TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Proteomik und Bioanalytik

Comprehensive characterization of the human proteome by multi-omic analyses

Dongxue Wang

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation

Vorsitzender: Prof. Dr. Dietmar Zehn

Prüfer der Dissertation: 1. Prof. Dr. Bernhard Küster
2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 24.09.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 31.10.2018 angenommen

Contents

Abstract

Quantitating the differential expression of human genes in different cell types, tissues and organs will significantly increase our knowledge and understanding of human biology and disease. Advances in genomic, transcriptomic and proteomic technologies have led to several large-scale studies that have described the landscape of protein-coding gene expression in human tissues. Most of these quantitative analyses of gene expression, however, have been performed at the transcript level. At the protein level, even fundamental aspects such as which proteins actually exist, where these are expressed and in what quantities, are still not fully resolved. To answer such questions, a quantitative proteomic analysis of 29 histologically-healthy human tissues was performed using high-resolution mass spectrometry. The 29 tissues cover many of the major tissues and organs of the human body and include a broad range of cell types. In this thesis, a total of 15,210 proteins were quantitated with an average of 11,000 proteins per tissue; thus representing the deepest quantitated proteome to date. Via integration with sample-matched transcriptomic data, a systematic, quantitative and deep proteome and transcriptome abundance atlas was generated. In total, 17,615 and 13,664 human genes are represented by transcripts and proteins, respectively. These analyses revealed that few proteins have a truly tissue-specific expression, vast differences exist between mRNA and protein quantities within and across tissues, and the expression level of proteins are often more stable across tissues than the complementary transcripts. Beyond transcript abundance, the data suggested that multiple post-transcriptional regulation mechanisms govern protein expression levels in cells; albeit the quantitative contribution of these factors to observed protein to mRNA ratios (PTR) is still lacking. Therefore, multivariate regression analysis was performed to quantitate the effects within and across tissues of 16 known sequence features. These consisted of 172 post-transcriptional regulatory elements and a further 15 novel motifs that were identified in this study. To functionally understand the novel RNA motifs, a mRNA pulldown combined with a competition binding assay was developed to identify the proteins that bind to specific RNA motifs. In addition to two known motifs, a novel 5’ UTR motif (CUGUCCU) showed an enhanced effect on PTR. The data revealed that CUGUCCU may play a role in the initiation of translation by recruiting ribosomal and mitochondrial ribosomal proteins. In addition to protein expression regulation, these large-scale proteomic data sets can be used to provide peptide evidence and improve genomic annotation. In general, an overall low sequence coverage for all proteins is obtained with the single enzyme trypsin. Thus, an ultra-deep tonsil proteome with a median sequence coverage of 54% was generated from a multi-protease digestion combined with additional mass spectrometry fragmentation methods. Comprehensive proteogenomic analyses were performed to discover protein isoforms, single amino acid variant (SAAV) peptides derived from single nucleotide polymorphisms, and novel peptides from unannotated protein coding loci (long noncoding RNA, lncRNA and alternative translation start sites, aTIS). The analysis revealed that very few genes have secondary protein isoforms and many proteins were still identified as groups even in the ultra-deep tonsil proteome. In the tonsil tissue, only 238 SAAV sites were confidently detected at the protein level, comprising ~2% and ~7% of all exome and mRNA variants, respectively. Although 117 aTIS peptides were identified, including 36 N-terminal peptides; not a single lncRNA peptide was observed. The results indicate that proteogenomics remains challenging, requiring rigorous validation with synthetic peptides and more sophisticated computational methods. This human proteome data will be available to the scientific community to complement available human genomic, transcriptomic, and proteomic data and thus facilitate biomedical research.

Zusammenfassung

Die Quantifizierung der unterschiedlichen Expression humaner Gene in verschiedenen Zelltypen, Geweben und Organen verbessert unser Wissen und unser Verständnis der menschlichen Biologie in Gesundheit und Krankheit. Fortschritte in der Genom-, Transkriptom- und Proteomforschung haben zu vielen wichtigen Studien geführt, die die Expression von Protein-codierenden Genen in menschlichen Geweben untersucht haben. Die meisten dieser Studien haben die Expression der Boten-RNA gemessen. Auf Ebene der Genprodukte – der Proteine – sind noch viele grundlegende Aspekte ungeklärt, unter anderem welche Gene als Proteine existieren, wo und wann diese exprimiert werden und in welchen Mengen. Um diese Fragen zu beantworten wurde in der vorliegenden Studie ein umfassender quantitativer Atlas der Proteinexpression von 29 histologisch unterschiedlichen gesunden humanen Geweben erstellt. Diese Proteomkarten wurden für die wichtigsten humanen Gewebe und Organe erzeugt und decken eine Vielzahl von Zelltypen ab. Mit einer Abdeckung von insgesamt 15,210 quantifizierten Proteinen und 11,000 Proteinen pro Gewebe stellt dieser Atlas einen der umfassendsten seiner Art dar, der durch Transkriptomkarten derselben Gewebe komplettiert wurde. Auf diese Weise wurde die Expressionslandschaft von 17,615 Boten-RNAs und 13,664 Proteinen kartiert. Eine Analyse dieser Karten hat gezeigt, dass nur sehr wenige Proteine wirklich Gewebe-spezifisch exprimiert werden, dass große Unterschiede in den Mengen von Proteinen und Boten-RNAs innerhalb und zwischen Geweben bestehen, und dass Proteine oft gleichmäßiger exprimiert werden als ihre komplementären Boten-RNAs. Die Mengen der exprimierten Proteine werden nicht nur von der Menge der exprimierten Boten-RNAs bestimmt, sondern auch von einer Vielzahl an post-transkriptionalen Regulationsmechanismen – auch wenn der Beitrag der einzelnen Faktoren zu dem beobachteten Protein-mRNA-Verhältnis bislang noch nicht quantifiziert wurde. In einem ersten Versuch, die Effekte von 16 bekannten Sequenz-Merkmalen innerhalb und zwischen unterschiedlichen Geweben zu bestimmen, wurde eine multivariate Regressionsanalyse von 172 post-transkriptionalen regulatorischen Elementen und 15 neu in dieser Studie identifizierten Sequenzmotiven durchgeführt. Um die neu entdeckten mRNA-Motive funktionell zu verstehen, wurde ein Kompetitions-Bindungs-Assays entwickelt, mit dessen Hilfe die Protein-Bindungspartner identifiziert und die Bindungsstärke des Protein-mRNA-Komplexes quantifiziert werden konnte. Neben zwei bekannten Motiven wurde ein neues 5‘ UTR-Motiv (CUGUCCU) entdeckt, dass einen verstärkenden Effekt auf das Protein-mRNA-Verhältnis zu haben scheint. Die Ergebnisse des Kompetitions-Bindungs-Assay deuten darauf hin, dass das genannte Motiv eine Rolle bei der Translationsinitiation spielt, indem es ribosomale und mitoribosomale Proteine rekrutiert. Diese umfassenden Proteomkarten können auch verwendet werden, um die Annotation des menschlichen Genoms zu verbessern. Proteommessungen, die mit einem einzigen Enzym erzeugt wurden, zeichnen sich durch eine niedrige Sequenzabdeckung aus. Um diese zu verbessern, wurde mittels mehrerer Proteasen und massenspektrometrischer Fragmentierungsmethoden eine detaillierte Karte eines einzelnen Gewebes erzeugt. Anhand dieser Daten konnten umfassende proteogenomische Analysen durchgeführt werden, um Isoformen, Aminosäureaustausch-Mutationen und neue Peptide von bislang nicht als protein-codierend bekannten Bereichen des Genoms zu identifizieren. Diese Untersuchungen haben vor allem gezeigt, dass proteogenomische Analysen rigorose Validierungen benötigen, beispielsweise mittels synthetischer Peptide oder ausgereifterer Computerverfahren. Dieser Proteomatlas wird der wissenschaftlichen Community zur Verfügung stehen, um vorhandene Genom-, Transkriptom- und Proteomdaten zu komplementieren und damit die biomedizinische Forschung weiter voranzubringen.

Chapter 1 General Introduction

1.1 Human proteome

Proteins are the main structural, functional and regulatory components in living cells. Resolving the complexity and variability of the human proteome in different cells and tissues would significantly increase and enhance our knowledge of health and disease. The number of proteins that constitute the human proteome, however, is still controversial. As a result of human genome sequencing and annotation[1,2], the number of human protein-coding genes has been estimated at approximately 20,000. When only the single representative protein of each gene is considered, this number is 20,230 (nextProt v.2.16.0). If different proteoforms (individual molecular forms of proteins) are taken into account, the number increases to millions (Figure 1)[3,4]. As described in the central dogma of biology, the DNA sequence of a gene is first transcribed into RNA transcripts. Through this process, several transcripts (isoforms) can be generated from one gene resulting in a vast increase in complexity that mainly arises from alternative splicing of RNA, alternative use of promoters or translation start sites. According to GENCODE (v28), there are 82,335 protein-coding transcripts that, in turn, could be translated into 82,335 functional proteins. After translation, the post-translational modifications (PTMs) of proteins further extend the diversity, complexity and heterogeneity of proteins by proteolytic cleavage or by addition of a modifying group to one or more amino acids[5]. More than 200 types of PTMs have been identified, such as phosphorylation, glycosylation and acetylation. These determine the activity state, localization, turnover and interaction with other proteins and thereby influence many aspects of cellular function. There are 188,594 PTM sites with protein evidence in the current release of nextProt. Beyond PTMs, proteins such as immunoglobulins and T-cell receptors also undergo somatic recombination which significantly increases the number of protein variants in certain cell types[6,7]. If the variation at the DNA levels across the human population and existing in an individual human beings are taken into consideration, genetic variations such as single nucleotide polymorphisms (SNPs), insertions and deletions give another tremendous increase in the diversity of human proteome[8,9]. Such genetic variations not only change protein sequence but also influence transcription and PTM processes. If only alternative splicing, single amino acid polymorphisms (SAPs) encoded by SNPs, and PTMs are taken into consideration, the number of human protein species can be as high as 1 to 6 million[10].

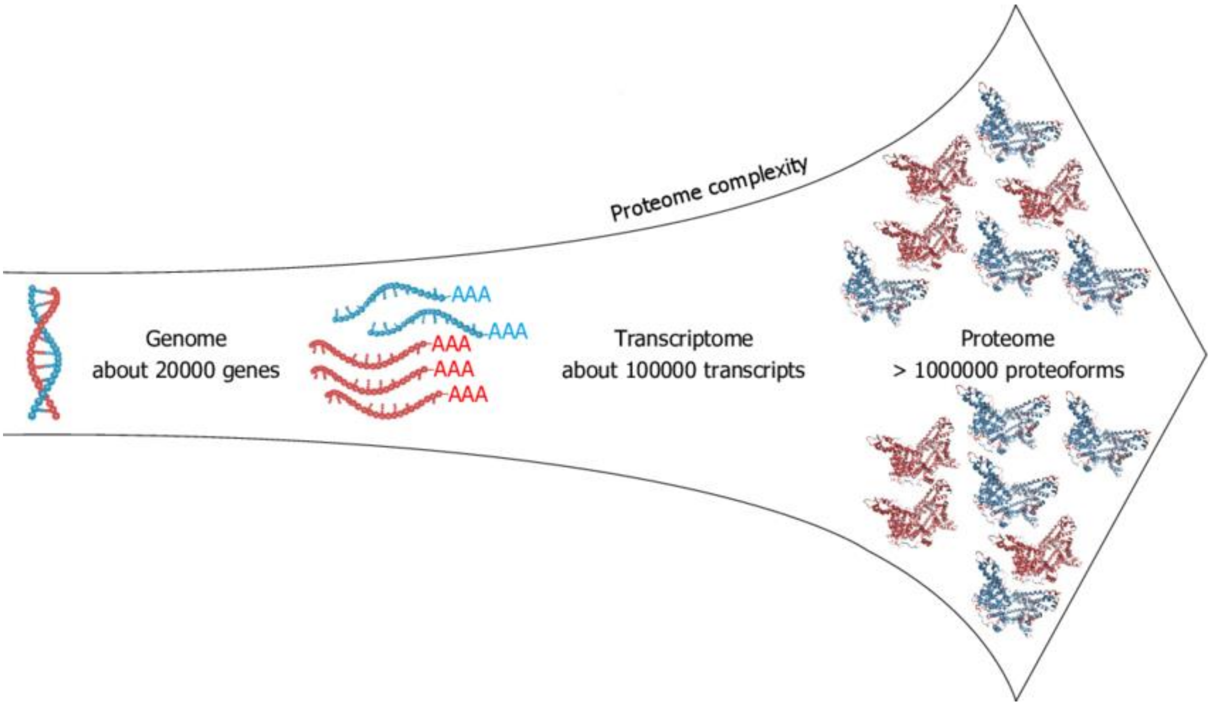


Figure 1. The human proteome is much more complex than the genome. Changes at the transcriptional and mRNA levels increase the size of the transcriptome relative to the genome, and the myriad of different post-translational modifications exponentially increases the complexity of the proteome relative to both the transcriptome and genome (Figure from[11]).

To explore the diversity of the human proteome at a genome-wide level, several initiatives have focused on describing the landscape of human protein expression. The Human Protein project[12] aimed at mapping the entire human proteome in a systematic effort using currently-available and emerging techniques; and the Human Protein Atlas Project[13] investigated the human proteome by antibody-based protein expression profiling. In 2014, the first draft maps of the human proteome based on mass spectrometry were published with publicly-available databases, Proteomics DB[14] and the Human Proteome Map[15]. These both confirmed that most protein-coding genes are translated into function proteins. In Wilhelm et al., the available raw MS data from databases plus contributions from colleagues (~60%) were combined with internally-generated MS data generated. In total, consisting of 60 human tissues, 13 body fluids, 147 cancer cell lines and ~13,00 affinity purifications. A total of 18,097 human genes comprising 92% genes annotated in SwissProt were identified by 770,000 unique peptides. Meanwhile, Kim et al. analyzed 30 histologically-normal human samples and revealed protein evidence for 17,294 genes that accounted for approximately 84% of the total annotated protein-coding genes in humans. In addition, both studies identified evidence to suggest that there is translation from non-coding DNA regions. Data included 403 long intervening noncoding RNAs (lincRNA) and transcripts of uncertain coding potential (TUCP)[14] and 808 novel annotations (*e.g.*, novel ORF, pseudogenes, novel coding regions, novel N-Terminal)[15]. Subsequently in 2015, the Human Protein Atlas generated an integrated human proteome covering 16,975 genes based on 32 healthy tissues. The study combined high-throughput mRNA sequencing (RNA-Seq) with antibody-based protein expression profiling using standardized immunohistochemistry[16]. A key finding was that, in most cases, one of the isoforms dominated across all tissues. Note, however, that the quantitation information for the proteins was derived from RNA-Seq which may not truly reflect actual protein expression levels. Recently, a subcellular map of the human proteome, including the *in situ* location of 12,003 human proteins, was reported by the Cell Atlas[17]. In this map, more than 50% of the analyzed proteins localizing to more than one compartment at the same time were identified. This suggested that proteins localize at multiple sites, thereby increasing the complexity of the cell from a systems perspective. In addition to a comprehensive human proteome map across multiple tissues and cells, detailed maps of a specific tissue or organ have also been generated, *e.g.*, a quantitative heart proteome map studying different regions and cell types of health and disease heart[18]. Despite these great achievements, the human proteome is still far from being completely described. 2,186 and 574 protein entries in neXtProt (2018-1) (PE 2-4, proteins based on evidence at transcript level, homology and prediction) and uncertain proteins (PE5), respectively, are still considered 'missing'[19]. These proteins are very difficult to detect because of either low abundance and/or transient expression, or expression in only a few cell types, or excluded by current computational tools[20]. Furthermore, a quantitative proteome map of different cell lines (~200 cell types in human) and tissues that relates to which human proteins actually exist, where these are expressed and in what quantities is still missing. Considering the complexity of human proteome, the ultimate goal is to generate a comprehensive quantitative map of protein species translated from all protein-coding genes with the information of splicing isoforms, variants and PTMs.

1.2 Mass spectrometry-based proteomics

Mass spectrometry-based proteomics has emerged as the main method for protein identification, quantitation, and characterization in complex biological samples. Currently, there are two prevalent approaches in MS-based proteomics, the 'top-down' and 'bottom-up' approaches[21,22]. The top-down approach analyzes intact proteins and enables the identification of proteoforms; although it is difficult to apply a proteome-wide analysis due to difficulties with protein fractionation, protein ionization and fragmentation in the gas phase. A more common approach is bottom-up proteomics. Here, proteins are characterized by analyzing peptides released by proteolysis; often referred to as 'shotgun proteomics'[23]. This method enables the parallel acquisition of quantitative information on thousands of proteins and post-translational modifications (PTMs) from a minimal quantity of input material. With advances in mass spectrometry and liquid chromatography, more than 10,000 proteins covering up to seven dynamic range of protein expression can be quantitated from 100 µg cell lysates[24]. Due to the scope of this thesis, the main focus is bottom-up proteomics. A typical workflow covers proteins extraction from a tissue or cell line, proteolytic digestion into peptides by site-specific proteases such as trypsin, high-resolution peptide separation (most commonly using liquid chromatography), peptide ionization (typically by electrospray ionization, ESI), tandem mass spectrometry analysis (MS/MS), and bioinformatic data analysis (Figure 2). The following paragraphs will encompass the techniques employed in the respective workflow steps and highlight the techniques related to the work in this thesis.

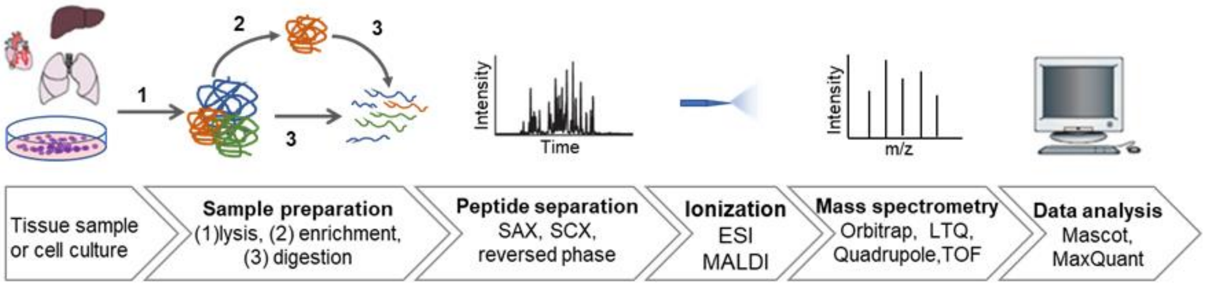


Figure 2. Typical bottom-up proteomic workflow including sample preparation, peptide separation, sample ionization, mass spectrometry analysis, and data analysis to identify and quantitate proteins in the samples (modified from[5,15]).

1.2.1 Sample preparation

Protein extraction from cells or tissue samples is the first and crucial step for MS-based proteomic experiments. The choice of extraction procedure determines the population of proteins and the state of proteins that will be analyzed and therefore sample complexity. The lysis buffer should be carefully chosen to enable rapid, efficient cell lysis and solubilization of the proteins, and ensure compatibility with downstream protein fraction or proteolytic digestion. For a deep proteome analysis, radio-immunoprecipitation assay buffer (RIPA buffer) and urea lysis buffer (applied in this thesis) are the two most effective and widely-used lysis buffers for extracting proteins from cells and tissues. Some studies, however, require functional proteins whereby the native three-dimensional protein structure, post-translational modifications and/or stable protein complexes are retained, *e.g.*, kinobead pulldowns or other affinity enrichments. To avoid protein denaturation in these experiments, detergents such as urea should not be used in the lysis buffer. Additionally, protease and phosphatase inhibitors must be included. To increase the protein extraction efficiency, additional physical homogenization can be applied[27,28].

Subsequently, proteins are digested into peptides by proteases. The digestion can be performed in solution, from a gel or with the aid of a filter[29]. The latter two methods allow the use of detergents such as SDS. Solution digestion was chosen in this study as it is robust, reproducible, and effective; and compared to *in situ* gel digestion, it is much less-time consuming. To digest proteins, trypsin is the most widely-used enzyme. High specificity proteolytic cleavage occurs after lysine and arginine residues[30]. There are also many alternative proteases frequently used in proteomics. These include Lys-C, Glu-C, Arg-C, Lys-N, Asp-N, and chymotrypsin, which have been reported to increase protein coverage and protein identification when combined with tryptic digestion[31–33]. Digestion with multiple proteases can be performed by using each protease individually or in tandem with trypsin[34]. Furthermore, nonspecific enzymes (such as pepsin) have also been investigated, with the expectation of increasing protein sequence coverage by overlapping peptides; however, this typically results in an increase in complexity and subsequent data analysis is difficult[35].

1.2.2 Peptide separation

Deep proteomic analysis of cell lines and tissue samples is still very challenging because of the large number of peptides that are generated by proteolytic digestion and the broad dynamic range of protein expression. Chromatographic peptide separation prior to mass spectrometric analysis is often used to reduce sample complexity and thus improve proteome coverage. A common approach is to separate peptides by ion-pairing reversed-phase liquid chromatography (LC). Silica beads with hydrophobic C18 alkyl chains are usually used as the stationary phase; and a mixture of water and acetonitrile supplemented with an amphiphilic acid (usually formic acid or trifluoroacetic acid) is most often used as the mobile phase. The separation power is driven by interactions of the nonpolar side chains of peptides with the nonpolar stationary phase. By increasing the concentration of the organic solvent acetonitrile, the peptides are separated based on hydrophobicity. When the LC is coupled on-line to a mass spectrometer, the peptides are directly ionized and then enter the mass spectrometer[36]. To further boost proteome coverage, peak capacity and quantitation, an additional dimension of off-line peptide separation prior to on-line LC-MS/MS can be introduced[37]. Ideally, the off-line separation should be orthogonal to the ion-pairing reversed-phase and offer high chromatographic resolution. Widely-used techniques include peptide separation by charge, such as strong cation exchange (SCX)[38] or strong anion exchange (SAX)[39], and peptide separation by hydrophily/hydrophobicity such as hydrophilic interaction liquid chromatography (HILIC)[40] and high-pH reversed-phase separation[41]. Hydrophilic strong anion exchange (hSAX) chromatography, where peptides are separated by both hydrophilicity and charge, has shown both good chromatographic resolution and orthogonality to on-line low pH reversed phase chromatography. hSAX has enabled the identification of more than 9,000 proteins in a mouse cell line[42] and more than 10,000 proteins in human brain tissue by pooling fractions[43]. For these reasons, hSAX was also the off-line peptide separation performed in this thesis. In addition, there are some efforts to extend beyond two-dimensional separation to gain an even greater depth in proteome coverage[23,44].

1.2.3 Mass spectrometry

At the simplest level, a mass spectrometer can be considered a molecular scale where the mass-to-charge (*m/*z) ratio of ions are measured to subsequently enable the identification and quantitation of molecules from simple and complex mixtures. Mass spectrometers consist of 3 major components: an ion source, a mass analyzer and an ion detector. Samples are introduced into the mass spectrometer directly via a solid probe, a syringe, or via a chromatographic device (*e.g.*, a HPLC system). The analytes are ionized in the instrument source and the generated ions separated by the mass analyzer according to *m/z* ratio. The relative abundance of each of the resolved ion species is recorded by the ion detector.

Ionization. The biomolecules need to be ionized to acquire either positive or negative charges. Once charged the ions travel through the mass spectrometer and are analyzed by a detector according to the *m/z* ratios of each of the charged species. Initially, it was very difficult to analyze peptides and proteins by mass spectrometry because of the inability to ionize and vaporize such labile molecules. The development of soft ionization technologies such as matrix-assisted laser desorption/ionization (MALDI)[45] and electrospray ionization (ESI)[46] was a huge breakthrough, and thus mass spectrometry-based proteomics became possible. In MALDI, analyte molecules are co-crystallized with suitable matrix molecules and subsequently irradiated by a pulsed laser to trigger ablation and desorption. The analyte molecules are ionized via protonation or deprotonation in the hot plume of ablated gas[47]. MALDI typically produces singly-charged ions. As ionization from the liquid phase and efficient, robust and highly automatable on-line coupling to peptide separation, ESI is the most prevalent soft ionization technique in MS-based proteomics (Figure 3). ESI performs well with large, polar molecules because of high-charge capacity. The sample solution is sprayed through an ESI capillary that is highly-charged at atmospheric pressure. Once the solvent arrives at the tip of the capillary, the droplets form a cone shaped stream (Taylor cone)[48] and a fine mist of droplets is generated that are charged on the surface. Desolvation gas (heated or a dry gas) is often applied to the charged droplets to cause solvent evaporation[49]. As the solvent evaporates, the charge density on the surface of the droplets increases to the Rayleigh limit[50]. At this point, electrostatic repulsion causes the larger droplets to disintegrate into smaller charged droplets. With repeated fission and/or evaporation events, highly-charged nanodroplets are generated. The final charged analyte can be achieved either in a passive process where the charge resides on the analyte due to subsequent solvent evaporation (charge residue model), or by charge repulsion on the droplet surface that leads to an active ion-generation process (ion evaporation model). ESI primarily results in doubly- or more highly-charged peptides, thereby facilitating the detection of large molecules with mass spectrometers of limited *m/z* range and enabling ion dissociation for tandem MS measurements. The efficiency of ionization can be enhanced by the addition of a low percentage of dimethylsulfoxide (DMSO)[51].

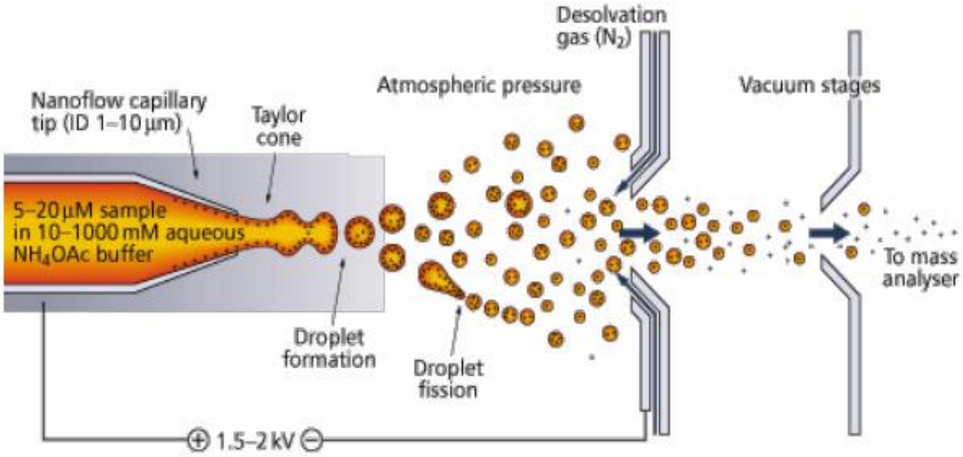


Figure 3. Schematic diagram of the electrospray ionization process (Thermo Fisher Scientific)

Mass analyzer. Then ionized peptides enter the vacuum chamber of the mass spectrometer and are separated according to *m/z* ratio by the mass analyzer. Mass analyzers should be able to separate two adjacent *m/z* signals (resolution), determine the true mass of the analyte with high accuracy and detect low-abundance ions (sensitivity). There are 4 commonly-used mass analyzers, time-of-flight (TOF), quadrupole, linear ion trap and orbitrap. Time-of-flight mass analyzers measure the time that accelerated ions with the same initial kinetic energy, but different masses take to travel from the ion source to the detector[52]. As the start time is critical for this calculation, ideally ions should form via a pulsed technique such as MALDI or rapid switching of electric fields. Quadrupole mass analyzers consist of four parallel metal rods. A radio frequency (RF) voltage with a DC offset voltage is applied between one pair of rods and the other. Only ions with a specific *m/z* value have stable trajectories through the rods for a given ratio of voltages, while other ions will collide with the rods. It is mostly used as mass filter in hybrid instruments[53]. By the application of electrodes at the end of the rods, the ions can be trapped in the quadrupole, a concept implemented in linear (2D) ion trap mass analyzers (LIT). The linear ion trap can be used as a mass filter or as a trap by creating a potential well for the ions along the axis of the trap[54]. The four quadrupolar rods of the 2D ion trap are cut into three sections. Each segments has an own DC to create a potential well and confine ion packets. Ions can spread axially to increase ion capacity. The main RF (AC) is applied to all rods and segments to provide the radially-confining force. The main RF also induces 'secular motion' that is proportional to the main RF amplitude and the mass of the ion. Whether an ion is trapped or expelled is dependent on the *m/z* ratio of the ion at a given main RF. Current ion traps apply 'resonance ejection' scanning by adding additional AC at the exit rods. To perform a scan, the main RF steadily increases to measure different ions. Another development is the orbitrap, an ion trap mass analyzer that consists of two outer barrel-like electrodes and a coaxial inner spindle-like electrode (Figure 4)[55,56]. It can act as both an analyzer and a detector. Electrostatic forces lead to the ions oscillating in the z-axial dimension while also orbiting around a central electrode. Ions with different *m/z* ratios oscillate at different frequencies, resulting in ion separation. Orbitraps measure the frequency of oscillatory motion in the axial dimension of each ion. In a scan, all *m/z* species are simultaneously measured. The time domain oscillation data is converted to *m/z* spectral data using Fourier transformation to produce a mass spectrum. The longer the ions are measured, the better the resolution.

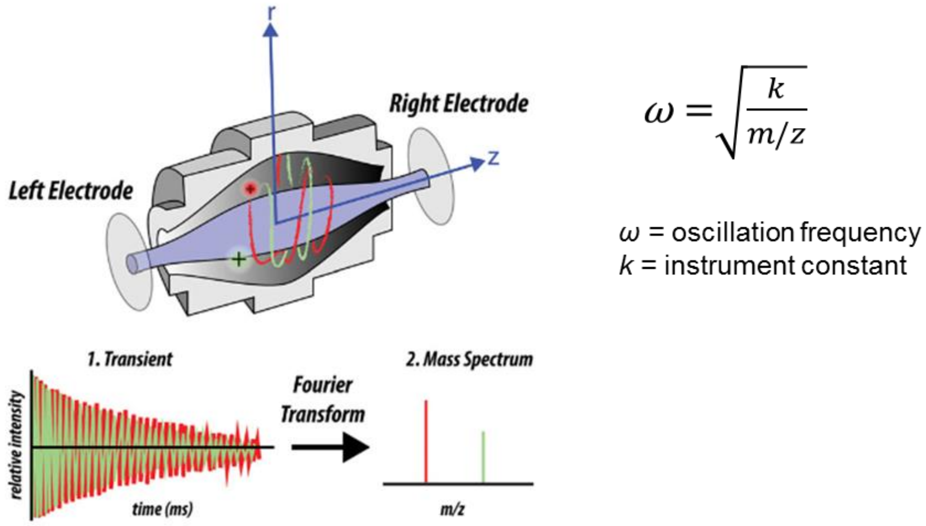


Figure 4. The orbitrap FT mass analyzer. In the orbitrap, ions oscillate around a central spindle-like electrode whilst also oscillating in the axial dimension. Axial oscillation is detected by electrodes to yield a transient that is transformed into the resultant mass spectrum by Fourier transformation. The transient is the frequency of ion oscillation in the axial dimension and will be unique for each *m/z* that is simultaneously detected (Figure modified from[56,57]).

Tandem mass spectrometry. Modern mass spectrometers often combine different mass analyzers to achieve different ion separation, resolution and detection. These include quadrupole-TOF, quadrupole-ion trap, linear ion trap-orbitrap and quadrupole-orbitrap instruments (Figure 5, Q Exactive Plus). Such hybrid mass spectrometers are widely-used in tandem mass spectrometry. Also known as MS/MS, this involves multiple steps of scanning and selection with intermediate fragmentation and enables the determination of peptide sequences. In a typical MS/MS experiment, a survey scan (MS^1) is initially performed to record the *m/z* ratios and intensities of intact peptides (precursor ions). This is then followed by selection of specific precursor ions, fragmentation, and acquisition of the MS/MS spectra (MS^2). This can be achieved by either one mass analyzer used for both operations in a consecutive manner ('tandem-in-time') or two mass analyzers within the same mass spectrometer acquiring parallel spectra ('tandem-in-space'). Most commonly, the precursor ions are automatically selected in a real-time strategy from the ions detected in the MS^1 scan; referred to as data dependent analysis (DDA). In DDA, the top N most intense precursors are selected for subsequent fragmentation. DDA is extremely powerful for measuring samples of unknown composition[58]. The implementation of dynamic exclusion enables fine tuning of the data acquisition. Peptide ions are excluded from selection and repeated MS/MS analysis for a predetermined time[59]. An obvious drawback of DDA, however, is the stochastic nature involved in precursor selection. This limits reproducibility of detection particularly for low-abundance peptides.

The principle of tandem mass spectrometry analysis is dependent on peptide fragmentation (or dissociation). The types of fragment ions (charge >= 1) observed in an MS/MS spectrum depend on many factors, for example, primary sequence, the amount of internal energy, how the energy was introduced, and charge state. When the charge is retained on the N-terminus, the ion is classified as an a-, b- or c- ion. Conversely, when the charge is retained on the C-terminus, the ion is classified as an x-, y-, or z- ion. A subscript indicates the position of the amino acid residue in the fragment (Figure 6)[25,60]. Multiple fragmentation methods such as the commonly-applied collision induced dissociation (CID), higher energy C-trap dissociation (HCD), electron transfer dissociation (ETD) and the more recent electron-transfer/higher-energy collision dissociation (EThcD) have been developed. In CID, the molecular ions are fragmented by collision with a chemically-inert molecule (often helium, nitrogen or argon) in the gas phase. CID can occur in an ion trap to predominantly produce b-, and y-ions[61]. HCD is a technique developed from CID and available in the hybrid orbitrap mass spectrometers. In HCD, ions are fragmented in a collision cell (HCD cell) rather than in an ion trap and are then return to the C-trap before analysis at high resolution in the orbitrap (Figure 6, QE+). HCD mainly generates b- and y- ions, but also produces internal and immonium fragments that can assist in high-confidence localization of labile modifications such as phosphorylation[62,63]. ETD was developed as an alternative fragmentation method to preserve more labile modifications[64]. In ETD, the peptide backbone dissociates into c- and z-ions by transferring an electron, while leaving labile post-translational modifications (PTM) intact. It is most suitable for longer peptide or polymer ions with more than 2 charges. EThcD is a novel method that employs dual fragmentation after ion selection. Both c-/z- (by ETD) and b-/y- ions (by HCD) are generated and recorded in a single MS/MS spectrum. The dual-fragment ion series enables highly-confident peptide assignment and localization of post-translational modifications (PTMs)[65]. In current proteomic studies, different fragmentation methods can be alternated during an LC-MS/MS analysis. Such an approach leads to higher sequence coverage, *e.g.*, alternating between CID and ETD according to charge state[66,67].

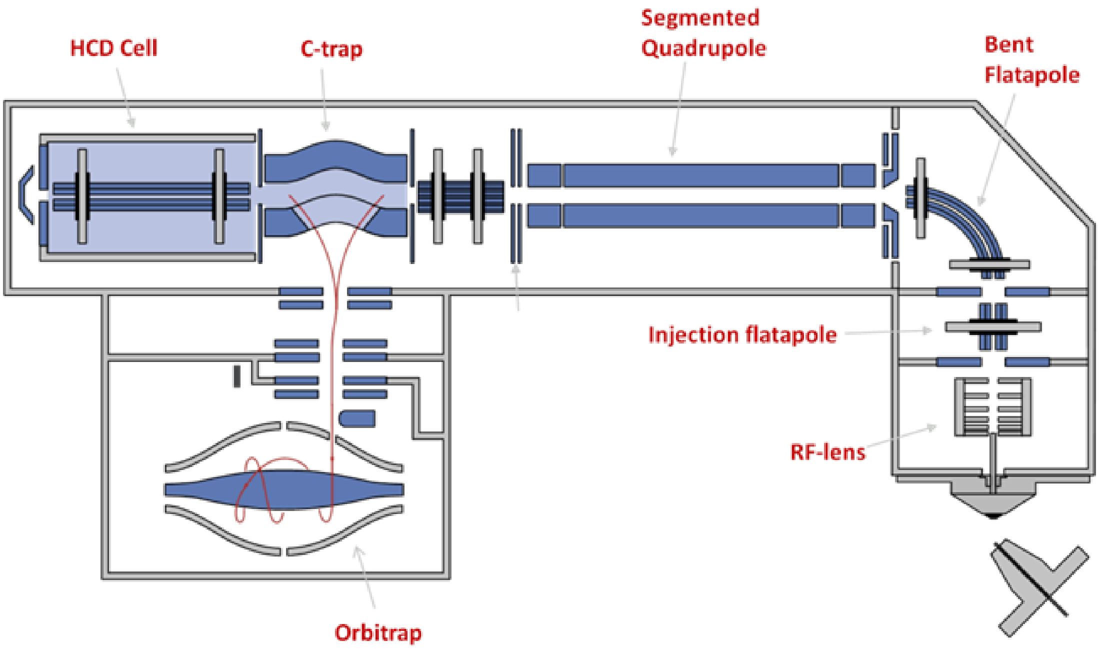


Figure 5. Schematic of the Q-Exactive Plus hybrid quadrupole orbitrap mass spectrometer. Note that the drawing is not to scale (Thermo Fisher Scientific[53]).

