blocks (intron)).

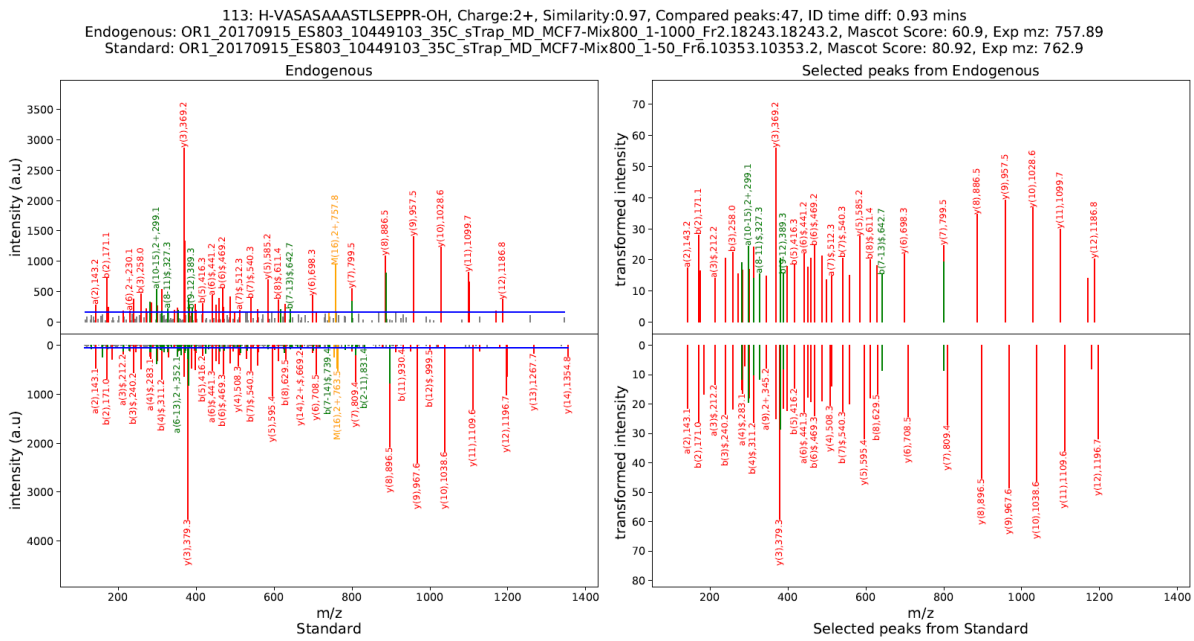

Figure 53. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from the N-terminally extended isoform of STARD10. Peptide "***K.VASASAAASTLSEPPR.R***", charge 2+. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a-type*, *b-type*) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 47 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.97.

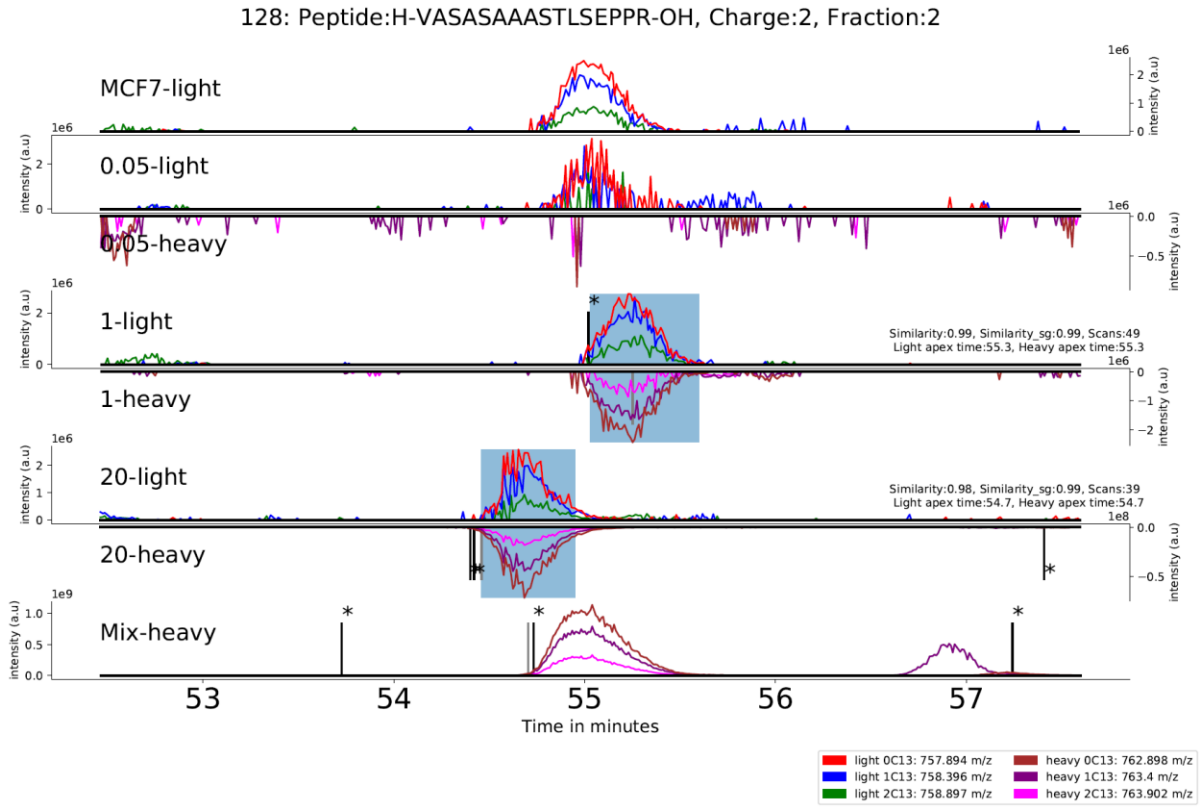

Figure 54. Extracted ion chromatograms (EIC) of the endogenous and SIS peptide from the N-terminal extended isoform of STARD10 in five validation runs. Peptide "***K.VASASAAASTLSEPPR.R***", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.99. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest).

The proteogenomics analysis provided evidence for the expression of a novel exon in Talin-1 (TLN1). We detected 2 peptides from the GNOMON predicted protein model of TLN1 (Figure 55). The peptides span exon 17, a novel exon, and exon 18 of transcript "ENST00000314888". The first peptide, "***R.SPPDSPT****DALMQLAK.A*", spanned the novel exon (black) and exon 18 (grey), and was validated at tier 1 and tier 2 (Figure 56 and Figure 57). The other peptide, "R.QNLLQAAGNVGQASGELLQQIGESDTDPHFQICASR.G", spanned exon 17 (grey) and the novel exon (black) but was too long to synthesize the isotopically labeled standard peptide needed for validation.

A blast search with the NCBI non-redundant protein sequences demonstrated that this protein model has 100% sequence similarity to protein "AQN67632.1", an isoform of TLN1. This isoform is not included in the reference protein datasets of RefSeq or GENCODE. The RNA-seq data from NCBI includes a putative exonic region (black arrow in Figure 55) between exon 17 and exon 18 of transcript "ENST00000314888". Our RNA-seq data contains a significant number of reads in this region, aligned by both TopHat and BWA, Figure 58. The validated proteogenomics peptides and the RNA-seq data confirm the translation of the GNOMON predicted isoform of TLN1 in MCF7 cells.

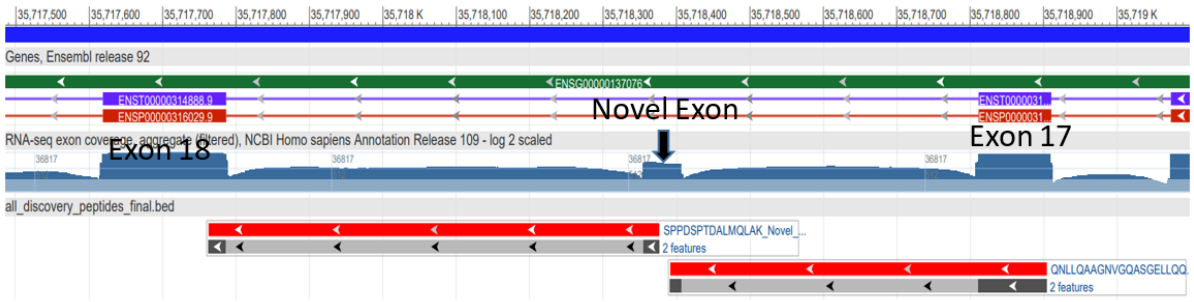

Figure 55. Visualization of the non-canonical peptides spanning a novel exon in the TLN1 gene ENSG00000137076. The peptide "R.**SPPDSPT**DALMQLAK.A" spans the novel exon and exon 18; the peptide "R.QNLLQAAGNVGQASGELLQQIGESDTDPHFQ**ICASR**.G" spans exon 17 and the novel exon. RNA-seq exon coverage from the NCBI includes a putative exonic region (black arrow) between exons 17 and 18. Expression of the novel exon in MCF7 cells would cause the insertion of 17 extra amino acids "**ICASR**GAGVR**SPPDSPT**" at position 666 in reference protein ENSP00000316029. The tryptic parts of this insertion are included in the peptides identified by the proteogenomics search (indicated in bold text). Green bars denote reference genes. Purple bars denote reference transcripts, and red bars proteins. Colored blocks represent exons, and lines introns. The locations of the peptide sequences identified by the proteogenomic search are shown in the bottom track (dark grey blocks overlapping with exons 17, 18 and the novel exon).

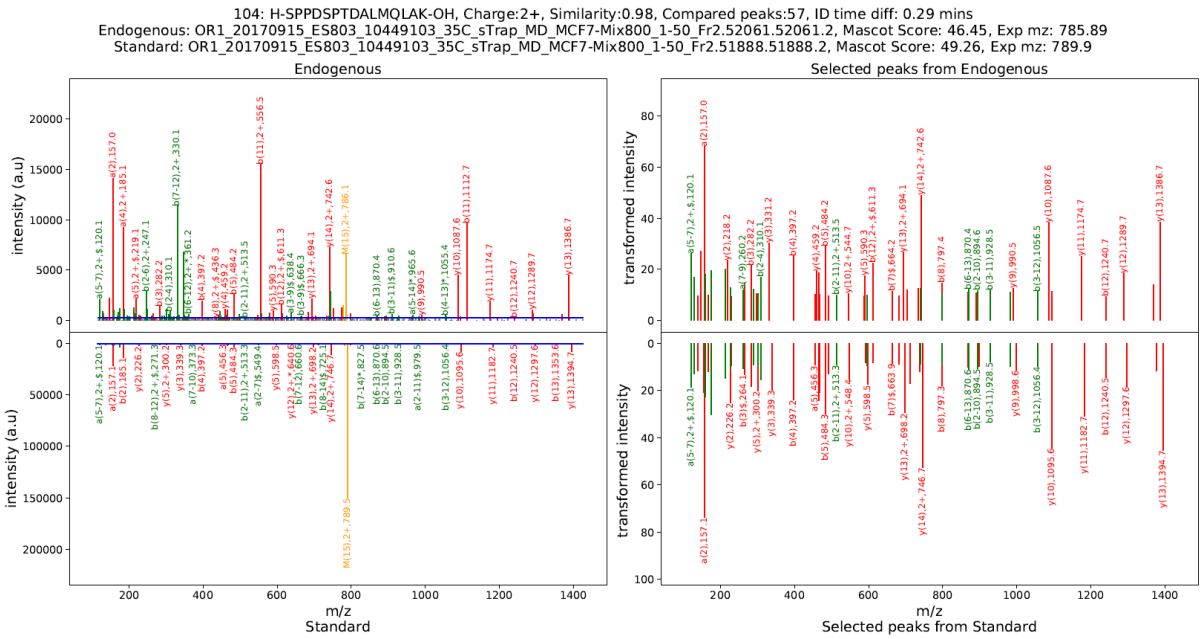

Figure 56. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from the novel isoform of TLN1. Peptide "***R.SPPDSPT****DALMQLAK.A*", charge 2+. The N-terminal (a/b) and C-terminal (y) fragment ions are shown in red, internal fragments (*a*-type, *b*-type) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 57 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.98.