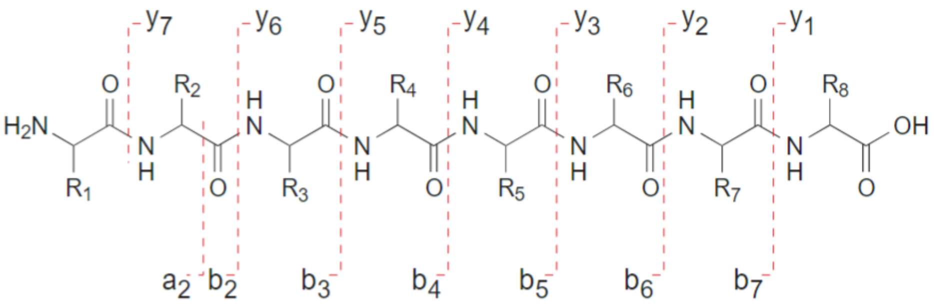


Figure 6. The peptide fragmentation process (Figure from[25]).

1.2.4 Data processing

Peptide identification. After data acquisition, the central element of proteomic data analysis is interpreting and converting MS/MS spectra to identified peptides. The strategies of peptide identification can be classed into the following categories (Figure 7): (1) database search approach, peptides are assigned by comparing experimental MS/MS spectra with theoretical spectra predicted for each peptide in a protein sequences database such as Ensembl or Uniprot; (2) spectral library searching, peptides are assigned by matching against a constructed spectral library based on previous experiments[68,69]; (3) *de novo* sequencing, which is independent of database but requires intensive computation and high-quality MS/MS spectra[70]; and (4) a hybrid approach combining *de novo* sequencing and database searching, for example, the sequence tag approach[71]. Sequence database searching is still the predominant method, and the most popular computational tools include Mascot[72], MaxQuant/Andromeda[73] and Sequest[74]. The basis of this approach is to generate a set of peptide-to-spectrum matches (PSMs) by comparing observed spectra with the theoretical fragment ion spectra generated for the candidate peptides from searched protein sequence database. The list of candidate peptides is generated by *in silico* database digestion and application of several criteria including the precursor mass tolerance, enzyme specificity, the allowed modifications and other parameters. A search score is calculated to determine the quality of each PSM. Only the highest scoring PSM for each MS/MS spectrum is taken as a candidate for the next statistical validation step. Due to incomplete fragmentation and noise, the highest-scoring PSM may still be a false positive. To estimate the false-positive rate (FDR), a target-decoy search strategy is commonly applied by searching decoy sequences created by reversing, shuffling or randomizing the target protein sequences[75,76]. This approach can also be utilized to estimate both the global FDR and posterior error probability (the statistical significance of each PSM, also referred as the local FDR)[77,78]. The main assumption underlying this strategy is that random matches (false positives) should occur with a similar likelihood in the target database and the decoy database. Therefore, the number of matches to the decoy database provides an estimate of the number of random matches expected in the target database.

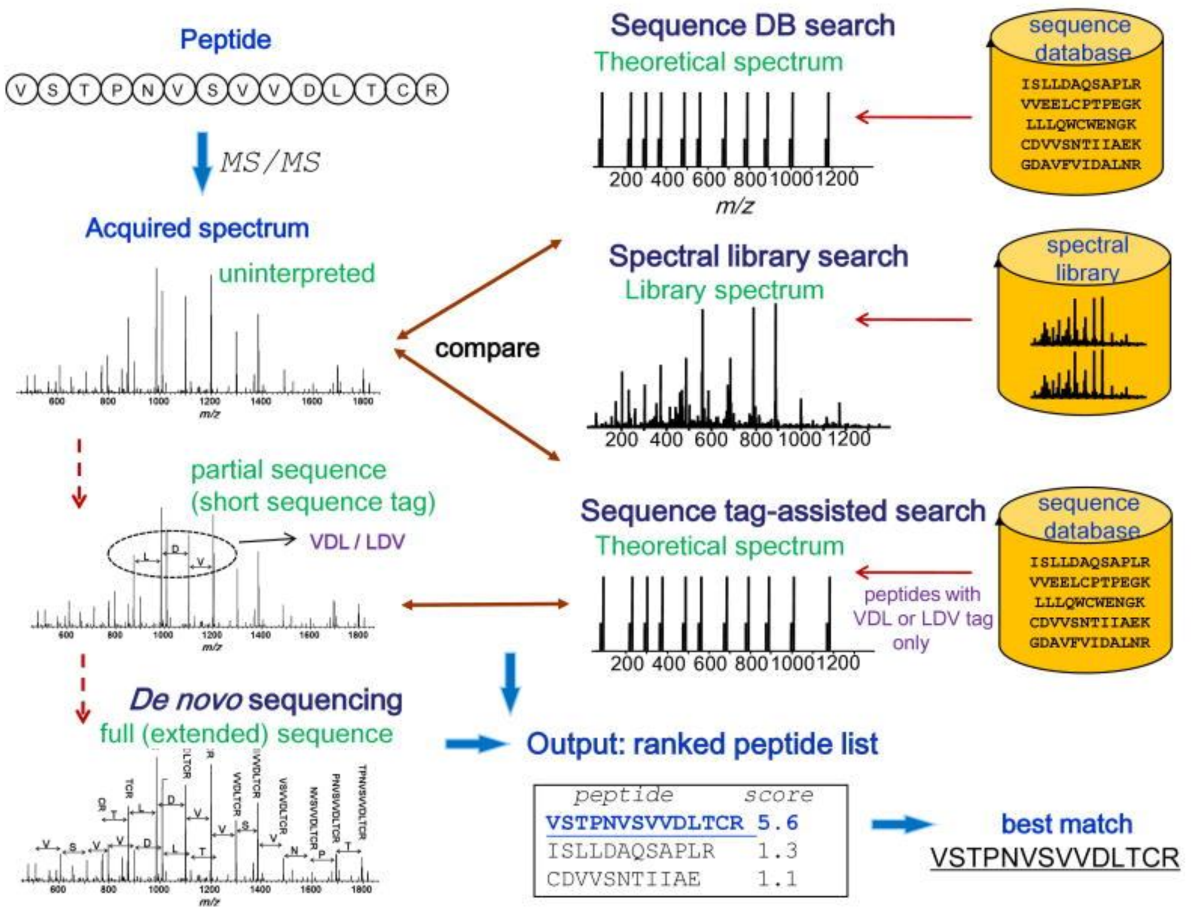


Figure 7. Conventional database search strategy for peptide identification (Figure from [79]).

Protein identification. As the peptides derived from proteolysis of proteins in bottom-up proteomics are measured, the connection between a peptide and the protein of origin is lost. To biologically interpret generated data, the identified peptides must be assembled into proteins; a challenging task especially for higher eukaryotic organisms. Some peptides can be present in more than one protein sequence in a database which will lead to ambiguous identification and quantitation. Such cases usually result from proteins that have the same domain, homologous proteins, splicing variants or redundant entries in the database. Figure 8 illustrates the different situations that can occur during protein inference[80]. Figure 8a shows the simplest situation, where the distinct proteins A and B are identified by unique peptides. Figure 8b shows when both unique and shared peptides are identified; only the unique peptides can be used to conclude presence of the protein. Figure 8c is a case where proteins A and B are impossible to distinguish as all peptides are shared. Figure 8d, e and f show that all peptides identified in protein B are shared by another protein or several other distinguishable proteins, thus it is difficult to conclude the existence of protein B. Figure 8f is a special case where all the peptides can be explained by protein A but are also shared by other proteins. Therefore, in proteomics, an identified protein is often reported as a group of proteins rather than as a single protein. The Occam's razor approach[81] is commonly applied to provide a minimal list of proteins that is sufficient to explain all identified peptides (Figure 9)[80]. The assembly of peptides into protein is accompanied by the propagation of a false positive rate. FDR control is usually consecutively applied at the PSM, peptide and protein level[82]. Both the global and local protein FDR can be calculated by the target-decoy strategy, although the FDR estimation is biased with increasing size of the data set and database[83].

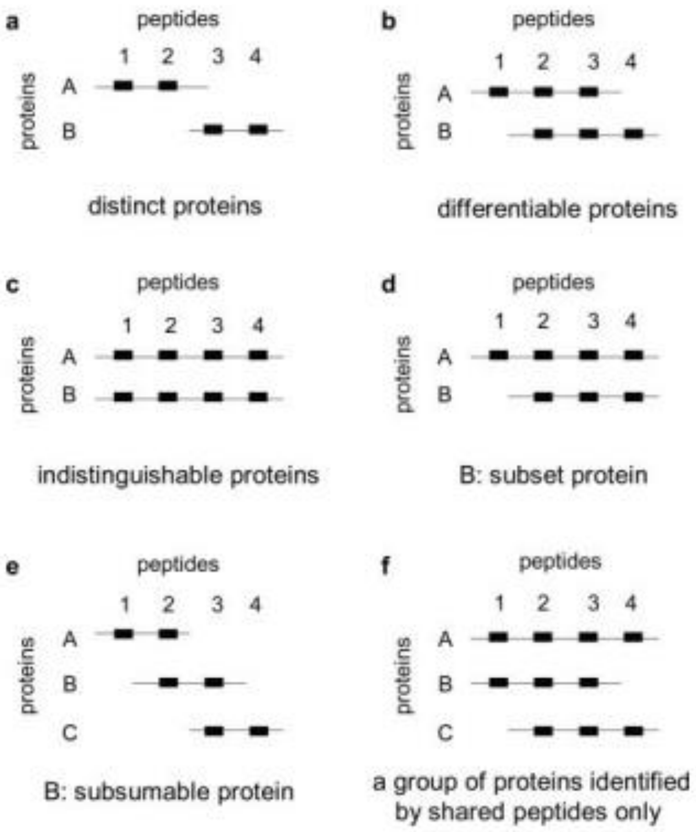


Figure 8. Basic peptide grouping scenarios. a), distinct protein identifications. b), differentiable protein identifications. c), indistinguishable protein identifications. d), subset protein identification. e), subsumable protein identification. f), an example of a protein group where one protein can explain all observed peptides, but identification is not conclusive (Figure from[80]).

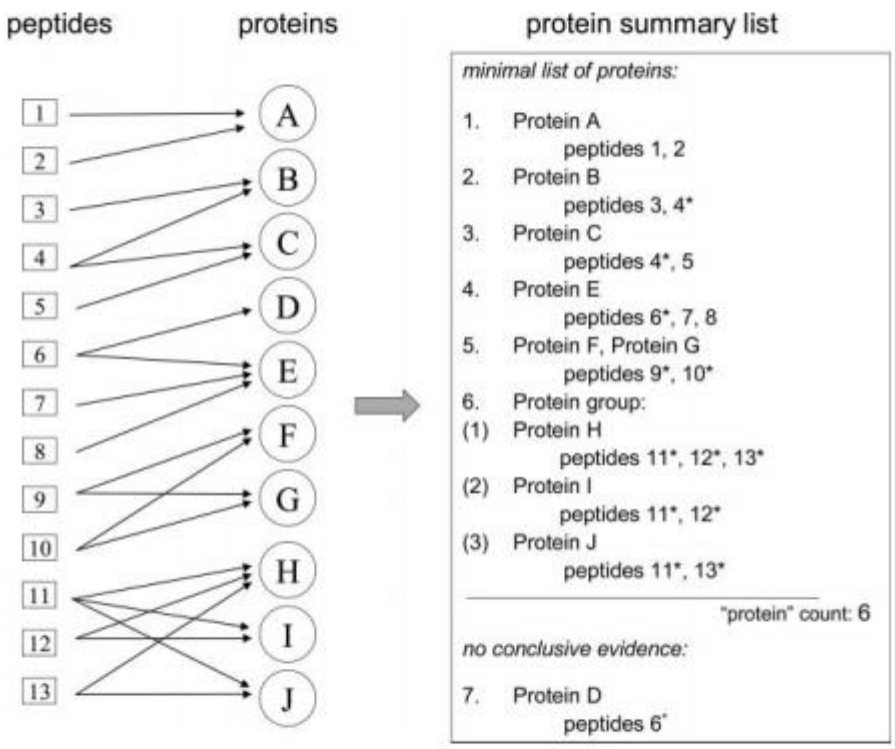


Figure 9. A simplified example of a protein summary list. Peptides are apportioned among all their corresponding proteins, and the minimal list of proteins is derived that can explain all observed peptides (Figure from[80]).

Protein quantitation. MS-based proteomics can not only identify proteins but can also enable global quantitation of the proteins. The quantitative profiling of proteome variation in different cell lines, tissues or response to stimuli can markedly increase our knowledge of biology and diseases. Quantitative proteomics can be performed by either absolute or relative quantitative approaches. While absolute quantitation aims at determining the exact concentration or copy number of proteins in a biological sample, relative quantitation measures the relative ratio change in protein concentrations. Both strategies can be achieved with the aid of labels or label free. Figure 10 shows an overview of quantitative proteomic methods[84]. In label-based quantitation, metabolic labels (*e.g.*, SILAC, stable isotope labeling by amino acids in cell culture) or chemical tags (*e.g.*, TMT, iTRAQ) are introduced into the sample either at the protein or peptide level[85–87]. The resultant mass increment of the labeled over the non-labeled version can be discriminated in the mass analyzer (SILAC by MS^1, TMT and iTRAQ by MS^2) and is used for relative quantitation within the same spectrum. The multiplexing of samples enables measurement of the labeled to unlabeled ratio in the same LC-MS/MS run, thus increasing the accuracy of quantitation. The introduction of labels early in the workflow assists in decreasing the loss of quantitative accuracy. SILAC is the most accurate quantitation method that introduces labeling at the level of living cells. Nevertheless, MS^1 complexity is markedly increased and there is a limitation in applicability (*e.g.*, tissues). While TMT (tandem mass tags) or iTRAQ (isobaric tags for absolute and relative quantitation) can be applied up to 11(TMT) or 8 (iTRAQ) different experimental conditions without increasing MS^1 complexity, both show a decrease in dynamic quantitative range due to co-elution and co-isolation of peptides[88]. With advances in resolution and accuracy, LC-MS/MS label free quantitation has developed as a fast and low-cost alternative method to measure protein expression levels in complex biological samples[89]. It is ideal for large-scale analysis in clinical screening and biomarker discovery. As such, the label-free approach was chosen in this study. Unlike labeling methods, label-free samples are processed separately and data acquisition time is thus increased. Careful control from sample preparation to LC-MS/MS analysis is critical to ensure accuracy and comparability between samples. Label-free quantitation can be performed using spectral counting or MS^1 intensities[90]. The first approach compares the sum of the MS/MS spectra from a given peptide across multiple samples and does not require specialized algorithms or other tools. As this method relies on the number of spectra, the robustness of the quantitation may be affected by the stochastic nature of DDA. Additionally, there is a bias towards highly-abundant proteins and a limitation caused by detector saturation effects[91]. In contrast to spectral counting, MS^1 intensity-based quantitation utilizes the peptide peak area as a direct read-out of the peptide abundance. iBAQ is one of the popular MS^1 intensity-based methods that are used to summarize peptide intensities to the protein level in label-free quantitation[92,93]. It is now implemented into MaxQuant. Since the long proteins are likely to have more identified peptides than short proteins, to compensate this issue, the iBAQ value of a protein is calculated by dividing the sum of all peptide intensities by the number of theoretical peptides. Another frequently used MS^1-based quantitation method is label-free quantitation (LFQ) using MaxLFQ algorithms in MaxQuant[94]. For compatibility with any upstream separation, it features a ‘delayed normalization’ that makes the assumption that the abundance of the majority of proteins does not change between different biological samples, and extracts the maximum ratio information from peptide signals to achieve accurate quantitation. To increase the number of peptides that can be used for quantitation beyond those identified by an MS/MS spectra, peptide identifications can be transferred to unsequenced or unidentified peptides by matching mass and recalibrated retention times between different LC-MS/MS analyses[94,95]. The method is referred as match-between-runs in MaxQuant, and can boost protein identification, decrease the number of missing values and reduce the variation in quantitative values.

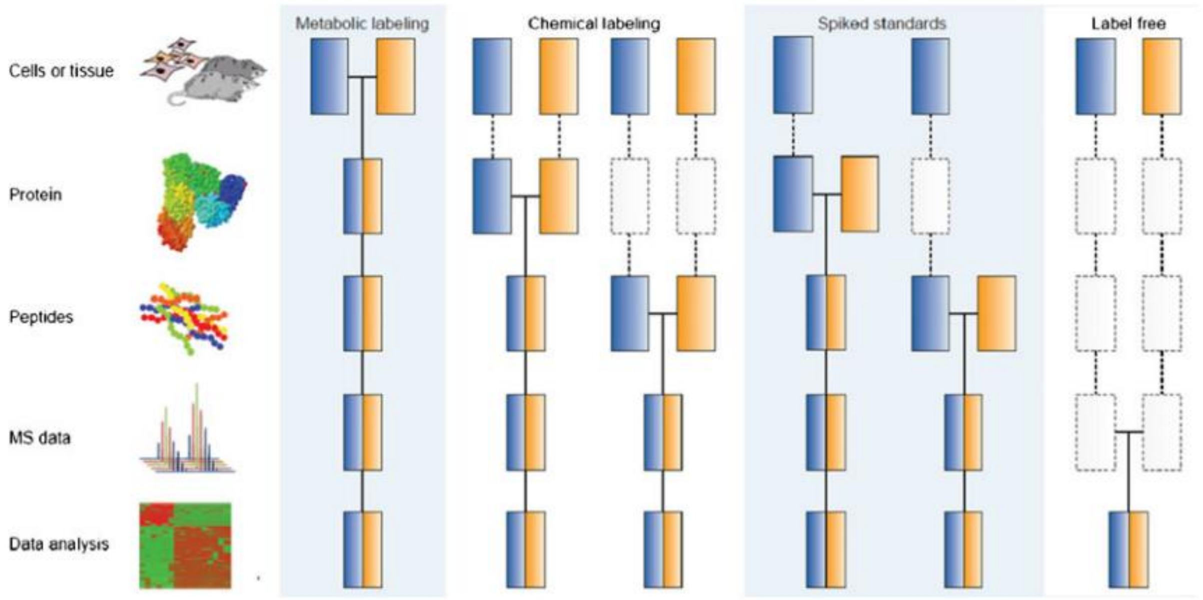


Figure 10. Overview of quantitative mass spectrometry workflows. Blue and yellow boxes represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate the points where experimental variation and thus quantitative errors can occur (Figure from[84]).

1.3 The relationship of protein and mRNA abundance

DNA, mRNA, and proteins are tightly linked by the central dogma of molecular biology. The DNA sequence of a gene determines the subsequent mRNA sequence, and the mRNA encodes the amino acid sequence that results in the proteins that determine cellular function and cell identity. Considering the relationship between the concentration of mRNA of a given gene and the concentration of the corresponding protein, the RNA concentration has been used as a surrogate measure of the final protein concentration in the cell in biological and clinical research[96,97]. Numerous studies, however, have concluded that the correlation between mRNA and protein levels are insufficient to predict protein expression levels from quantitative mRNA data[93,98–104]. The often-observed correlation coefficient ranges from 0.4 to 0.6, implying only 40% to 60% of the variation in protein abundance can be explained by mRNA abundance. In the literature, an increasing consensus is that multiple biological processes beyond mRNA concentration contribute to establish the protein expression level by regulating translation and protein degradation (Figure 11)[102].

During eukaryotic translation, multiple regulation can control protein expression through initiation, elongation, and termination by mediated via different sequence features. Most regulation is exerted at the initiation step, where the ribosome binds to the 5′ end of the mRNA and scans through the 5′ untranslated region (5′ UTR) for the start codon[105]. Generally, the start codon is AUG. This is usually located in highly-conserved short sequences known as Kozak sequences[106]. The AUG sequence context can influence the efficiency of AUG recognition. Several features in 5' UTR can influence the translation initiation, for example, the secondary structures in 5' UTR can reduce the efficiency of initiation by impeding the ribosome scan[107,108], and internal ribosome entry sites (IRESs), can regulate translation by stimulating a cap-independent translation[109]. Besides, many AUG and non-AUG start codons exist in 5' UTR, often with lower initiation efficiency, which can reduce the translation of the main open reading frame (ORF) by initiating a translation of a different isoform or an upstream reading frame (uORF)[110–113]. The rate of elongation is thought to be maximal under most conditions, but can be interfered with by several factors, such as the occurrence of rare codons, secondary structure in the coding region of a mRNA, and the folding of the nascent protein[114–118]. Termination is triggered by recognition of a stop codon and is not rate-limiting under most circumstances for protein synthesis. But it can be suppressed leading to frame-shifting or read-through due to the unfavorable sequence context of stop codon[119–121]. Furthermore, mRNA processes and dynamic mRNA modifications can regulate the mRNA stability and translation efficiency, for example, the length of poly A tail and N6-methyladenosine (m6A)[122,123]. Finally, translation rates can be modulated through the binding of RNA-binding proteins (RBPs) or non-coding RNAs (*e.g.*, miRNA) to regulatory elements on the transcript, thereby influencing the stability of mRNA, translation initiation or elongation[124–126].

Complementary to translation, protein degradation is another important process that determines final protein abundance. The ubiquitin–proteasome system (UPS) is responsible for the degradation of 80% to 90% proteins in eukaryotic cells[127]. Such degradation has high selectivity, with individual protein half-lives ranging from minutes to years[128]. Proteins with shorter half-lives often have a specific residue at the N terminus, which is known as N-degron[129]. Degrons can be a degradation signal alone and can be implicated in the ubiquitination process for targeting a protein to the proteasome. In many cases, post-translational modifications such as phosphorylation are required to create a functional degron[130]. Another a degradation signal is the 'PEST' sequence, which directs protein to the ubiquitin-proteasome pathway resulting in rapid turnover[131]. In addition, disordered protein regions can also destabilize a protein[132].

Given the diverse regulatory mechanisms that govern protein abundance in living cells, a straightforward correlation between protein and mRNA abundances cannot really be expected. To better understand the factors governing protein expression, it is of tremendous value to define relative contribution. Vogel et al., analyzed ~200 sequence features including length, nucleotide composition, amino acid composition, RNA secondary structure, uORFs, putative target sites of miRNAs, codon usage, and protein degradation signals in the medulloblastoma Daoy cell line[133]. The findings from this study were that mRNA abundance and sequence features can explain ~ 60% of the protein abundance variation in a human cell line. Due to the limits of mRNA and protein-profiling technologies, however, their analysis only covered 476 protein-coding genes and was confined to known sequence elements. With the advent of high-throughput techniques for genomic, transcriptomic and proteomic analyses, over the last years several more studies have explored the contribution of RNA- and protein-level regulation to set steady-state protein concentration[93,134,135]. Nevertheless, a comprehensive categorization of the regulatory elements or the contributions thereof in determining protein abundance at a genome-wide level are still not fully known.

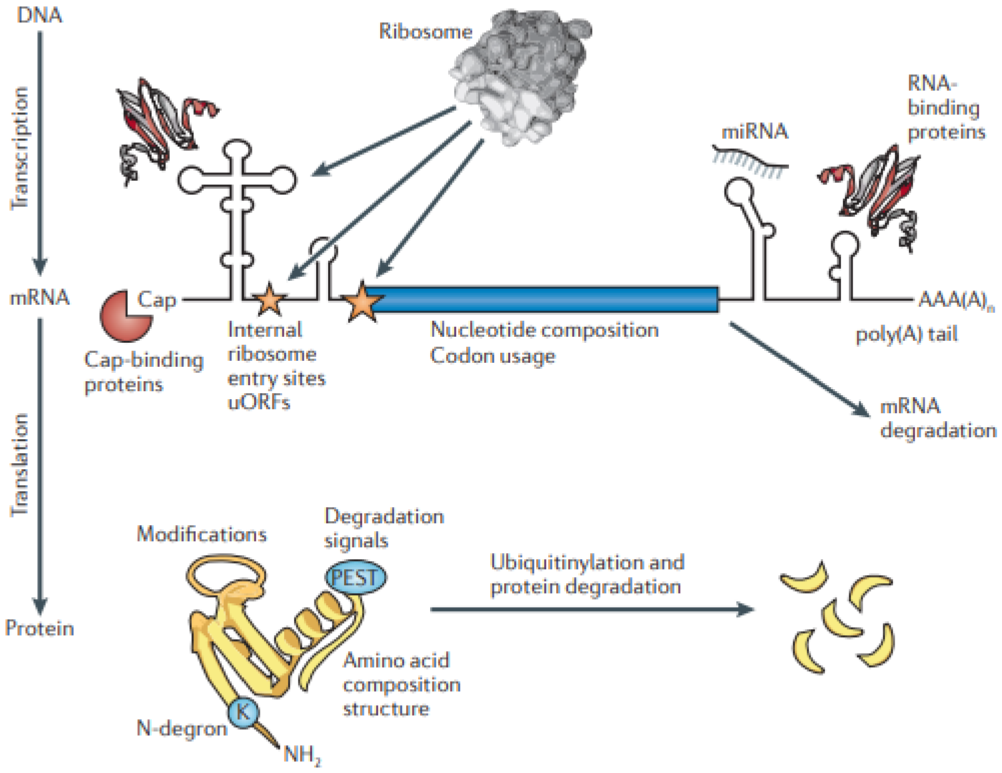


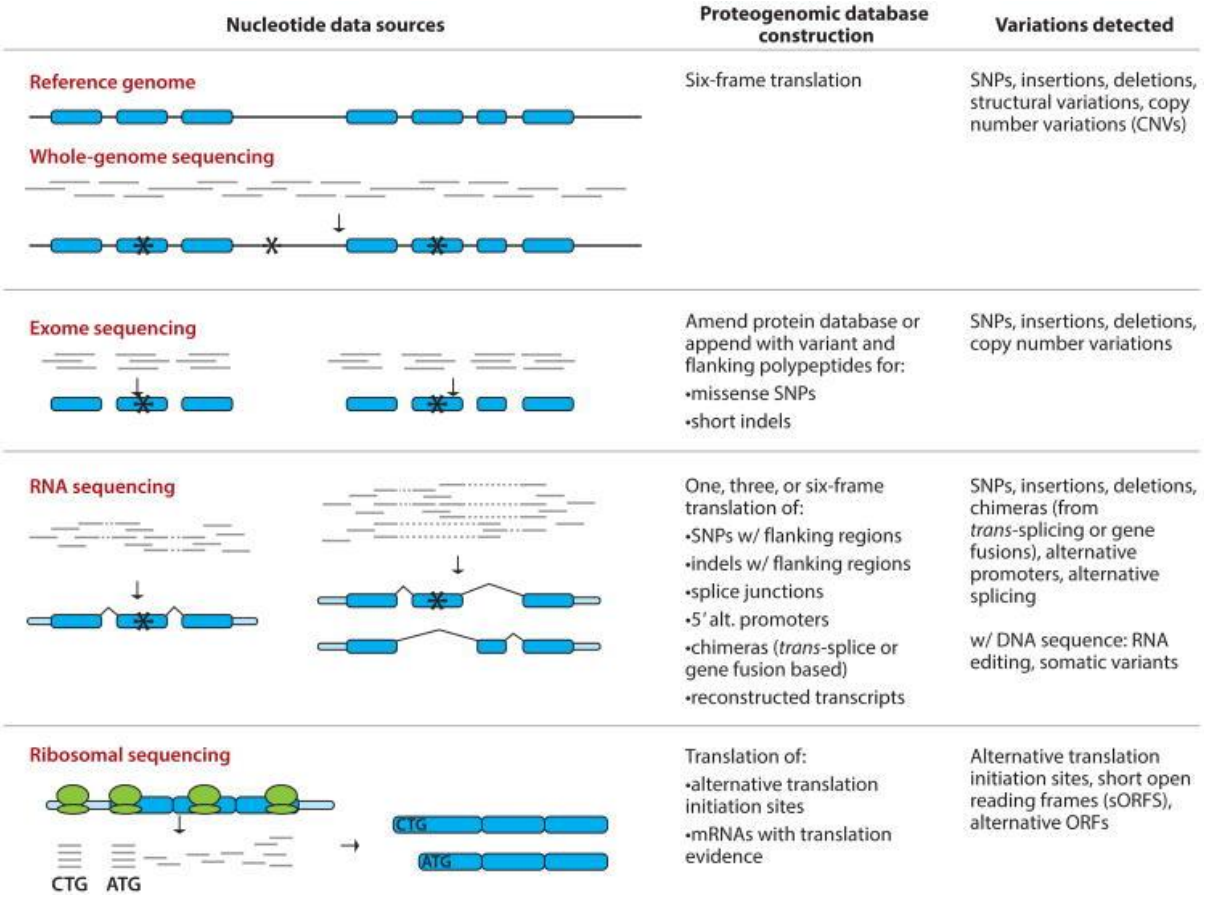
Figure 11. Protein abundances are determined by a balance of regulation of both RNA and protein production and turnover, and some of the major determinants of protein abundance are illustrated here. The figure focuses on major mechanisms of the regulation of translation and transcript stability (upper panel) and protein degradation (lower panel). Transcription is covered in this figure (Figure from[102]).

1.4 Proteogenomics

Proteogenomics is a rapidly-evolving research field that integrates proteomics with genomics and transcriptomics. The term 'proteogenomic' was first introduced by George Church in

2004 and used to improve the genome annotation of *Mycoplasma pneumonia* based on protein level validation[136]. In this integrative approach, protein databases containing the potential translation products were constructed based on genome sequencing, exome sequencing RNA-Seq, and ribosome profiling, to match spectra from tandem mass spectrometry analysis. Beyond prokaryotes, the proteogenomic approach has also provided solid evidence to aid human gene annotation. These identified novel peptides represent single amino acids variants[137–139], splice isoforms[140,141], gene fusions[142], novel protein-coding genes[143], alternative translation initiation (TISs)[15,144,145], short open-reading frames (sORFs)[146], long intergenic noncoding RNAs (lincRNA)[14], pseudogenes[15,147], and many others. In turn, this approach has also improved protein identification through the construction of comprehensive databases[148,149]. In typical proteomics, the peptides and proteins are identified by a search against a reference protein database (*e.g.*, Uniprot, Ensembl, and RefSeq). As these do not include translation products of novel-coding loci and mutations, novel or variant proteins remain unidentified[149]. Additionally, the combination of proteogenomic identifications with quantitative genomic, transcriptomic and proteomic data provide novel insights into gene expression regulation, signaling pathways, and disease subtypes[138,139,150–152]. With these advantages, proteogenomics has now been widely-applied in the following studies: genome annotation, cancer/disease-related genomic variants detection at the protein level, biomarker discovery, antibody characterization, and metaproteomic investigation[153,154].

The key step in proteogenomics is constructing customized protein sequence databases. The effectiveness of the databases is dependent on how close the predicted protein sequence is to real translation products. The sequences for databases construction can be from genome, exome, transcriptome, and translatome (Figure 12)[155]. Constructing databases from sample-matched genomic and transcriptomic data are advocated as low false-positive and false-negative rates for novel findings is reduced. Genome sequences from whole genome sequencing or a reference genome that contains the original backbone of all protein sequences can be translated to protein sequences by six- or three- frame translation (6-FT or 3 FT). This strategy has been utilized to identify alternative translation start sites (aTIS) and translation from non-coding regions[15,143,156]. While 3-FT and 6-FT genome search is associated with a large search space and higher false-positive rates, this strategy has enabled the identification of novel peptides that may be missing from protein or transcriptome-based databases. The exome sequence, comprising 1% of the genome, is often used to construct variant protein sequence databases with SNPs and indels[157]. Similar to whole exome sequencing, RNA-Seq can be used to extract variant information by aligning to a reference, and then translating to a variant-containing protein sequence[158,159]. The transcript abundance can be a reference to remove noise from the database (proteins that exist in the database but are not expressed in the sample). The protein sequence databases from RNA-Seq are promising to identify splice isoforms and simplify the protein grouping problem in bottom-up proteomics[160]. Furthermore, RNA-Seq provides protein sequences from novel splice junctions and chimeric transcripts[161,162]. This type of database, however, is lacking information on proteins from mRNA without a poly A tail[163] and proteins with a low mRNA abundance. Another popular technique is ribosome profiling. Here, information on which mRNA is actively translating and the translation start sites are provided [164]. This approach can be utilized to create a highly-specific proteogenomic database for novel translation start sites and coding regions[165].



Nucleotide data sources	Proteogenomic database construction	Variations detected
Reference genome Whole-genome sequencing	Six-frame translation	SNPs, insertions, deletions, structural variations, copy number variations (CNVs)
Exome sequencing	Amend protein database or append with variant and flanking polypeptides for: •missense SNPs •short indels	SNPs, insertions, deletions, copy number variations
RNA sequencing	One, three, or six-frame translation of: •SNPs w/ flanking regions •indels w/ flanking regions •splice junctions •5′ alt. promoters •chimeras (*trans*-splice or gene fusion based) •reconstructed transcripts	SNPs, insertions, deletions, chimeras (from *trans*-splicing or gene fusions), alternative promoters, alternative splicing w/ DNA sequence: RNA editing, somatic variants
Ribosomal sequencing	Translation of: •alternative translation initiation sites •mRNAs with translation evidence	Alternative translation initiation sites, short open reading frames (sORFS), alternative ORFs

Figure 12. Schematic of the sources of nucleotide data for proteogenomics, database construction methods, and discoverable variations. Shown are noncoding regions (black lines), exons (dark blue boxes), and 5′ and 3′ untranslated regions (light blue boxes). Asterisks represent small nucleotide variations, such as single-nucleotide polymorphisms (SNPs) or indels (Figure from [166]).

Recently, several bioinformatic tools and stringent workflows have been developed to facilitate proteogenomic studies, *e.g.*, CustomProDB[167], Galaxy-P[158], PGTools[168], and IPAW[169]. These pipelines mainly focus on constructing databases from genomic and transcriptomic data and visualization of peptide data on a genomic scale, rather than the curation and validation of novel findings. There is an increasing concern about the reliability of these novel findings in proteogenomics due to the high FDR[149,170]. Similar to proteomics, target-decoy strategy is typically used to estimate the global FDR and the local FDR (posterior error probability) for novel peptide identification[171]. It has been observed, however, that false positives are likely to be enriched in a search space of novel peptides because it is much larger than the known search space[172]. To better estimated the FDR, some researchers have suggested separating peptides into different groups according to the difference in the likelihood of identifying different classes of peptides during an FDR estimation[149,172,173]. Still, the accurate estimation of FDR remains challenging in proteogenomics. To increase the specificity of identification and the sensitivity of FDR estimation, the size of search database could be reduced as much as possible, by taking the balance between the completeness and the size of database into consideration[149]. Furthermore, to increase the reliability of novel findings, more efforts need to be made to curate and validate these novel peptides. Zhu et. al integrated the curation and validation of single amino acid-substituted peptides by requiring fragment ions to directly support the residue substitutions in MS/MS spectra into the IPAW workflow[169]. Nevertheless, there is still no efficient pipeline to curate and validate novel peptides translated from novel coding regions. In previous studies, MS/MS spectra of novel/variant peptides were validated using reference spectra from synthetic peptides by manual inspection[14,15,172]. This may provide the most confident validation, however, this strategy has not been not widely implemented for all novel/variant peptides. Disadvantages are the high costs, the labor-intensive manual inspection and the lack of system comparison methods.