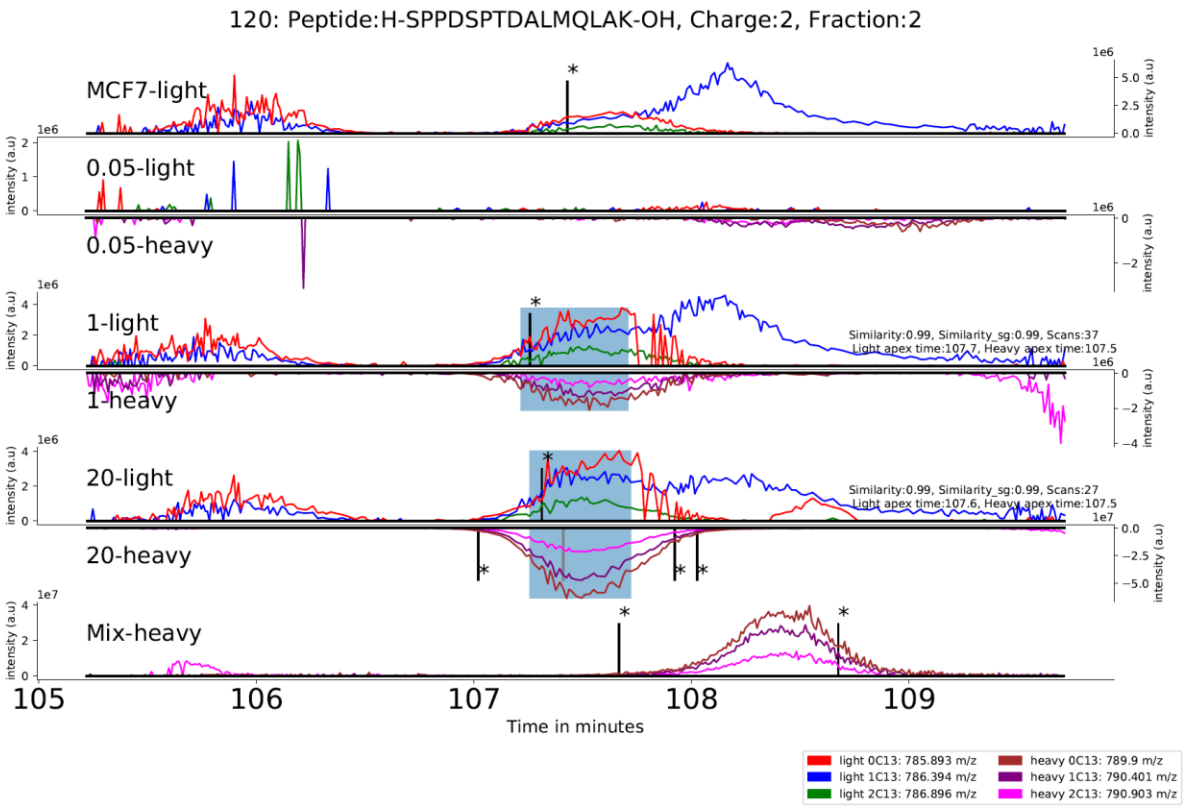

Figure 57. Extracted ion chromatograms (EIC) of endogenous and SIS peptides from a novel isoform of TLN1 in five validation runs. Peptide "***R.SPPDSPT****DALMQLAK.A*", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.99. Apex elution times for the SIS and endogenous peptides were near identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

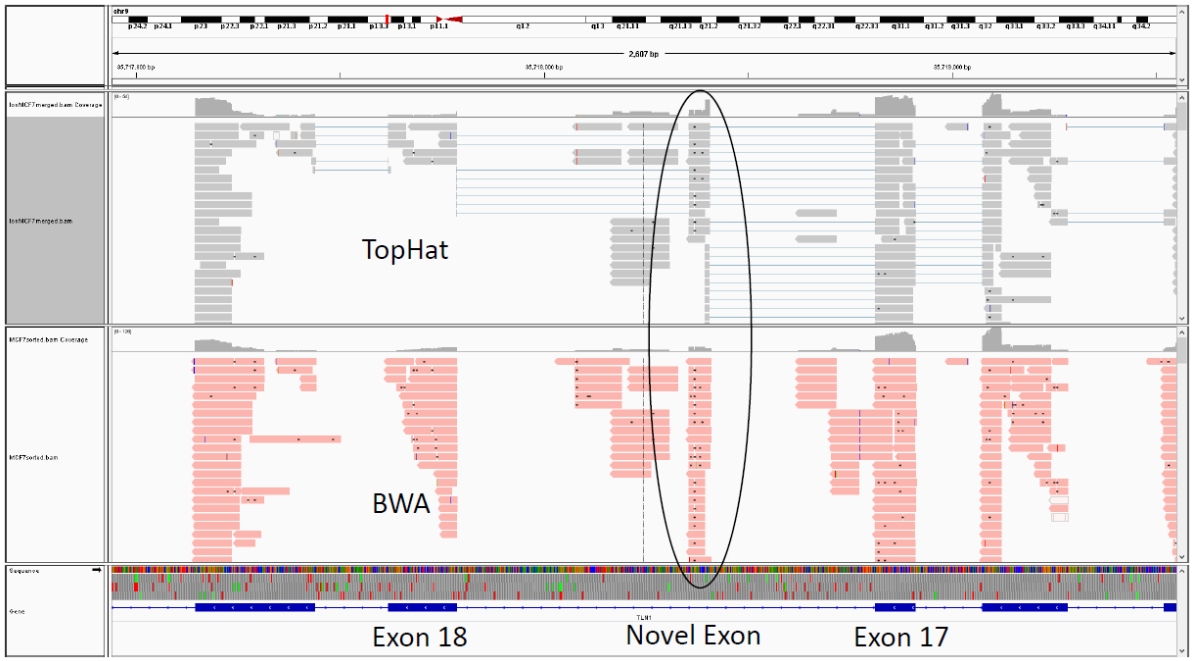

Figure 58. Visualization of the RNA-seq data of MCF7 cells analyzed by TopHat and BWA. A significant number of reads (indicated by an oval), aligned by both TopHat and BWA, were recorded at the novel exon between exon 17 and exon 18 of TLN1.

4. Discussion

We utilized customized protein sequence databases for the identification of non-canonical peptides in MCF7 cells. These were:-

i) SNVs and Insertion/Deletion mutations identified in Exome-Seq, RNA-seq and publicly available through COSMIC were utilized for the identification of peptides from variant proteins.

ii) Transcript sequences from GENCODE and RefSeq were utilized for discovering peptides from uORFs, dORFs, and alternate-reading-frame encoded proteins in protein coding genes.

iii) Transcripts with biotypes “lncRNA” and “retained intron” from GENCODE were searched in three frame for the identification of peptides from non-coding transcripts.

iv) Protein coding gene sequences and CDS sequences including flanking sequences of up to 100 base pairs were translated in three frames in order to identify peptides spanning exon-intron boundaries and introns.

v) GNOMON predicted proteins were utilized for the identification of novel isoforms and variants that were not identified in Exome-seq and RNA-seq experiments.

vi) A six-frame translation search of the transcriptome (GENCODE and RefSeq) and full human genome was also performed in order to identify peptides from novel coding DNA sequences.

A reference proteome database was constructed by merging reference protein sequences from GENCODE, RefSeq and UniProt. We first performed extensive searches with the reference proteome database to identify all peptides that could be assigned to reference proteins. Thus, the results concern peptides that could not be assigned to a reference proteome.

The analysis led to the identification of 1227 non-canonical peptides (Table 2 and online appendix 2) after an exhaustive quality control analysis of the MS/MS spectra and search engine identification results to ensure all non-canonical peptides originated from high quality spectra, and which could not be explained by known modifications. Specifically,

i) Non-canonical peptides were removed if they differed from reference peptides by only Leucine/Isoleucine or deamidation;

ii) Non-canonical peptides were removed if they were only detected in a modified form;

iii) Non-canonical peptides from ORFs, GNOMON, transcriptome and genome searches were accepted only if they possessed sufficient sequence diversity (at least 3 amino acid difference) from any peptide within the reference proteome database.

iv) Non-canonical peptides were removed if their precursor ion isolation contained greater than 70% interference.

v) Non-canonical peptides were removed if their MS/MS spectrum was characterized by a low number of peaks or low signal-to-noise.

To validate the results and confirm the presence of the non-canonical peptides stable isotope-labeled standard (SIS) peptides were synthesized and used as reference standards. We then followed a two-tier validation strategy. In tier 1 we quantified the similarity of the MS/MS spectrum of the endogenous peptide with that of its isotopically labeled reference standard (using only those fragment ions that could be assigned). Endogenous peptides with cosine similarity greater that 0.9 were considered validated at tier 1. For tier 2 validation we examined the extracted ion chromatogram of the tier 1 peptides. Peptides that eluted at identical elution times, and that had elution *and* isotopic profile similarities greater than 0.9 were considered validated at tier 2.

4.1. Choice of reference proteomes can impact which proteoforms are identified

The LC-MS/MS data was searched with all three common reference proteomes from GENCODE, UniProt and RefSeq. It was found that utilizing different reference proteome databases could impact which *canonical* protein isoforms were identified. For example, two peptides "GYATDESTVSSVQGSR", and "EKGYATDESTVSSVQGSR" were identified from a predicted splice isoform (XP_005247870) of FMR1 autosomal homolog 1 (FXR1) in the RefSeq database. This predicted isoform is not present in the UniProt or GENCODE reference protein databases. The peptides spans the boundary of intron 12-13 (grey text above) and exon 13 (black text above) of the pre-mRNA transcript NM_005087.3 (Figure 59). The retention of intron 12-13 in the protein-coding transcript would cause the insertion of 28 amino

acids “MGFRPSSTRGPEKEKGYATDESTVSSVQ” at position 379 in the reference FXR1 protein NP_005078.2. The splice isoform is predicted to be produced by RefSeq annotation but not by GENCODE. The RNA-seq data of MCF7 cells contains alignments at intron 12-13, and the proteogenomic analysis reported here identified the associated proteolytic peptide, indicating that this predicted splice isoform of FXR1 is expressed in MCF7 cells at the gene and protein level.

Figure 59. Visualization of the RNA-seq data of MCF7 cells analyzed by TopHat and BWA. RNA-seq reads aligned between exon 12 and exon 13 of FXR1 protein coding transcript NM_005087.3. A significant number of reads (black circle) aligned to intron 12. The identified peptides “GYATDESTVSSVQGSR” “EKGYATDESTVSSVQGSR” contain amino acids from intron 12 (grey amino acids) and exon 13 (black amino acids).