1.5 Aim of this work

Resolving the protein expression variation in different cell types, tissues and organs, and delineating the factors that govern protein expression and activity are of central importance in biological and medical research. Although the number of protein-coding genes in the human genome stabilizes at approximately 20,000, high-quality protein evidence has not been found for all (~13% are absent). While it is generally accepted that quantities of proteins vary greatly within and between cell types, tissues, organs and body fluids; for many tissues proteins have not yet been quantitated. mRNA concentration are the major determinants of protein abundance, and large-scale transcriptomic data of tissues have generated as proxies of proteins. Systematic studies of transcripts and protein quantitative profiling, however, have revealed that multiple post-transcriptional regulations also play important roles in determining protein expression levels. Yet these studies mainly focused on a single cell type or tissue from an associated disease. Given these points, the purpose of this thesis was to generate quantitative deep proteomes of 29 healthy tissues that are representative of the major human organs. These were then to be integrated with sample-matched transcriptomes from the Human Protein Atlas project to ultimately describe the landscape of protein expression, to explore the relationship between mRNA and protein abundance and to study protein expression control. Since typical proteomic analyses are dependent on predefined protein sequences derived from reference protein-coding genes to identify protein, proteins from unannotated coding regions or with variants cannot be identified, and the complexity of the reference database together with a large number of shared peptides leads to ambiguous protein inference. To overcome this, another aim of this work was to improve protein isoform identification, to identify genetic variation at the protein level, and to aid gene model refinement by detecting peptides from previously-assumed non-coding gene regions and transcripts by proteogenomic analysis.

Chapter 2 Materials and Methods

2.1 Human tissue specimens

The 29 human tissue samples used for mRNA and protein expression analysis were obtained from the Department of Pathology, Uppsala University Hospital, Uppsala, Sweden as part of the sample collection governed by the Uppsala Biobank (www.uppsalabiobank.uu.se/en/). All tissue samples were collected and handled using standards developed in the Human Protein Atlas (www.proteinatlas.org) and in accordance with Swedish laws and regulations. Tissue samples were anonymized in agreement with approval and advisory reports from the Uppsala Ethical Review Board (References # 2002-577, 2005-338 and 2007-159 (protein) and # 2011-473 (RNA)). The need for informed consent was waived by the ethics committee.

The two pituitary samples were provided by Professor Marily Theodoropoulou, Ludwig Maximilian University of Munich (LMU). In total, 31 histological human tissues (50 samples, including replicate tissues) were analyzed by LC-MS/MS. The list of all tissues together with the corresponding sample preparation and measurement information is provided in Table 1.

Table 1. Tissue sample and corresponding LC-MS/MS measurement number

	Tissue name	Tissue ID	RNA-Seq ID	MS ID	Proteases	Fragment methods	Instrument
1	Adrenal gland	V122	P282_105	P015424	Trypsin	HCD	QE+
2	Appendix	V154	P282_120	P013677	Trypsin	HCD	QE+
3	Bone marrow	V364		P013674	Trypsin	HCD	QE+
4	Brain	V102	P262_159	P013129	Trypsin	HCD	QE+
5	Colon	V269	P973_107	P013387	Trypsin	HCD	QE+
6	Colon*	V291		P010693	Trypsin	HCD	QE+
7	Duodenum	V145	P282_126	P013187	Trypsin	HCD	QE+
8	Endometrium	V200	P414_119	P013386	Trypsin	HCD	QE+
9	Esophagus	V184	P414_107	P013198	Trypsin	HCD	QE+
10	Fallopian tube	V277	P973_112	P013559	Trypsin	HCD	QE+

11	Fat	V315	P973_140	P015159	Trypsin	HCD	QE+
12	Gallbladder	V179	P414_104	P013197	Trypsin	HCD	QE+
13	Heart	V191	P414_111	P013201	Trypsin	HCD	QE+
14	Kidney	V359	P1973_516	P013560	Trypsin	HCD	QE+
15	Liver	V358	P1973_515	P012502	Trypsin	HCD	QE+
16	Liver	V362	P1973_519	P013127	Trypsin	HCD	QE+
17	Liver*	V289		P010738	Trypsin	HCD	QE+
18	Liver*	V290		P010748	Trypsin	HCD	QE+
19	Lung	V133	P282_114	P013163	Trypsin	HCD	QE+
20	Lung*	V299		P010740	Trypsin	HCD	QE+
21	Lymph node	V193	P414_113	P015160	Trypsin	HCD	QE+
22	Ovary	V233	P415_106	P013679	Trypsin	HCD	QE+
23	Pancreas	V229	P415_104	P013678	Trypsin	HCD	QE+
24	Pituitary ▼			P018021	Trypsin	HCD	QE+
25	Pituitary ∆			P018020	Trypsin	HCD	QE+
26	Placenta	V223	P415_102	P013680	Trypsin	HCD	QE+
27	Placenta*	V262		P010695	Trypsin	HCD	QE+
28	Prostate	V128	P282_109	P013675	Trypsin	HCD	QE+
29	Rectum	V321	P973_143	P013681	Trypsin	HCD	QE+
30	Salivary gland	V240	P415_112	P015161	Trypsin	HCD	QE+
31	Small intestine	V151	P282_136	P013188	Trypsin	HCD	QE+
32	Smooth muscle	V266	P973_105	P013562	Trypsin	HCD	QE+
33	Spleen	V82	P262_150	P013114	Trypsin	HCD	QE+
34	Stomach	V90	P262_154	P013128	Trypsin	HCD	QE+
35	Stomach*	V296		P010739	Trypsin	HCD	QE+
36	Testis*	V134	P282_115	P013164	Trypsin	HCD	QE+
37	Thyroid	V196	P414_115	P013385	Trypsin	HCD	QE+
38	Tonsil*	V287	P973_118	P010747	Trypsin	HCD	QE+
39	Tonsil	V287	P973_118	P010747	Trypsin	CID	Lumos
40	Tonsil	V287	P973_118	P010747	Trypsin	EThcD/ETD	Lumos

41	Tonsil	V287	P973_118	P018440	Lys-C	HCD	HF
42	Tonsil	V287	P973_118	P018441	Glu-C	HCD	HF
43	Tonsil	V287	P973_118	P018442	Arg-C	HCD	HF
44	Tonsil	V287	P973_118	P018443	Asp-N	HCD	HF
45	Tonsil	V287	P973_118	P018444	Lys-N	HCD	HF
46	Tonsil	V287	P973_118	P018699	Chymotrypsin	HCD	HF
47	Tonsil	V287	P973_118	P018699	Chymotrypsin	CID	Lumos
48	Tonsil*	V293		P010694	Trypsin	HCD	QE+
49	Tonsil	V301		P013107	Trypsin	HCD	QE+
50	Urinary bladder	V176	P414_102	P013196	Trypsin	HCD	QE+

Tissues samples marked with '*' were generated by Li-hua Li (former group member); Pituitary▼, a sample from a male adult with bipolar suicide; Pituitary ∆, a sample from female adult with cardiac tamponade; QE+, Q exactive plus mass spectrometer; HF, Q Exactive HF mass spectrometer; Lumos, Orbitrap Fusion Lumos mass spectrometer.

2.2 RNA sequencing

The RNA-Seq data were provided by Human Protein Atlas project. Procedures for RNA extraction from tissues, library preparation, and sequencing have been previously described[16]. Briefly, pieces of frozen human tissue were embedded in Optimal Cutting Temperature (OCT) compound and stored at -80°C. Cryosections were cut and stained with hematoxylin-eosin for microscopical confirmation of tissue quality and proper representativity. 5-10 cryosections (10 µm) were transferred to RNAse free tubes for extraction of total RNA using the RNeasy Mini Kit (Qiagen). RNA quality was analyzed with an Agilent 2100 Bioanalyzer system with the RNA 6000 Nano LabChip Kit (Agilent Biotechnologies). Only samples of high-quality RNA (RNA Integrity Number ≥7.5) were used for mRNA sample preparation and sequencing. The mRNA strands were fragmented using Fragmentation Buffer (Illumina) and the templates were used to construct cDNA libraries using a TruSeq RNA Sample Prep Kit (Illumina). Gene expression was assessed by deep sequencing of cDNA on Illumina HiSeq 2000/2500 system (Illumina) for paired-end reads with a read length of 2 × 100 bases. RNA sequencing data was aligned against the human reference genome (GRCh38, v83) using Tophat2.0.8b. FPKM (fragments per kilobase of exon model per million mapped reads) values were calculated using Cufflinks v2.1.1 as a proxy for transcript expression level. The FPKM values of each gene were summed in an individual sample and median normalization was applied to evaluate gene expression levels between tissues. A cut-off value of 1 FPKM was used as the lower limit of detection across all tissues.

2.3 Whole proteome analysis

2.3.1 Tissue lysis

5-20 mg fresh frozen tissues were added in 250 µL precooled tissue lysis solution containing 50 mM Tris/HCl (pH 7.6), 8 M urea, 10 mM TCEP-HCl, 40 mM CAA,1× EDTA free protease inhibitor mixture (complete mini, Roche), 5× phosphatase inhibitors 1 and 2, and 1× phosphatase inhibitor 3 (Sigma Aldrich). The mixtures were transfer into Precellys tubes containing ceramic beads and incubated for 5 min on ice. Precellys tubes were mounted in the Precellys 24 bead-milling device and tissue lysis and homogenization were performed at 5500 r.p.m. for 1× 25 s with 5 s pause. The lysates were incubated on ice for 30 min to remove the protein foam generated during homogenization. Protein concentration was determined using the Bradford assay. Lysates were stored at −80°C or used immediately.

2.3.2 Solution digestion with trypsin and 6 additional proteases

The tissue lysates were digested with trypsin. To decrease the urea concentration to 1.6 M and prevent inhibition of protease activity, aliquots of lysate containing 300 µg protein were diluted with four volumes 50 mM Tris/HCl, pH 7.6. If there was less than 300 µg protein in the lysate, the whole lysate was diluted for digestion. The proteins were digested with sequence-grade trypsin (Roche) at a protease-to-protein ratio of 1:50 (w/w) for 4 h in a thermoshaker at 37°C and 700 r.p.m. Another 1:50 (w/w) trypsin was added and the digest mixture was incubated overnight in a thermoshaker at 37°C and 700 r.p.m.

For ultra-deep proteomic analysis of the single tonsil tissue, Lys-C, Glu-C, Arg-C, Asp-N, Lys-N and chymotrypsin were also used to digest the lysate. Six aliquots of tonsil lysate, each with 300 µg protein, were used for digestion and diluted with 8× corresponding buffer for each protease to decrease the urea concentration to <1. For digestion with Lys-C, the lysate was diluted with 50 mM Tris/HCl, pH 8.5. 1:50 (w/w) Lys-C(Wako) was added for 4 h pre-digestion at 37°C and 700 r.p.m., followed by overnight digestion with another 1: 50 (w/w) Lys-C at 37°C and 700 r.p.m. in a thermoshaker. For digestion with Glu-C and Asp-N, 50 mM Tris/HCl, pH 7.5 was added to dilute the lysate. 4 h pre-digestion and overnight digestion were performed by adding 2× 1:50 (w/w) Glu-C (Promega) or 1:100 (w/w) Asp-N at 37°C and 700 r.p.m. in a thermoshaker. For digestion with Arg-C, the lysate dilution buffer contained 50 mM Tris/HCl (pH 7.8), 5 mM DTT, 2 mM EDTA (pH 8), 5 mM $CaCl_2$. 1:60 (w/w) Arg-C (Promega) was added 2× for 4 h pre-digestion and overnight digestion at 37°C and

700 r.p.m. in a thermoshaker. For digestion with Lys-N, 50 mM Tris/HCl (pH 8.0) was used to dilute the lysate and 1:100 (w/w) Lys- N (Promega) was added for 4 h pre-digestion and overnight digestion at 37°C and 700 r.p.m. in a thermoshaker. For digestion with chymotrypsin, the lysate was diluted with 100 mM Tris/HCl, 10mM $CaCl_2$ (pH 8.0) and 1:50 (w/w) chymotrypsin (Promega) was added for 4 h pre-digestion and overnight digestion at 25°C and 700 r.p.m. in a thermoshaker.

2.3.3 Desalting with Sep-Pak

Desalting of peptide mixtures was achieved using Sep-Pak columns (C18 cartridges Sep-Pak Vac 1cc, 50 mg, Waters) on a vacuum manifold. Briefly, samples were cooled to room temperature and acidified to pH ~2 by the addition of ~0.5% (v/v) TFA to quench digestion. To precipitate insoluble matter, the acidified peptides were centrifuged at 14,800 × g at 4°C. The Sep-Pak columns were primed with 3× 1 mL solvent B (0.07% TFA in 50% ACN) and equilibrated with 3× 1 mL solvent A (0.07% TFA). The acidified supernatants were then transferred onto columns and slowly loaded. The flow-through was collected and re-loaded twice more. After five washing steps with 5× 1 mL solvent A, the peptides were slowly eluted with 0.3 mL solvent B and lyophilized in a Speed-Vac (samples were frozen at -80 °C prior to lyophilization). The peptides were stored at -80 °C under required for further processing.

2.3.4 Hydrophilic strong anion exchange chromatography (hSAX)

The desalted peptides were dissolved in 105 µL solvent A (5 mM Tris/HCl, pH 8.5) in 1.5 mL reaction vessels, mixed on a vortexor and then centrifuged to spin down all drops. To fully solubilize the peptides, the mixtures were sonicated 3×1 min with 2× 1 min incubation on ice, following 30 s centrifugation at 5,000×*g* to remove insoluble debris. 100 µL dissolved peptides were injected onto an IonPac AS24 strong anion exchange column (Thermo Fisher Scientific) with an IonPac AG24 guard column (Thermo Fisher Scientific) coupled to a Dionex Ultimate 3000 HPLC system (Thermo Fisher Scientific). The peptides were eluted with a 50 min two-step linear gradient at a flow rate of 0.25 mL/min flow (0-3 min: 100% solvent A; 3-27 min: 0% to 25% solvent B, 5 mM Tris/HCl, pH 8.5, 1 M NaCl; 27-40 min: 25% to 100% solvent B; 40-44 min: 100% solvent B; 44-50 min: 100% solvent A). The chromatography was detected with UV absorption at fixed wavelengths of 280 and 214 nm. 40 fractions were collected with a 96-well plate starting from 2 min in 1 min intervals, and subsequently combined into 36 fractions according to the chromatographic graph (fractions 6-8 and 38-40 were normally pooled).

2.3.5 Desalting of hSAX fractions

All 36 fractions were desalted with StageTips as previous described, with minor modifications[174]. In short, fractions were cooled to room temperature and acidified with TFA to pH 2-3. Five C18-discs (Empore Octadecyl C18 47 mm Solid Phase Extraction Disks, 3 M Purification, Eagan, MN, USA) were excised with a round punch (approx. 1.5 mm diameter) and inserted in 200 µL pipette tips. The tips were placed in 1.5 mL Eppendorf tubes through a pre-cut hole in the lid. Unless otherwise stated, 250 µL solvent was used and forced through the StageTips by centrifugation at 3,000 r.p.m. for 15 min, emptying the tubes after every second centrifugation step. Care was taken to ensure that the C18 material was not drying out and that there was no air trapped between the applied liquid and the packed C18 material during each centrifugation. The StageTips were primed with MeOH and then with solvent B (0.1% FA in 60% ACN), following by 2× equilibration with 0.1% FA. All tubes were replaced prior to loading the samples. Acidified samples were slowly loaded onto the StageTips at 2,000 r.p.m. The loading procedure was repeated for samples with volumes exceeding 250 µL. The flow-through were collected and reapplied twice. Subsequently, the StageTips were washed 2× with 0.1% FA. Any remaining solvent A was forced through the StageTips with a syringe. Finally, the tubes were again replaced and the desalted peptides were eluted with 100 µL 0.1% FA in 60% ACN at 2,000 r.p.m. for 10 mins. The remaining solvent was forced through the StageTips with a syringe to ensure that all solvent and peptides were eluted. The eluted fractions were transferred to 96-well plates and frozen at −80 °C before lyophilization in a vacuum centrifuge.

2.3.6 LC-MS/MS analysis

Unless otherwise stated, peptides were dissolved in 50 µL solvent A (0.1% FA in HPLC grade water), and 5 µL were injected per measurement.

Quantitative label-free LC-MS/MS analysis was performed on a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled on-line to a nanoflow LC system (NanoLC-Ultra 1D+, Eksigent, USA). Peptides were delivered to a trap column (100 µm inner diameter × 2 cm, packed with 5 µm C18 resin, Reprosil PUR AQ, Dr. Maisch GmbH, Germany) at a flow rate of 5 µL/min for 10 min in 100% solvent A. After 10 min of loading and washing, peptides were transferred to a 40 cm reversed-phase C18 column (75 µm inner diameter, packed with 3 µm C18 resin, ReproSil-Pur C18-AQ, Dr. Maisch GmbH, Germany) and separated over a 110 min gradient from 2% to 32% solvent B (0.1 FA%, 5% dimethyl sulfoxide in ACN) in A_1 (0.1% FA and 5% DMSO in HPLC grade water) at a flow rate of 300 nL/min [51]. Full scans (*m/z* 360-1,300) were acquired at a resolution of 70,000 using an AGC target value of 3e6 and a maximum ion injection time of 100 ms. Internal calibration was performed using the signal of a DMSO cluster as a lock mass [51]. Tandem mass spectra were generated for up to 20 precursors by HCD with a normalized collision energy of 25%. The dynamic exclusion was 35 s. Fragment ions were acquired at a resolution of 17,500 with an AGC target value of 1e5 and a maximum ion injection time of 50 ms.

To achieve an ultra-deep tonsil proteome, a single tonsil was separately digested with trypsin, Lys-C, Arg-C, Glu-C, Asp-N, Lys-N and chymotrypsin as described above. Excluding the tryptically-digested tonsil sample that was analyzed by HCD on a Q Exactive Plus, the other samples were processed on a Q Exactive HF mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled on-line to a nanoflow LC system (NanoLC-Ultra 1D+, Eksigent, USA). Full scan MS spectra were acquired at 60,000 resolution and a maximum ion injection time of 25 ms. Tandem mass spectra were generated for up to 15 peptide precursors and fragments detected at a resolution of 15,000. The MS2 AGC target value was set to 2e5 with a maximum ion injection time of 100 ms. The other settings were the same as for the Q Exactive Plus.

The tryptically-digested tonsil sample was also analyzed by CID and EThcD/ETD fragmentation methods, and the chymotryptically-digested sample was measured with CID on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled on-line to a nanoflow LC system (UltiMate™ 3000 RSLC nano System, Thermo Fisher Scientific). Full MS scans were performed at a resolution of 60,000, a maximum injection time of 50 ms and an AGC target value of 5e5, followed by MS^2 events with a duty cycle of 2 s for the most intense precursors and a dynamic exclusion of 60 s. CID spectra were acquired with 35% normalized collision energy and Orbitrap acquisition (1e5 AGC target, 0.25 activation Q, 20 ms maximum injection time, inject ions for all available parallelizable time enabled, 1.3 *m/z* isolation width). EThcD/ETD scan used charge-dependent parameters, 2+ precursor ions were fragmented by EThcD with 28% collision energy and 3+ to 7+ ions were fragmented by ETD. The MS2 scans were read out in the Orbitrap (1e5 AGC target, 0.25 activation Q, and 100 ms maximum injection time). The other settings were the same as the Q Exactive Plus and Q Exactive HF.

2.3.7 Peptide and protein identification and quantitation

For peptide and protein identification and label-free quantitation, the MaxQuant suite of tools was used (version 1.5.3.30)[175]. The spectra were searched against the Ensembl human proteome database (release-83, GRCh38) with cysteine carbamidomethylation as a fixed modification. Oxidation of methionine and *N*-acetylation of the protein were variable modifications A maximum of 5 modifications per peptide was allowed. Trypsin/P was specified as the proteolytic enzyme with a maximum of 2 missed cleavage sites. Only peptides with >6 amino acids were considered for protein identification and quantitation and the required FDR was set at 1% for both PSMs and proteins. The peptide tolerance for the initial Andromeda search was 20 ppm which was used for time and mass independent recalibration. The peptide mass tolerance after calibration for main search was 4.5 ppm. The MS/MS tolerance was 20 ppm. The second peptide function was activated to identify co-fragmented peptides. The match-between-runs function was enabled, with a match time window of 0.7 min and an alignment time window of 20 min, to transfer peptide identification between fractions (+/-1) based on accurate mass and retention times across liquid chromatography (LC)-MS runs to obtain peptide identity in cases where the precursor peptides were present in MS^1 but not selected for fragmentation and identification by MS^2 in a given run. Label-free quantitation was performed using the iBAQ approach (intensity-based absolute quantitation). For single tissue analyses, matching data between fractions was disabled.

For the ultra-deep sequenced tonsil data, the spectra were also processed using the MaxQuant suite of tools (version 1.5.3.30) by searching against the Ensembl human proteome database (release-83, GRCh38). These 10 samples were set to 10 different groups to enable set different proteases and maximum missed cleavages. Lys-C/P, Arg-C and Lys-N were specified with a maximum of 2 missed cleavage sites. Glu-C and Asp-N with 3 missed cleavages. Chymotrypsin cleavage at the C-terminus of F, Y, L, W, or M was allowed with a maximum of 4 missed cleavage sites. As the peptides generated by different proteases were quite different, the match-between-runs function was disabled.

2.4 Quantitative analysis of transcriptomes and proteomes

The quantitative analyses of proteomic and transcriptomic data were performed at the gene level. To evaluate gene expression level, the total abundance of each gene in all the individual samples was used. The data was log transformed (base 10) and normalized using median centering across tissues. The transcriptomic data was normalized as described in 2.2. For the proteomic data (proteingroup.txt generated by MaxQuant), protein groups were first filtered by ‘only identified by site’ and ‘reversed’. If all proteins were of human origin for protein groups referred to as ‘potential contaminants’, the proteins were not excluded. The first protein in each protein group was considered representative of that protein group. iBAQ values of each gene were summed in an individual sample and median normalization was applied to evaluate gene expression levels between tissues. A cut-off value of 0 iBAQ was used as the lower limit for protein identification and quantitation across all tissues.

The genes were classified into ‘Tissue enriched’, ‘Group enriched’, ‘Tissue enhanced’, ‘Expressed in all’ and ‘Mixed’ as described by Uhlén et al[16,176]. Gene ontology analysis of genes only identified in transcriptomes and proteomes, and the elevated proteins expressed in each tissue was performed using the R package ‘clusterProfiler’ and p-values were adjusted according to the method by Benjamini-Hochberg (BH)[177]. The resultant (redundant) gene ontology terms (biological process) of elevated genes were removed using the ‘simplify’ function in clusterProfiler based on GOSemSim[178]. The list of 1,158 mitochondrial genes was obtained from MitoCarta 2.0[179]. Essential genes (n=583) were assembled from three human essential genes studies using CRISPR-Cas9 and retroviral gene-trap genetic screens[180–182]. Disease-related genes (n=3,896) and kinase genes (n=504) were obtained from Uniprot. Cancer genes (n=719) were downloaded from Cosmic[183]. Drug target genes (n=784) were obtained from Drugbank[184] and restricted to proteins directly related to the mechanism of action for at least one of the associated drugs. GPCR genes (n=1,410) were obtained from HGNC and phosphatase genes (n=238) were from DEPOD[185]. Transcription factor genes (IF, n=1,639) were obtained from the HumanTFs collection[186].

The Spearman correlation coefficient was used to correlate transcriptome and proteome levels of single tissues. Using the R package ‘lmodel2’[135], the slopes were estimated by ranged major-axis (RMA) regression. This allows errors in both variables and is symmetric. The protein-mRNA Spearman correlation coefficients of 9,485 genes which were expressed in at least 10 tissues at both the mRNA and protein level were calculated. Based on the correlation coefficients, KEGG pathway enrichment analysis was conducted with the Kolmogorov-Smirnov test using the R package ‘fgsea’. The p-value of each pathway was adjusted by the Benjamini-Hochberg method and the cut-off significance was set to 0.05. The Co-inertia analysis (CIA) was performed using the ‘cia’ function in the ‘made4’ R-package[187]. 9,485 genes which were expressed in at least 10 tissues at both the mRNA and protein level were considered, and the remaining missing values were replaced with a positive value 1×10^4 times smaller than the lowest expression value in each data set.

2.5 RNA pulldowns

2.5.1 Cell culture and lysis

The HEK293FT cell line was cultured and lysed for RNA pulldowns. The culture conditions were according to the user guide from Thermo Fisher Scientific. The cells were grown in high glucose Dulbecco's Modified Eagle medium (DMEM, Biochrom GmbH) containing 10% fetal bovine serum (FBS, Biochrom GmbH), 0.1 mM MEM non-essential amino acids (NEAA), 6 mM L-glutamine, 1 mM MEM sodium pyruvate, and 500 µg/mL geneticin (G418, 50 mg/mL, Sigma) in a 37°C incubator with a humidified atmosphere of 5 –10% CO^2 in air using T-175 flasks (for cell passage) and culture dishes (15 cm diameter). The cells were harvested at approximately 80-90% confluency with 2× pre-washing with 10 mL Dulbecco's phosphate-buffered saline containing Ca^{2+} and Mg^{2+} (DPBS, calcium and magnesium). The cells were lysed with 400 µL CP lysis buffer (50 mM Tris/HCl, pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 1 mM Na_3VO_4, 25 mM NaF,1 mM DTT, 0.8% NP40, 1× phosphatase inhibitors 1, 2 and 3) followed by scraping the cells. After incubation on ice for 10 min, the lysates were transferred to 1.5 mL tubes and clarified by ultracentrifugation at 14,800 × g for 1 h at 4°C. Protein concentration was determined by the Bradford method (Coomassie Protein Assay Kit, Thermo Fisher Scientific) and the cleared lysates stored at -80 °C until required.

2.5.2 RNA probe design

To functionally understand how the RNA motifs affect the protein to mRNA ratio, 3 motifs were selected for RNA pulldowns. These included 2 known motifs (AAUAAA-polyadenylation signal, UAUUUAU- AU rich motif, as a positive control) and 1 novel motif (CUGUCCU). Motif selection was based on Basak Eraslan's sequence analysis of 29 healthy human tissues. Aiming at sequence affect, RNA probes were designed as 20 bp length without consideration of secondary *in vivo* influences. For each motif, 3 probes contained the motif and equal flanks from the real gene, and 1 random probe consisting of a random sequence with the same length as the motif and flanks from the real gene. The secondary structure and melting temperature (T_m) of candidate probe sequences were analyzed on an OligoAnalyzer 3.1 (https://eu.idtdna.com/calc/analyzer). Sequences with highly potential strong dimerization or hairpin were excluded. The final probe sequences for these 3 motifs are shown in Table 2. The 5'-amino modified RNA oligos were synthesized by IDT (Integrated DNA Technologies, Inc).

Table 2. The RNA probes designed for RNA pulldowns.

Motif	AAUAAA	UAUUUAU	CUGUCCU
probe-1	AGUAAACAAUAAAAGUGUCC	GUGAGUUAUUUAUUGGAAGC	CCCACGGCUGUCCUCCCGGU
probe-2	UUUCUCAAAUAAAGUUCAAA	UUCAAAUAUUUAUUGAGCAC	GUUAUUGCUGUCCUUGAUUG
probe-3	GCGCAUCAAUAAAAG	UGCAUAUAUUUAUAU	UCCACUGCUGUCCU

	CCGUC	UUUGC	CUUCAG
Random probe	AGUAAACGUACGUAGUGUCC	GUGAGUACAGUGAUGGAAGC	UCCACUGUACCGUACUUCAG

# The probes in grey were not included in the final motif-specific binding proteins analysis.

2.5.3 Coupling to NHS beads

Immobilization of 5'-amino modified RNA oligos was accomplished by reaction of the primary amine of the probe with the NHS-activated sepharose beads (GE Healthcare, Freiburg, Germany). The coupling procedure were developed from the kinobeads pulldown as described (Médard et al., 2015). Briefly, 1 mL NHS-beads (settled beads, stored in isopropanol) were washed 4× with 10 mL anhydrous dimethyl sulfoxide (DMSO). The beads were then resuspended in 1.8 mL anhydrous DMSO and 200 µL 2 nmol/µL RNA oligos (dissolved in RNAse-free water) were added to achieve a coupling density of 0.05 µmol of the probe per 1 mL settled beads (this coupling density was optimized). After mixing and centrifugation (at 1,200 r.p.m.), 40 µL supernatant were removed as a coupling control before the reaction and measured with a Nanodrop 2000 (Thermo Fisher Scientific). The reaction was initialized by the addition of 15 µL triethylamine (TEA) and incubated in the dark for 16-20 h at room temperature on an end-over-end-shaker. Subsequently, another 40 µL supernatant were removed as a coupling control after the reaction. 50 µL amino ethanol was added to block the remaining binding sites of the NHS beads. The mixture was further incubated in the dark for 16-20 h at room temperature on an end-over-end-shaker. Finally, the functionalized beads were washed 1× with 10 mL anhydrous DMSO and 3× with 10 mL ethanol. The beads were stored in 1 mL ethanol at 4°C in the dark until required.

2.5.4 Pulldown with competition assay

The RNA pulldown procedures were developed according to the kinobead pulldowns as previously described[188]. The experiments were conducted in 96-well filter plates. Unless otherwise stated, liquid was rinsed through the filter plates and residual liquid was removed via centrifugation at 1,200 r.p.m. for 2 min at 4°C. The cell lysate was diluted with 1:1 CPPI buffer (50 mM Tris/HCl, pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 1 mM Na_3VO_4, 25 mM NaF,1 mM DTT, 1× phosphatase inhibitors 1, 2 and 3) and CP-0.4 (50 mM Tris/HCl, pH 7.5, 5% glycerol, 1.5 mM MgCl2, 1 mM Na_3VO_4, 25 mM NaF,1 mM DTT, 0.4% NP40, 1× phosphatase inhibitors 1, 2 and 3) resulting in a final concentration of 5 mg/mL protein and a reduction of the NP40 concentration from 0.8% to 0.4%. The diluted cell lysates (2.5 mg total protein/well) were incubated for 1 h at 4 °C in an end-over-end shaker with 0 nM (water control), 3 nM, 10 nM, 30 nM, 100 nM, 300 nM, 1 µM, 3 µM, or 10 µM of the free RNA oligos dissolved in RNase-free water. 20 µL settled beads were washed twice with 1 mL CP buffer (50 mM Tris/HCl, pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 1 mM Na_3VO_4) and then equilibrated with 1 mL CP-0.4 buffer (50 mM Tris/HCl, pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 1 mM Na_3VO_4, 0.4% NP40). The equilibrated beads were combined with the cell lysates and incubated for 30 min at 4°C on an end-over-end shaker. The water control lysate was recovered and incubated similarly with RNA oligo beads as a pulldown of the pulldown experiment to calculate the depletion factor. The beads were then washed 3× with 1 mL CP-0.4 % and 2× with 1 mL CP-0.2 (50 mM Tris/HCl, pH 7.5, 5% glycerol, 1.5 mM $MgCl_2$, 1 mM Na_3VO_4, 0.2% NP40).The bound proteins were subsequently eluted by incubation for 30 min at 50°C with 60 µL 2× NuPAGE LDS sample buffer (Invitrogen, Darmstadt, Germany) containing 50 mM DTT and harvested via centrifugation at 1,200 r.p.m. for 2 min at 4°C. The eluates were alkylated using 55 mM chloroacetamide (CAA) for 30 min at room temperature in the dark. Prior to MS analysis, samples were desalted and concentrated by a short electrophoresis (about 0.5 cm) on a 4−12% NuPAGE gel (Invitrogen). *In situ* gel digestion of proteins after RNA pulldowns was performed by Andreas Klaus (technical assistant from the Küster research group) according to standard procedures. Dried peptides were stored at -20°C until required.

2.5.5 LC-MS/MS analysis

RNA pulldown peptides were analyzed using nanoflow LC-MS/MS by directly coupling a nanoLC-Ultra 1D+ (Eksigent) to an Orbitrap Elite mass spectrometer (Thermo Fisher). Peptides were dissolved in 20 µL solvent A (0.1% FA in HPLC-grade water). 10 µL peptides were delivered to a trap column (100 µm × 2 cm, packed in-house with Reprosil-Pur C18 AQ 5 µm resin, Dr. Maisch) at a flow rate of 5 µL/min in 100% solvent A (0.1% FA in HPLC-grade water). Peptides were separated on an analytical column (75 µm × 40 cm, packed in-house with Reprosil-Gold C18, 3 µm resin, Dr. Maisch) using a 100 min gradient ranging from 4% to 32% solvent B (0.1% FA and 5% DMSO in acetonitrile) in A_1 (0.1% FA and 5% DMSO in HPLC grade water) at a flow rate of 300 nL/min. The mass spectrometer was operated in data dependent mode, automatically switching between MS and MS^2 spectra. Full scans (*m/z* 360-1,300) were acquired at a resolution of 30,000 using an AGC target value of 1e6 and a maximum ion injection time of 100 ms. Internal calibration was performed using the DMSO cluster as a lock mass. Tandem mass spectra were generated for up to 15 precursor ions by HCD with a normalized collision energy of 30%. The dynamic exclusion was 30 s. Fragment ions were acquired at a resolution of 15,000 with an AGC target value of 2e4 and a maximum ion injection time of 100 ms.

2.5.6 Peptide and protein identification and quantitation

Peptide and protein identification and quantitation were performed using MaxQuant (v1.5.3.30). The tandem MS data was searched against the human Uniprot reference database (v22.07.13) with the embedded search engine Andromeda. Label-free quantitation (LFQ)[94] and match-between-runs were enabled within MaxQuant. All other settings were as described for the quantitative whole proteome. EC_{50} values were determined from this ratio with an internally-developed R-Script using nonlinear regression with variable slope. A K_d was then calculated by multiplying the EC_{50} by a correction factor for each protein. The correction factor (*r*) for a protein is defined as the ratio of the amount of protein captured from two consecutive pulldowns of the same water control lysate.

2.6 Proteogenomic analysis

2.6.1 Protein isoform identification

The sample-specific RNA-Seq databases were provided by the Human Protein Atlas project. In brief, RNA sequencing data was aligned to the human reference genome (GRCh38, v83) using Tophat 2.0.8b. FPKM values were calculated using Cufflinks v2.1.1 as a proxy for transcript expression level. Rvboost was used for variant calling. All transcripts with FPKM>1 were translated into protein sequences and included in the search database. Each tissue was searched against the matched RNA-Seq database using MaxQuant as described above. The match-between-runs function was disabled. The MaxQuant output data were used for the isoform analyses. The databases were constructed by Björn Hallström (Science for Life Laboratory, KTH-Royal Institute of Technology, Stockholm, Sweden).

2.6.2 Exome sequencing and variant calling

The exome of tonsil tissue was enriched using the Agilent SureSelectXT kit (v5) and sequenced on an Illumina HiSeq4000 sequencer. The raw data was aligned to the human reference genome (hg38) using bwa (v0.7.12) and duplicate reads were marked using Picard Tools (v2.4.1). Genomic variants were called and filtered using the GATK (v.3.6) HaplotypeCaller and VariantFitration modules, respectively, according to the best practice guide (https://software.broadinstitute.org/gatk/best-practices/). Furthermore, variants at sites with a read depth <10× were removed. All I/L variants were removed as these cannot be distinguished by mass spectrometry. The resultant variants were annotated using the Ensembl Variant Effect Predictor (v85). The RNA sequencing data was aligned to the human reference genome (hg38) using STAR aligner (v2.5.2) and duplicate reads were marked using Picard Tools (v2.4.1). Variants were called using the GATK (v.3.6) HaplotypeCaller module, according to the aforementioned best practice guide.

A variant fasta-formatted database was created by the 'customProDB' package from the exomic variants [167]. Searching of the ultra-deep mass spectrometry data with Mascot was performed against this database together with protein databases from UniProt and Ensembl using the following parameters: peptide mass tolerance of 10 p.p.m., MS/MS tolerance of 0.05 Da, carbamidomethylation of cysteine as a fixed modification, oxidation of methionine and acetylation as variable modifications. Up to 2 missed cleavage sites were allowed for peptides digested with trypsin, Lys-C, Arg-C and Lys-N; up to 3 missed sites for peptides produced from digestion with Asp-N and with Glu-C (V8-DE in the Mascot search engine); and up to 4 missed cleavage sites for peptides generated by digestion with chymotrypsin. The resultant PSMs were analyzed using Percolator (v3.01) and an overall FDR cut-off of 1% was applied.