4.2. Validation of peptides with missed cleavages

Peptides with missed cleavages are not routinely utilized for validation because the digestion efficiency in different experiments may differ; furthermore it has been reported that missed cleavage peptides are detected with lower sensitivity [105]. However, missed cleavage peptides span a larger fraction of a protein’s sequence, thereby enabling the identification of more

proteoforms. In our analysis we targeted peptides for validation that contained missed cleavages. The peptide "*K.GTAAAAAAAAAAAAKVPAKK.I*", with 2 missed cleavages, was identified in the proteogenomic discovery experiments with 3 extra Alanine residues (grey text above). The MS/MS spectra used to identify this peptide in the discovery experiments exhibited a high similarity (0.97) with the MS/MS spectrum of the SIS peptide standard (Figure 32). However, we did not detect this missed cleavage peptide in the validation experiments presumably because of the increased digestion efficiency of the validation experiments (determined by calculating the proportion of missed cleavage peptides). The terminal tryptic peptide "K.GTAAAAAAAAAAAAK.V*"* was detected in both the proteogenomic discovery and validation experiments but was omitted from the list of non-canonical peptides because the spectrum also matched a semi-tryptic peptide from the reference proteome, but with a lower score. As a conservative approach, we rejected all peptides whose spectra matched to peptides from the reference proteome. Although, we could not detect the endogenous missed cleavage peptide in the validation experiments the high similarity score in tier 1 validation indicates the proteoform is present in MCF7 cells.

4.3. Variant missed in next-generation sequencing

The variant calling programs used for the analysis of next-generation sequencing data reject mutations that do not meet a user-defined minimum number of reads. If a mutation does not generate sufficient reads they are rejected as artifacts. Or in other words, if the sequencing experiments are not performed with sufficient depth the exons on which the variant lies might not be sequenced, or the result deemed unreliable. We detected a total of 15 peptides from the GNOMON-small database that were classified as novel-isoform-pep and non-coding-pep but could be aligned to reference proteins by a single amino acid variation (online appendix 6). Accordingly, we considered if these peptides were variants that were not detected by NGS. Six of the peptides were targeted for validation, of which 5 passed tier 1 validation and 3 also passed tier 2 validation. The genomic coordinates of the peptides as undetected variants of reference proteins were obtained by utilizing the software tool PoGo [106] and are provided in online appendix 7. As an example, the peptide "***K.AEEPGDGPAEEWLGR.A***" was identified from the GNOMOM protein model of SFT2 domain containing 3 (SFT2D3) but with the Arginine (**R**) at position 38 replaced by a Glycine (**G**) (shown in grey for clarity). This peptide

was successfully validated at tier 1 and tier 2 (Figure 60 and Figure 61). A longer missed cleavage peptide, "K.AGGPAAAEPLLAAEKAEEPGDGPAEEWLGR.A", was also identified but was too long for our provider of stable isotope-labeled standard peptides, and so could not be validated.

Figure 62 shows the mapping of these 2 peptides onto the SFT2D3 gene on the alternate assembly. Close examination of the dbSNP database revealed a missense variant "rs10206957" that would cause this amino acid substitution. SFT2D3 is an intronless gene on chromosome 2. The variant was not detected in the NGS experiments due to absence of read coverage in the CDS region of the gene (Figure 63).

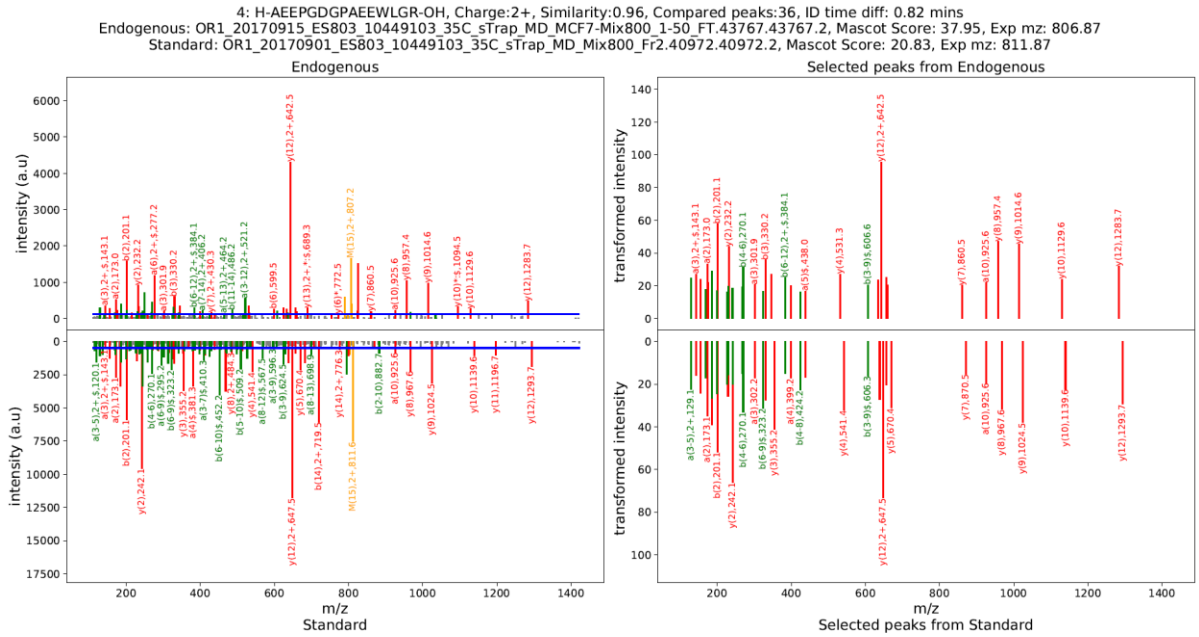

Figure 60. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from an undetected variant of SFT2D3. Peptide "***K.AEEPGDGPAEEWLGR.A***", charge 2+. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a*-type, *b*-type) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 36 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.96.

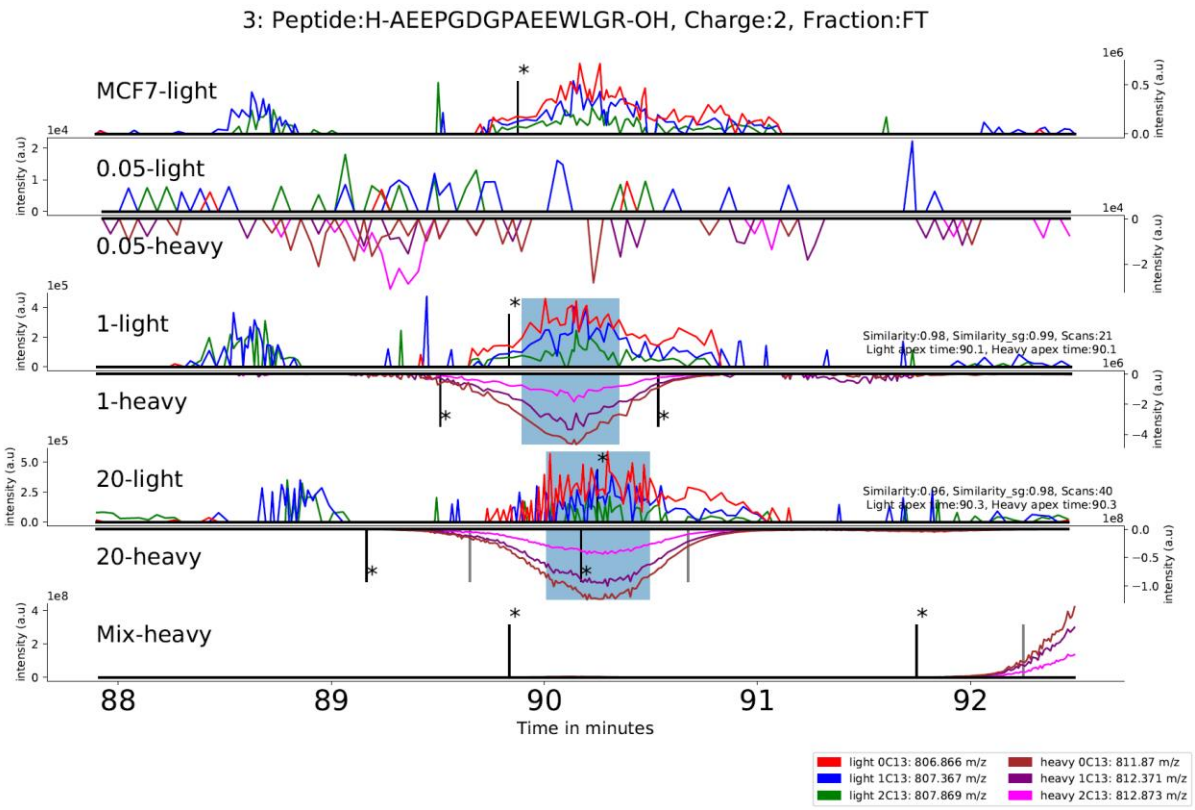

Figure 61. Extracted ion chromatograms (EIC) of endogenous and SIS peptides from an undetected variant of SFT2D3 gene from five validation runs. Peptide "***K.AEEPGDGPAEEWLGR.A***", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.98. Apex elution times for the SIS and endogenous peptides were identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9.The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

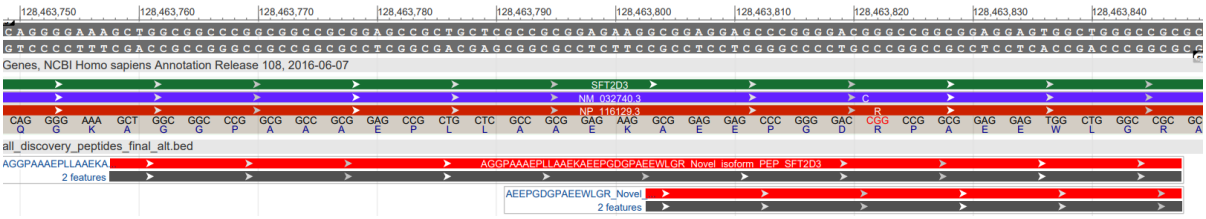

Figure 62. Peptide evidence for the SFT2D3 protein from the alternate genome assembly CHM1_1.1. The two peptides "***K.AEEPGDGPAEEWLGR.A***" and "K.AGGPAAAEPLLAAEKAEEPGDGPAEEWLGR.A" map onto the SFT2D3 gene. Nucleic acid at chromosome location chr2:128463821 on the alternate assembly is Guanine (G) whereas it is Cytosine (C) in the reference assembly (GRCh38) for transcript NM_032740.3. This nucleic acid change would cause the amino acid Arginine (R) at position 38 on the reference assembly protein NP_116129.3 to be replaced by Glycine (G), as was detected here (shown with grey text above). Green bars denote reference genes. Purple bars denote reference transcripts, and red bars reference proteins. The location of the peptide sequences are shown in the bottom track (dark grey blocks).

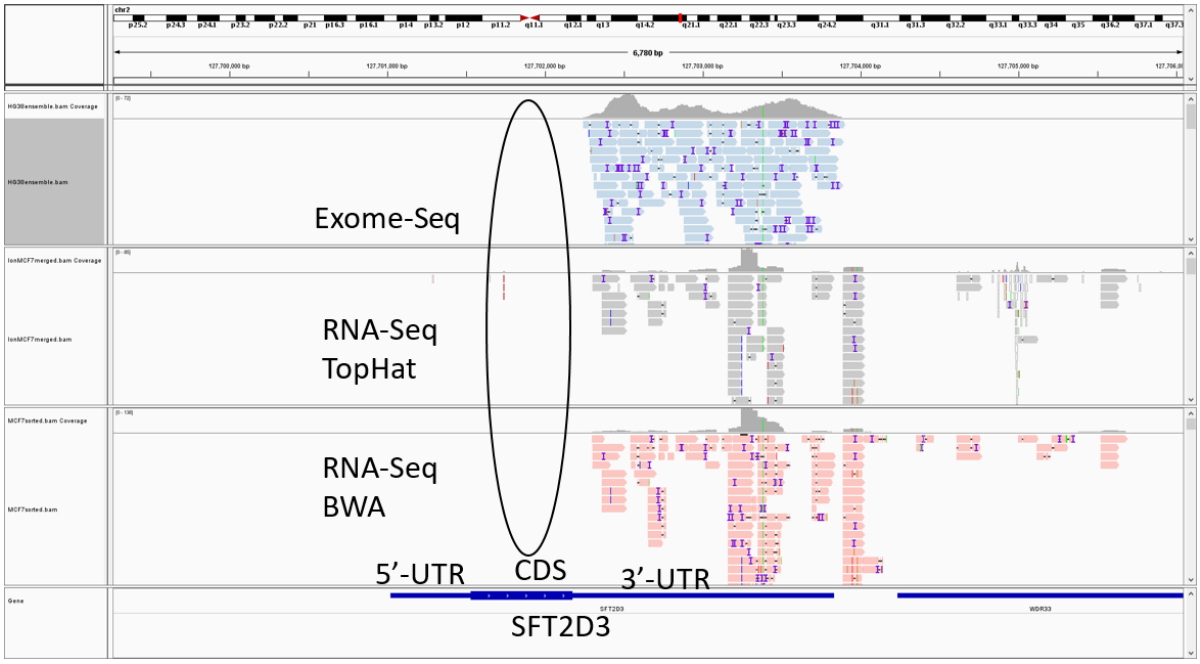

Figure 63. Low read coverage in the coding DNA sequence (CDS) of SFT2D3. Significant reads were aligned in the 3'-UTR region of SFT2D3 but almost no reads aligned with the CDS region where the variant "rs10206957" is located (annotated with an oval).

Two peptides "K.NSVTPDM-EEMYKK.A" and "***K.NSVTPDM-EEMYK.K***" were identified from GNOMON predicted protein of 60S ribosomal protein L5 (RPL5) gene, in which the methionine (M) at position 236 on reference protein ENSP00000495549 is deleted (shown with a hyphen above). The peptide "***K.NSVTPDM-EEMYK.K***" was validated at tier 1 and 2 (Figure 64 and Figure 65). An examination of the dbSNP database did not reveal any deletion variants at this position. The peptides span the boundary of exon 6 and exon 7 on transcript ENST00000644759, indicated with black and grey text above. The RNA-seq intron features from NCBI show that transcripts with both consensus (GT-AG) and non-consensus (GT-TG) splice sites have been reported at the junction of exons 6 and 7 in RPL5 (Figure 66). The canonical splicing encodes the reference RPL5 protein whereas the alternate splicing would cause the skipping of Methionine at position 236.

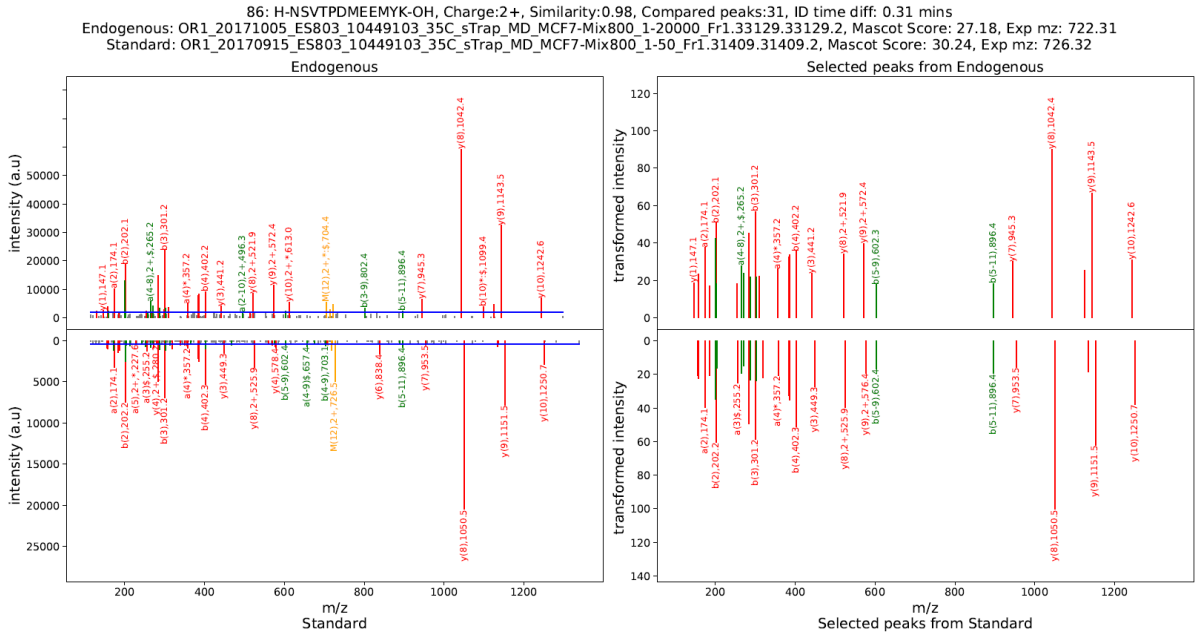

Figure 64. Annotated MS/MS spectra of the endogenous (top left) and SIS peptide (bottom left) from the GNOMON predicted protein of RPL5. Peptide "***K.NSVTPDM-EEMYK.K***", charge 2+. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments (*a*-type, *b*-type) in green and un-fragmented precursors in orange. Ammonia loss is shown with a * and water loss is shown with a $ sign. The spectra on the right show the MS/MS peaks selected for the similarity computation (top: endogenous peptide, bottom: SIS peptide). The intensities of the selected peaks were variance stabilized by square root transform and normalized to sum 1000 before similarity computation. The intensities of 31 common annotated peaks were compared. The peptide passed tier 1 validation with a similarity score of 0.98.

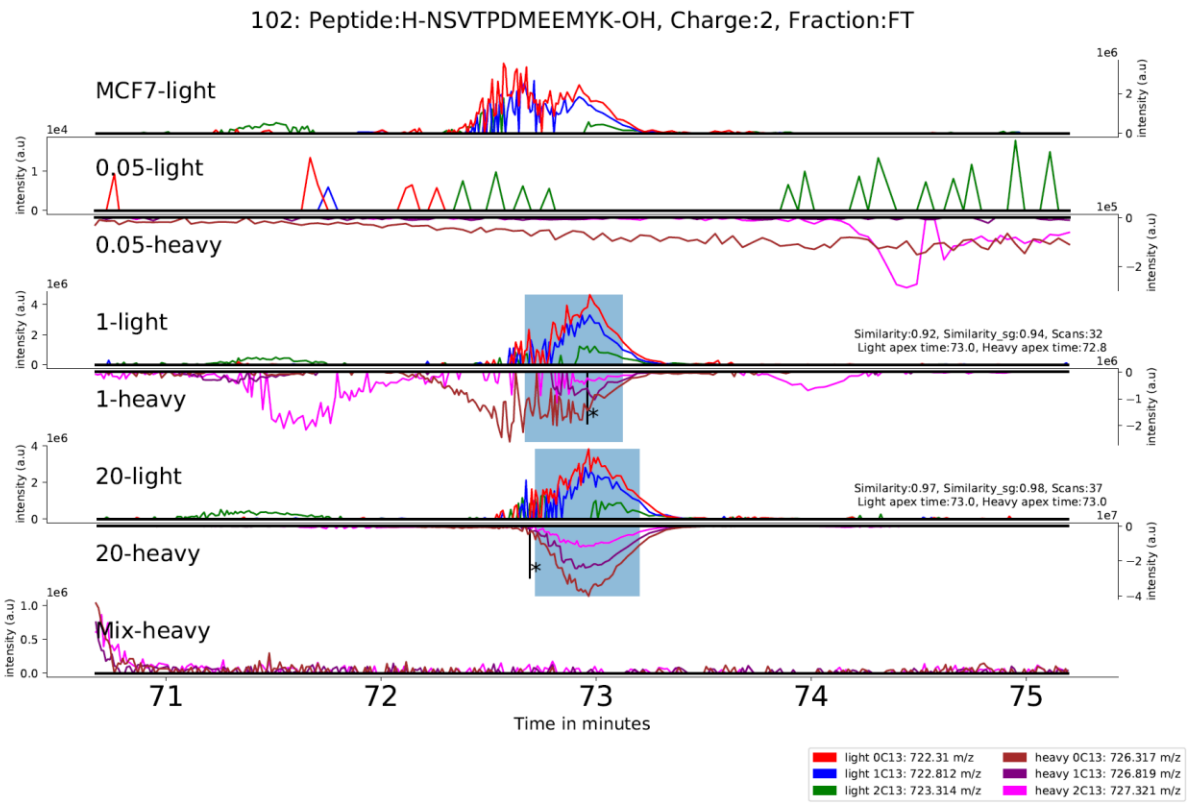

Figure 65. Extracted ion chromatograms (EIC) of endogenous and isotopically-labeled standard peptides from the GNOMON predicted protein of RPL5 gene in five validation runs. Peptide "***K.NSVTPDM-EEMYK.K***", charge 2+. The elution profiles were compared within the time window highlighted with a light blue box. The peptide passed tier 2 validation with a profile similarity of 0.97. Apex elution times for the SIS and endogenous peptides were nearly identical in all validation runs. The MS/MS identification time points are shown with black vertical lines. Asterisks (*) indicate the MS/MS spectra have also been validated at tier 1 with a cosine similarity above 0.9. The top EIC (MCF7-light) shows the EIC of the endogenous peptide (MCF7-light) without any spike. The bottom EIC (Mix-heavy) shows the EIC of the SIS peptide in the SIS peptide mixture. The six EICs in between (0.05-light, 0.05-heavy, 1-light, 1-heavy, 20-light and 20-heavy) show EIC's of the light and heavy (SIS) peptide from the three validation experiments, in which the MCF7 tryptic digest was spiked with the SIS peptide mix at three different concentrations (0.05, 1 and 20 fmol for every µg of tryptic digest). The EICs of the SIS peptide in the validation runs are inverted for ease of comparison with the EIC's of the endogenous peptide.