A custom python script was used to identify PSMs that covered variant sites and showed either the variant or the canonical genotype. All initial candidate variant peptides had met the following criteria: (i) Mascot ion scores of at least 25; (ii) a Mascot delta score of at least 10; (iii) the peptide must only map to the variant database; (iv) the peptide must map to a single genomic position only; (v) for missense variants, the peptide must either show the variant amino acid or it must be cleaved according to a novel protease cleavage site arising from the variant; (vi) for nonsense variants, the peptide must end at the novel C-terminus. For canonical genotypes, the same criteria were applied except: (i) at least one protein the peptide maps to must not be from the variant database; (ii) for missense variants, the peptide must show the wild-type amino acid; (iii) for nonsense variants, the end of the peptide must be after the novel C-terminus (after nonsense variant sites). The resultant candidate peptides were mapped against UniProt using BLAST to exclude other obvious explanations. To further consolidate the variant peptides and to reduce false positives, peptide identification by MaxQuant was performed in parallel. Using the customized exomic variant database with the same parameters used in Ensembl database searches described above. The list of candidate variant peptides for the spectra angle analysis required identification by both Mascot and MaxQuant. All the candidate variant peptides were manually inspected and compared to synthetic peptides.

Variant calling at exon and RNA, variant databases construction, and the python script for filtering were performed by Thomas Wieland and Thomas Hopf (OmicScouts GmbH, Freising, Germany).

2.6.3 Identification of peptides translated from non-coding regions

A database of products from possible alternative translation initiation sites (aTIS) was constructed by searching the 5’ UTR of GENCODE transcripts (v25) for putative alternative start codons of eukaryotes (AUG, CUG, GUG, UUG, ACG, CCG, GCG, UCG) and *in silico* translation of these ‘novel coding sequences’. This resulted in 474,991 aTIS ‘proteins’ with >6 amino acids. The lncRNA product database was generated by three-frame-translating the GENCODE (v25) lncRNA database and resulted in 29,524 sequences. The standard 29 tissue proteomic data sets were supplemented with two additional tissues for which only proteomic data was available (bone marrow, pituitary gland). In total, 48 samples (including replicates of some organs) was searched against concatenated sequence collections comprising the aTIS and lncRNA databases, GENCODE (v25), UniProt (downloaded 3 February 2017) and sample specific RNA-Seq based databases using Mascot to identify peptides from known proteins. The search parameters were the same as described for the exome variant peptide identification. The resultant PSMs were processed using Percolator and an overall FDR cut-off of 1% was applied. A custom python script was used to identify PSMs from putative translated lncRNAs or the aTIS database. Candidate peptides had to meet the following criteria: (i) the PSM must map to a single database only, *i.e.*, aTIS or lncRNA but not any other; (ii) the Mascot score must be at least 25; (iii) the Mascot delta score must be at least 10; (iv) the original underlying transcript must be expressed in at least one of the tissues (RNA-Seq FPKM >1). The resultant PSMs were then mapped against UniProt using BLAST to exclude other explanations for the novel peptide (*e.g.*, peptides arising from a novel tryptic cleavage site due to a genomic variant). To consolidate the list of candidate aTIS and lncRNA peptides and to reduce false positives, the raw MS data was also searched by MaxQuant (with the same parameters as described for searches using Ensembl). Only peptides identified by both Mascot and MaxQuant were allowed to pass to the stage of spectral contrast angle analysis.

lncRNA and aTIS databases construction and the python script for filtering were performed by Thomas Wieland and Thomas Hopf (OmicScouts GmbH, Freising, Germany).

2.6.4 Validation of variant and non-coding peptides by synthetic reference peptides

All peptides that passed the filter criteria for Mascot described above, were synthesized at JPT Berlin using Fmoc-based solid-phase synthesis. The details of peptide synthesis, sample preparation and MS measurement are as previously described[189]. RAW data were analyzed using MaxQuant and Mascot searching individual LC-MS/MS runs against pool-specific databases. Pool-specific databases were generated by an internally-developed shiny tool (Fasta builder) by concatenating peptide sequences into a single protein. For unmodified tryptic peptides (pool BC_1, 2,3,4), variable modification was set to oxidation on methionine, and the other settings were the same as for the tryptic Mascot search described for the exomic variants. For unmodified non-tryptic peptides (pool BC_5,6), non-enzyme was specified, whilst all other settings were the same as for the identification of unmodified tryptic peptides. For tryptic acetylated peptides (pool ac_1), variable modifications were acetylation at the peptide N-termini and oxidation on methionine, whilst all other settings were also identical to those used for the unmodified tryptic peptides. For non-tryptic acetylated peptides (pool ac_2), non-enzyme was specified, whilst all other settings were the same as for identification of tryptic acetylated peptides. Normalized spectral contrast angle (SA) analysis was performed to compare endogenous and synthetic peptides using internally-developed Python scripts [190]. Candidates passed if: (i) the SA value was ≥0.7 (Pearson of ~0.9); (ii) the endogenous peptide had a Mascot score of ≥50; or (iii) manual inspection of the spectra substantiated the candidate peptide sequence assignment. In parallel, the tandem MS spectra of all candidate peptides were also manually inspected.

Scripts of spectral contrast angle analysis were from Mathias Wilhelm and Daniel Zolg (Küster group).

Chapter 3 Human transcriptome and proteome map based on 29 healthy tissues

3.1 In-depth transcriptome and proteome

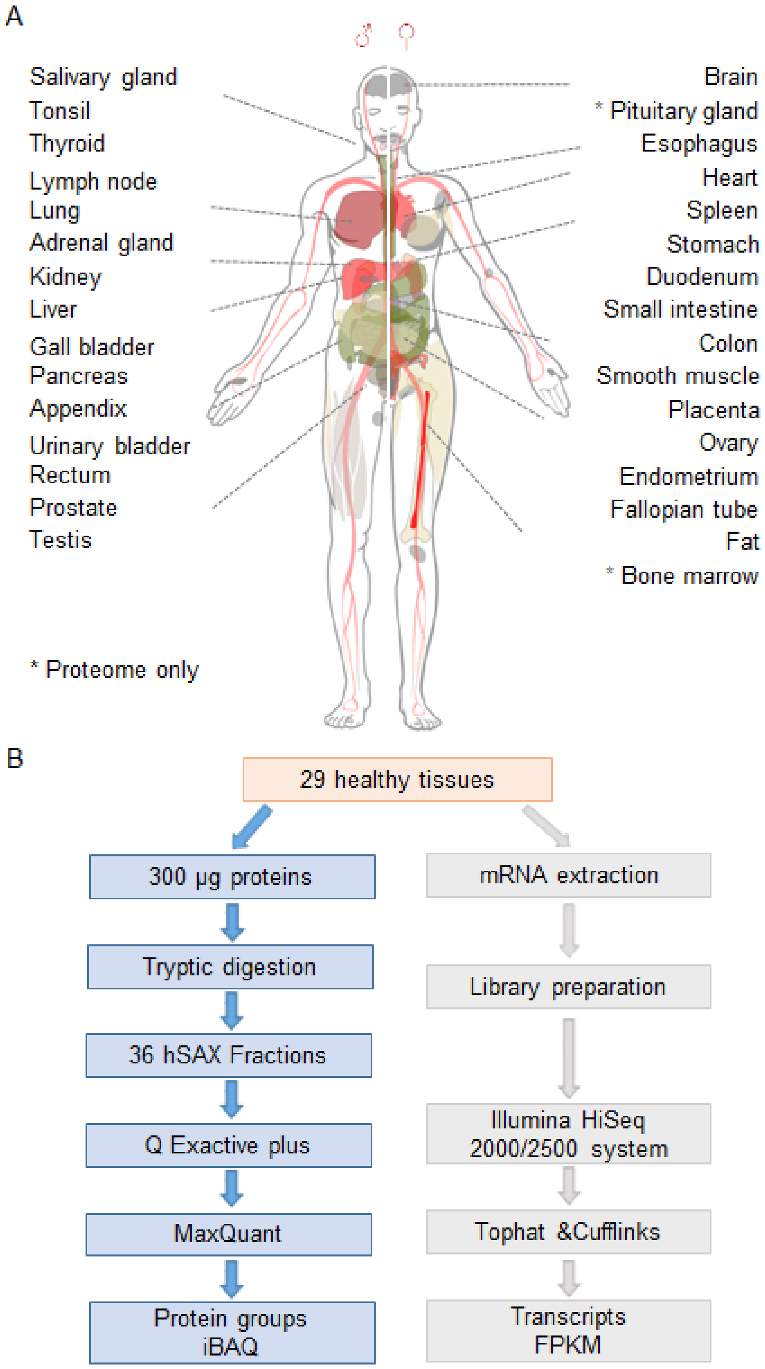


Figure 13. (A) Body map of analyzed tissues. (B) The cryo-sections of 29 tissues were analyzed by RNA-Seq and mass spectrometry-based proteomics.

The primary aim of this thesis was to generate a comprehensive human transcriptome and proteome map of 29 healthy tissues. Tissues were collected by the Human Protein Atlas project (HPA) and adjacent cryo-sections[191] were used for paired (allele specific) transcriptome and proteome analyses (Figure 13). The transcriptomic data had been collected by the Human Protein Atlas project[16], whilst the in-depth proteome data tissues were generated in this thesis. A total of 31 tissues represented by 50 samples were analyzed, including replicate samples of colon, liver, stomach, placenta, and tonsil, and a tonsil tissue digested with different proteases and fragmentation methods (n=10) (Figure 13 A body map, Table 1, Appendix Table 1). 48 samples representing 30 tissues were from HPA, and 2 pituitary samples provided by Marily Theodoropoulou, LMU. To achieve a great depth of proteome coverage, the proteins were extracted by urea lysis buffer with homogenization by Beadbeater, and the tryptic digestions were separated into 36 fractions by hSAX. Each fraction was measured on a Q Exactive Plus hybrid quadrupole-orbitrap mass spectrometer with a 2 h LC gradient. Via independent searches using the MaxQuant platform, 7,000 to 11,000 proteins were identified in each tissue. One exception was bone marrow as this tissue contains a large proportion of lipids and as such, only 4,500 proteins were identified[192]. A high correlation between the liver replicates was observed (spearman correlation, rho=0.8~0.9, Figure 14), to provide highly-reproducible quantitative data on thousands of proteins. These results suggested that the proteome of a tissue sample from one donor can be representative of the general protein expression for a specific tissue type. These data represent a catalogue of expressed human proteins and profiles thereof across human tissues. Importantly, these proteomes reflect the proteins expressed in each organ, but does not necessarily represent the protein expression of every type of cell type in that tissue.

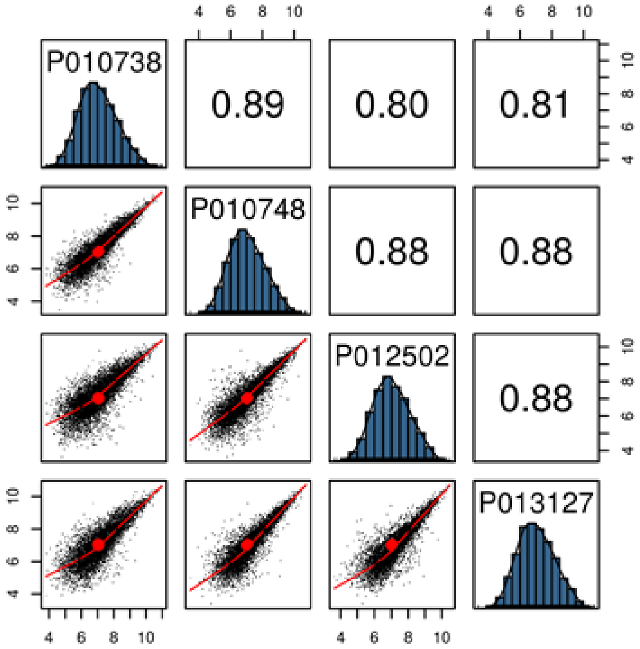


Figure 14. Reproducibility of proteomic data. Quadruplicate analyses of liver tissues revealed a high correlation between replicates. The numbers given in the top right panel are the Spearman correlation coefficients (rho).

To generate a comprehensive transcriptome and proteome map, 29 of the 31 tissues (sample=29) with both transcriptomic and proteomic data (generated by trypsin digestion and HCD fragmentation) were utilized. The proteomic data from 29 tissues was analyzed using the MaxQuant platform and resulted in the identification and intensity-based absolute quantitation (iBAQ)[93] of a total of 15,210 protein groups. An average of 11,005 (±680) protein groups per tissue were identified at a false discovery rate (FDR) of <1% at the protein, peptide, and peptide-to-spectrum match (PSM) level (Figure 15). Protein identification was based on 277,698 non-redundant tryptic peptides that represents a total of 13,664 genes. On average, this equates to 10,547 (±512) genes per tissue that, on average, covers 88% of the expressed genome (based on the RNA-Seq) in every tissue (Figure 16). The total number of confidently-identified proteins in this study is smaller than that of other (community-based) resources such as ProteomicsDB[193] and neXtProt[194] (coverage of 15,721 and 17,470 protein-coding genes respectively). Nevertheless, this study provides a highly-consistent collection of tissue proteomes including the deepest proteomes to date for many of the analyzed tissues. In addition, our data also provides evidence for 69 proteins (represented by at least one unique peptide with an Andromeda score ≥100) that are not yet included in neXtProt (release 2018-01-17).

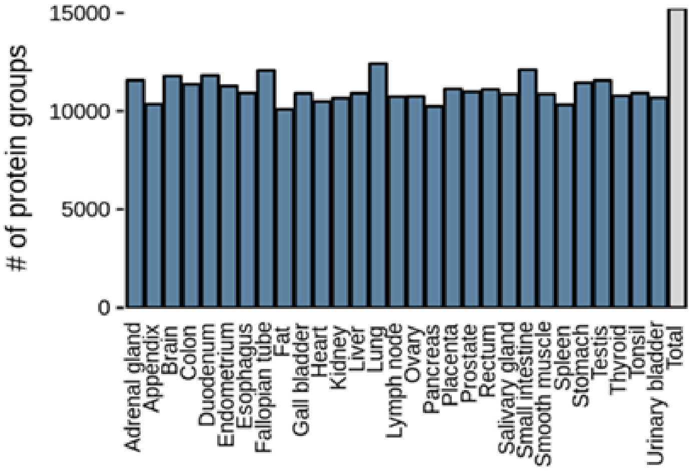


Figure 15. The number of proteins identified in 29 tissues by mass spectrometry-based proteomics.

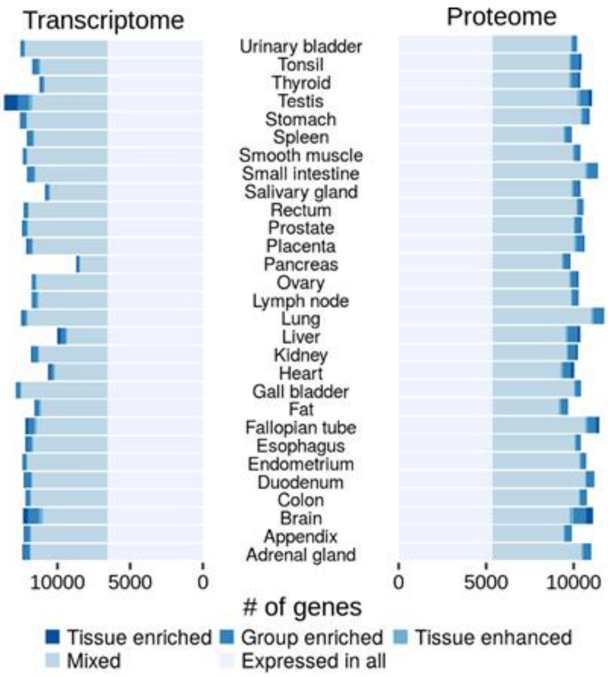


Figure 16. The number of genes identified as transcript and protein in 29 tissues.

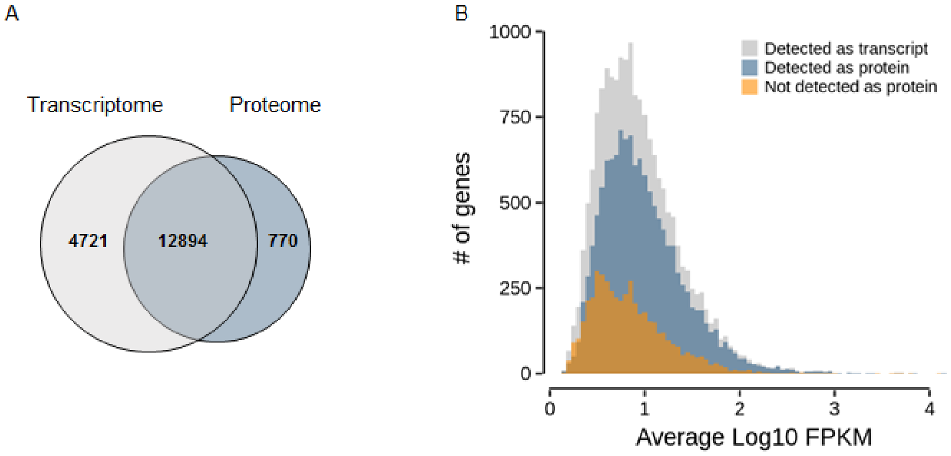


Figure 17. (A) The overlap of genes identified by transcriptomics and proteomics. (B) Abundance distribution of all transcripts detected in all tissues (grey); the fraction of detected proteins is shown in blue and the fraction of transcripts for which no protein was detected is shown in orange.

In total and when using a cut-off of 1 fragment per kilobase million (FPKM), RNA-Seq profiling detected and quantitated 17,615 protein-coding genes with an average of 11,927 (±937) genes per tissue (Figure 16)[16]. Overall, 12,894 protein-coding genes were detected at both the transcript and protein level (Ensembl gene annotation, an average of 8,500 (±670) genes per tissue, Figure 17A). The identified proteins spanned almost the entire range of mRNA expression, again indicative of a very substantial coverage of the expressed proteome (Figure 17B). Some proteins, however, were not observed even for highly-expressed mRNAs (*i.e.*, higher than the mean mRNA abundance). Approximately 1/3 of these mRNAs were observed in thetestis (486 from 1,574). No other tissue contained nearly as many highly-expressed mRNAs without protein evidence (Figure 18). Gene ontology analysis (clusterProfiler; n=88 genes; BH adjusted p-value = 5×10^{-12}) revealed that the ‘missing’ proteins in thetestis were statistically, significantly-enriched for processes related to spermatogenesis. The rich expression of mRNAs intestis has been known for a long time and exploited for, *e.g.*, the cloning of many genes from cDNAs. Nevertheless, the apparent absence of so manytestis proteins with high mRNA expression is still surprising. This was not due to, *e.g.*, poor coverage of the testis proteome (11,033 detected protein-coding genes) or other obvious technical factors. Interestingly, almost 300 of these ‘missing’ proteins have also not been detected by antibodies intestis (according to the Human Protein Atlas project) and nearly 200 have no ascribed molecular function. Despite high levels of mRNA, the inability to observe these proteins by mass spectrometry or antibodies poses a number of questions. For example, are these proteins rapidly degraded implying specialized (and perhaps transient) functions in thetestis or sperm functionality? Are the proteins perhaps stabilized in response to egg fertilization? Proteins that were absent from the lower end of the mRNA expression range (less than the mean mRNA abundance) are overrepresented in G-protein coupled receptor activity (n=170; BH adjusted p-value = 4×10^{-45}), ion channels (n=111; BH adjusted p-value = 5×10^{-7}) and cytokine-related biology (n=121; BH adjusted p-value = 3×10^{-10}). The abundance of these proteins may simply have been below the detection limit of the mass spectrometer. Alternatively, as has been described many times, these proteins can be difficult to extract from cells because of multi-pass transmembrane domains that give rise to few (if any) MS-compatible tryptic peptides after digestion. Interestingly, for 770 identified proteins, no corresponding mRNA was detected in any tissue (Figure 17A). These proteins were enriched for, *e.g.*, immune-related processes including major histocompatibility complexes (MHC; n=40; BH adjusted p-value = 1×10^{-41})

and antibodies (n=39; BH adjusted p-value = 2×10^{-31}) that are either produced (on and off) by certain cell types in a given tissue or originate from elsewhere in the body that was not covered by our proteomes and transcriptomes.

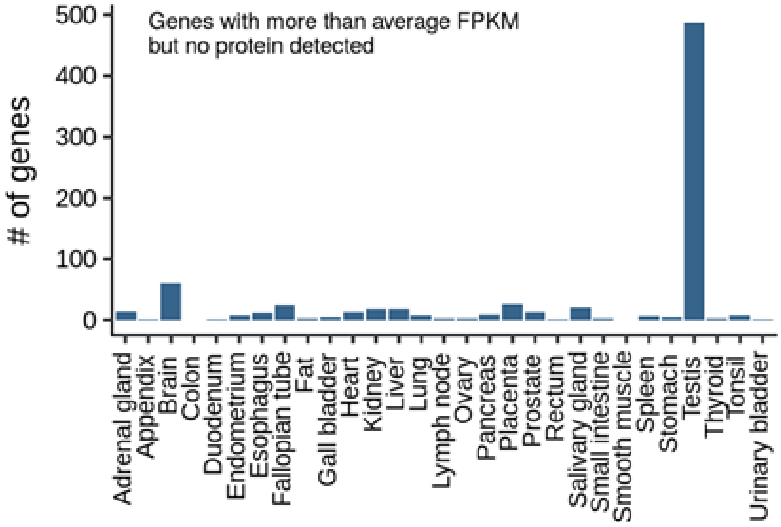


Figure 18. The number of genes in all tissues that were highly detected at the transcript but not at the protein level.

These quantitative transcriptomes and proteomes enabled us to investigate associations between gene-expression level and tissues. Unsupervised hierarchical clustering of the transcriptomic and proteomic data both recapitulated the similarities between related tissues (Figure 19). For example, gastrointestinal tissues (stomach, colon, rectum, small intestine and duodenum) tightly cluster together. Interestingly, appendix clustered with the immune system tissues at both the transcriptome (spleen, lymph node and tonsil) and proteome (spleen, lymph node, tonsil and lung) rather than with the gastrointestinal tract. This observation is in accordance with previous findings that have shown quantitative transcriptomics and proteomics can link anatomically-distinct tissues with related functions[16,195].

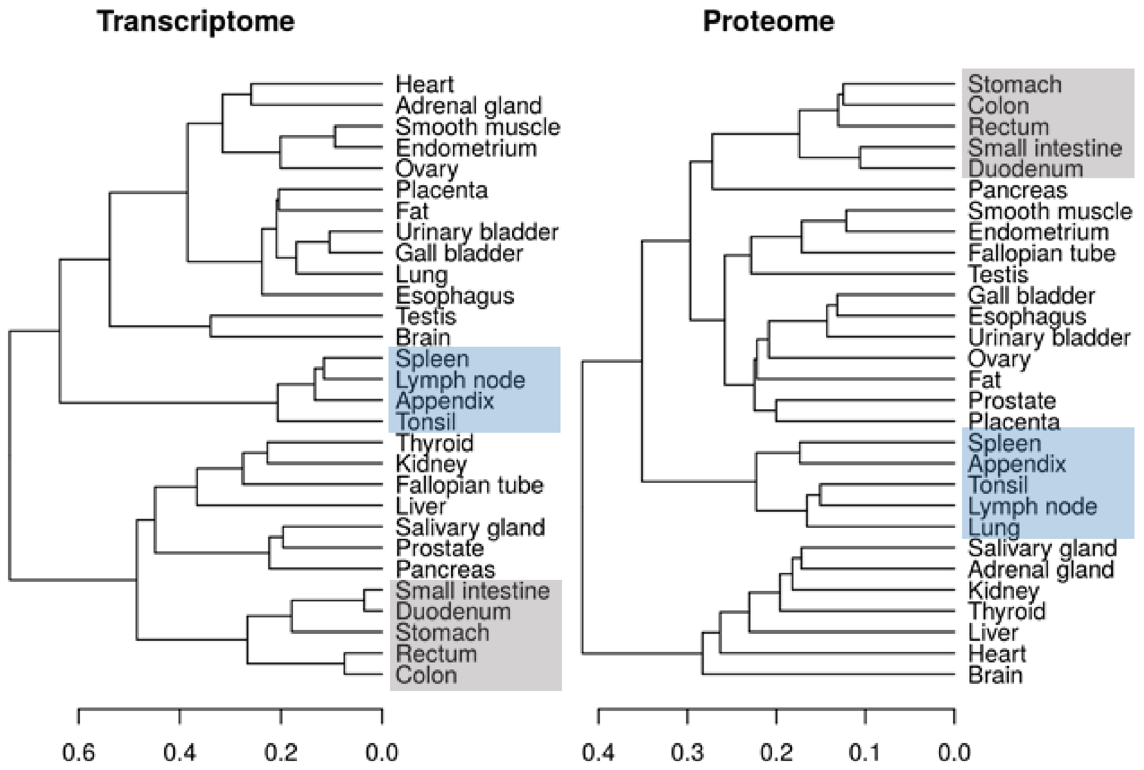


Figure 19. Unsupervised hierarchical clustering of the transcriptome and proteome of 29 tissues.

3.2 Classification of identified human protein-coding genes

To explore which proteins (and how many) show a tissue-specific expression profile, the classification scheme of Uhlén et al.[16,176] was applied. This approach was previously developed for mRNA profiling and stratifies genes into five classes: 'tissue enriched' (5-fold above any other tissue); 'group enriched' (5-fold above any group of 2-7 tissues); 'tissue enhanced' (5-fold above the average of all other tissues); 'expressed in all' (similarly expressed in all tissues); and 'mixed' genes (that do not match the other categories). Overall, a large fraction of all represented genes was expressed in all tissues: 37% (6,562) at the transcript level and 39% (5,394) at the protein level (Figure 20). These ubiquitously-expressed genes encode histones, ribosomes, proteasomes, structural proteins; and are often referred to as 'housekeeping' genes. 43% (7,516) of all transcripts and 53% (7,272) of all proteins showed elevated expression in one or more tissues ('tissue enriched', 'group enriched' or 'tissue enhanced'). Only 4.3% (on average) of all transcripts and 5.4% of all proteins showed a tissue-enriched profile. Two notable exceptions were brain (transcript level n=340 and protein level n=392) and testis (transcript level n=921 and protein level n=164) that exhibited a higher percentage of tissue-enriched proteins and transcripts. This finding was in accordance with a recent analysis of RNA-Seq data from the Human Protein Atlas and GTEx projects[196]. In comparison with other tissues, one explanation for the large abundance of tissue-enriched genes in brain andtestis is that brain is the only tissue that contains a large amount of neuronal and glial cells; and testis harbors the only male cell type that undergoes meiosis to generate haploid cells.

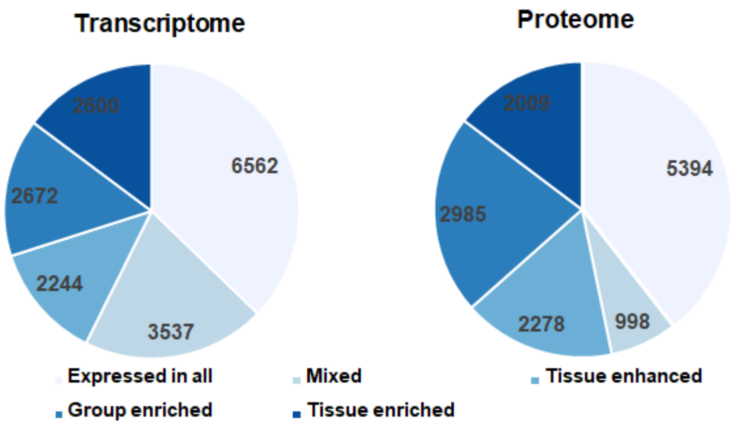


Figure 20. The number of genes classified in each expression category.

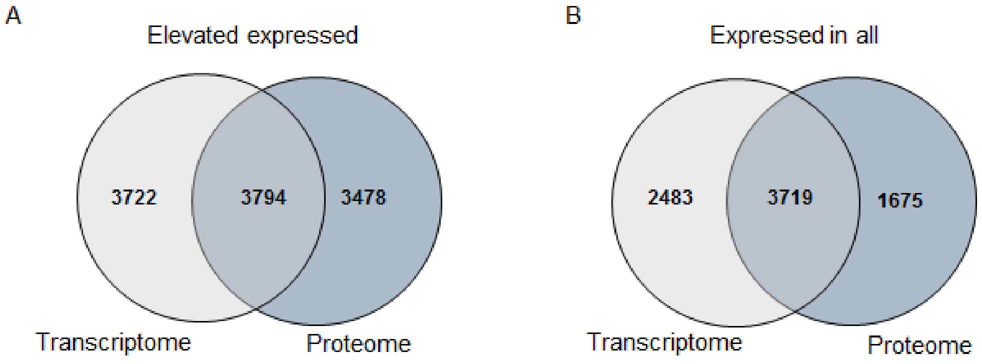


Figure 21. (A) Comparison of elevated expressed genes at the transcript and protein level. (B) Comparison of genes expressed in all at both the transcript and protein level.

It is interesting, albeit not unexpected, that the genes with elevated expression at the transcript and protein level are not always identical (Figure 21). One explanation is that some tissue-enriched genes were not detected as a protein. For example, the ‘missing’ proteins intestis. Another typical example is albumin: which is enriched at the transcript level in the liver; but detected as a protein in all 29 tissues at a similar expression level. Albumin is expressed in the liver and transported to the other tissues via the blood. Although there are discrepancies, most of the elevated expressed proteins or transcripts still represent the function of the respective tissue. A functional gene ontology (GO) biological process analysis of the elevated expressed genes is summarized in Figure 22. In the appendix, for example, genes related to immune functions are elevated at both the transcript and protein level. This is in agreement with the hierarchical clustering results and that the appendix functions as an important immune tissue rather than a rudimentary part of the intestine. In brain, genes are enriched for neuron-related functions, *e.g.*, synaptic vesicle cycle and neuron migration. In heart, elevated genes are related to cardiac muscle and energy metabolism; and in thetestis, elevated genes are associated with functions related to the reproductive process and spermatogenesis. Interestingly, proteins with a more tissue-restricted expression tended to have a somewhat lower abundance; while genes expressed in all tissues tended to have a higher abundance (Figure 23, an example of the brain proteome). Despite the differences in detail, our data set confirms that at the protein level there is a core set of ubiquitously-expressed genes/proteins. In addition, individual tissues are not strongly-characterized by the categorical presence or absence of mRNAs or proteins, but rather by quantitative differences[195].

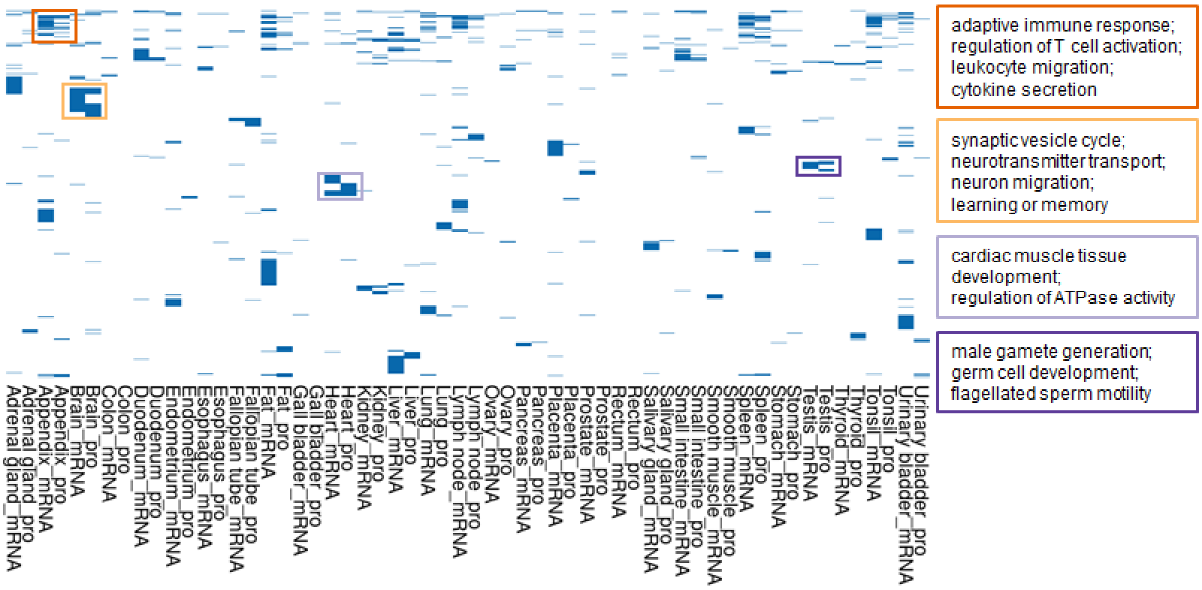


Figure 22. Clustering of gene ontology terms (biological process) for proteins and transcripts that show the most divergent expression across all tissue. Boxes give examples of GO terms for four different tissues (appendix, brain, heart, and testis).

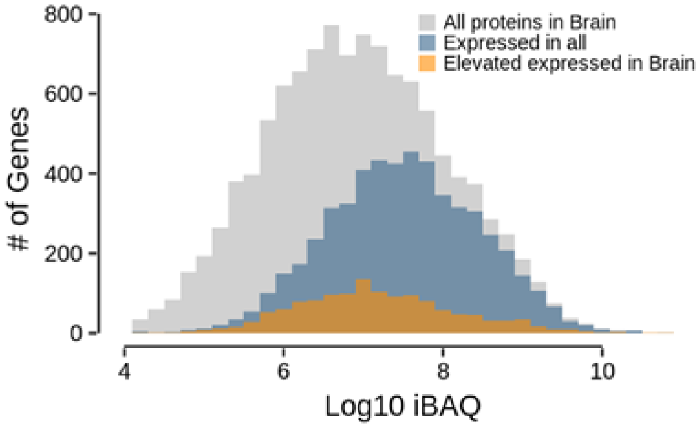


Figure 23. Abundance distribution of all proteins detected in human brain (grey). Proteins in blue are expressed in all 29 tissues and proteins in orange show elevated expression in brain.

3.3 Functional classes of genes in 29 tissues

With the transcriptome and proteome profiling of 29 tissues, 9 functional classes of genes were examined. These included mitochondrial genes, essential genes, disease-related genes, cancer genes, drug targets, transcription factors, GPCRs, kinases, and phosphatases. The dysregulation or dysfunction of such genes are tightly-related to human health and disease. Intriguingly, the global trends in transcript and protein tissue expression distributions were also mirrored by functional categories of genes with some interesting findings that may be of general value for drug discovery (Figure 24).

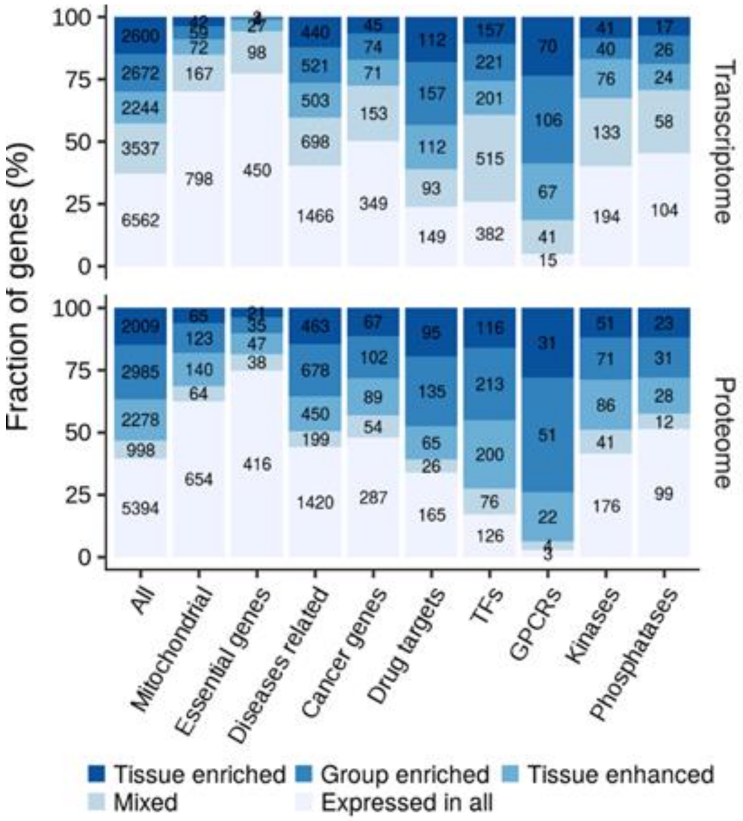


Figure 24. Distribution of selected functional classes of transcripts and proteins across the expression categories.