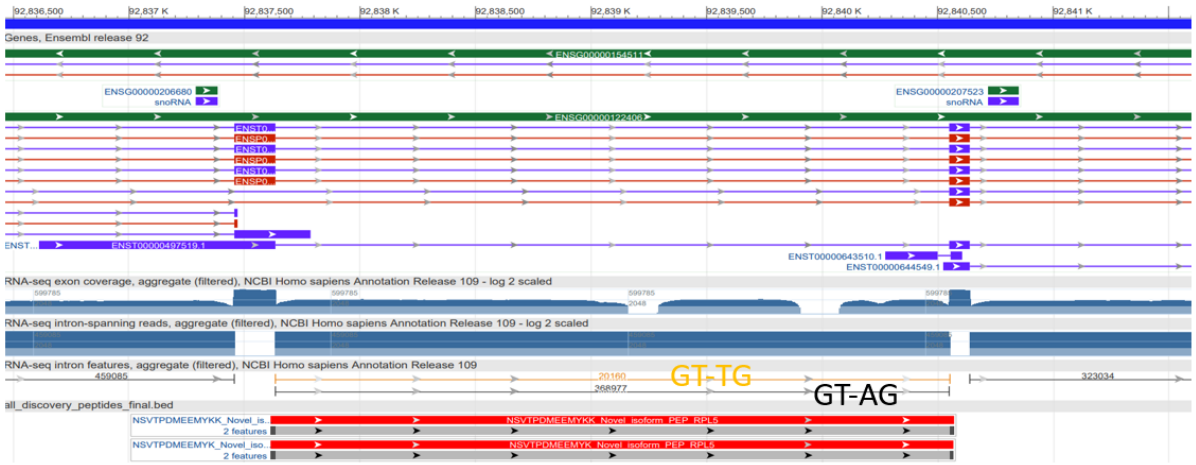

Figure 66. Peptide evidence of alternative splicing in RPL5. Visualization of two peptides "K.NSVTPDM-EEMYKK.A" and "***K.NSVTPDM-EEMYK.K***" on the RPL5 gene, Ensembl id: ENSG00000122406. The NCBI reference transcripts (shown in purple) with the canonical splicing GT-AG (black arrow line) would lead to the generation of the canonical protein sequence (with methionine), whereas the alternate splicing GT-TG (orange arrow line) would lead to the detected protein, without methionine at position 236. Green bars denote reference genes. Purple bars denote reference transcripts, and red bars reference proteins. Colored blocks represent exons, and lines introns. The genomic locations of the peptides are shown in the bottom track (short dark grey blocks at the beginning and end of the light grey block (intron), and overlapping with the exons).

4.4. Ambiguous proteogenomics peptides

A total of 65 non-canonical peptides were classified as ambiguous (Table 2). For example, the non-canonical peptide "R.CHLGPGHQAGPGLHRPPSPR.C" was identified with 2 PSMs from protein coding transcript ENST00000611571 of Mucin 1 (MUC1) in a non-canonical frame (Figure 67). Proteogenomic mapping revealed the peptide could be mapped onto 22 genomic co-ordinates, all within the MUC1 mRNA (Figure 68). The peptide is classified as ambiguous due to multiple coordinate hits on the transcriptome but it is produced unambiguously by the MUC1 gene.

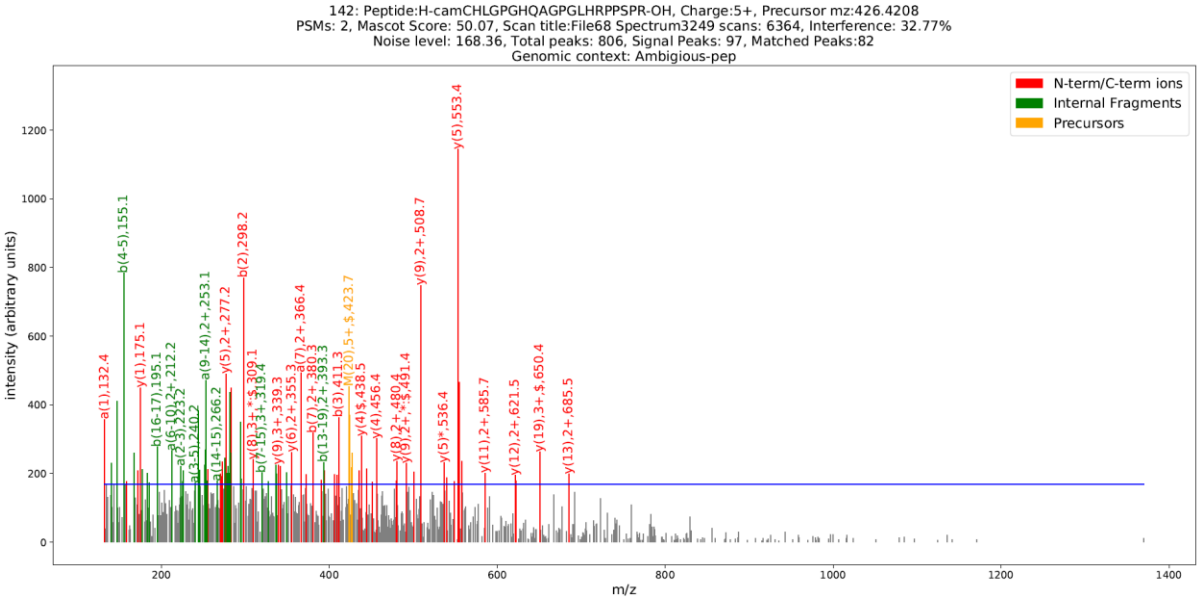

Figure 67. MS/MS spectra of peptide "R.CHLGPGHQAGPGLHRPPSPR.C", charge 5+, *m/z* 426.42. The noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments are shown in green and un-fragmented precursors in orange. Ammonia loss is shown with a * sign and water loss is shown with a $ sign.

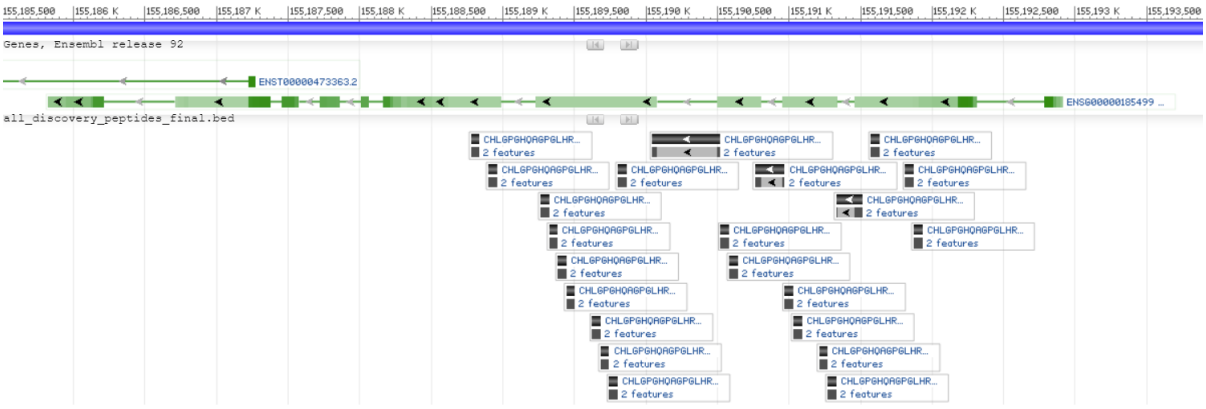

Figure 68. Ambiguous mapping of a peptide. The peptide "CHLGPGHQAGPGLHRPPSPR" can be mapped on to 22 different coordinates all within the MUC1 gene ENSG00000185499.

Some non-canonical peptides could be classified into different classes if different genome annotation systems were used. Peptide "R.VDPQGQCPEHGACPSLLAHVSAEGR.R" was identified with a single PSM from the HELZ2 gene (Figure 69). Proteogenomic mapping revealed that the peptide has a unique coordinate within the HELZ2 gene (Figure 70). However, the peptide is classified as ambiguous because it could be mapped onto two non-coding transcripts of HELZ2 from GENCODE with different biotypes; ENST00000370082 - "retained intron" and ENST00000479540 - "processed transcript". Note: The GENCODE annotation was preferentially used over NCBI; if the NCBI annotation was given priority the peptide would have been unambiguously classified as an uORF-peptide because it maps onto the 5'-UTR region of RefSeq protein coding transcripts NM_001037335, XM_024452007 and XM_024452006 (Figure 70). Utilizing NCBI annotation over GENCODE would cause some proteogenomic peptides to be classified differently due to the differences in the transcript and gene models produced by these two annotation systems.

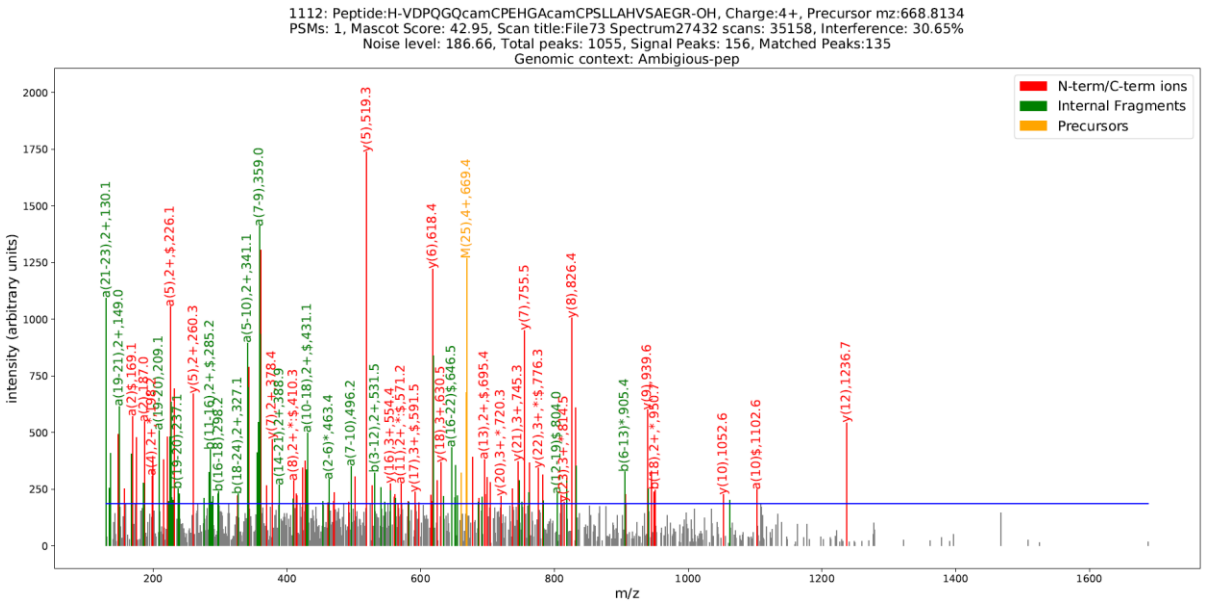

Figure 69. MS/MS spectra of peptide "R.VDPQGQCPEHGACPSLLAHVSAEGR.R", charge 4+, *m/z* 668.81. The noise level (blue horizontal line) was determined by DNL. The N-terminal (*a/b*) and C-terminal (*y*) fragment ions are shown in red, internal fragments are shown in green and non-fragmented precursors in orange. Ammonia loss is shown with a * sign and water loss is shown with a $ sign.

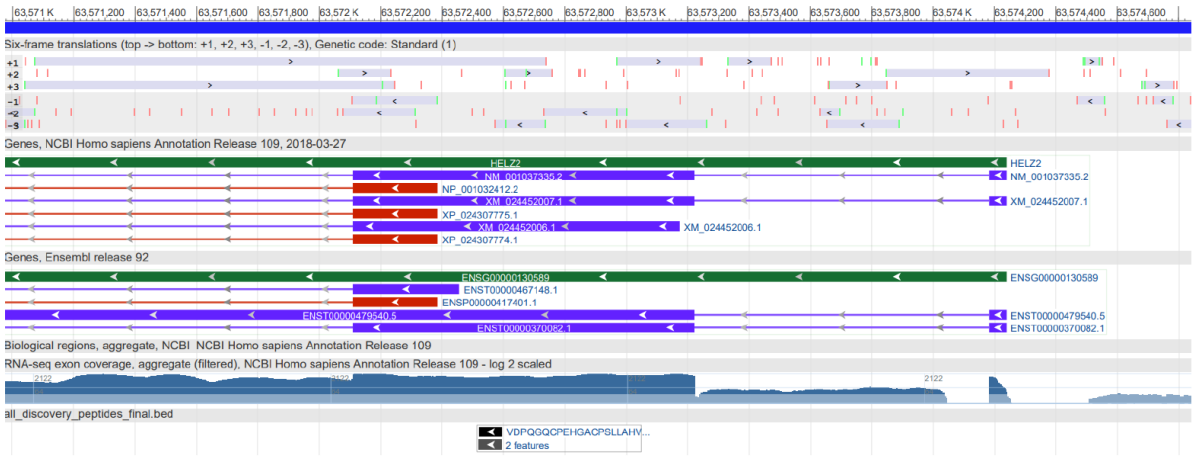

Figure 70. Mapping of "R.VDPQGQCPEHGACPSLLAHVSAEGR.R" onto HELZ2 is dependent on which annotation system is used (Ensembl or NCBI). The peptide maps onto two Ensembl transcripts ENST00000479540 and ENST00000370082 with biotypes "processed transcript" and "retained intron", respectively. Based on NCBI RefSeq annotation the peptide maps onto the 5'-UTR region of protein coding transcripts NM_001037335, XM_024452007 and XM_024452006. Green bars denote reference genes. Purple bars denote reference transcripts, and red bars reference proteins. Colored blocks represent exons, and lines introns. The location of the non-canonical peptides identified by the proteogenomic search is shown in the bottom row.

4.5. Unmapped proteogenomics peptides

A total of 55 peptides were classified as unmapped in our analysis (Table 2). Unmapped peptides were identified in searches conducted with CDS extensions, 6 frame transcriptome and GNOMON databases. Some peptides remained unmapped due to differences between the annotations of the genomic features utilized to generate the databases (GENCODE version 25) and the one used to map them (GENCODE version 27). Most unmapped peptides either had a stop codon replaced with a standard amino acid or were produced from the N-term/C-terminal of the ORFs from CDS extension databases. During database search with Mascot (version 2.5) stop codons in nucleic acid searches and unknown amino acids (X) in amino acid searches are replaced by all standard amino acids. Peptides identified from these databases may have a stop codon or unknown amino acid replaced by a standard amino acid. We considered such peptide matches to be spurious results and so they were not utilized for assignment of genomic context.

4.6. Novel proteoforms are expressed in MCF7 cells

We successfully validated the protein level expression of variants identified in next-generation sequencing. Our analysis demonstrated that besides SNVs many other proteoforms could be detected in MCF7 cells. Peptide "***K.AGGAAD*MTDNIPLQPVR.Q***" was identified from the uORF of ATP9A mRNA. Expression of uORFs have been shown to regulate the expression of main ORF genes [50]. The peptide "***R.GLQLLQPHQLLQGR.G***" was validated from the last exon of the KRT8 gene in a non-canonical frame. An undetected frame-shift or ribosomal frame-shifting may be responsible for the production of the resulting truncated KRT8 protein in MCF7 cells. KRT8 is the major component of intermediate filament cytoskeleton and its high expression has been linked to tumor progression and metastasis of gastric cancer [107]. KRT8 expression is enhanced in MCF7 cell lines [108].

A peptide "***R.GAEVPGEAAPGAR.A***" was validated from the long non-coding RNA gene GATA3-AS1. The peptide was identified from an ORF of length 125 amino acids within the GATA3-AS1 gene (Figure 71). The translation of GATA3-AS1 in MCF7 cells is particularly interesting because the gene coding for the protein GATA3 is on the opposite strand and harbors an insertion mutation that leads to the translation of a truncated proteoform of GATA3 [41]. Recent evidence demonstrates that peptides produced from long non-coding RNA genes can have important biological and functional roles, for example the small peptide Myoregulin (MRLN) that regulates muscle performance [53].

MEPDFLHSVGVKLPHTPPNTCPRASPSHPPSQGRRDPVPVEVGKPSRVQKAEAMAQS
GGAAFWGSALGLQTQGAEMLAAGPPTRARQPALPGELRGAEVPGEAAPGARALPD
LGNRQSGAPGSKS

Figure 71. GATA3-AS1 ORF showing the identified peptides (red).

We validated multiple peptides from an N-terminally extended novel isoform of STARD10 protein (Figure 72). STARD10 protein is over-expressed in breast cancer [109] and the MCF7 breast cancer cell line used here[110,111]. STARD10 functions as a phospholipid transporter, and a loss of expression has been reported as indicative of poor prognosis [112]. The 77 amino acid extension identified here represents an extension of 26% over the normal length protein.

MEEELALGPRGQGGASLAGRDGRSAGAGSYGALANSAWGGPRKVASASAAASTLS
EPPRRTQESRTRTRALGLPTLPMEKLAASTEPQGPRPVLGRESVQVPDDQDFRSFRSE
CEAEVGWNLTYSRAGVSVWVQAVEMDRTLHKIKCRMECCDVPAETLYDVLHDIEY
RKKWDSNVIETFDIARLTVNADVGYYSWRCPKPLKNRDVITLRSWLPMGADYIIMN
YSVKHPKYPPRKDLVRAVSIQTGYLIQSTGPKSCVITYLAQVDPKGSLPKWVVNKSS
QFLAPKAMKKMYKACLKYPEWKQKHLPHFKPWLHPEQSPLPSLALSELSVQHADS
LENIDESAVAESREERMGGAGGEGSDDDTSLT

Figure 72. Peptides detected from a GNOMON predicted N-terminal extended isoform of STARD10. The N-terminal extended part is shown in black and the reference STARD10 protein is shown in grey. The 6 non-canonical peptides identified by the proteogenomics search, namely "MEEELALGPR.G", "R.SAGAGSYGALANSAWGGPR.K", "R.SAGAGSYGALANSAWGGPRK.V", "R.KVASASAAASTLSEPPR.R", "R.KVASASAAASTLSEPPRR.T", and "K.VASASAAASTLSEPPR.R", were all identified from the extended N-terminal region (peptides underlined for clarity).

The peptide "***R.SPPDSPT****DALMQLAK.A*" mapped onto a novel exon situated between exons 17 and 18 in the gene TLN1. This isoform of TLN1 gene has 17 amino acids "ICASRGAGVRSPPDSPT" coded by the novel exon, which is inserted at amino acid position 666 in the reference TL1 protein. TLN1 codes for a cytoskeletal protein that is concentrated in areas of cell-substratum and cell-cell contacts [113,114].

The peptide "***K.NSVTPDM-EEMYK.K***" was validated from an isoform of the protein RPL5 due to non-canonical splicing of exon 6 and 7. The variant was not detected in NGS experiments but was identified from the GNOMON database. The splicing caused deletion of a methionine residue at the start of exon 7. RPL5 protein is a component of the large ribosomal subunit. Mutations in RPL5 have been associated with defects in the maturation of ribosomal RNAs in the small or large ribosomal subunit production pathways [115].

The peptide "*R.ASAA**EGVGEPGASAGR.A***" was validated from an upstream extension of exon 1 within the gene WDR26. The peptide spans the boundary of 5'-UTR and the genomic region upstream of 5'-UTR, suggesting the expression of a novel CDS in this region. The protein belongs to the WD repeat protein family, and is involved in a variety of cellular processes including cell cycle progression, signal transduction, apoptosis and gene regulation [116,117] . It has been shown that WDR26 is overexpressed in highly malignant breast cancers and has been proposed as a potential therapeutic target for breast cancer [118].

We performed a comprehensive proteogenomics analysis of MCF7 cells with customized protein sequence database searches. In addition to confirming the protein forms of variants identified by next-generation sequencing, multiple novel proteoforms were also validated. Bottom-up LC-MS/MS is widely used for profiling of proteomic landscape of complex biological samples. It is well known that a large proportion of the acquired spectra in LC-MS/MS experiments cannot be assigned. Some of these spectra remain unassigned due to absence of the proteoforms in the compact reference proteome databases utilized in proteomics data analysis pipelines. Our results demonstrates that a subset of these unassigned spectra originate from genomic mutations and pervasive translations from outside of the known protein coding regions of the genome. The results demonstrated how the reference databases commonly utilized in proteomics workflows do not fully capture the complexity of the oncoproteome space, and if possible should be supplemented with sample specific variant and novel proteoforms.

5. Conclusions

In this work I have developed a comprehensive proteogenomics identification and validation pipeline. The pipeline was utilized to discover and validate proteogenomics peptides in MCF7 cells (Figure 73). I used the Python programming language to develop scripts for: construction of customized databases, noise-detection in MS/MS spectra, comprehensive annotation of matched MS/MS spectra, proteogenomic mapping of all classes of peptides, and the quantitative comparison of endogenous non-canonical peptides with their isotopically labeled analogues (SIS peptides).

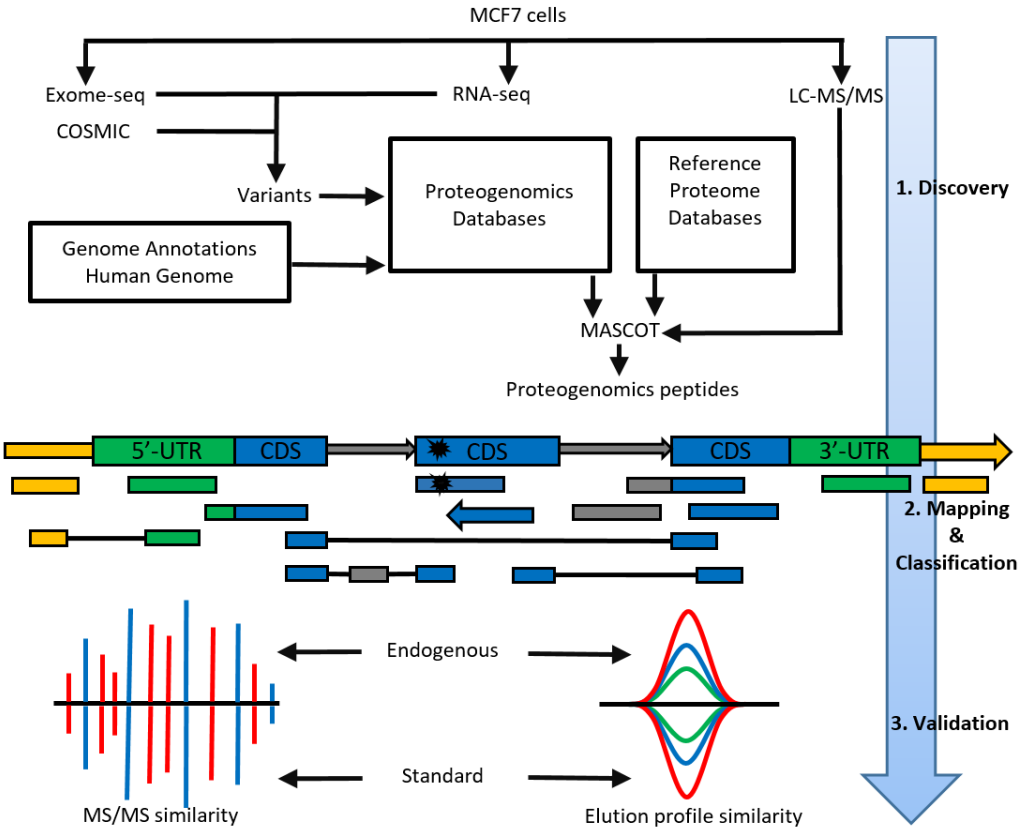

Figure 73. Proteogenomics analysis and validation pipeline.