Mitochondrial proteome. Mitochondria are complicated organelles involved in essential pathways such energy metabolism, ion homeostasis, signaling and apoptosis. Human mitochondria contain proteins encoded by 1,158 genes[179]. The proteins encoded by the 13 mitochondrially-located genes are all core components of oxidative phosphorylation. The other >1,000 proteins are nuclear-encoded and imported into the mitochondria. Either mutation of nuclear or mitochondrial DNA can cause dysfunction of the mitochondrial respiratory chain and result in mitochondrial disorders. Due to the housekeeping role of mitochondria, it is not surprising that most of these genes lack tissue specificity. The highest fraction of transcripts is found in heart (~80%), and the highest fraction of proteins are found in heart, kidney and liver (~30%), illustrating the high-energy demand of these tissues[197–199]. Transcripts encoded by mitochondrial genes are the highest in the majority of tissues, while in the proteome only a few fractions were detected (Figure 25). This may be due to technical reasons in that our proteomic protocol was not sufficient for multi-transmembrane proteins, but a biological reason cannot be excluded.

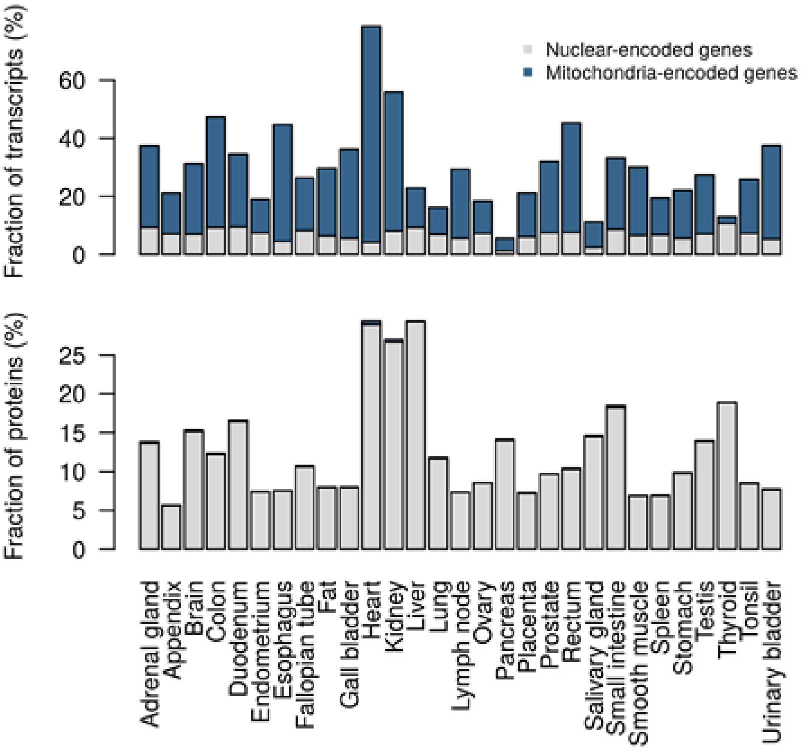


Figure 25. The fraction of transcripts (upper panel) and proteins (lower panel) encoded by mitochondrial genes for each tissue, subdivided by genes encoded by the mitochondrial genome and chromosomes, respectively.

Essential proteome. The genes ubiquitously-expressed in all cell types and tissues are often defined as housekeeping genes. It is still unclear, however, how the loss of each gene affects cell viability. In 2015, three landmark studies revealed ~2,000 essential genes by the CRISPR-Cas9 and gene-trapping approaches[180–182]. To exclude dependency on cell type, only the essential genes identified in all three studies (n=583) were compiled for further analysis. As expected, more than 95% of the genes were identified in the 29 tissues (581 and 557 detected as transcripts and proteins, respectively). Most were highly-expressed in every tissue, and are related to core cellular processes such as DNA replication, DNA repair, mitosis, mRNA processing, translation and protein degradation. One of the essential genes is superoxide dismutase (SOD1); expressed in all 29 tissues and responsible for destroying radicals produced within the cells. Single amino acid mutations of SOD1 are associated with amyotrophic lateral sclerosis that causes the death of neurons controlling voluntary muscles[200].

Disease related proteome. There are 3,896 reviewed and curated genes that are assigned as disease-related in UniProt. These represent approximately 20% of all human genes. In the 29 tissue data, 3,628 and 3,210 genes were identified at the transcript and protein level, respectively. It is a very comprehensive gene set that covers the majority of basic metabolism pathways in cells. Therefore, the expression distribution at both levels are most similar to the whole transcriptome and proteome in 29 tissues. One example of gene-enriched expression in the brain at both levels is SYN1. Mutations in this protein may be associated with X-linked disorders with primary neuronal degeneration such as Rett syndrome[201].

Cancer proteome. A large proportion of human cancer is caused by the acquisition of somatic mutations across the lifetime of an individual. Genes that contain mutations that have been causally-implicated in cancer are referred to as cancer genes. According to the catalog of somatic mutations in cancer (COSMIC), there are 719 cancer genes including 227 hallmark genes[202]. In this work, 96% of the cancer genes were identified as transcripts and 85% were identified as protein. Expression analysis revealed that ~50% of the identified genes have a similar expression level across all 29 tissues at both the transcript and protein level. The lack of tissue specificity is not surprising as these genes are often essential for normal growth and cell cycle regulation. This emphasizes the potential adverse effect of treating cancer by targeting proteins expressed in all tissues.

Druggable proteome. Most of the drugs for treating human diseases target proteins and regulate activity thereof. The largest fraction of drugs target GPCRs, ion channels, kinases and nuclear receptors[203]. In Drugbank, there are 784 target proteins of FDA-approved drugs with pharmacological activity. These are directly-related to the mechanism-of-action for at least one of the associated drugs. In the 29 tissues, 623 and 486 drug targets were detected at the transcript and protein level, respectively. Only 12 of these drug target proteins are in essential gene list and include PSMB1, PSMB5, PSMB2, GGPS1, RRM1, RRM2, POLA1, POLE2, TOP1, TOP2A, DNMT1, and TUBG1. It was appeasing to observe that most of the essential genes are not used as a drug target (Figure 26). In addition to known targets, the expression map of the 29 tissues can used to evaluate the tissue specificity of novel findings and to discover novel drug targets.

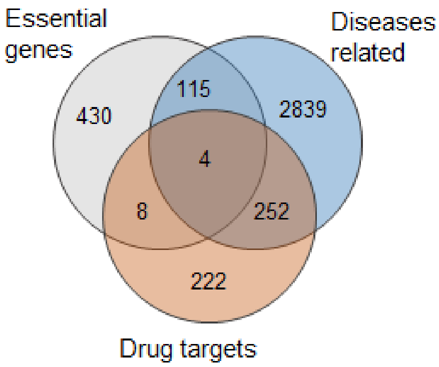


Figure 26. The overlap of proteins encoded by essential genes, disease-related genes and drug target protein coding genes.

Transcription factors. Transcription factors (TFs) are an important class of regulatory proteins that control gene expression in the first step of DNA decoding. There are 1,639 known or likely human TFs[186]. Most contain at least one C2H2-ZFs (n=747) and homeodomains (n=196) DNA binding domain. In the 29 tissues, 1,476 and 731 TFs were identified at the transcript and protein level, respectively; and the overlap of transcripts and proteins was 724. 135 from 752 'missing' TF proteins were from the testis and comprised almost half of the testis genes that have transcript but no protein evidence. This indicated that these TFs might have specific functions in the differentiation program of spermatogenesis and degrade rapidly. The testis-specific TFs that were only identified as transcripts included RHOXF2, ZBTB32, HSFY1, CTCFL and SPZ1. Another reason for 'missing' TFs at the protein level may be that these were present at a relatively low expression level of mRNA. Figure 27 shows the expression profiles of TFs at the transcript and protein level. Interestingly, the proportion of tissue elevated expressed TFs at the protein level was higher than at the transcript level (~70% vs. ~30%). Tissues with similar or related functions, such as the immune system tissues, gastrointestinal tract, and female tissues, had a similar expression pattern for TFs. These results tend to suggest that TFs may regulate gene expression in a tissue-dependent fashion.

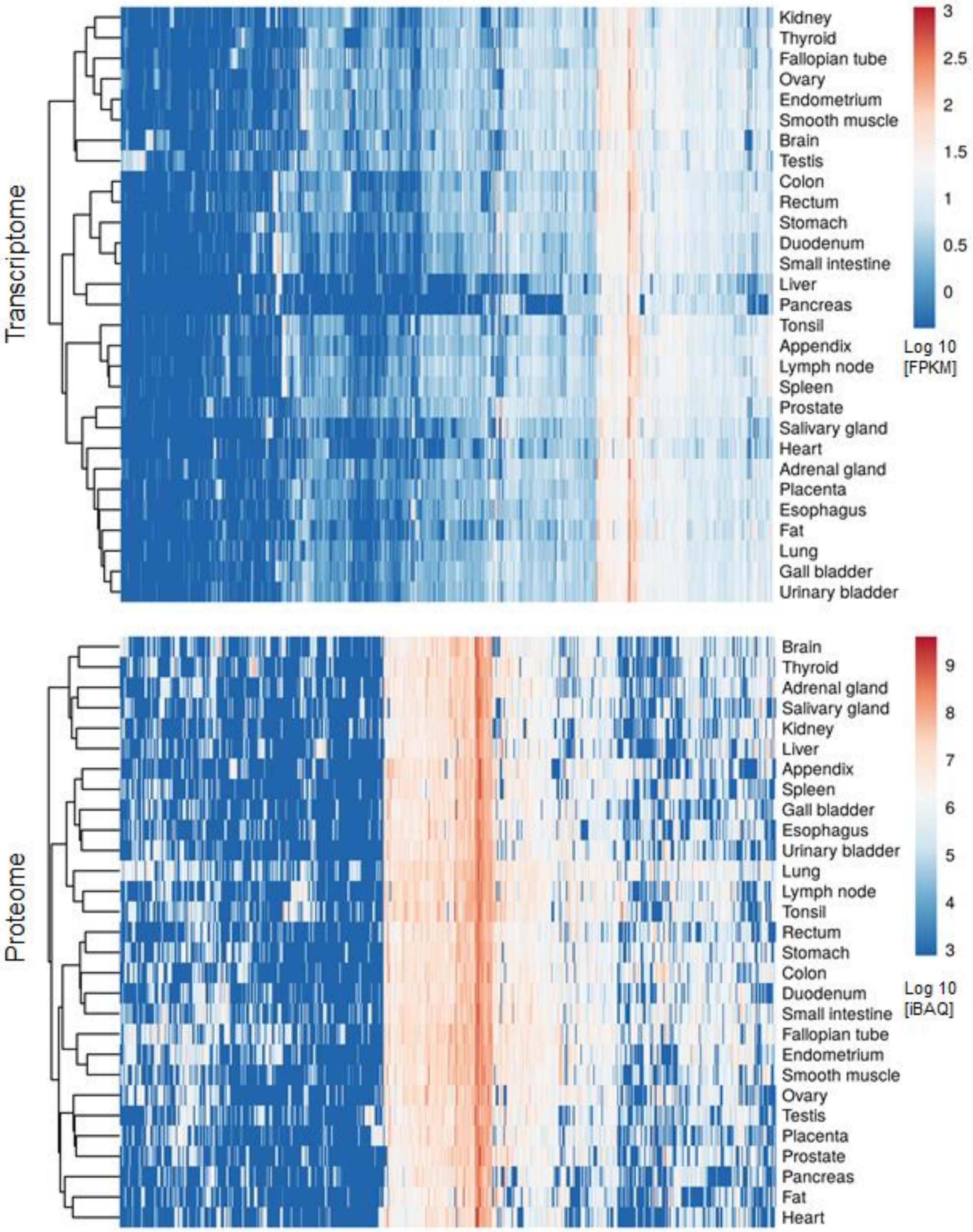


Figure 27. RNA-Seq gene expression profiles of 1,476 TFs and proteomic profiles of 724 TFs across the 29 human tissues. Tissues using hierarchical clustering by Pearson correlation and TFs by Euclidean distance. The missing values were imputed with ½minimum FPKM or iBAQ.

GPCRs. G protein-coupled receptors (GPCRs) transduce extracellular signals across the cell membrane. This class are the largest family of membrane proteins in the human proteome and the most important pharmaceutical targets. According to HGNC, there are

1,410 GPCRs that belong to 7 different subsets including 7TM orphan receptors, rhodopsin-like (Class A), secretin-like (Class B), metabotropic glutamate/pheromone (class C), frizzled (FZD, class F), taste receptors and vomeronasal receptors[204]. It is interesting that GPCRs are much more tissue-restricted compared with other functional classes[205]. If not ubiquitously expressed, these are potentially better targets for drug therapy[206]. It was also noticed that only ~20% GPCRs (a total of 300: 299 at the transcript level, 115 at the protein level) were identified in the 29 tissues. Figure 28 shows the number of GPCRs identified as transcripts and proteins in each tissue. Most of the detected GPCRs belong to rhodopsin (Class A). A total of 220 and 58 where observed at the transcript and protein level, respectively. The rhodopsin-like family, that also contain the olfactory receptors, still comprise the largest proportion of 'missing' proteins (~15%)[19]. While multi-transmembrane proteins are difficult to detect, the absent GPCR proteins may also be a consequence of low expression.

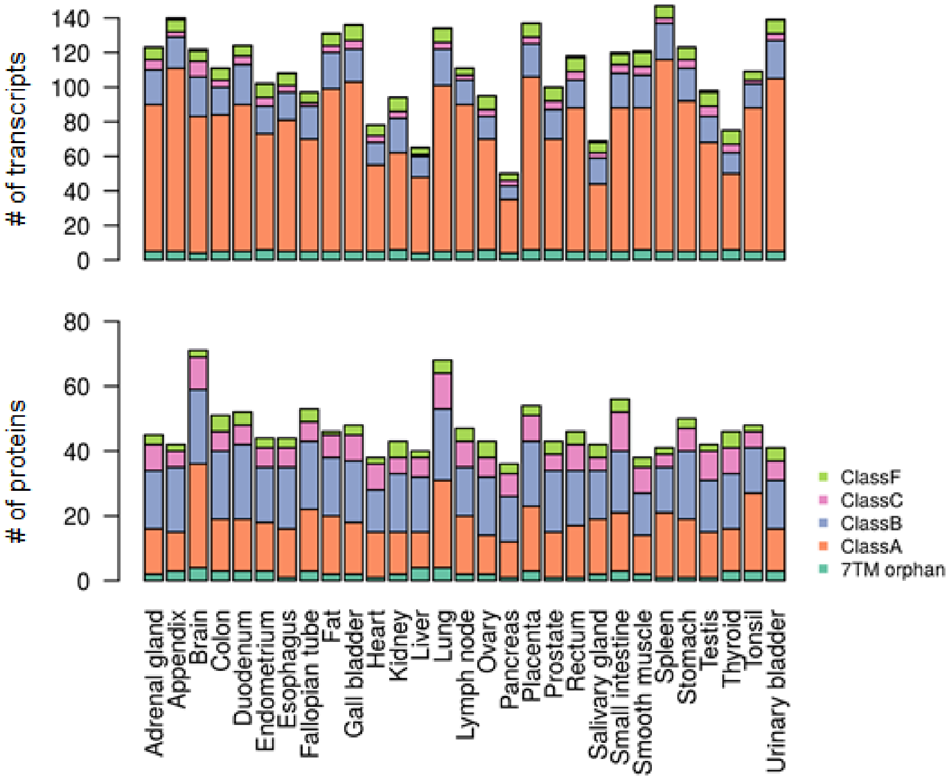


Figure 28. The identification of GPCRs in the transcriptome and proteome of the 29 tissues.

Kinome. Kinases, collectively referred to as the kinome, catalyze protein phosphorylation. Aberrant expression and/or activation/deactivation deregulate signal transduction networks, leading to diseases including cancer and inflammation. In the last years, kinases became one of the most popular enzyme super-families as drug targets for anti-cancer therapy. More than 250 kinase inhibitors (KIs) are currently undergoing clinical trials, and 37 have been approved for human use[207]. To explore the expression pattern of kinases in the 29 tissues, a list containing 504 kinase genes was used as a reference (downloaded from UniProt). The transcript (n=484) and protein (n=425) expression distribution of kinases had a very similar global trend. The number of kinases detected as proteins were quite similar across each tissue, ranging from 322 (rectum) to 386 (lung). In general, more kinases were detected as transcripts than proteins in each tissue, but this trend was reversed in the pancreas, liver and heart (Figure 29). An example of a kinase expressed in all tissues is PDK2, which plays a key role in the regulation of glucose and fatty acid metabolism and homeostasis. There were more kinases with enriched expressed in the brain than many other tissues. Similar to previous observations, most are CAMK kinases, and include CAMK 2A which is a prominent kinase in the central nervous system and may function in long-term potentiation and neurotransmitter release[208].

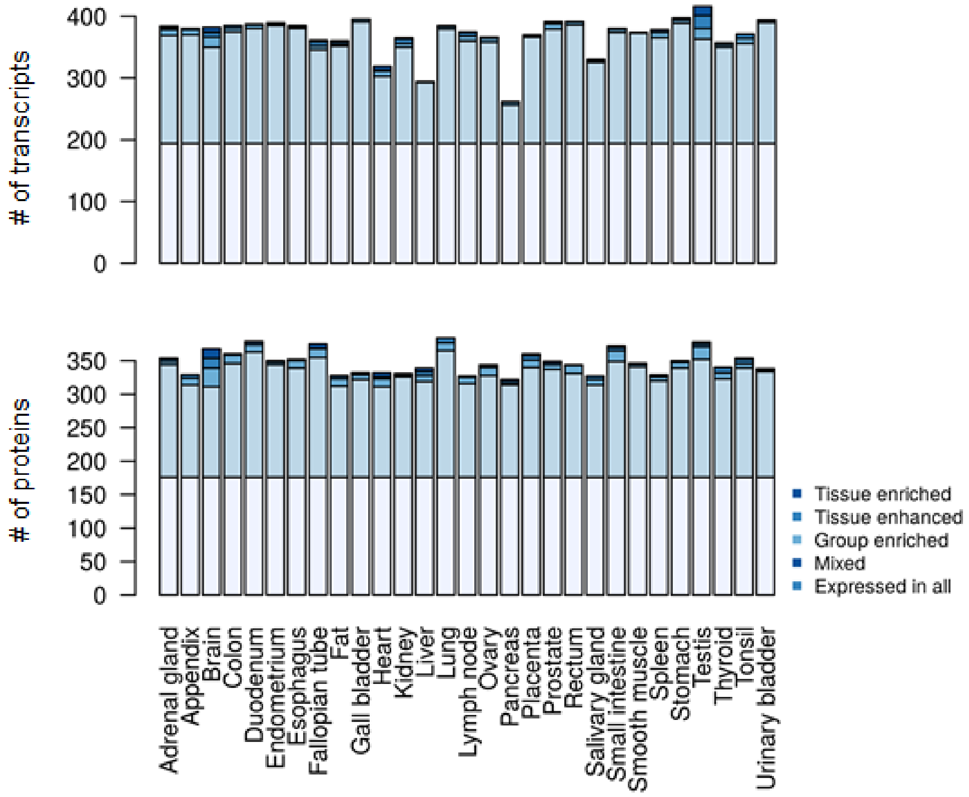


Figure 29. The identification of kinases in the transcriptome and proteome of the 29 tissues.

Phosphatases. The concerted activities of phosphatases and kinases regulate the phosphorylation levels that have diverse roles in cellular regulation and signaling. Unlike kinases, however, knowledge on phosphatases is still limited, especially the phosphatase–substrate relationship. According to DEPOD database, there are 238 genes coding phosphatases in humans[185]. In this work, most were identified as transcripts (n=229) and proteins (n=193). A similar trend to kinases was observed. Each tissue had approximately from 150 to 175 proteins, although less were detected as transcripts than as proteins in the pancreas and liver (Figure 30). These results suggest that some kinases and phosphatases act in concert with one another. Among the very few phosphatases with tissue-enriched expression, ACPP is the only one in prostate that acts as a tumor suppressor of prostate cancer through dephosphorylation of ERBB2 and deactivation of MAPK-mediated signaling. Expression level increases proportionally with prostate cancer progression[209] .

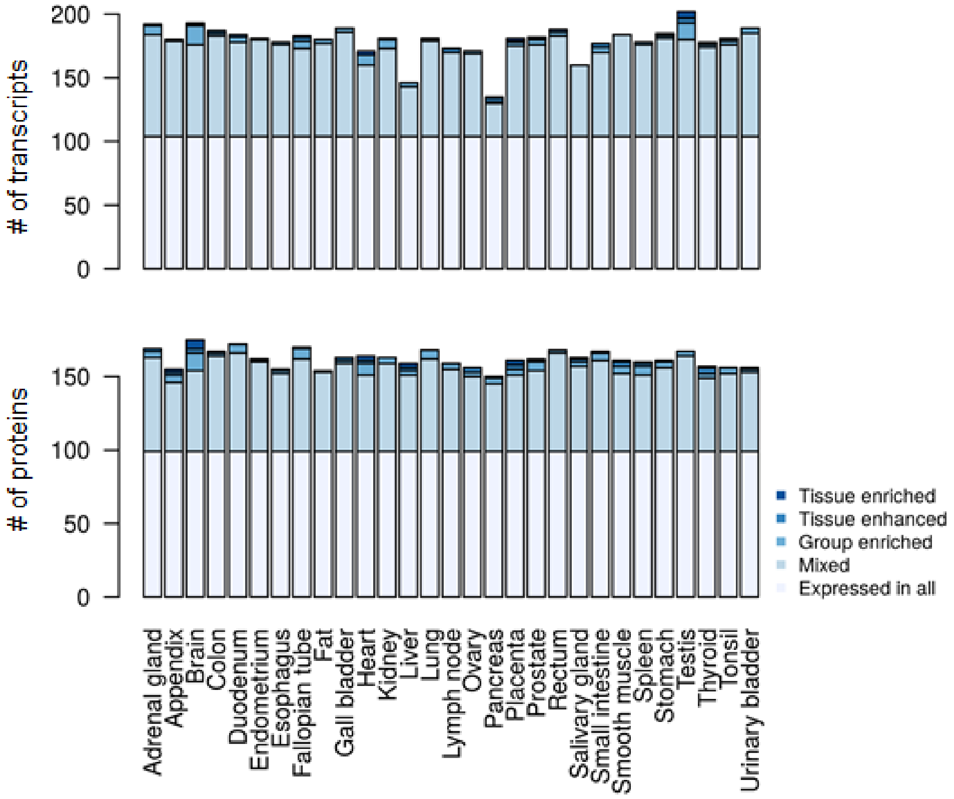


Figure 30. The identification of phosphatases in the transcriptome and proteome of the 29 tissues.

Chapter 4 Relationship between mRNA and protein expression

4.1 mRNA and protein expression

The matched proteomic and RNA-Seq measurements enabled the global analysis of transcript-protein relationships in healthy human tissues. The dynamic range of proteins detected by mass spectrometry spanned eight orders of magnitude which was much broader than that of the transcripts detected by RNA-Seq which spanned about four orders of magnitude (Figure 31). This difference alone explains (at least in part) the overall higher coverage of the expressed proteome by RNA-Seq compared to that of LC-MS/MS. The much wider dynamic range at the protein level implies that protein synthesis and protein stability play an important role in determining protein levels beyond mRNA levels.

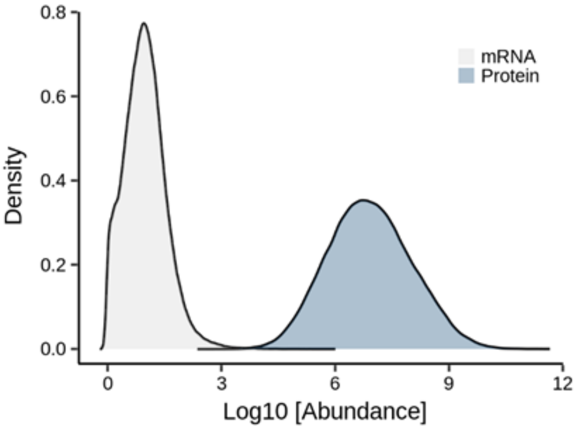


Figure.31 Distribution of the global transcript and protein abundance in all tissues. It is apparent that the dynamic range of protein expression vastly exceeds that of mRNA expression.

The often vast differences in mRNA and protein expression within a tissue can also be visualized by plotting the ranked order of relative intensities of transcript and protein (Figure 32). For example, in the heart (Figure 32 A), 41% of the total mRNA quantity (by FPKM) represents a single protein (MT-ATP8) and nearly 60% of all mRNA covers just five transcripts (MT-ATP8, MT-CO1, MT-ND3, MT-CO3 and MT-ATP6, all coded by mitochondrial genes). In contrast, about 13% of the total protein quantity (by iBAQ) is contributed by five proteins. Four of these five proteins are myosins that are the major protein that comprises the cardiac muscle thick filament and has a primary function in cardiac muscle contraction. One protein (hemoglobin) is a ‘contamination’ from blood that was present in the tissue and is also observed at high abundance across all other tissues. It is not unexpected that the heart is rich in both protein families because of the contractile

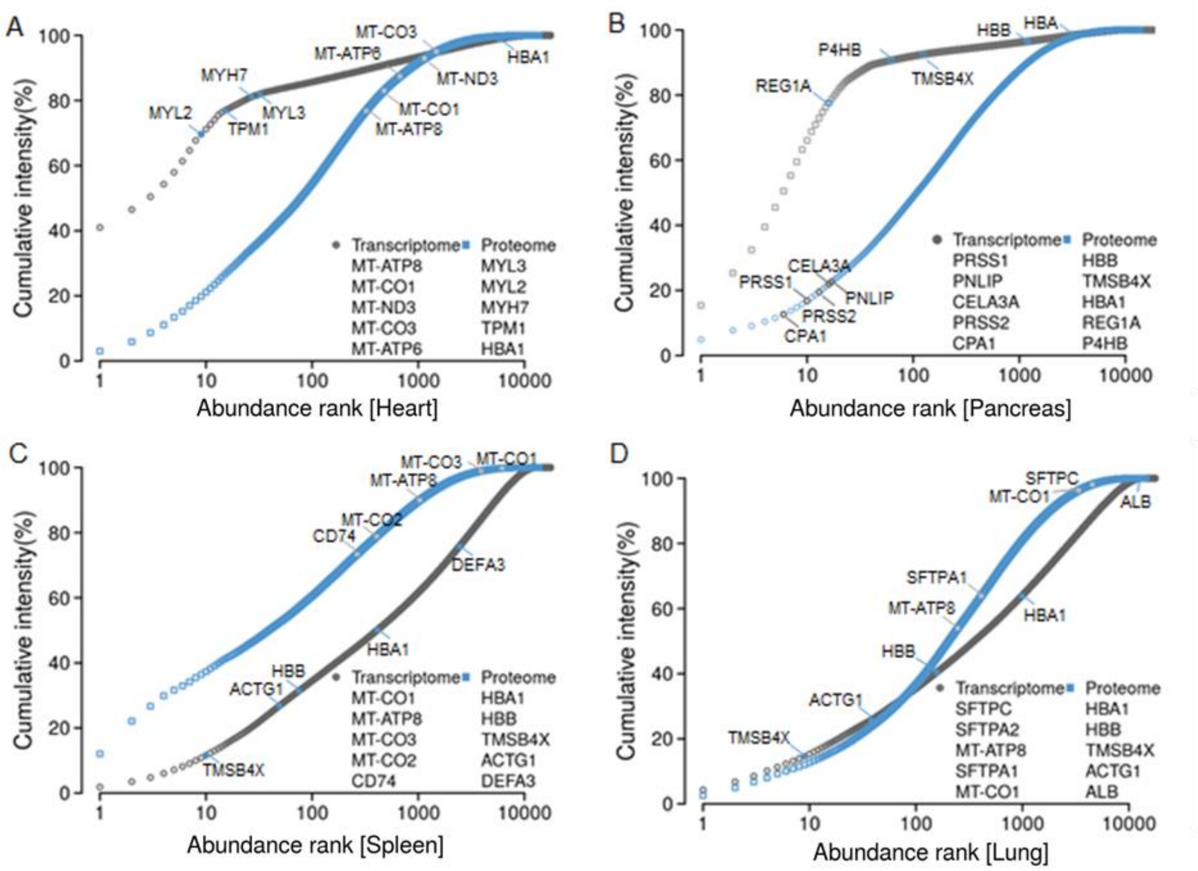


Figure 32. Cumulative protein mass from highest to the lowest abundance proteins in heart, pancreas, spleen and lung. The gene names of the top 5 mRNA transcripts and proteins are listed in the plots.

function of the organ which requires a lot of energy. It is possible that mitochondrial proteins are underrepresented in quantitative terms (these are not underrepresented merely by counting presence/absence) because the chosen lysis conditions may not have fully solubilized this organelle with high efficiency (discussed in the mitochondrial sub-proteome section) (Figure 32 B-D, for other tissues see Appendix Figure 1). The abundance distribution of transcripts and proteins is also quite different between tissues. The pancreas showed a similar trend to the heart; while the spleen showed opposite characteristics compared to the heart. The lung showed a more even distribution of transcript and protein levels. In the pancreas, spleen and lung, the top five most abundant genes represented by transcript and protein were different. In addition to the technical issues discussed above, this may also be a consequence of a protein inference problem that results from low sequence coverage in bottom-up proteomics. This may explain the absence of SFTPA2 at the protein level in lung. To systematically visualize the difference, genes coding the top 100 abundant mRNA and proteins were compared between each tissue pair (Ensembl gene annotation). It was surprising that only about 20% are the same and that the overlap only increased to about 60% for the 5,000 most abundant proteins and transcripts (Figure 33 A-B). Across all tissues, more than the top 100 proteins showed a similar expression level than that observed for the mRNAs (350 from 536 vs. 299 from 680). The top 100 mRNAs showed a more tissue-enriched expression (242/680 vs. 109/536) indicating the importance of protein abundance to maintain basic functional stability in cells (Figure 34).

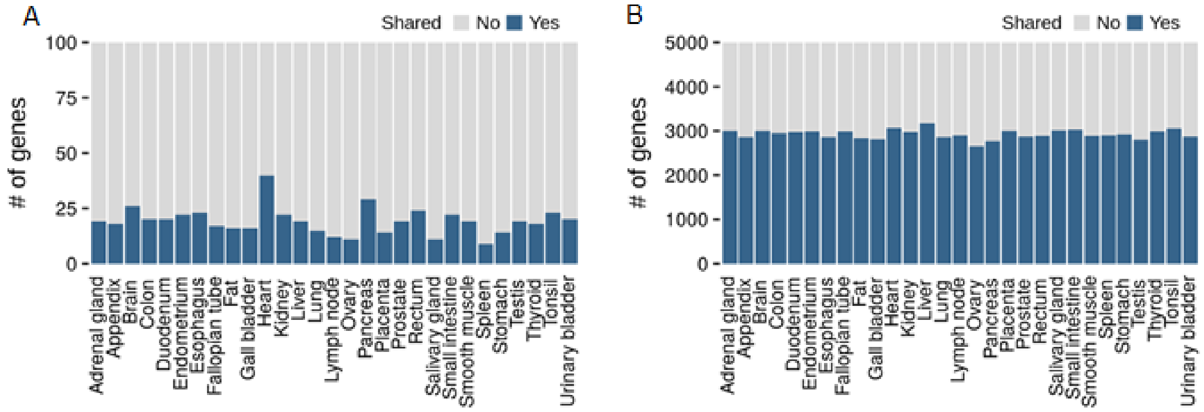


Figure 33. (A)The overall expression distribution of the top 100 mRNAs and proteins across the 29 tissues. (B) The overall expression distribution for the top 5,000 mRNAs and proteins across the 29 tissues.

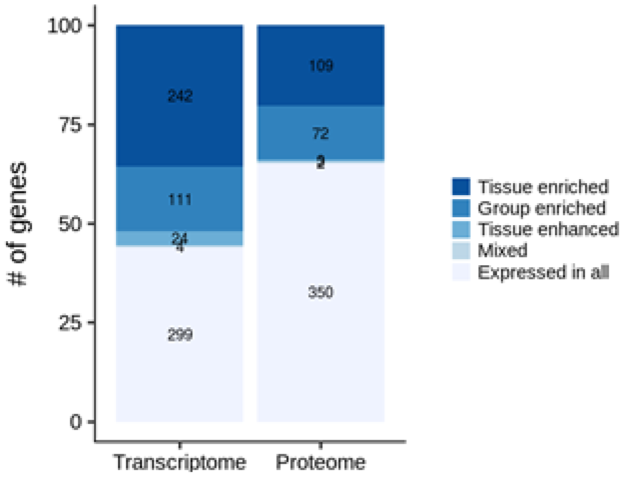


Figure 34. Distribution of the top 100 transcripts and top 100 proteins across the expression categories.

It is noteworthy that the proteomic data has a stronger correlation between tissues (median of 0.77) than for the transcriptomic data (median of 0.67) (Figure 35 A). This may be due to the fact that the dynamic range of protein levels is larger and thus small biological or technical variations of individual genes have a negligible impact on the overall rankings. It may also imply that there are mechanisms in the cells that buffer the protein quantities against changes in mRNA abundance[104,210]. For both transcripts and proteins, the strongest correlations were observed for the anatomically-adjacent small intestine and duodenum. At the proteomic level, the brain and heart showed clear differences to other proteomes; while the gastrointestinal organs appeared to be more similar to one another. Despite changed mRNA levels, another observation that supports the buffering of the protein level is that stable protein complexes tend to maintain constant protein expression[195]. As shown in Figure 35B, the 7α and 7β subunits of the 20S proteasome complex have a similar protein expression in all 29 tissues. This is despite the fact that PSMA5 has a much low mRNA abundance than other members of the complex. Thus suggesting that protein abundance is much more conserved than mRNA abundance[211]. Notably, the reduced protein level variance by protein buffering can dampen the mRNA-protein correlation[104].

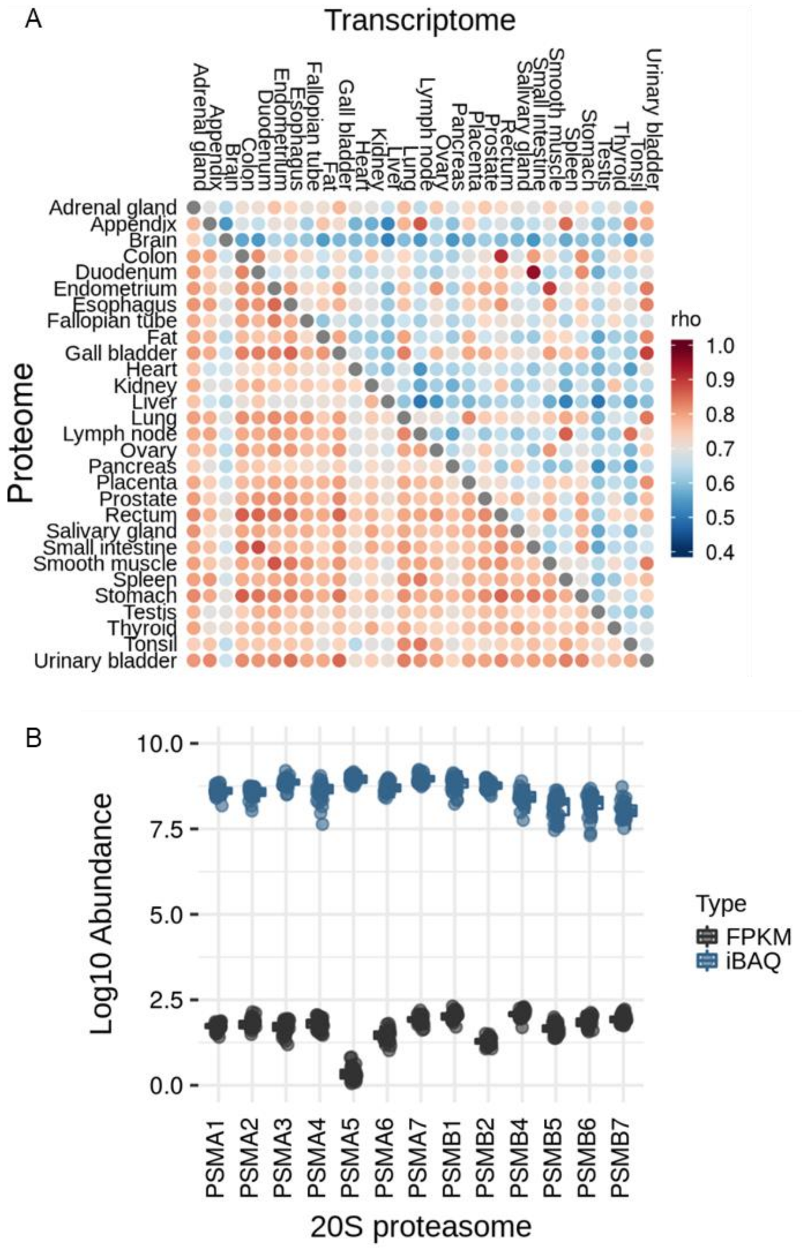


Figure 35. Protein levels buffer changes at the mRNA level. (A) Global correlation analysis of proteomes and transcriptomes across human tissues. It is apparent that proteomes correlate

stronger between tissues than transcriptomes. (B) The expression distribution of 20S proteasome members at the transcript and protein level.