First several customize databases were generated and used to analyze the LC-MS/MS data to discover novel proteoforms. Variants detected in Exome-seq, RNA-seq and publicly available in COSMIC were utilized to identify SNVs and InDel peptides. Exon-skipped peptides were identified by generating a database of novel exon-skipped proteins from reference transcripts. Peptides from uORFs, dORFs and alt-frame translations were identified using a database of ORFs generated from a reference transcriptome. Peptides spanning exon-intron boundaries were identified using a database of ORFs generated from 100 base-pair extensions of CDS and gene sequences. Peptides from non-coding transcripts were identified using a database of ORFs generated from non-coding transcript sequences. Peptides from novel protein isoforms were identified using a database of Gnomon proteins.

The LC-MS/MS data was first interrogated with all common reference proteomes from UniProt, GENCODE and RefSeq. I demonstrated that the choice of reference proteome database can affect the identification of non-canonical peptides. This problem was avoided by utilizing non-redundant protein sequences from all common reference proteomes. A merged

reference proteome database was created incorporating the reference proteomes in UniProt, GENCODE and RefSeq.

Peptides that were identified exclusively from the proteogenomics searches were subjected to rigorous quality control. First I applied a noise detection algorithm to filter out PSMs whose MS/MS spectra were of low signal-to-noise, and which did not contain sufficient signal peaks. Then, I developed a spectrum annotation tool to ensure the non-canonical peptides could account for the majority of fragment ions contained in their MS/MS spectra. These QC steps ensured all reported non-canonical peptides were identified using high quality MS/MS spectra that could describe the full MS/MS spectrum, and thus reduced the likelihood of false-positive identifications.

I performed genomic mapping of the QC controlled peptides. I was able to successfully map all classes of peptides to their genomic co-ordinates. I assigned a genomic context to the peptide based on their mapping which was later used for their classification.

Finally I validated many of the non-canonical peptides with SIS peptides. I developed a two tier validation scheme: in tier 1 I compared the fragmentation pattern of endogenous and SIS peptides. Peptides that had a similarity score greater than 0.9 were subjected to tier 2 validation, in which I compared their elution profiles. To guard against co-eluting ions, profile similarity was also computed for each peptide isotope (monoisotopic, $^{1}C_{13}$ and $^{2}C_{13}$). Peptides that had elution profile similarity and isotopic composition similarity greater than 0.9 were considered validated at tier 2.

6. References

1. Faulkner S, Dun MD, Hondermarck H. Proteogenomics: Emergence and promise. *Cell Mol Life Sci*. 2015;72(5):953-957. doi:10.1007/s00018-015-1837-y.
2. Avery OT, Macleod CM, McCarty M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med*. 1944;79(2):137-158. https://www.ncbi.nlm.nih.gov/pubmed/19871359.
3. Venter JC, Adams MDM, Myers EEW, et al. The sequence of the human genome. *Sci …*. 2001;291(February):1304-1352. doi:10.1126/science.1058040.
4. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062.
5. Collins FS, Lander ES, Rogers J, Waterson RH. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931-945. doi:10.1038/nature03001.
6. Schneider VA, Graves-Lindsay T, Howe K, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 2017;27(5):849-864. doi:10.1101/gr.213611.116.
7. https://www.ncbi.nlm.nih.gov/grc.
8. Ostell J, McEntyre J. The NCBI Handbook. *NCBI Bookshelf*. 2007;(Md):1-8. doi:10.4016/12837.01.
9. Cordero F, Botta M, Calogero RA. Microarray data analysis and mining approaches. *Brief Funct Genomic Proteomic*. 2007;6(4):265-281. doi:10.1093/bfgp/elm034.
10. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. *Nature*. 2008;456(7218):60-65. doi:10.1038/nature07484.
11. Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T., and Lipman D. Gnomon – NCBI eukaryotic gene prediction tool. *Ncbi*. 2010:1-24. http://www.ncbi.nlm.nih.gov/core/assets/genome/files/Gnomon-description.pdf.
12. https://www.ncbi.nlm.nih.gov/.
13. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res*. 2012;22(9):1760-1774.

doi:10.1101/gr.135350.111.

14. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: Algorithms for computing spliced alignments with identification of paralogs. *Biol Direct*. 2008;3:1-13. doi:10.1186/1745-6150-3-20.
15. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biol*. 2007;5(10):2113-2144. doi:10.1371/journal.pbio.0050254.
16. Sherry ST. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311. doi:10.1093/nar/29.1.308.
17. https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi.
18. Metzker ML. Sequencing technologies the next generation. *Nat Rev Genet*. 2010;11(1):31-46. doi:10.1038/nrg2626.
19. Yi X, Liang Y, Huerta-Sanchez E, et al. Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. *Science (80-)*. 2010;329(5987):75-78. doi:10.1126/science.1190371.Sequencing.
20. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484.
21. Sassi C, Guerreiro R, Gibbs R, et al. Exome sequencing identifies 2 novel presenilin 1 mutations (p.L166V and p.S230R) in British early-onset Alzheimer's disease. *Neurobiol Aging*. 2014. doi:10.1016/j.neurobiolaging.2014.04.026.
22. Huber C, Faqeih EA, Bartholdi D, et al. Exome sequencing identifies INPPL1 mutations as a cause of opsismodysplasia. *Am J Hum Genet*. 2013. doi:10.1016/j.ajhg.2012.11.015.
23. Do R, Stitziel NO, Won HH, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*. 2015. doi:10.1038/nature13917.
24. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*. 2010;42(1):30-35. doi:10.1038/ng.499.
25. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. *G3:Genes|Genomes|Genetics*. 2015;5(8):1543-1550. doi:10.1534/g3.115.018564.
26. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by

microarray hybridization. *Nat Methods*. 2007. doi:10.1038/nmeth1111.

27. Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B. A beginners guide to SNP calling from high-Throughput DNA-sequencing data. *Hum Genet*. 2012;131(10):1541-1554. doi:10.1007/s00439-012-1213-z.
28. Steijger T, Abril JF, Engström PG, et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods*. 2013. doi:10.1038/nmeth.2714.
29. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet*. 2013;93(4):641-651. doi:10.1016/j.ajhg.2013.08.008.
30. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003;422(6928):198-207. doi:10.1038/nature01511.
31. Cañas B, López-Ferrer D, Ramos-Fernández A, Camafeita E, Calvo E. Mass spectrometry technologies for proteomics. *Brief Funct Genomic Proteomic*. 2006;4(4):295-320. doi:10.1093/bfgp/eli002.
32. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007. doi:10.1038/nmeth1019.
33. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*. 2004;4(1):59-77. doi:10.1002/pmic.200300511.
34. Garin-Muga A, Corrales FJ, Segura V. Proteogenomic Analysis of Single Amino Acid Polymorphisms in Cancer Research. *Adv Exp Med Biol*. 2016;926:93-113. doi:10.1007/978-3-319-42316-6_7.
35. Xing X Bin, Li QR, Sun H, et al. The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. *Genomics*. 2011;98(5):343-351. doi:10.1016/j.ygeno.2011.07.005.
36. Helmy M, Sugiyama N, Tomita M, Ishihama Y. Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing. *Genome Biol*. 2010;11(Suppl 1):P17. doi:10.1186/gb-2010-11-s1-p17.
37. Rivers RC, Kinsinger C, Boja ES, Hiltke T, Mesri M, Rodriguez H. Linking cancer genome to proteome: NCI's investment into proteogenomics. *Proteomics*. 2014;14(23-

24):2633-2636. doi:10.1002/pmic.201400193.

38. Ellis MJ, Gillette M, Carr SA, et al. Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium. *Cancer Discov*. 2013. doi:10.1158/2159-8290.CD-13-0219.
39. Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods*. 2012. doi:10.1038/nmeth.2227.
40. Morin RD, Mendez-Lago M, Mungall AJ, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*. 2011;476(7360):298-303. doi:10.1038/nature10351.
41. Adomas AB, Grimm S a, Malone C, Takaku M, Sims JK, Wade PA. Breast tumor specific mutation in GATA3 affects physiological mechanisms regulating transcription factor turnover. *BMC Cancer*. 2014;14(1):278. doi:10.1186/1471-2407-14-278.
42. Shapiro IM, Cheng AW, Flytzanis NC, et al. An emt-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet*. 2011;7(8). doi:10.1371/journal.pgen.1002218.
43. Sendoel A, Dunn JG, Rodriguez EH, et al. Translation from unconventional 5′ start sites drives tumour initiation. *Nature*. 2017;541(7638):494-499. doi:10.1038/nature21036.
44. Mehta A, Trotta CR, Peltz SW. Derepression of the Her-2 uORF is mediated by a novel post-transcriptional control mechanism in cancer cells. *Genes Dev*. 2006;20(8):939-953. doi:10.1101/gad.1388706.
45. Wang H, Wang Y, Xie S, Liu Y, Xie Z. Global and cell-type specific properties of lincRNAs with ribosome occupancy. *Nucleic Acids Res*. 2017;45(5):2786-2796. doi:10.1093/nar/gkw909.
46. Müeller-Pillasch F, Lacher U, Wallrapp C, et al. Cloning of a gene highly overexpressed in cancer coding for a novel KH-domain containing protein. *Oncogene*. 1997;14(22):2729-2733. doi:10.1038/sj.onc.1201110.
47. Barbosa C, Peixeiro I, Romão L. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genet*. 2013;9(8):1-12. doi:10.1371/journal.pgen.1003529.
48. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5′UTRs, and pseudogenes are

translated and some are likely to express functional proteins. *Elife*. 2015;4:1-21. doi:10.7554/eLife.08890.