Despite the discrepancy of mRNA and protein expression, visualizing the transcriptome and proteome profiles in a plane using co-inertia analysis (CIA)[187] indicated that mRNA and protein levels have higher similarity to one another within tissues than between tissues (Figure 36). This observation is also reflected by an RV coefficient of 0.78 (a multivariate generalization of the squared Pearson correlation coefficient). Moreover, the CIA grouped several tissues according to similarities in physiological function; with tissues from the immune system and the gastrointestinal tract representing the largest groups. These results are consistent with independently performed hierarchical clustering analysis (Figure 19). It is interesting to note that this clustering appears to be driven by the cellular composition of individual tissues. For instance, the appendix co-clusters with the spleen, lymph node and tonsil and all four tissues contain a large fraction of lymphocytes (Figure 37, blue panel). Similarly, the stomach, duodenum, small intestine, colon and rectum all comprise a large proportion of (intestinal) glandular cells. These are important determinants of the molecular composition of these tissues (Figure 37, grey panel).

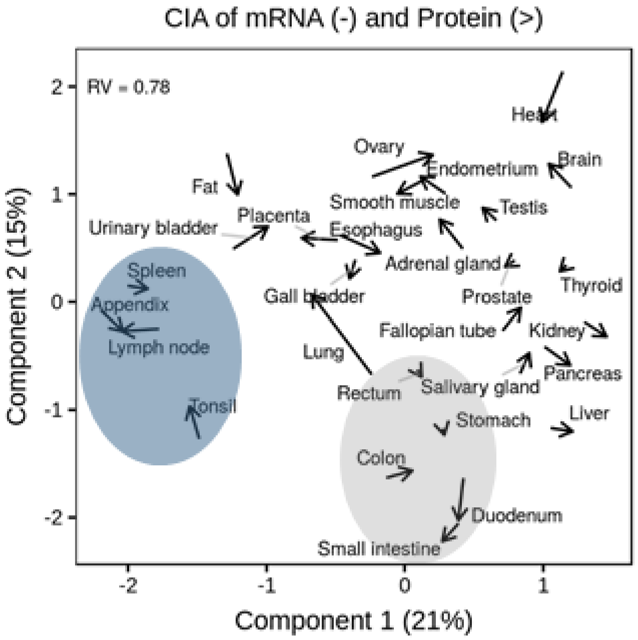


Figure 36. Co-inertia analysis of transcriptome and proteome levels of all 29 tissues (arrow base: transcriptome; arrow head: proteome) showing that some tissues have similarities in transcript and protein expression profiles.

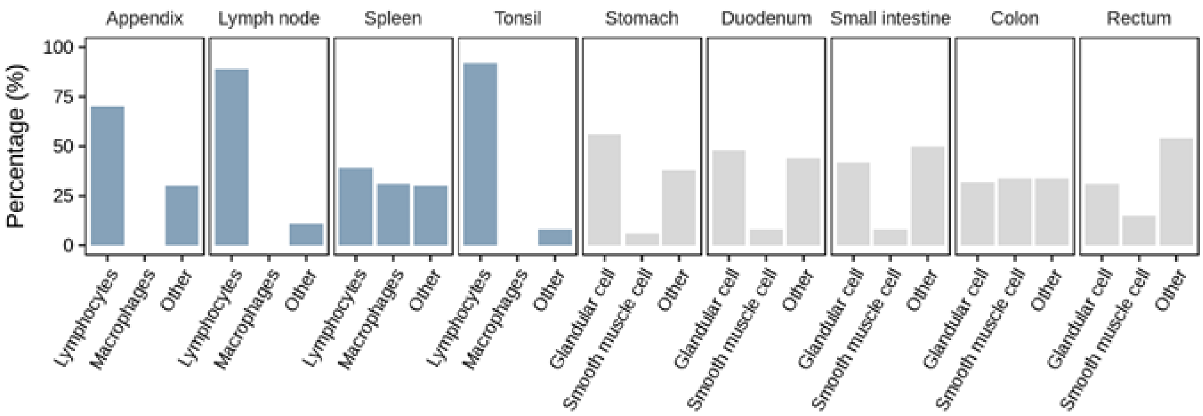


Figure 37. Average cellular composition of selected tissues showing that the similarities in CIA analysis are largely driven by similarities in cell type.

4.2 mRNA and protein correlation

To explore to what extent the differences in mRNA level are reflected at protein level, the mRNA and protein abundance for each gene within individual samples were compared. For example, the brain showed a significant protein-mRNA correlation of 0.53 (spearman correlation coefficient, p-value<0.01) and revealed a markedly linear relationship with a log-log slope of 2.6 (Figure 38A). Similar to brain, all tissue pairs showed a significant protein-mRNA correlation ranging from 0.41 to 0.58 (Appendix Figure 2), which are comparable to previous reports[212], and shows a nearly quadratic relationship between mRNA levels and protein levels in every tissue (between 1.8 and 2.7 for all 29 tissues, Figure 38B). The data implies that the number of protein copies produced per molecule of mRNA appears to be much larger for high- than for low-abundance transcripts. This may be rationalized by cellular economics such that genes encoding highly-abundant proteins not only express high mRNAs levels, but also encode regulatory elements that favor high translation efficiency and high protein stability[133].

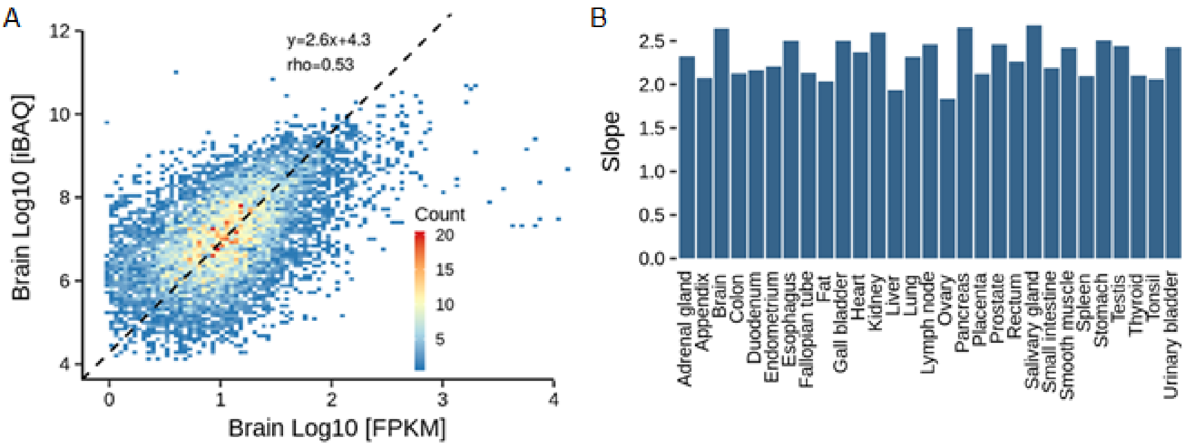


Figure 38. (A) Protein to mRNA abundance plot for brain tissue. The slope of the regression line indicates that high-abundance mRNAs give rise to more protein copies per mRNA than low-abundance mRNAs. (B) Slopes of protein vs. mRNA abundance plots for all tissues.

Next, the concordance of the variation of mRNA and protein concentration for each single gene across the 29 tissues was examined. To ensure the correlation quality, only the genes expressed in both mRNA and proteins level in at least 10 tissues (n=9,845) were studied. Due to the fact that a majority of the proteins are expressed at similar levels across human tissues, it was not surprising that the correlation of mRNA/protein ratios across the tissues was, in general, not very strong (Figure 39; median 0.35). Still, there is a positive correlation in almost 90% of all cases and almost half were also statistically significant. These results are similar to the observations in human tumors[138,139,150]. Considerable care must to be taken when interpreting this distribution. It is generally observed that proteins that are highly (low) expressed in one tissue are also highly (low) expressed in many (but not always all) other tissues. A significant correlation requires variation at protein and mRNA level, the larger the variation the more significant. As shown in Figure 39B, the transcript and protein levels of the tyrosine kinase SYK are highly-correlated across tissues reflecting the specialized function of the protein in T- and B-cell biology. In contrast, other proteins such as EIF4A3 (a DEAD-box RNA helicase involved in translation initiation) showed no such correlation. This is merely the result of similar expression levels in most tissues that also reflects the role of such proteins in central biological processes across all tissues[213]. All the above illustrates that there must be multiple molecular factors and mechanisms determining the quantitative expression of a protein. This particular aspect of the present mRNA/protein expression resource may be particularly useful for the community as it provides a rich data source for the study of protein expression control (see Section 4.4).

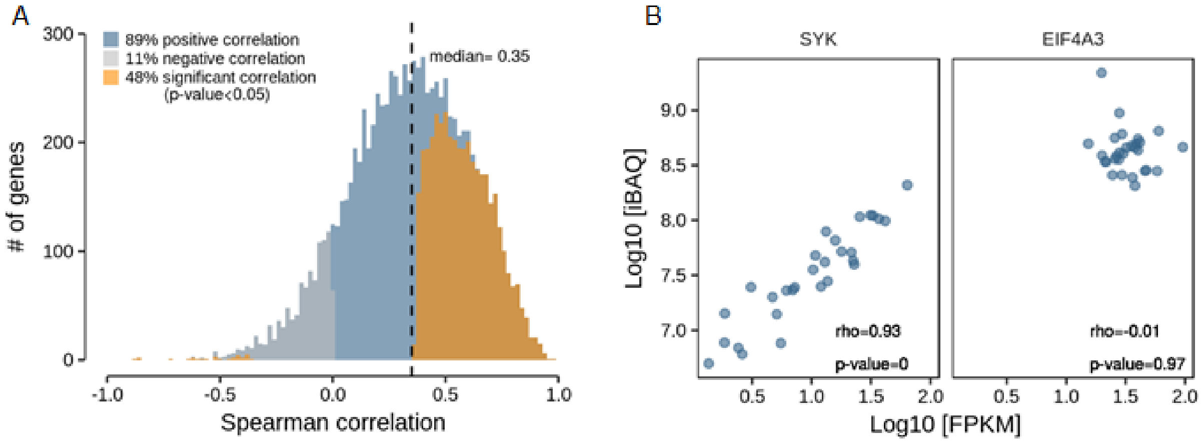


Figure 39. (A) Correlation analysis of the expression of proteins across all tissues. Almost 90% of all proteins show a positive correlation across tissues. Different biological pathways and processes showed significantly-different levels of correlation (bottom panel). Metabolic

pathways and the interferon response displayed high mRNA-protein correlation. (B) Examples of proteins that show high (SYK, left panel) or no (EIF4A3, right panel) correlation of protein expression across tissues. While the former indicates that different tissues express different quantities of SYK, EIF4A3 expression appears to be similar across all tissues.

4.3 protein to mRNA ratio

Due to a similar protein and/or mRNA expression level across cell lines or tissues, protein and mRNA abundances originating from the same gene are not sufficiently correlated to act as proxies for one another in steady state cells or tissues. Several genome-wide studies have suggested, however, that the protein abundance to mRNA abundance ratio (PTR) are rather constant in human cell lines[100,214] and tissues[14,215]. Consistent with previous findings, the PTR across the 29 tissues are rather stable. Figure 40 shows the PTR distribution of 8 example genes, and the distribution of the coefficients of variation for log scale protein/mRNA ratios (n=9,485 genes, protein/mRNA pairs exist in at least 10 tissues), most of which are less than 15%. The stability of PTR across tissues and the difference between different genes enables the systematic exploration of the quantitated contribution of known post-transcriptional regulatory elements and the discovery of novel elements based on PTR.

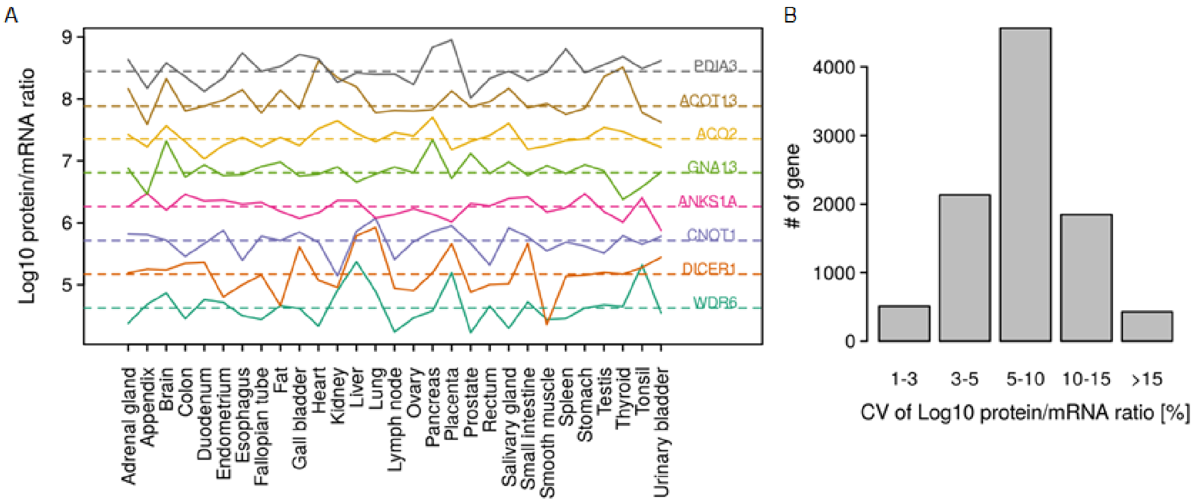


Figure 40. Analysis of mRNA and protein levels across at least 10 tissues (n=9,485) showed that the protein/mRNA ratios are largely constant. (A) Log_{10} PTRs of 8 example proteins across 29 tissues. (B) The distribution of coefficient of variance of PTR in $\log_{10}$ scale across tissues.

4.4 Sequence determinants of protein-to-mRNA ratio

To identify and quantitate sequence determinants of the protein-to-mRNA ratio (PTR), a model predicting tissue-specific PTR ratio (n=11,575) was derived from mRNA and protein sequence alone. These model-based analyses were performed by Basak Eraslan (Computational Biology, Technical University of Munich). Briefly, the model is a multivariate linear model that included 16 mRNA-encoded and protein-encoded sequence features representing 172 post-transcriptional regulatory elements that are known to modulate translation initiation, elongation, and termination, and protein stability (Figure 41, generated by Basak Eraslan). The model also included 15 novel motifs that were *de novo* identified through a systematic testing of the association between median PTR ratio across tissues and presence of k-mers, *i.e.*, subsequences of a predefined length *k*, in the 5' UTR, the coding sequence, the 3' UTR, and the protein sequence. Altogether the model predicted the PTR ratio of individual genes at a median precision of 3.2-fold from sequence alone, while the PTR ratio spans about 200-fold across 80% of the genes. This model explained 22% of the variance in median across tissues (min. 18%, max. 26%, Figure 42A, generated by Basak Eraslan). Codons alone explained 17% of the variance in the median across tissues, followed by the CDS length, the linear protein motifs, and our *de novo*-identified CDS amino acid motifs. The amount of explained variance is driven by the combination of effect size, frequency and variability of the feature across genes. Hence, sequence features that play a crucial role in translation, such as the Kozak sequence, the 5' and 3' UTR motifs, explain a small fraction of the variance between genes although the effect size can be large. Besides, the positive correlation between the predicted PTR ratios based on sequence features and the mRNA levels supports the hypothesis that highly transcribed genes also have optimized sequences for post-transcriptional up-regulation, hence yielding higher amounts of proteins, consistent with earlier work by Vogel et al.(Figure 42 B, generated by Basak Eraslan)[133]. This model-based analysis showed that high protein-expression levels are achieved by the joint effect of high mRNA levels and genetically-encoded elements favoring the synthesis and stability of proteins. A further fraction of this apparent amplification effect may be explained by regulatory elements that affect both the mRNA levels and protein per mRNA copy numbers.

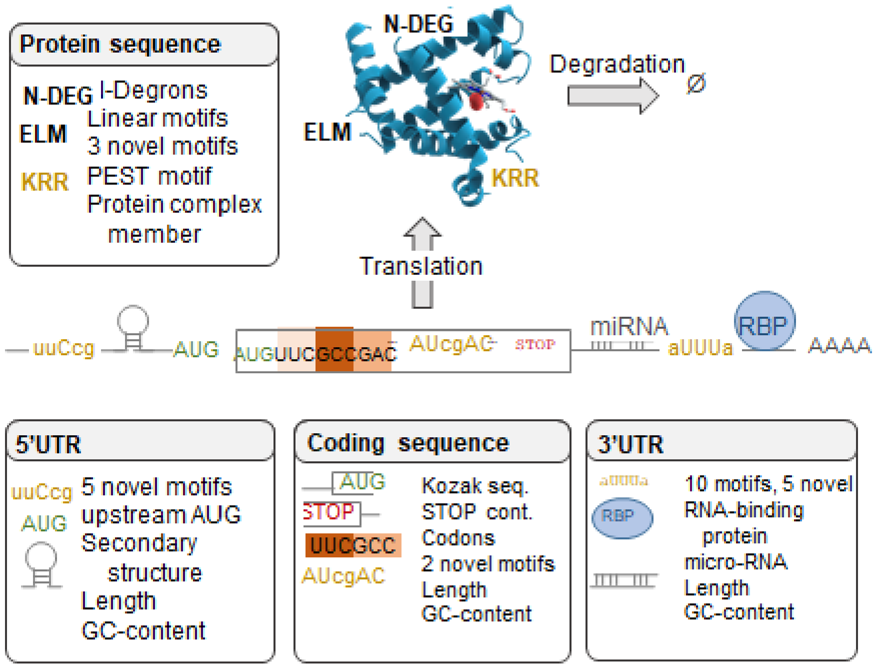


Figure 41. Analysis of sequence determinants of protein-to-mRNA ratio. Sequence features of 5' UTR, coding sequence, 3' UTR, and protein sequence were considered in the model.

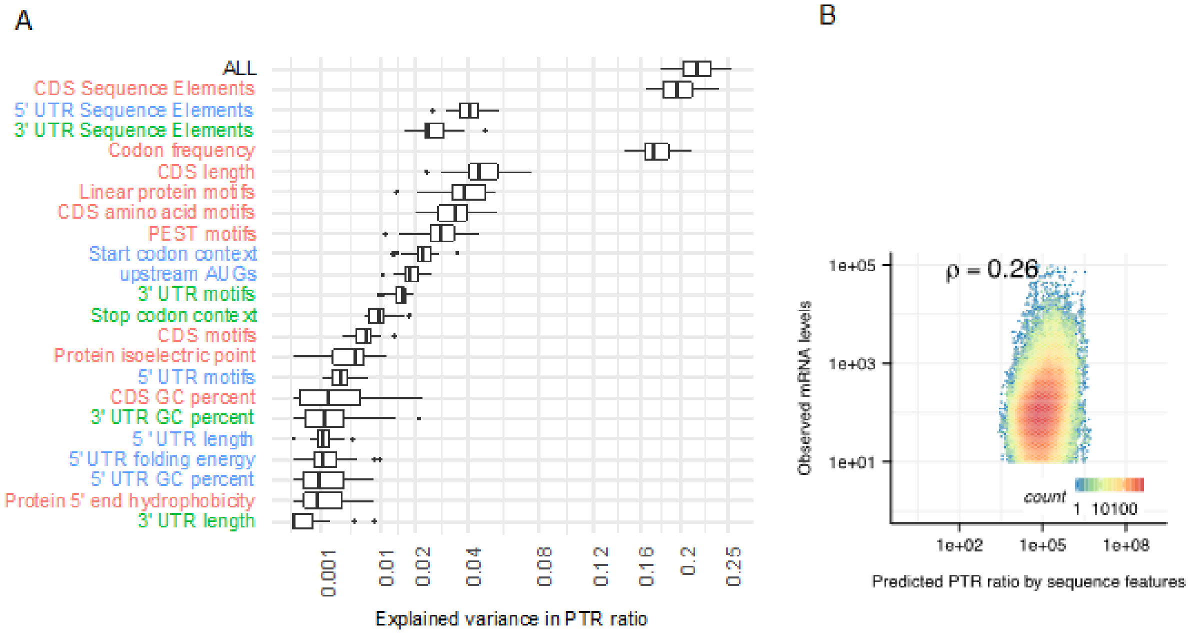


Figure 42. (A) Distribution of the explained variances (R2) in tissue-specific PTR ratios only by each of the individual sequence features, and by the models combining these based on the mRNA region types (5' UTR (blue), CDS (green) and 3' UTR (green)) and combined (ALL). Features are sorted based on median-explained variances. (B) mRNA levels (y-axis) versus predicted PTR ratios by sequence features for all genes in all tissues. Colors represent the density of genes.

4.5 Motif-specific binding protein characterization by RNA pulldown

In order to investigate the functional understanding of sequence elements, an RNA oligo competitive-binding assay was developed to systematically identify motif-specific RNA binding proteins and to measure interaction strength thereof (Figure 43, Methods and Materials). This method was developed from the kinobead pulldown[188,216]. Briefly, the RNA probes (length=20 bp) were immobilized on NHS-activated sepharose beads through the 5' amino group on the C6 linkers, thus enabling the affinity capture of the motif-binding proteins from HEK293FT cell lysates. Lysate competition using the free RNA probes in increasing concentrations (0 nM, 0.3 nM, 1 nM, 3 nM, 10 nM, 30 nM, 100 nM, 300 nM, 1000 nM, 3000 nM, 10,000 nM, 30,000 nM) led to a concentration-dependent loss of specific protein binding to the beads. For each concentration of competitor, proteins bound to the beads can be identified and quantitated using a label-free quantitative mass spectrometry (LFQ[94]). The more protein that binds to the free RNA oligos in solution, the less that can bind to the beads and the decrease in peptide intensity is detected by the mass spectrometer. The relative quantity of each protein per concentration of competitor can then be used to derive EC_{50} values from a concentration response plot using nonlinear regression analysis. The EC_{50} value for each protein can be converted to a binding constant K_d by applying a correction factor that accounts for the depletion of a protein from the lysate in the affinity pulldown. The RNA pulldown profile of AAUAAA (using a HEK 293FT cell lysate incubated with free RNA probes that are identical to those immobilized on the beads) showed a low K_d value for cleavage stimulation factor subunit 2 (CSF2, K_d=1.9). CSF2 has a documented function of recognizing polyadenylation signal (PAS). A much higher K_d was calculated for heterogeneous nuclear ribonucleoprotein U-like protein 1 (HNRNPUL1, K_d=339). No curve was generated for DDX5, a non-specific binding protein of the AAUAAA probe (Figure 43C-F).

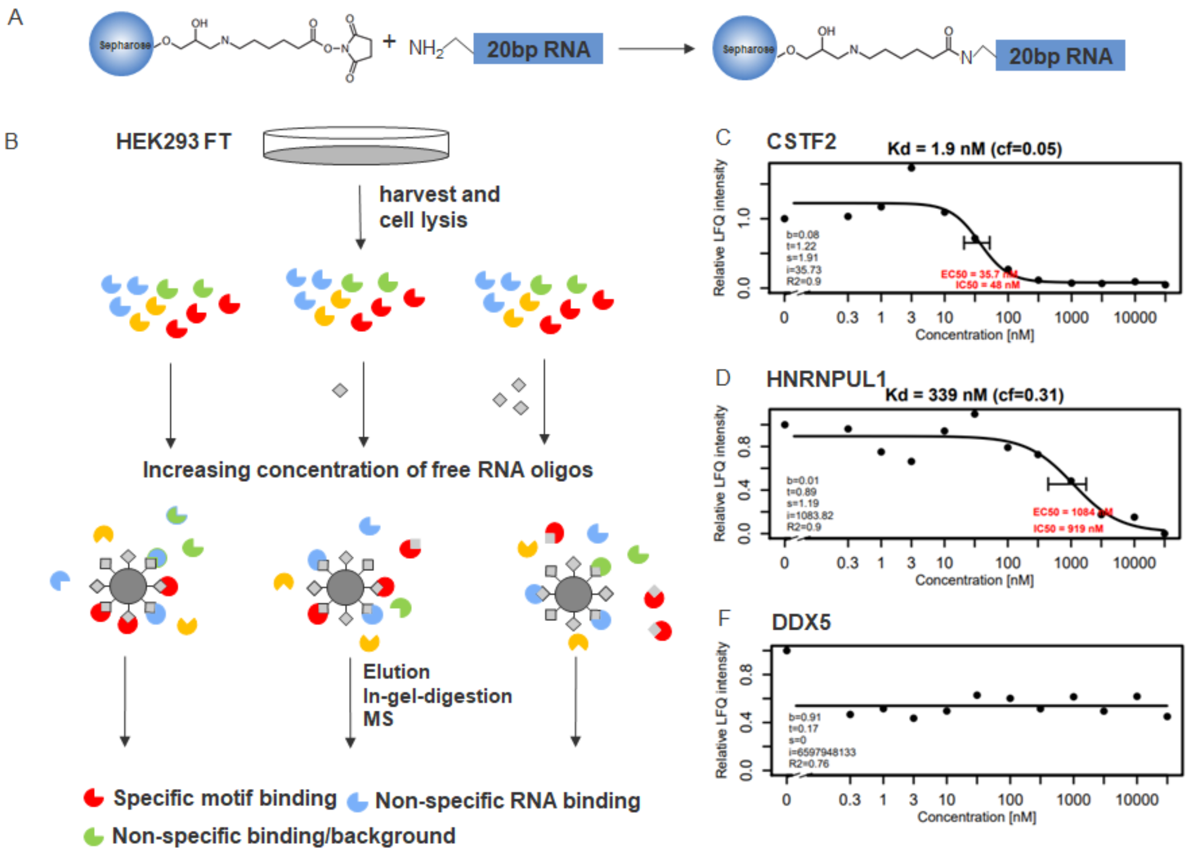


Figure 43. (A) Coupling RNA oligos with 5' amino modifier to sepharose beads. (B) RNA oligo competitive-binding assay workflow. The protein repertoire from lysates of the HEK293FT cell lines was incubated with increasing concentration of RNA oligos. The affinity matrix captured RNA-binding proteins, and other proteins with binding that is not influenced by the given RNA oligo concentration. LC-MS/MS analysis quantitates the amount of captured protein at each RNA concentration, thus enabling the computation of concentration-dependent affinity curves. (C-F) Concentration response curves for AAUAAA_1 probe and cleavage stimulation factor subunit 2 (CSF2), heterogeneous nuclear ribonucleoprotein U-like protein 1 (HNRNPUL1), and probable ATP-dependent RNA helicase DDX5 (DDX5).

Once the specificity of the chemical proteomic assay was established, 3 motifs from sequence determinants of protein-to-mRNA ratio analysis and corresponding randomized motifs were investigated. These included the polyadenylation signal AAUAAA, the AU-rich element UAUUUAU and a 5' UTR novel motif CUGUCCU. The known motifs were also used as positive controls. Randomized sequences were used as negative controls. In total, 253 proteins were identified. These directly or indirectly (*e.g.*, as complex members) interacted with the assessed RNA motifs in a sequence-specific or sequence-independent manner. Of these, 88 proteins were annotated as RNA-binding proteins (RBPDB v1.3.1[217]) and 247 proteins were annotated as RNA-processing or -binding proteins, nuclear localized proteins, or (mitochondrial) ribosomal proteins (according to DAVID v6.7[218]). For further analyses, the sequence-specific binding proteins were defined based on the estimated K_d values. In order to be motif-specific, a K_d of the protein-RNA interaction must be at least 10 times more potent than the next best motif or negative control (Appendix Table 2). In this way, general RNA-binding proteins such as HNRNPUL1, could be excluded.

AAUAAA is the central sequence motif of polyadenylation signal (PAS) and is a defined feature of eukaryotic protein-coding genes. PAS are required for polyadenylation of pre-mRNA, cleavage in 3′UTRs and the promotion of downstream transcriptional termination[219]. The AAUAAA motif exists in 8,653 from 11,575 genes investigated in sequence determinants analysis and shows a positive effect on PTR (1.17). In the competition binding experiments, 50 motif-specific interactors of the consensus polyadenylation signal sequence AAUAAA were identified; 32 are known to bind poly(A) tails and 12 have an annotated RNA-binding domain. The results unambiguously recapitulate that the motif is tightly bound by the cleavage and polyadenylation specificity factor (CPSF) complex[220] (Figure 44). As subunits of the CPSF complex: CPSF1, CPSF2 CPSF3, and FIP1L showed similar affinities to AAUAAA (Figure 45). Other known binders with high affinity are subunits of cleavage stimulating factor (CSTF) complex, CSTF1, CSTF2 and CSTF3. These are required for polyadenylation and 3'-end cleavage of mammalian pre-mRNAs[221,222]. It is also not surprising that many spliceosomal proteins were enriched, as PAS also functions essentially as a splicing enhancer and proteins have dual roles in both polyadenylation and splicing[223].

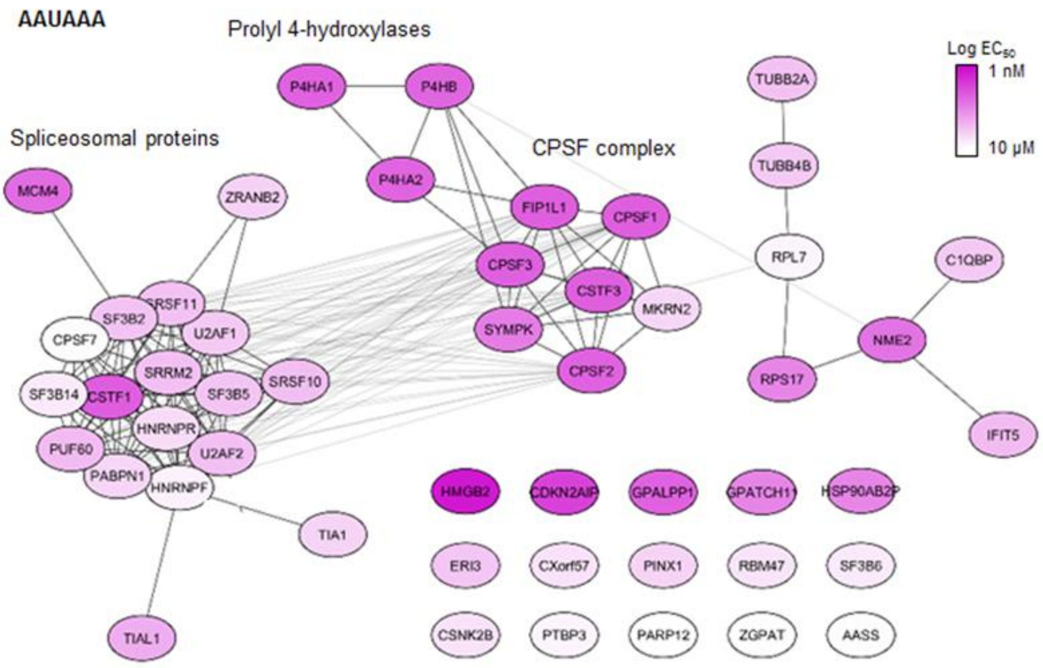


Figure 44. AAUAAA motif-specific RNA binding proteins (and complex partners) and the interaction strength to the free RNA probe; node color: pEC50; physical and functional interactions of proteins derived from STRING.

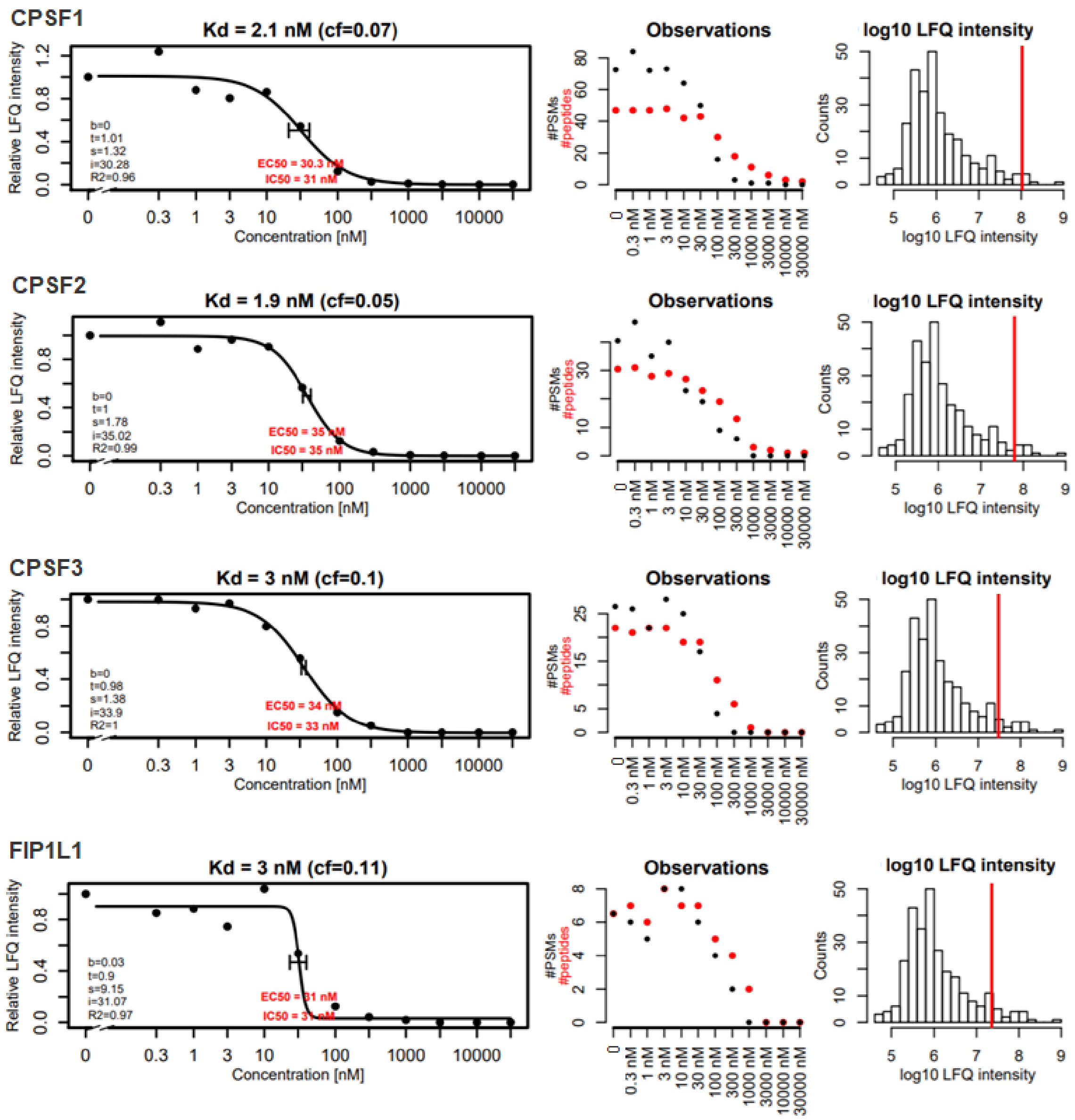


Figure 45. Concentration response curves for the polyadenylation signal motif AAUAAA and the cleavage and polyadenylation specificity factor (CPSF) complex (enriched by probe AAUAAA_1).

The second positive control was the 3' UTR AU-rich motif UAUUUAU [224]. This motif occurred in 3,158 (27%) of the investigated genes and was associated with lower PTR ratios (0.78). The assay identified 38 motif-specific interacting proteins (Figure 46). These include the zinc-finger RNA-binding proteins ZFP36L1 and ZFP36L2 (K_d=24 nM, and 11 nM, respectively, Figure 47). These are known to destabilize several cytoplasmic AU-rich element (ARE)-containing mRNA transcripts by promoting poly(A) tail removal or deadenylation, and hence provide a mechanism for attenuating protein synthesis[225,226]. The competition assay also revealed interaction of many other proteins including ZCCHC11, MEX3C, MEX3D, CNBP, SKIV2, and TTC37 that are involved in mRNA decay consistent with the primary function of the AU-rich element. Although lacking the motif specificity, the well-studied AU-rich binding protein ELAVL1/HuR (stabilizer of a wide variety of mRNAs) was also highly-enriched by the UAUUUAU probe (also enriched by the AAUAAA_2 probe) and showed a much lower affinity (K_d=440 nM). It was interesting to note that mitochondrial ribosomal proteins show high affinity. This may not influence the expression of most of the UAUUUAU-containing genes due to limited access in the cell. Nevertheless, some unknown functions may be implied.