49. Somers J, Pöyry T, Willis AE. A perspective on mammalian upstream open reading frame function. *Int J Biochem Cell Biol*. 2013;45(8):1690-1700. doi:10.1016/j.biocel.2013.04.020.
50. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci*. 2009;106(18):7507-7512. doi:10.1073/pnas.0810916106.
51. Banfai B, Jia H, Khatun J, et al. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res*. 2012;22:1646-1657. doi:10.1101/gr.134767.111.
52. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol*. 2007;5(5):1052-1062. doi:10.1371/journal.pbio.0050106.
53. Anderson DM, Anderson KM, Chang CL, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*. 2015;160(4):595-606. doi:10.1016/j.cell.2015.01.009.
54. Matsumoto A, Pasut A, Matsumoto M, et al. MTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*. 2017;541(7636):228-232. doi:10.1038/nature21034.
55. Breuza L, Poux S, Estreicher A, et al. The UniProtKB guide to the human proteome. *Database*. 2016;2016:1-10. doi:10.1093/database/bav120.
56. Pruitt KD, Brown GR, Hiatt SM, et al. RefSeq: An update on mammalian reference sequences. *Nucleic Acids Res*. 2014;42(D1):756-763. doi:10.1093/nar/gkt1114.
57. Alfaro J a, Sinha A, Kislinger T, Boutros PC. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nat Methods*. 2014;11(11):1107-1113. doi:10.1038/nmeth.3138.
58. Chan SH, Lim WK, Ishak NDB, et al. Germline Mutations in Cancer Predisposition Genes are Frequent in Sporadic Sarcomas. *Sci Rep*. 2017;7(1):1-8. doi:10.1038/s41598-017-10333-x.
59. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med*. 2015;373(24):2336-2346.

doi:10.1056/NEJMoa1508054.

60. Clamon GH, Bossler AD, Hejleh TA, Furqan M. Germline mutations predisposing to non-small cell lung cancer. *Fam Cancer*. 2015;14(3):463-469. doi:10.1007/s10689-015-9796-x.
61. Zhang H, Feng M, Feng Y, et al. Germline mutations in hereditary diffuse gastric cancer. *Chinese J Cancer Res*. 2018;30(1):122-130. doi:10.21147/j.issn.1000-9604.2018.01.13.
62. Baysal BE, Willett-Brozick JE, Lawrence EC, et al. Prevalence of SDHB, SDHC, and SDHD germline mutations in clinic patients with head and neck paragangliomas. *J Med Genet*. 2002;39(3):178-183. doi:10.1111/j.1399-0004.2004.00328.x.
63. Astuti D, Douglas F, Lennard TWJ, et al. Germline SDHD mutation in familial phaeochromocytoma. *Lancet*. 2001;357(9263):1181-1182. doi:10.1016/S0140-6736(00)04378-6.
64. Tomlinson IPM, Alam NA, Rowan AJ, et al. Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer the multiple leiomyoma consortium. *Nat Genet*. 2002;30(4):406-410. doi:10.1038/ng849.
65. Amary MF, Damato S, Halai D, et al. Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2. *Nat Genet*. 2011;43(12):1262-1265. doi:10.1038/ng.994.
66. Cohen AL, Holmen SL, Colman H. IDH1 and IDH2 mutations in gliomas. *Curr Neurol Neurosci Rep*. 2013;13(5):345. doi:10.1007/s11910-013-0345-4.
67. Branca RMM, Orre LM, Johansson HJ, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods*. 2014;11(1):59-62. doi:10.1038/nmeth.2732.
68. Ruggles K V., Tang Z, Wang X, et al. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol Cell Proteomics*. 2016;15(3):1060-1071. doi:10.1074/mcp.M115.056226.
69. Slavoff SA, Mitchell AJ, Schwaid AG, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol*. 2013;9(1):59-64. doi:10.1038/nchembio.1120.Peptidomic.
70. Pauli A, Valen E, Schier AF. Identifying (non-)coding RNAs and small peptides:

Challenges and opportunities. *BioEssays*. 2015;37(1):103-112. doi:10.1002/bies.201400103.

71. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*. 2014;11(11):1114-1125. doi:10.1038/nmeth.3144.
72. Askenazi M, Ruggles K V., Fenyö D. PGx: Putting Peptides to BED. *J Proteome Res*. 2016;15(3):795-799. doi:10.1021/acs.jproteome.5b00870.
73. Risk BA, Spitzer WJ, Giddings MC. Peppy: Proteogenomic search software. *J Proteome Res*. 2013. doi:10.1021/pr400208w.
74. Sanders WS, Wang N, Bridges SM, et al. The proteogenomic mapping tool. *BMC Bioinformatics*. 2011;12. doi:10.1186/1471-2105-12-115.
75. Ferro M, Tardif M, Reguer E, et al. Pep line: A software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. *J Proteome Res*. 2008. doi:10.1021/pr070415k.
76. Ghali F, Krishna R, Perkins S, et al. ProteoAnnotator - Open source proteogenomics annotation software supporting PSI standards. *Proteomics*. 2014;14(23-24):2731-2741. doi:10.1002/pmic.201400265.
77. Zickmann F, Renard BY. MSProGene: Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. In: *Bioinformatics*. ; 2015. doi:10.1093/bioinformatics/btv236.
78. Sheynkman GM, Johnson JE, Jagtap PD, et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics*. 2014;15(1):1-9. doi:10.1186/1471-2164-15-703.
79. Schlaffner CN, Pirklbauer GJ, Bender A, Steen JAJ, Choudhary JS. A Fast and Quantitative Method for Post-translational Modification and Variant Enabled Mapping of Peptides to Genomes. *J Vis Exp*. 2018. doi:10.3791/57633.
80. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.
81. https://ioncommunity.thermofisher.com/welcome.
82. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36. doi:10.1186/gb-2013-14-4-r36.

83. Bayat A, Gaëta B, Ignjatovic A, Parameswaran S. Improved VCF normalization for accurate VCF comparison. *Bioinformatics*. 2017;33(7):964-970. doi:10.1093/bioinformatics/btw748.
84. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923.
85. Broad Institute. Picard tools. *https://broadinstitute.github.io/picard/*. 2016. http://broadinstitute.github.io/picard/.
86. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509.
87. Forbes SA, Tang G, Bindal N, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res*. 2010;38(Database issue):D652-7. doi:10.1093/nar/gkp995.
88. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330.
89. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res*. 1997;25(1):31-36. doi:10.1093/nar/gkq1020.
90. Cock PJA, Antao T, Chang JT, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163.
91. McKinney W, PyData Development Team. Pandas - Powerful Python Data Analysis Toolkit. *Pandas - Powerful Python Data Anal Toolkit*. 2015. doi:10.1073/pnas.1803154115.
92. Oliphant TE. *Guide to NumPy*. 2nd ed. USA: CreateSpace Independent Publishing Platform; 2015.
93. Jones E, Oliphant T, Peterson P, others. SciPy: Open source scientific tools for Python. *Comput Sci Eng*. 2007. doi:10.1143/JJAP.31.4110.
94. Goloborodko AA, Levitsky LI, 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*. 2013;24(2):301-304. doi:10.1007/s13361-012-0516-6.
95. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng*. 2007;9(3):90-95.

doi:10.1109/MCSE.2007.55.

96. Xu H, Freitas MA. A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra. *BMC Bioinformatics*. 2010;11(1):436. doi:10.1186/1471-2105-11-436.
97. Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE. *Electrophoresis*. 1999;20:3551-3567.
98. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem*. 1964;36(8):1627-1639. doi:10.1021/ac60214a047.
99. Brocke KS, Neu-Yilik G, Gehring NH, Hentze MW, Kulozik AE. The human intronless melanocortin 4-receptor gene is NMD insensitive. *Hum Mol Genet*. 2002;11(3):331-335. doi:10.1093/hmg/11.3.331.
100. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*. 2007;4(11):923-925. doi:10.1038/nmeth1113.
101. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389-3402. doi:10.1093/nar/25.17.3389.
102. rs11136334. http://www.ensembl.org/Homo_sapiens/Variation/Explore?r=8:143926920-143927920;v=rs11136334;vdb=variation;vf=381436094.
103. rs2669761. http://www.ensembl.org/Homo_sapiens/Variation/Explore?r=10:50129423-50130423;v=rs2669761;vdb=variation;vf=73540178.
104. rs369485042. http://www.ensembl.org/Homo_sapiens/Variation/Explore?r=3:40461530-40462529;v=rs369485042;vdb=variation;vf=619087928.
105. Brownridge P, Beynon RJ. The importance of the digest: Proteolysis and absolute quantification in proteomics. *Methods*. 2011;54(4):351-360. doi:10.1016/j.ymeth.2011.05.005.
106. Schlaffner CN, Pirklbauer GJ, Bender A, Choudhary JS. Fast, Quantitative and Variant Enabled Mapping of Peptides to Genomes. *Cell Syst*. 2017;5(2):152-156.e4.

doi:10.1016/j.cels.2017.07.007.

107. Fang J, Wang H, Liu Y, Ding F, Ni Y, Shao S. High KRT8 expression promotes tumor progression and metastasis of gastric cancer. *Cancer Sci*. 2017;108(2):178-186. doi:10.1111/cas.13120.
108. data available from v18.proteinatlas.org/ENSG00000170421-KRT8/cell. v18.proteinatlas.org/ENSG00000170421-KRT8/cell.
109. Olayioye MA, Vehring S, Müller P, et al. StarD10, a START domain protein overexpressed in breast cancer, functions as a phospholipid transfer protein. *J Biol Chem*. 2005;280(29):27436-27442. doi:10.1074/jbc.M413330200.
110. Uhlen M, Fagerberg L, Hallstrom BM, et al. Tissue-based map of the human proteome. *Science (80-)*. 2015;347(6220):1260419-1260419. doi:10.1126/science.1260419.
111. data available from v18.proteinatlas.org/ENSG00000214530-STARD10/cell.
112. Murphy NC, Biankin A V., Millar EKA, et al. Loss of STARD10 expression identifies a group of poor prognosis breast cancers independent of HER2/Neu and triple negative status. *Int J Cancer*. 2010;126(6):1445-1453. doi:10.1002/ijc.24826.
113. Burridge K, Connell L. A new protein of adhesion plaques and ruffling membranes. *J Cell Biol*. 1983;97(2):359-367. doi:10.1083/jcb.97.2.359.
114. Rees DJG, Ades SE, Singer SJ, Hynes RO. Sequence and domain structure of talin. *Nature*. 1990;347(6294):685-689. doi:10.1038/347685a0.
115. Gazda HT, Sheen MR, Vlachos A, et al. Ribosomal Protein L5 and L11 Mutations Are Associated with Cleft Palate and Abnormal Thumbs in Diamond-Blackfan Anemia Patients. *Am J Hum Genet*. 2008;83(6):769-780. doi:10.1016/j.ajhg.2008.11.004.
116. Zhu Y, Wang Y, Xia C, et al. WDR26: A novel gβ-like protein, suppresses MAPK signaling pathway. *J Cell Biochem*. 2004;93(3):579-587. doi:10.1002/jcb.20175.
117. Zhao J, Liu Y, Wei X, Yuan C, Yuan X, Xiao X. A novel WD-40 repeat protein WDR26 suppresses H2O2-induced cell death in neural cells. *Neurosci Lett*. 2009;460(1):66-71. doi:10.1016/j.neulet.2009.05.024.
118. Ye Y, Tang X, Sun Z, Chen S. Upregulated WDR26 serves as a scaffold to coordinate PI3K/ AKT pathway-driven breast cancer cell growth, migration, and invasion. *Oncotarget*. 2016;7(14):17854-17869. doi:10.18632/oncotarget.7439.