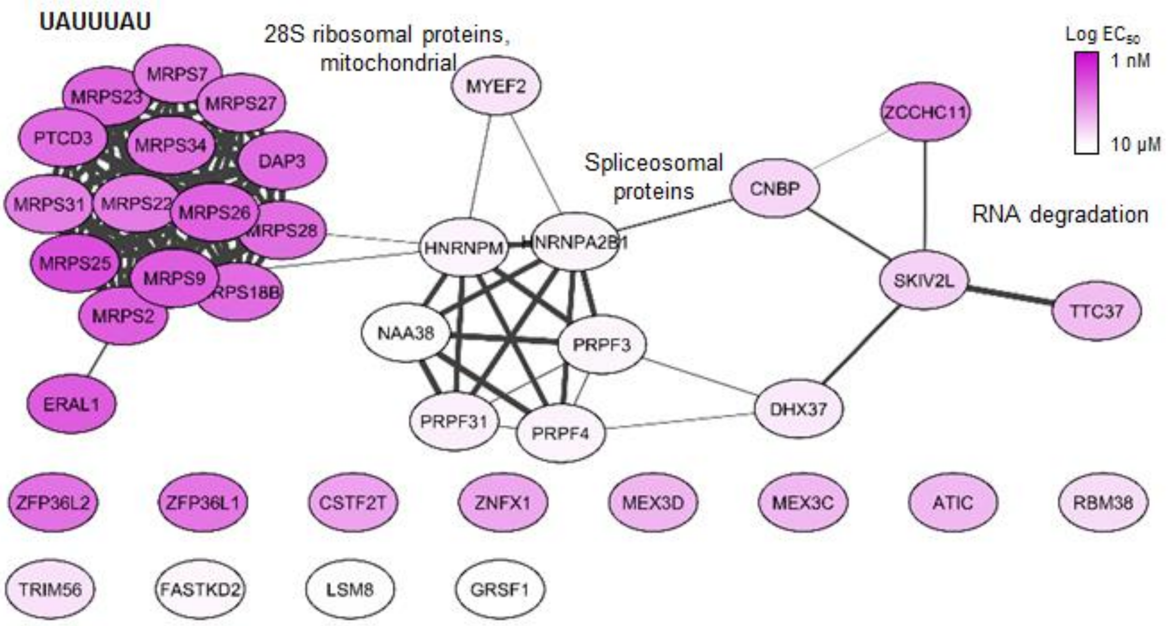


Figure 46. UAUUUAU motif-specific RNA binding proteins (and complex partners) and the interaction strength to the free RNA probe; node color: pEC$_{50}$; physical and functional interactions of proteins derived from STRING.

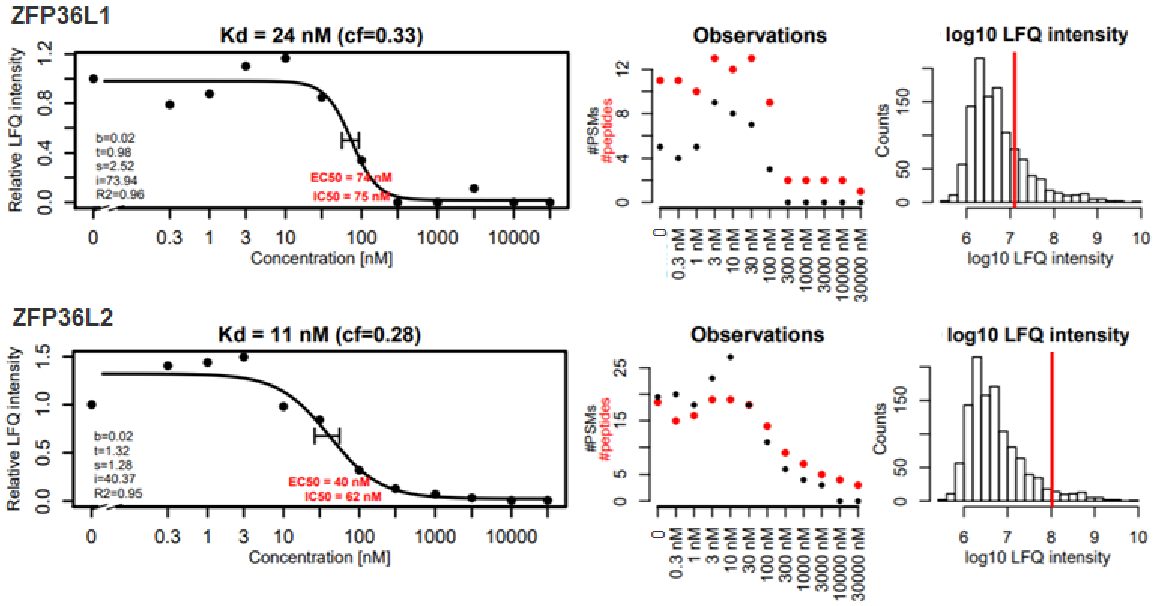


Figure 47. Concentration response curves for the polyadenylation signal motif UAUUUAU and the mRNA-decay activator protein 1 and 2 (enriched by probe UAUUUAU_1).

The 5' UTR motif CUGUCCU is one of the novel motifs t with a predicted positive effect on the PTR ratio (1.33) that was observed in 323 (3%) of the investigated genes. For this motif, the interaction with 30 binding partners was identified and quantitated. Interactors included 19 proteins from the 39S mitochondrial ribosomal subunit, 5 proteins from the 40S ribosome complex, and 4 proteins with a KH domain known to be involved in splicing (Figure 48). The ribosomal proteins tightly bind the 5' UTR motif with an affinity of 68 nM (±31 nM std. dev.). It can be speculated that the presence of this motif enhances the interaction of the 5' UTR with the mito-ribosome and also with the small subunit of the cytoplasmic ribosome. This subunit plays a key role in translation initiation[227], leading to a higher efficiency in translation initiation. Figure 49 shows examples of concentration response curves for the mitochondrial ribosomal proteins, MRPL11, MRPL12 and MRPL13. Two other proteins bound by the motif are ANGEL2 and IGF2BP3. ANGEL2 is known to bind the 3' UTR of mRNAs resulting in stabilization[228], and IGF2BP3 may recruit and cage target transcripts to cytoplasmic protein-RNA complexes. Like other IGF2BPs, IGF2BP3 may thereby modulate the rate and location at which target transcripts encounter the translational apparatus and shield these from endonuclease attack or microRNA-mediated degradation[229,230].

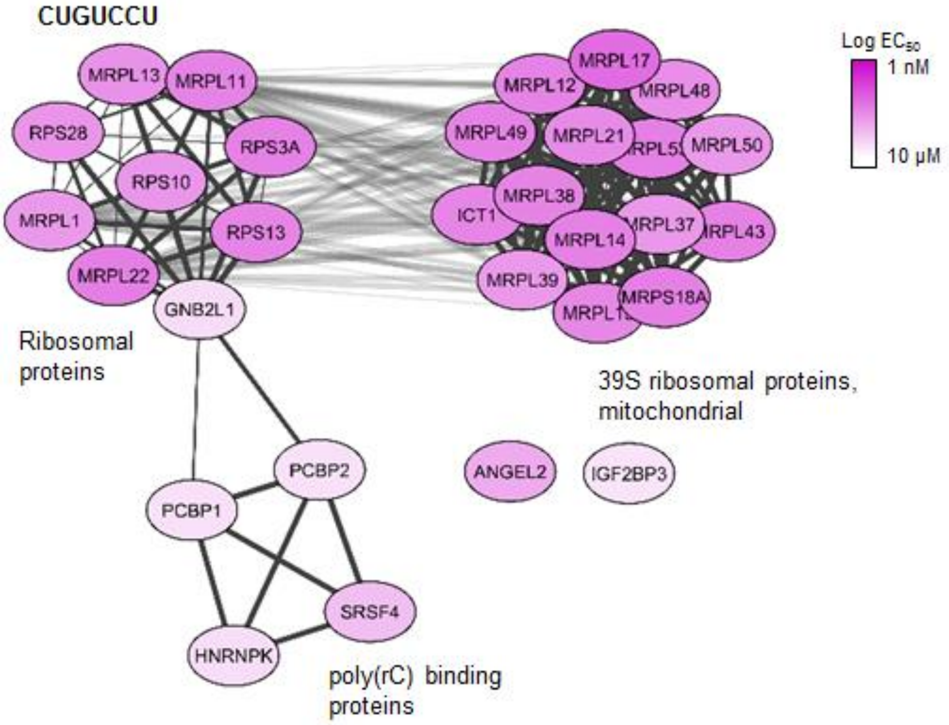


Figure 48. CUGUCCU motif-specific RNA binding proteins (and complex partners) and the interaction strength to the free RNA probe; node color: pEC$_{50}$; physical and functional interactions of proteins derived from STRING.

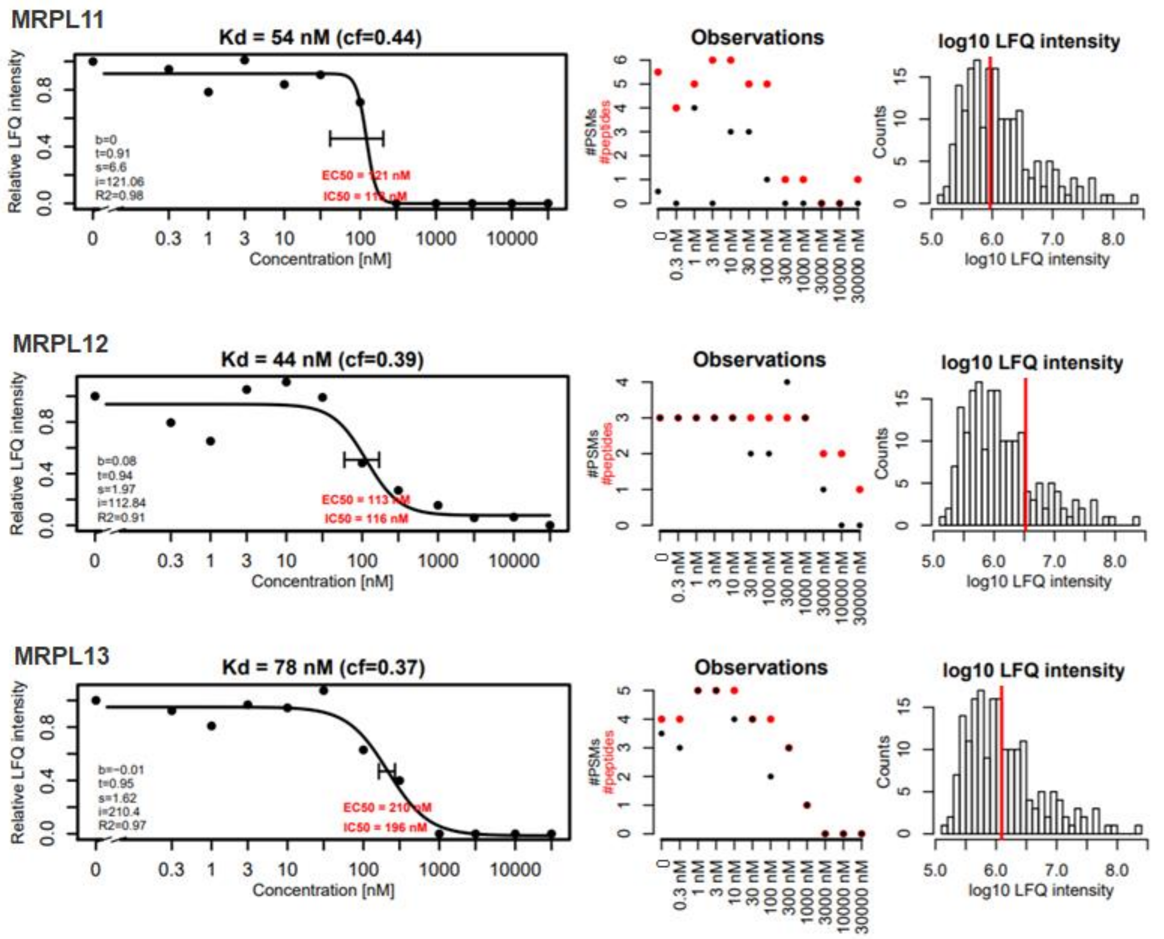


Figure 49. Concentration response curves for the polyadenylation signal motif CUGUCCU and three mitochondrial ribosomal proteins (MRPL11, MRPL12, MRPL13) (enriched by the probe CUGUCCU_1).

Chapter 5 Proteogenomic characterization of human tissues

5.1 Ultra-deep tonsil proteome

With the peptide fractionation by hSAX prior to online LC-MS/MS and match-between runs function in data processing, comprehensive proteomes of 29 tissues with a median overall protein sequence coverage at 42% were achieved. However, the median sequence coverage achieved for each tissue was limited and ranged from 14% to 25% (Figure 50). This is because tryptic digestion generates a large number of short peptides (~56%) that are not in the suitable length (7-35 aa) for mass spectrometry-based sequencing technology[31,32]. In order to improve the sequence coverage and protein discovery in a single tissue, an ultra-deep proteomic analysis of tonsil tissue was performed in this study by utilizing seven different proteases (trypsin, Lys-C, Lys-N, Glu-C, Arg-C, Asp-N, and chymotrypsin) and three peptide fragmentation techniques (HCD, CID and EThcD/ETD). Similar to the tryptic proteome, aliquots of 300 µg proteins extracted from the tonsil tissue by urea lysis were digested with each protease and separated into 36 fractions via hSAX chromatography. Peptides in each fraction were separated by a 2 h LC gradient and analyzed by MS/MS on a Q Exactive Plus or a Fusion Lumos mass spectrometer (CID and EThcD/ETD) to generate a total of 360 analyses. The peptides were identified at a 1% FDR by MaxQuant and increased from 104,538 in the tryptically-digested tonsil tissue alone (HCD) to 421,073 (Figure 51A, search against Ensembl database, without match-between-runs). Consequently, the median protein sequence coverage increased by up to 54% (Figure 51B). When comparing this data with the tryptic proteomes of the 29 tissues, the identification of peptides increased 1.5 times. In this ultra-deep tonsil proteome, the chymotrypsin data set comprised the largest number of exclusive peptide identifications (n=56,432), followed by Lys-N (n=56,068), Asp-N (n=38,767), Lys-C (n=37,179), Glu-C (n=33417) and Arg-C (19252). Notably, most peptides generated by Glu-C, Asp-N and Lys-N are exclusive peptides. Trypsin, however, still performed the best among these seven proteases. The trypsin-HCD data set contained the largest number of peptides (n=104,538), followed by chymotrypsin-HCD (n=88,687), and trypsin-EThcD (n=65,310); while Glu-C (n=33,659), Asp-C (n=36,347), Arg-N (n= 39,532) contained the lowest number of peptides. This result is not surprising as trypsin not only generates shorter peptides with a basic arginine or lysine residue at the C-terminus, but these peptides are then ideally amenable to chromatographic separation, HCD fragmentation and algorithm-based identification methods[32]. These data sets also clearly illustrated why HCD is typically chosen for peptide fragmentation. The small gain achieved with the EThcD/ETD method (based on charge decision) can potentially be explained by the fact that the majority of tryptic peptides are doubly-charged. Nevertheless, the additional proteases and dissociation methods contributed 75% of the identified peptides in this ultra-deep tonsil proteome, which can be used to improve protein isoform identification and single amino acid variants profiling at the protein level.

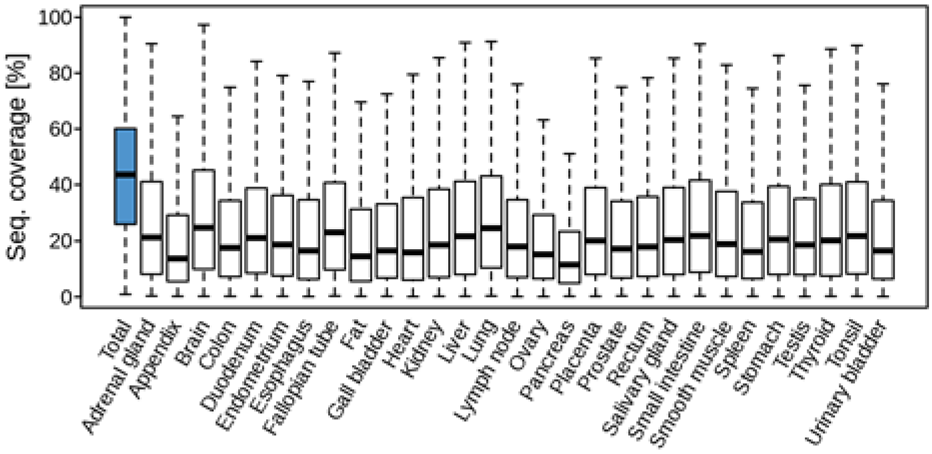


Figure 50. Distribution of peptide sequence coverage obtained for proteins by mass spectrometry in all tissues.

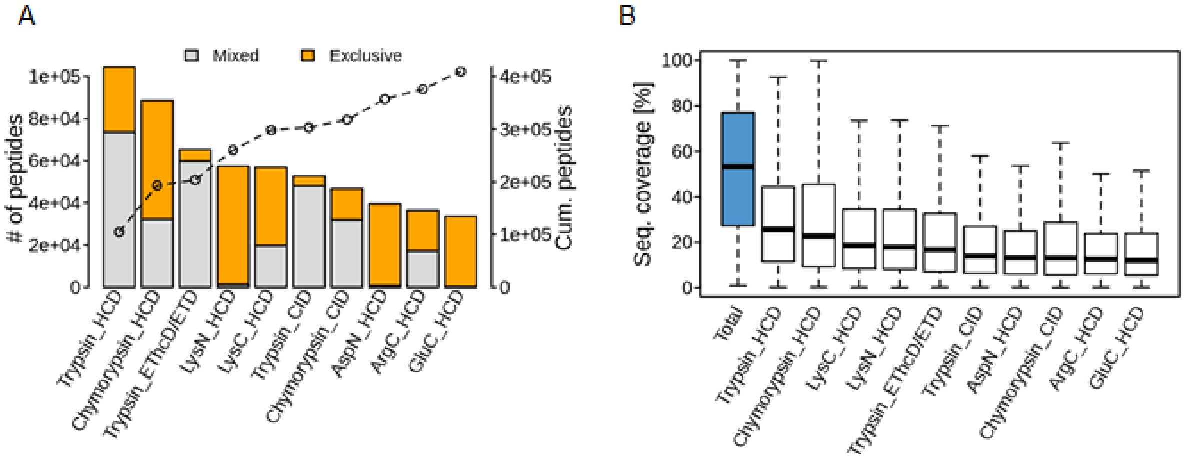


Figure 51. (A) Number of identified peptides segregated by protease and fragmentation method in the ultra-deep tonsil proteome (cut-off peptide intensity>0). Those covered by more than one workflow are marked in gray, and those exclusive to one workflow are marked in orange. The line indicates the cumulative number of peptides when adding data from the individual workflows. (B) Distribution of peptide sequence coverage obtained for proteins by mass spectrometry in tonsil tissue segregated by protease and fragmentation method.

At the protein level, the identification increased from 9,673 in the tryptically-digested tonsil proteome acquired with HCD fragmentation to 11,569, and very few of these proteins/genes were only identified by a single protease (Figure 52A). Subsequently, the number of genes increased from 8,568 to 10,288. The proteins identified by the additional proteases and fragmentation methods were primarily those with a low mRNA abundance in tonsil (Figure 52B); thus indicating an increased dynamic range in this deep proteome. 345 genes, which were only identified as transcripts but not as proteins in the 29 tissues, were identified as proteins in the ultra-deep tonsil proteome (Figure 53A). Another 81 genes were identified at the protein level but were not detected at the mRNA level and these proteins subsequently enriched for antigen processing and presentation (n=9; BH adjusted p-value = 7.5×10^{-7}). Interestingly, there were 895 proteins that were not detected as mRNA transcripts in the tonsil tissue; but were expressed in other tissues. 70 such proteins were enriched in the liver at the mRNA level (Figure 53B). In addition, these also enriched as blood microparticle (n=46; BH adjusted p-value = 2.8×10^{-27}) or for serine hydrolase activity (n=28, BH adjusted p-value = 1.5×10^{-4}). This not only indicated the functional coordination and cooperation between different tissues, but also the significance of investigating gene expression at both the mRNA and protein levels.

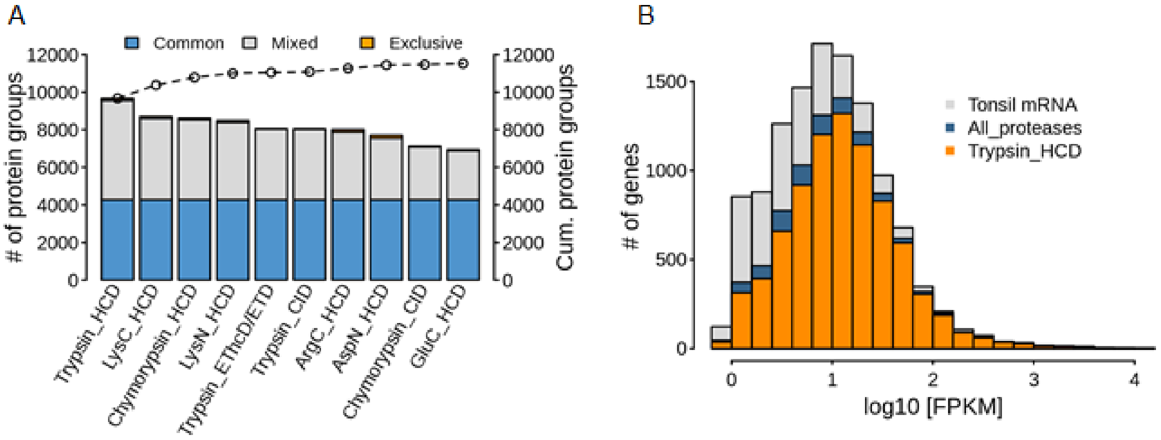


Figure 52. (A) Number of identified proteins segregated by protease and fragmentation method in the ultra-deep tonsil. Proteins observed by all workflows are marked in blue, these observed by more than one workflow are marked in grey, and those exclusive to a single workflow are marked in orange. The line indicates the cumulative number of proteins when the data is added from the individual workflows. (B) The coverage of genes identified in the tonsil proteome. Abundance distribution of all transcripts detected in tonsil (grey); the fraction of detected proteins in the ultra-deep tonsil proteome is shown in blue; and the fraction of proteins detected in tonsil with trypsin and HCD fragmentation is shown in orange.

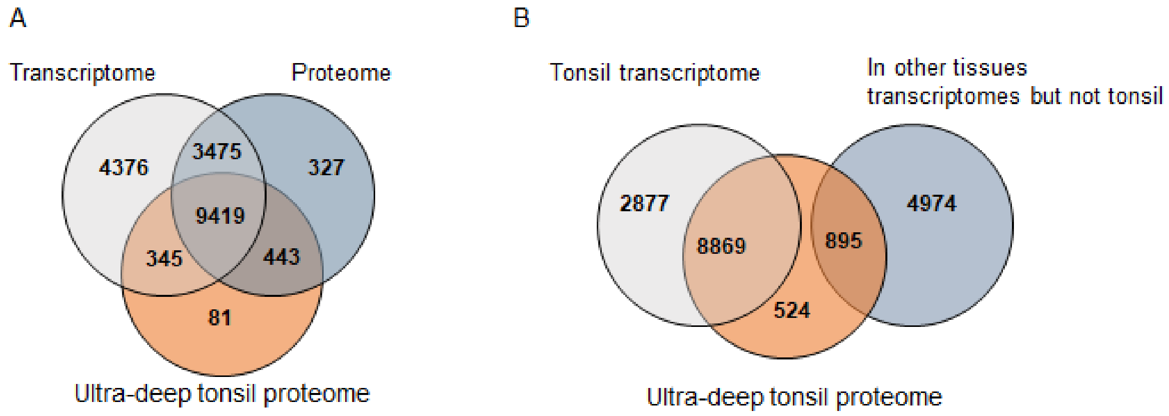

A
Transcriptome
Proteome
4376
3475
327
9419
345
443
81
Ultra-deep tonsil proteome
B
Tonsil transcriptome
In other tissues transcriptomes but not tonsil
2877
8869
895
4974
524
Ultra-deep tonsil proteome

Figure 53. (A) The overlap of genes identified by the transcriptomic data from the 29 tissues, proteomic data of the 29 tissues and the ultra-deep tonsil proteome. (B) The overlap of genes identified in the tonsil transcriptome, ultra-deep tonsil proteome and in other tissue transcriptomes.

Furthermore, protein quantitation between different proteases and fragmentation method were compared (Figure 54). Since the more peptides identified, the more accurate of quantification, trypsin-HCD was supposed to have the best protein quantification. Thus, similarity to the trypsin-HCD data set could indicate the quality of quantitation. Tryptic data sets generated by EThcD and CID showed the highest similarities with the trypsin-HCD data (Spearman correlation coefficients >0.9). This is most likely due to the fact that many of the identified peptides are shared between the data sets. The Lys-C data set also showed a high correlation to the trypsin-HCD data, followed by Lys-N and chymotrypsin-HCD. Although Lys-C and Lys-N digestion generates slightly different peptides, quantitation of the peptides still resulted in a similar level. Notably, the Asp-N digestion showed the least similarity to the trypsin-HCD data and the other data sets. This may be because overall, less peptides were identified, and these peptides were exclusive to the Asp-N data set. Considering that protein sequence coverage improved by digestion of the tonsil tissue with different proteases, it follows that combining all the peptides will also improve protein quantitation. Unfortunately, a computational algorithm to achieve such data integration has not yet been developed. This ultra-deep tonsil proteome data is nevertheless a valuable resource for such method development.

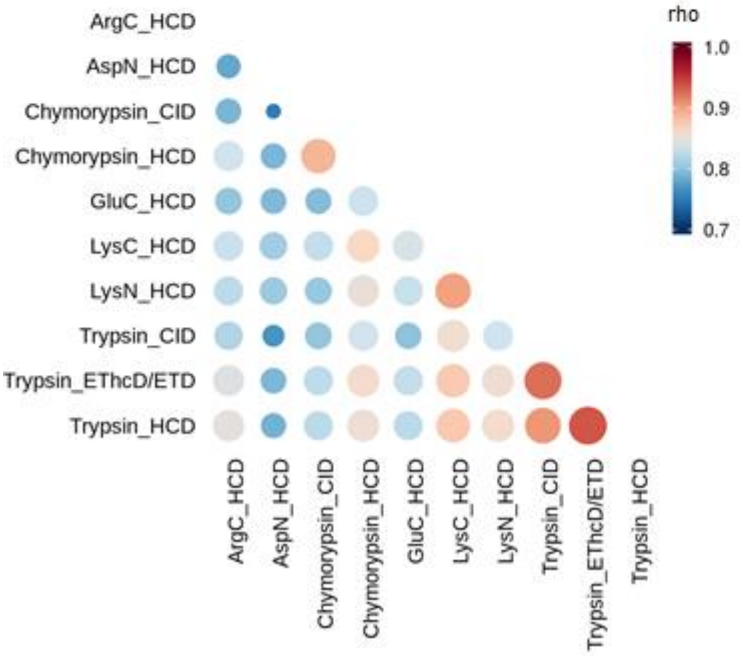


Figure 54. Comparison of protein quantitation using iBAQ values between the tonsil proteome generated by different proteases and fragmentation methods. The colors represent the Spearman correlation coefficient (rho).

5.2. Isoform identification

One aspect of the data covered in more detail in this study is the considerable interest in the community to use proteomic data to annotate genome, often referred to as proteogenomics. With sample matched RNA-Seq and proteomic data available, the merits of proteogenomics was assessed at several levels. Firstly, the identification of protein isoforms was investigated. 92% to 94% of human genes undergo alternative splicing and ~86% have a minor isoform frequency of >15%[231] that markedly increases the complexity of transcriptome. Previous studies illustrated that many alternatively-spliced protein isoforms show tissue-specific expression pattern[232,233] and vastly different interaction profiles[234]. Genome-wide investigations of protein isoform expression will significantly increase the knowledge on human biology. Yet, in proteomics, isoforms are much more difficult to unambiguously distinguish because the identification of proteins is inferred from the underlying peptide data. Given that the median sequence coverage achieved for each protein is limited (14 and 25%; Figure 50), many potential isoforms are not covered by unique peptides. This problem is particularly confounded for low-abundance proteins. In addition, some peptides match multiple entries in comprehensive sequence collections such as Ensembl (102,450 entries); and thus a so-called protein group is usually identified rather than one specific protein or isoform thereof. Illustrated by the proteomic data obtained from tryptically-digested tonsil (Figure 55A), only 14% of all protein groups (n=1,284) contained one single protein when searching the MS data against Ensembl. When the same data was searched against a protein sequence database constructed from the tissue specific RNA-Seq data (cut-off at FPKM>1, 22,899 entries), the proportion of single-entry protein groups increased to 53% (n=4,538, see Appendix Figure 3 for all tissues). Thus, the efficiency and accuracy of protein identification was largely improved by the RNA-Seq-generated database that only contained protein sequences with expression evidence[160]. It is important to mention that some protein sequences were excluded from the RNA-Seq based database due to such factors as low expression, transcripts lacking a poly-A tail, some long-lived proteins, or mature proteins transported from other tissues and this did result in a loss in protein identification (Figure 55B). As this work focused on the ability of proteomics to identify protein isoforms, potential missing sequences were not combined into the database. The proteins that were only identified via searching the RNA-Seq database were identified by variant peptides. By searching the MS data against the tissue-specific RNA-Seq-derived database, 53,219 non-redundant protein isoforms were identified for 17,673 genes and 15,257 confirmed by proteomics for 11,833 genes across the 29 tissues (Figure 56). Unlike the transcriptome, where about half of the genes have minor isoforms identified, very few genes have more than 2 isoforms identified by proteomic approach. Excluding the protein inference problem, it may also be that the most highly-expressed protein-coding genes only have a single dominant isoform and the alternative isoforms may be expressed infrequently or have very short half-lives[16,235–238].

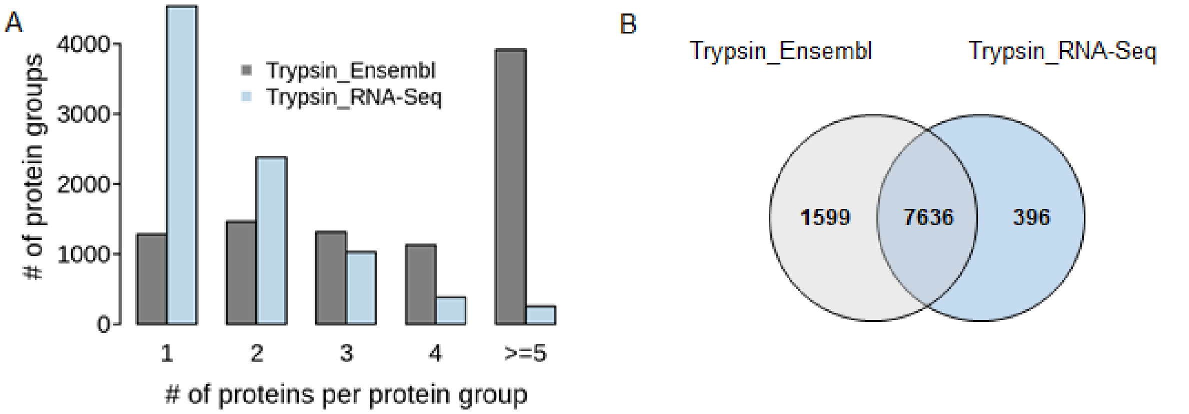


Figure 55. (A) Searching the tryptic tonsil proteomic data against a tissue-specific sequence database constructed from RNA-Seq data drastically reduced the number of individual protein sequences in the protein groups compared to searches against Ensembl, thus allowing a more efficient detection of protein isoforms. (B) The number of genes identified in the tonsil tissue by database search against Ensembl (data from 29 tissues combined search with match-between-runs) and RNA-Seq derived database.

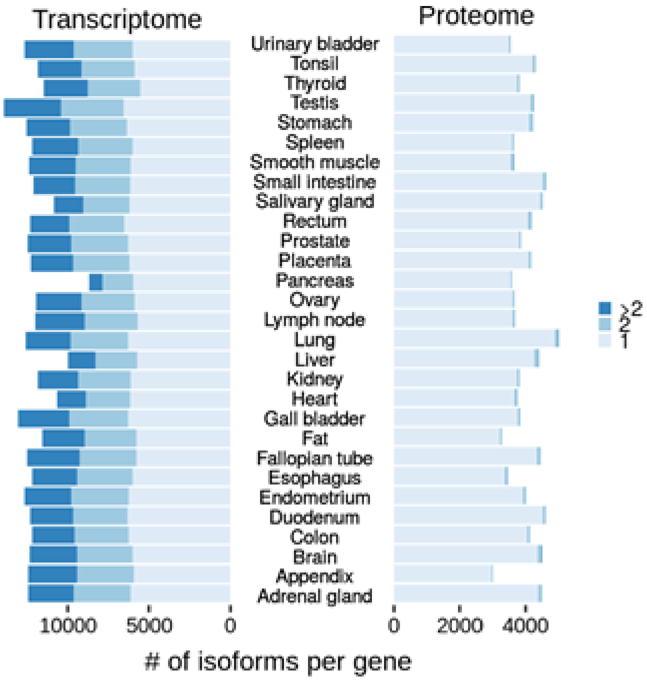


Figure 56. Analysis of the number of isoforms detected by transcriptomics or proteomics in all tissues. Transcript isoforms of a gene included all transcripts with FPKM>1 in a single tissue. Protein isoforms were determined based on protein groups with 1 protein.

One way to improve the detection of isoforms is to increase the sequence coverage. By searching the ultra-deep tonsil data set against the Ensembl database, 2,201 protein groups were unambiguously linked to a single isoform. This is 1.7 times more than the number of isoforms identified from the tryptically-digested tonsil tissue with HCD fragmentation, resulting from an increase in protein unique peptides from 21,726 to 86,853. When searched against protein sequences derived from the tonsil-specific RNA-Seq data, 10,592 protein groups were identified; of which 6,293 represented isoforms identified by unique peptides (Figure 57A). 1,755 more protein isoforms were identified than from the tryptically-digested tonsil data searched against the RNA-Seq-derived database. Subsequently, the number of isoforms increased from 4,304 to 5,551; although the number of genes with more than 2 identified isoforms was still small (n=339, Figure 57B). Thus, isoform identification at the protein level is feasible, but to systematically do so remains challenging. It is further noted that because most protein isoforms were detected by very few isoform-specific peptides, confident quantitation of the different protein isoforms from the same gene in the same tissue is currently very difficult and may require targeted MS assays rather than ‘shotgun’ approaches [239]. In addition, there is currently no tractable means of determining which allele of a gene gave rise to a detected protein or isoform thereof.

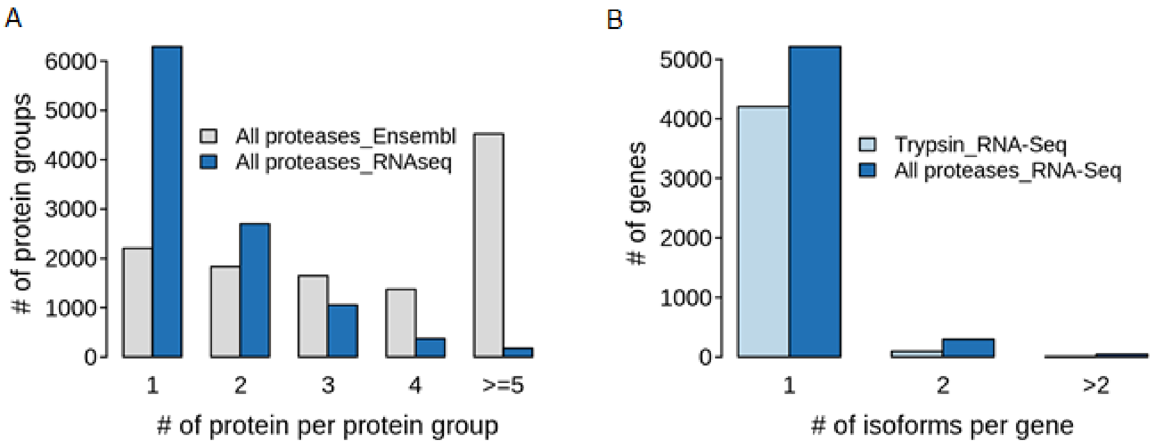


Figure 57. Integrated multi-protease approach and database search against tissue-specific sequences constructed from RNA-Seq data drastically increased the identification of protein isoforms. (A) The number of protein groups identified in the ultra-deep tonsil MS data by searching against Ensembl (light gray) and RNA-Seq based database (dark blue), respectively. (B) The number of protein isoforms identified from HCD analysis of the tryptically-digest tonsil tissue (light blue) and ultra-deep tonsil tissue (dark blue) MS data by searching against the RNA-Seq based database.

5.3 Variant identification in the ultra-deep tonsil proteome

With the development and widespread availability of high-throughput next generation sequencing technologies, genetic variation has been characterized by sequencing the whole genome and exome of thousands of humans. Information includes single nucleotide variants (SNVs), insertions/deletions (indels), and structure variants[9,240]. Between two genomes, SNVs occur on average once every 860 base pairs and are the most intensively-studied variation and many are associated with various diseases or traits[241]. The direct detection of single amino acid variants (SAAVs) derived from SNVs will aid our understanding of the functional significance of these variants. It has been noted, however, that the identification of SAAVs by proteomics is challenging and plagued by false positives in standard database search regimens because: (i) the tandem mass spectra used for database searching are often poor quality; (ii) these spectra often do not contain the complete amino acid sequence information of the underlying peptide; and (iii) the current FDR statistics for peptide/protein identification do not translate well to variant calling at the peptide level. As a result, random matches can, and will, frequently occur, raising substantial concern about a proportion of the currently-available proteogenomic literature[149].

To assess the ability of proteomics to detect genetic variants, whole exome sequencing (WES), RNA-Seq and ultra-deep proteomic data of the tonsil tissue were analyzed. In the WES data, the average exon coverage was 98× and 97% of the exons were covered >20×, thus providing a sound basis for the identification of single amino acid variants (SAAVs).

Variant calling and filtering of WES data resulted in 9,848 high-quality, nonsynonymous point mutations (*i.e.*, nonsense and missense variants excluding I>L and L>I variants that cannot be distinguished by mass spectrometry). These represent 5,527 human genes and include 6,112 heterozygous and 3,736 homozygous variants (Figure 58). In the RNA-Seq data, 3,524 of the 9,848 genomic variants (36%; 2,171 heterozygous and 1,353 homozygous cases; representing 2,428 genes) were sufficiently covered (≥10×) to assess genotype. The reason for the substantial loss in coverage in the RNA-Seq vs. exome data is because: (i) not all genes are expressed from both alleles in a given tissue; and (ii) even at a sequencing depth of 50 million reads, the dynamic range of mRNA abundance is too high to cover all transcripts and variants multiple times.

In order to detect the variant peptides, a variant database was constructed from the high-quality WES data and contained 16,635 protein sequences. When searching our proteomic data against the concatenated sequences from the variant database, Ensemble and UniProt and requiring an identification by both Mascot and Andromeda as well as a number of further criteria (for details see in Materials and Methods), 1,942 candidate peptides mapping to 724 of the 9,848 (non-canonical) exome variants (7.4%; 400 heterozygous and 324 homozygous cases) were identified. These peptide variants are all missense mutations). For 41% of the heterozygous cases (165 out of 400), peptide level evidences for the canonical and alternative variant were obtained, while for the remaining cases, only the alternative variant (235 out of 400) was identified.

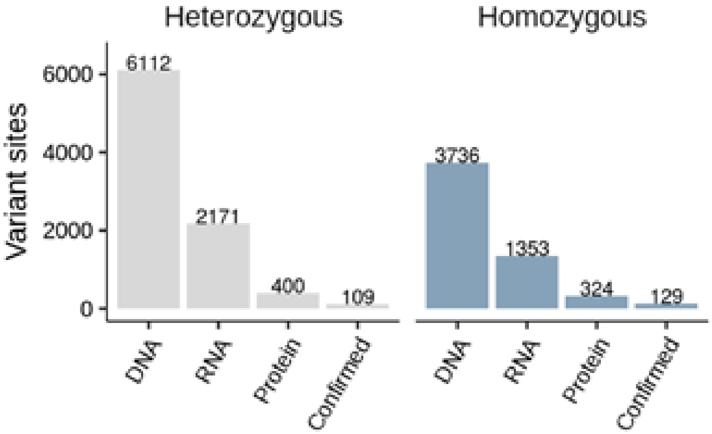


Figure 58. Number of single amino acid variants detected by whole exome sequencing, RNA-Seq, mass spectrometry and validation using synthetic peptide spectra comparisons. It is apparent that only a very small fraction of all variants detected at the DNA or RNA level can be detected at the proteome level.

For validation, candidate peptide spectra were compared to those of synthetic reference standards[189]. To this end, the synthesis of reference peptides for all 724 alternative variants was attempted and such peptides for 574 cases were obtained. Automated spectral angle analysis[190] provided evidence for 238 variants (SA≥0.7, Mascot ion score ≥50) including 109 heterozygous and 129 homozygous cases (variant peptides are provided in the Appendix Table 3, while all mirror plots are available in ProteomeXchange). Manual inspection of the above 724 candidate peptides identified 204 unique alternative variant sites of which 158 were also found in the SA analysis. The variants that passed the (conservative) filtering criteria merely represent 2.4% of all variants detected at the exome level, 6.7% of the variants detected at the mRNA level and 32% of the candidates suggested by database searching. When tracing the confidently identified peptide variants back to the proteomic workflow, it became clear that the vast majority of all variants are represented by peptides generated by trypsin, Lys-C and Arg-C cleavage and using the standard HCD fragmentation technique. In addition, the confirmation rate (using synthetic peptide reference standards) for tryptic peptides was also much higher than that of the other enzymes (Figure 59, synthetic peptides were not measured by EThcD/ETD). This can be attributed to the fact that trypsin-like peptides generally show easily-predictable fragmentation behavior and that most bioinformatic tools are optimized for use with data generated from tryptic digestion of proteins. In addition, the number of final confirmed variant peptides will increase if the candidates identified by Mascot are taken into consideration (Figure 60). There are discrepancies between the Mascot and Andromeda searches caused by the inaccurate scoring for variant peptides by both search engines. To be conservative, only the overlapping identifications from both search engines were taken into consideration.

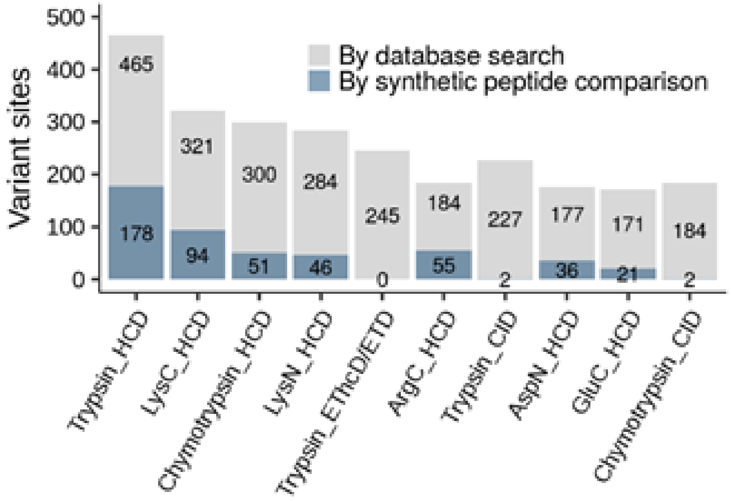


Figure 59. Analysis of the proteomic workflow that contributed to the confident detection of single amino acid variants.

While the above shows that some of the variants detected on the nucleotide level could be confirmed at the protein level, the overall success rate was low. It is notable here that this was not due to lack of expression of the underlying gene because the proteomic data covered 76% of all expressed tonsil genes (8,869 of 11,746 mRNA-Seq genes), 47% (2,615 of 5,527) of all the genes for which variants were detected by exome sequencing and 75% (1,822 of 2,428) by RNA-Seq, respectively. Instead, the main reasons for poor coverage of variants at the proteome level are still the limited sensitivity and dynamic range of detection, limited peptide coverage of a protein and insufficient coverage of amino acids in peptide mass spectra along with shortcomings in peptide identification algorithms.

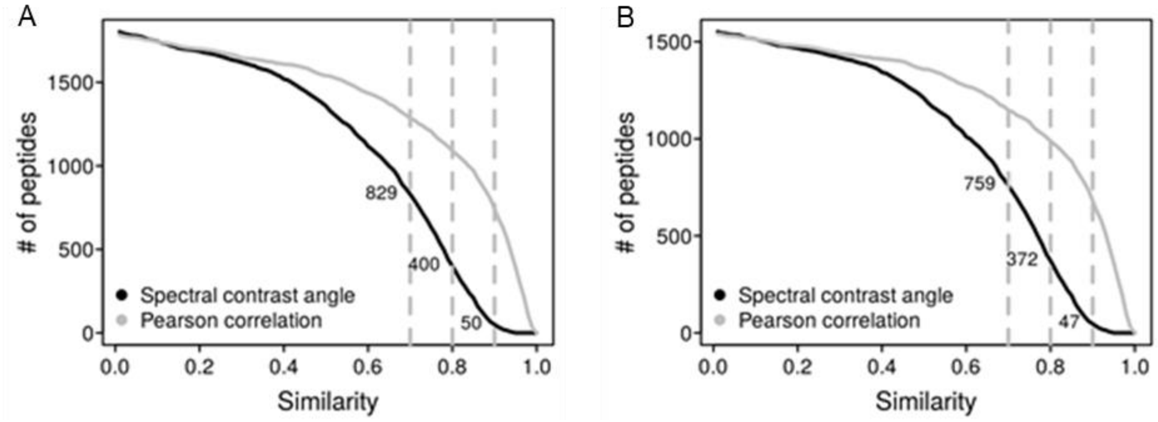


Figure 60. (A) Number of experimental vs. synthetic peptide reference spectra comparisons for candidate variant peptides (only the spectra with highest spectral angle for each peptide was plotted) after database searching using Mascot as a function of the spectral angle and Pearson correlation coefficient. Dotted lines mark spectral angles of 0.7, 0.8 and 0.9. (B) Same as panel A but showing only candidate peptides that were identified by both Mascot and Andromeda.

5.4 Identification of peptides translated from non-coding regions

Recent research showed that there is more heterogeneity in gene models than previously anticipated, as a result of e. g. alternative translation initiation sites (aTIS)[144] and there is an ongoing debate in the community whether or not long non-coding RNAs (lncRNAs) can be translated into proteins[242]. Ribosomal profiling has shown that thousands of potential aTIS may exist and that ~40% of all lncRNAs can at least engage the ribosome[243]. It will be interesting to see whether lncRNAs and aTIS contribute *bona fide* polypeptides in the human proteome. In order to explore if our resource can provide protein evidence for such cases, a database search strategy was used[244] whereby all LC-MS/MS files were searched against combined sequences from: (i) a curated lncRNA database (GENCODE v.25); (ii) a database containing protein sequences derived from alternative translation initiation sites (see methods); (iii) GENCODE; (iv) UniProt; and (v) the tissue-specific RNA-Seq data. Any potential lncRNA or aTIS peptide must: (i) originate from just one single sequence collection (*i.e.*, lncRNA or aTIS and no other database); (ii) be identified by both Mascot and

Andromeda; (iii) fulfil stringent score cut-offs (see methods); (iv) be confirmed by expression of the underlying transcript in at least one of the tissues (FPKM >1); and (v) fail a BLAST search against UniProt to exclude obvious alternative explanations. This approach yielded 5 lncRNA and 344 aTIS peptides, respectively. Implementation of Mascot filtering criteria alone returned 125 lncRNA and 1,307 aTIS peptides, thus indicating that neither Andromeda nor Percolator applies a proper FDR estimation. This could be explained by the imbalance between the small amount of expected PSMs and the large database (aTIS: 474,991 entries; lncRNA: 29,524 entries).

To confirm the presence of as many of these peptides as possible, all peptides identified by Mascot that passed the filter criteria were synthesized, and achieved 70 lncRNA and 800 aTIS synthetic peptides, respectively. Interestingly, not a single lncRNA peptide could be substantiated by synthetic peptides indicating that lncRNA are rarely if at all translated[245]. To validate the candidate aTIS peptides, spectra of endogenous and synthetic peptide reference standards were compared as described in the variant peptide detection section. Only 66 aTIS peptides (including 8 N-terminally acetylated peptides) covering 53 genes and 57 alternative translation start sites could be confirmed. Manual spectrum interpretation yielded 96 aTIS peptides (overlap of 45 to the SA analysis) mapping to 76 genes and 81 alternative translation start sites. In total, 117 aTIS peptides mapping to 89 genes and 99 alternative translation start sites were confirmed (Appendix Table 4). This included 14 peptides from 12 genes reported in previous studies, for example FXR2, RPA1 and CDV3[15,246]. Fifty-five of the above aTIS peptides represent 5' N-terminal extensions of the original gene, 32 peptides represent novel (acetylated) N-termini downstream of the canonical start site, 17 represent frame-shifts potentially leading to an entirely new sequence, 5 peptides were likely to represent upstream ORFs (uORF) with a stop codon before the canonical start site and 8 peptides with mixed annotation (Figure 61, left panel). The mirror mass spectra in Figure 62 for the endogenous (top) and synthetic (bottom) peptide (ac)ATTQISKDELDELKEAFAK from the actin-binding protein plastin-3 (PLS3) provides an example for the detection of a novel N-terminal sequence. Considering the distance between the novel and reference terminus of PLS3, the translation start site identified in this study is more likely to be the actual start site during translation rather than a novel start site. All the mirror plots for aTIS peptides are available in the Appendix S1. For 36 of the peptides representing aTIS, the exact start site was identified as the peptide was N-terminally acetylated (Figure 61, right panel). Among these, 18 contained an AUG start codon (Met), 8 contained a GUG start codon (Val), 5 a UUG and 4 a CUG start site (both Leu), and one GCG start site (Ala). This confirms the emerging notion that non-AUG translation initiation events are not as infrequent as previously believed and may represent a mechanism to regulate protein expression[144,243].

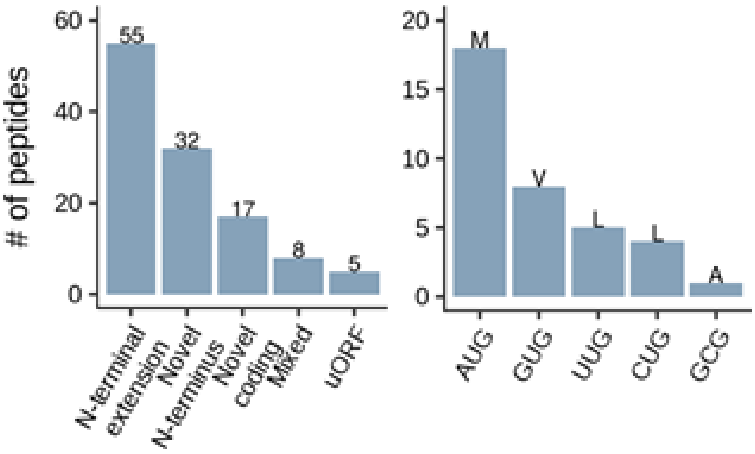


Figure 61. Results of the detection of non-canonical coding regions using proteomic data (left panel) and the alternative start codon identified by acetylated N-terminal peptides (right panel). The majority are N-terminal extensions of annotated genes. All but one of the detected alternative translation start sites correspond to point mutations of the first base of the classical AUG codon.

This study only identified a relatively small number of aTIS events compared to others [144] implying that enrichment of N-terminal peptides[247,248] is a more efficient means to systematically detect such events. Furthermore, the lncRNA and aTIS analysis showed the high FDR that is estimated with commonly-used search engines. Thus emphasizing that existing literature may not be completely free of a substantial number of errors. As shown Figure 63, 124 of the 800 aTIS peptides returned by the Mascot search had >70% similarity to the synthetic reference spectra, while about half of peptides showed a similarity of <40% and were not identified with MaxQuant. If only the identifications common to both search engines were taken into consideration, 32 of the 124 peptides that have high similarity to the synthetic peptides would have been excluded. These data suggest that the implementation of different search engine algorithms could indeed assist in removing false positive identifications, but the true identifications could also be removed. Sophisticated search engines that enable accurate estimation of FDR are thus urgently required. Comparing the small discrepancy of variant peptide identification by Mascot and Andromeda (Figure 60), it was suggested that decreasing the search space may improve the sensitivity of identification. To minimize the available search space for aTIS peptide detection, a sample-specific database can be derived from ribosome profiling data[165].

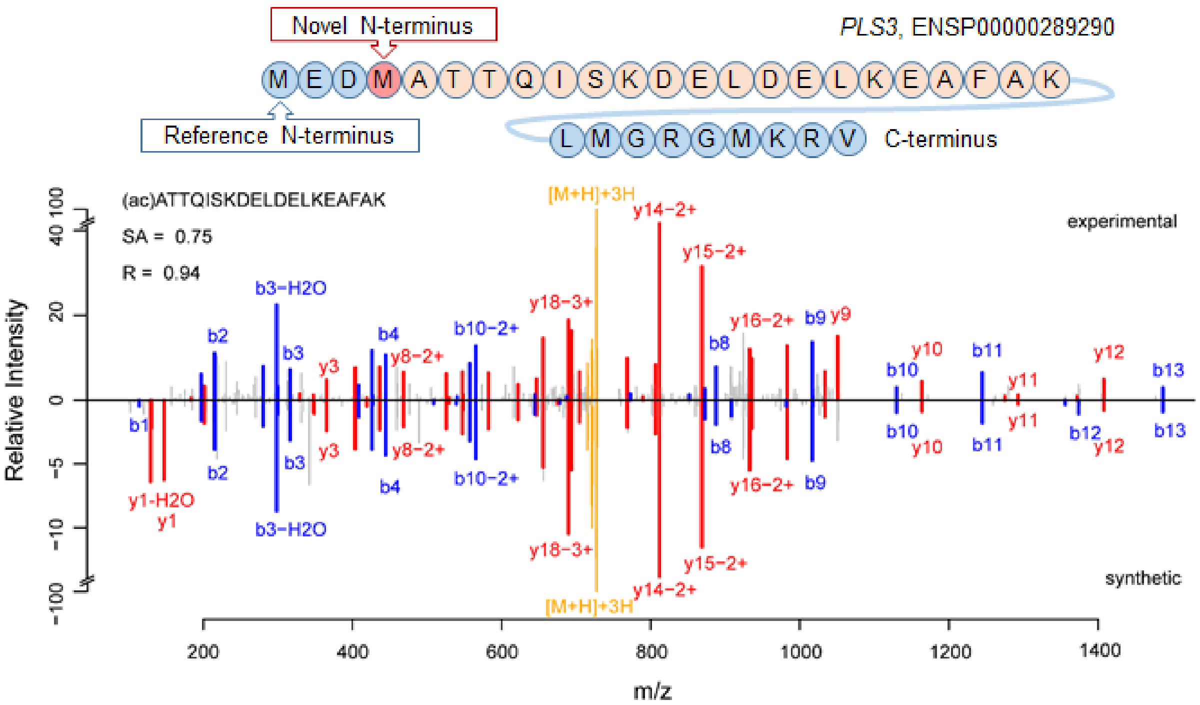


Figure 62. Validation of a novel translation start site for the protein PLS3. The upper panel shows the novel translation site position within the amino acid sequence and the lower panel shows the tandem mass spectra of the endogenous N-terminally acetylated peptide and the corresponding synthetic peptide spectrum.

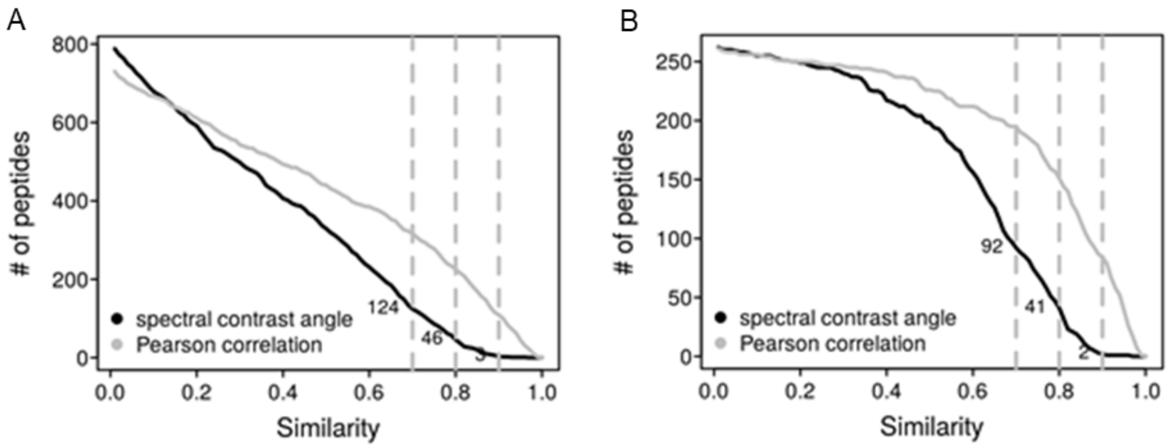


Figure 63. (A) Number of experimental vs. synthetic peptide reference spectra comparisons for candidate aTIS peptides (only the spectra with highest spectral angle of each peptide was plotted) after database searching using Mascot as a function of the spectral angle and Pearson correlation coefficient. Dotted lines mark spectral angles of 0.7, 0.8 and 0.9. (B) Same as panel A, but showing only candidate peptides that were identified by both Mascot and Andromeda.

Chapter 6 General discussion

In the last years, several landmark studies have established valuable resources for the human proteome across the majority of the major tissues and organs in the human body. These resources have been generated by mass spectrometry-based proteomics[14,15], by integrating RNA-Seq and antibody-based profiling [16], or by integrating all three techniques[17]. Meanwhile, many efforts have focused on a single (diseased) tissue or cell-type resolved analysis of protein expression[18,138,139,150]. Despite these advances, however, the number of protein-coding genes that are expressed as proteins and expression levels across different tissues, especially in healthy tissues, are still not clear. Therefore, one of the main aims of this thesis was to generate a quantitative proteome map based on different healthy tissues to provide a comprehensive baseline of protein expression across the human body.

In Chapter 3, 29 histologically-healthy human tissues representing most of the major tissues and organs of the human body were studied by 'bottom-up' proteomics using intensity-based absolute quantitation (iBAQ)[93]. These were profiled to a depth of 10,000 to 12,400 proteins at 1% protein FDR per tissue. A total of 15,210 proteins representing 13,664 protein-coding genes were quantitated in 29 tissues, including 69 neXtProt-annotated 'missing' proteins. Although the absolute coverage of protein-coding genes was much lower than the first two human proteome drafts (13,664 vs.18,907 and 17,294 in Wilhelm et al. and Kim et al., respectively), thanks to advances in mass spectrometry, the number of proteins and genes identified in each tissue was higher (3,000 to 4,000 proteins per tissue in the human proteome draft studies)[14,15]. These identifications per sample are similar to those obtained in single tissue analyses with modern mass spectrometry, *e.g.*, the region and cell-type resolved proteome of heart[18] and the brain proteome of Alzheimer's and Parkinson's disease[249]. This work represents by far the deepest quantitated proteome for most tissues and the most comprehensive human proteome map of multiple healthy tissues.

Integration of the proteomic data with sample-matched RNA-Seq data provided by the Human Protein Atlas project enabled the generation of a systematic, quantitative and integrative deep transcriptomic and proteomic map of the 29 healthy tissues. The data consisted of 17,615 human genes represented by transcripts and 13,664 genes represented proteins; and 12,894 protein-coding genes were identified at both the transcript and protein level. Intriguingly, most of the highly-expressed transcripts without protein evidence were testis specific. This observation may be explained by the transient expression of such proteins for a tissue-specialized function. Note too, that testis and sperm cells are a potential resource for characterizing 'missing' proteins because of specific and local functions such a spermatogenesis and germ cell-specific expression[250,251]. To explore which proteins, and how many, have a tissue-specific expression profile, all detected genes were classified into 5 categories. According to Uhlen et al., these categories are based on expression level and tissue distribution and include 'tissue enriched', 'group enriched', 'tissue enhanced', 'expressed in all' and 'mixed' expression[16]. Consistent with independent data sets, most of the genes expressed in all the studied tissues and only a very small proportion of genes were exclusively-expressed in one or few tissues[252]. Moreover, various sub-proteomes that correspond to functional groups of genes were presented. For example, GPCRs were much more tissue-restricted, suggesting that such proteins may be better as drug targets as these are not ubiquitously expressed. As a baseline map of protein expression across the human body, the transcriptomic and proteomic data generated here may have general value in drug discovery as, *e.g.*, the expression of a protein target of interest can be rapidly evaluated. Thus, based on the tissue expression profile of a protein, this map can aid in understanding adverse clinical effects and off-target mechanisms-of-action of drugs. For instance, a recent study revealed phenylalanine hydroxylase (PAH) as an off-target of the pan-HDAC inhibitor panobinostat. This protein expression map showed that PAH is abundantly-expressed in liver (and kidney). The liver is the major site of hydroxylation in the human body[253], thus indicating that panobinostat exerts detrimental effects in this organ, *i.e.*, leading to decreased tyrosine levels and eventual hypothyroidism in affected patients.

Although mRNA levels are a major determinant of protein abundance at steady state, it has been debated for years whether mRNA levels of a given gene can be used as proxies for the corresponding protein levels[14,93,104,133,215,254,255]. With the high-quality quantitative transcriptomes and proteomes from this study, the relationship between mRNA and protein expression was systematically explored in four aspects of this thesis (Chapter 4). Firstly, the expression discrepancy between protein and mRNA level was compared by ranking the order of relative intensities of transcripts and proteins. It is noteworthy that the overlap of genes encoding the top 100 most abundant transcripts and proteins is rather small (~20%). Comparison of the gene classification for the top 100 transcripts with the top 100 proteins showed that high-abundance proteins tend to be encoded by housekeeping genes[195]. Although blood 'contaminants' such as hemoglobin subunits were highly-abundant in the proteomes, these discrepancies may be primarily a result of post-transcriptional regulation. Next, the correlation between the transcriptome and proteome was explored. The results supported the frequent reports that only ~40% to 60% of the variation observed at the protein level can be explained by mRNA abundance[102,135,195]. These relatively low correlations also illustrated the expression discrepancy between protein and mRNA levels. More interestingly, was that a nearly quadratic relationship exists between mRNA and protein levels in tissues. Thus revealing a much larger protein copy number per mRNA for high-abundance mRNAs than for low-abundance mRNAs. This finding is consistent with observations made in yeast[135,256,257], mouse[93], and human[14]. The implication is that genes encoding highly-abundant proteins not only express mRNAs at high levels but also encode high translation and protein stability-favored sequence elements[133]. Furthermore, the correlation of mRNA and protein concentrations originating from the same genes across 29 tissues was examined. The absence of a significant correlation or low correlations between protein and mRNA (the median of spearman correlation is 0.35 in this thesis) may be due to the fact that similar proteins and/or abundances thereof are comparable across tissues. No conclusion, however, can be made with respect to the influence of variations at the mRNA level variation on protein abundance. Some good examples are members of protein complexes such as the ribosome and proteasome. In previous studies, these have also shown little protein-mRNA correlation[138,139,150]. In addition, the correlation between protein and mRNA is dampened by protein buffering. Here, changes at the mRNA level are counterbalanced to prevent any overall change at the protein level[104,258]. Such results imply that protein-mRNA correlation across tissues may not be the most ideal indicator to study how the variation of mRNA level results in corresponding protein abundance changes. Last but not least, the protein to mRNA ratio (PTR) was also investigated. The PTRs are rather constant across tissues, although significant variation between different genes have shown a dynamic range of about 5-fold[14,100,215]. This observation thus enables the use of PTR as a measure for studying sequence elements that function in post-transcriptional regulation. Whether PTR can be directly used to predict protein concentration from mRNA levels in steady-state tissue/cells still requires more accurate data and deeper studies[14,213,215,255].

All the observations on the protein and mRNA relationship analysis suggested that multiple processes beyond mRNA abundance contribute to determine the final protein concentration [102,104,135,258,259]. To explore the factors that govern protein expression and quantitate the contributions thereof, the matched proteome and transcriptome expression levels for 11,575 genes across the 29 human tissues were exploited to predict protein-to-mRNA ratios (PTR ratios) from sequence based on multivariate regression analysis (cooperative work). The model-based analysis quantitated the contribution within, and across, tissues of 16 known sequence features representing 172 post-transcriptional regulatory elements, and a further 15 novel motifs that were identified as predictive of PTR ratios. Consistent with Vogel et al., the coding sequence is the major contributor, followed by 3'UTR and then 5'UTR[133]. Novel candidate regulatory elements in 5'UTR and 3'UTR, with estimated effects in the range of well-known canonical motifs, were identified in this analysis. Due to the smaller size of the data sets used (n=476), these were not covered by Vogel et al. Moreover, this model confirmed the hypothesis that high protein expression levels are obtained by the joint effect of high mRNA levels and genetically-encoded elements favoring the synthesis and stability of proteins.

To functionally understanding the RNA motifs found in the sequence determinants analysis, a mRNA pulldown with a competitive binding assay was developed during this thesis to identify motif-specific binding proteins. The basic idea underlying the method is to combine the interaction strength between mRNA motifs and proteins (K_d) with the knowledge in databases or literature to explain the effect of the motifs. The method enabled the simultaneous discovery of proteins interacting with specific mRNA motifs and the measurement of the interaction strength of all interactors. The method was implemented on 96-well filter plates and combined with LC-MS/MS label-free quantitation, thus enabling the parallel investigation of several mRNA motifs. With this approach, 253 proteins interacting with the polyadenylation signal AAUAAA, and the AU-rich element UAUUUAU and a 5' UTR novel motif CUGUCCU were identified. Filtering on the K_d of the protein-RNA interaction, 50 AAUAAA specific binding proteins, 38 UAUUUAU specific binding proteins, and 30 CUGUCCU binding proteins were identified. The results suggested that the CUGUCCU motif may improve translation initiation by enhancing the interaction between the 5' UTR with mitochondrial ribosomal proteins and the small subunit of the cytoplasmic ribosome. Furthermore, this was the first report of the immobilization of RNA sequences via the 5' amino modifier on NHS-activated sepharose beads. In comparison with biotinylated RNA pulldowns[260], this method was more suitable for short RNA sequences, *i.e.*, for sequences that are too short to clone, such as miRNAs. An added advantage is that the immobilized beads can be stored for months, thereby increasing the time flexibility of the workflow. Although electrophoretic mobility shift assay (EMSA) can be used to measure the interaction between proteins and mRNA, it cannot be performed in a high-throughput manner[261]. Like all other *in vitro* detection for protein-mRNA interactions, EMSA is not conducted under physiological conditions[262]. To shorten our entire procedure, the reaction time to couple the RNA probes to the beads still requires a degree of optimization. Nevertheless, a simple and alternative high-throughput method to study RNA-binding proteins with a specific sequence was described. The approach can be readily adapted and also applied to DNA-binding proteins.

During the last decade, proteogenomics has emerged as an active field that combines mass spectrometry-based proteomics with genomic and transcriptomic data to accurately annotate and reciprocally refine genomic and proteomic models. In chapter 5, the identification of protein isoforms, variant peptides and novel peptides from lncRNA and alternative translation initiation was investigated. As the protein sequence coverage is quite low in typical bottom-up proteomics, to facilitate proteogenomic analysis, an ultra-deep tonsil proteome was generated by multi-protease digestion and different peptide fragmentation techniques. In the tonsil proteome, more than 400,000 unique peptides were identified resulting in a median sequence coverage of 54%. This tonsil data represents the deepest tissue proteome to date. By searching the MS data against a protein sequence database generated from sample matched RNA-Seq data, 15,257 protein isoforms derived from 11,833 genes were identified in the tryptic MS data of 29 tissues, which only consisted 28% of the identified transcript isoforms. Unlike the transcriptome, a second protein isoform was only detected for very few genes, even in the ultra-deep tonsil data set. It was further noted that very few isoform-specific peptides were identified, suggesting protein isoforms are still quite difficult to detect due to the low sequence coverage in proteomic analyses. Importantly, RNA-Seq data can aid in improving the identification of protein isoforms[160], however, information is lost for proteins transported from other tissues. From the ultra-deep tonsil proteome, around 900 proteins were not detected as transcripts in the tonsil tissue but were detected in other tissues. To detect variant and novel peptides, stringent workflows were applied. These included the construction of variant and novel protein databases, database searching, strict filtering criteria of variant peptides, and validation by synthetic peptides and manual inspection. In total, 238 from 9,848 variant sites on the exome were detected at the protein level in the ultra-deep tonsil data. Additionally, 117 aTIS peptides were identified in all MS data generated in this study, but not a single lncRNA peptides was observed. Thus implying that identifying protein variants or novel coding sequences by proteomics is possible but remains very challenging. The low number of variant and novel peptides identified by synthetic peptides and the number of peptide candidates from database search, confirmed the increasing concern on the reliability of previously-reported novel proteins in proteogenomic studies. It is problematic to accurately estimate the FDR for variant and novel peptides in proteogenomics, which is further exacerbated by the imbalance in probability of correct PSMs in databases with different search spaces[263,264]. Meanwhile, there were large discrepancies between the results of the two database search engines used (Mascot and Andromeda) implying that the underlying scoring schemes are not yet optimized to match variants and novel-coding regions (Figure 60 and Figure 63). Synthetic peptide reference spectra appear to be mandatory for validation and manual spectral comparison still must play a central role. Neither approach has been systematically followed in the literature to date. Obviously, both are not without error but are clearly more powerful than purely relying on statistical criteria with largely arbitrary cut-off values[14,149,172,265]. It appears that even with the latest proteomic technology, proteogenomics currently offers relatively small gains compared to the very massive efforts in data generation, analysis and validation and that large improvements will be required to substantially change this situation in the future. It is possible in this study that the filtering criteria were perhaps too stringent and additional variants may be present in the. Since no convincing FDR estimation has yet been published for spectral angle analysis (let alone for manual data analysis), it is better to remain more conservative. Nevertheless, the resource provided by this work should be of considerable value for scientists that wish to develop more sophisticated approaches for proteogenomics. Undoubtedly there is considerable future potential in the use of synthetic peptide references in conjunction with spectral angle analysis. This is key for the many chimeric spectra present in classical data-dependent proteomic data sets but also more so for the rising use of data-independent data acquisition regimes.

Outlook

The comprehensive transcriptomic and proteomic map that was generated in this thesis from 29 healthy tissues is an invaluable resource for the scientific community. The data will be useful for many studies in basic and clinical research and will have an impact on the pharmaceutical industry. Many uses of this resource can be envisaged, ranging from the study of gene/protein expression regulation, evaluation of protein biomarker specificity, reconstruction of the genome-scale metabolic models (GEMs) for the analysis of metabolic processes across tissues[266], mining PTM distribution in human tissues (particularly for PTMs lacking methodology to robustly enrich, detect and localize the modification[265]), investigating the proteome difference between human and model animals for human diseases (*e.g.*, mouse, rat, and monkey), and to further proteogenomic analyses (*e.g.*, sORF) to name a few. It is somewhat difficult to obtain healthy material, thus some tissues and organs (*e.g.*, skin, eye, and different regions of the brain) and fetal tissues were not covered in this study. In addition, the number of samples limited any gender and/or age comparisons. Finally, these tissue proteomes reflect the sum of proteins expressed in all the cell types in the tissue, not necessarily of the individual cell types. With future advances in single-cell transcriptomics[267] and proteomics[268] and spatial proteomics[17,269], it will become necessary to generate a cell-type resolved human proteome map with both quantitative and spatial information covering all the organs. Such a map will reveal cell-type specific cellular composition, cell-specific functions and the cross talk between cells. Furthermore, as alternative splicing, genomic variation, and PTMs significantly enhances the complexity and diversity of the human proteome, the analysis of human proteoforms will become increasingly important in order to gain deeper insights into protein function and fundamental mechanisms of the human body[4,270].

Abbreviations

ac Acetylation

ACN Acetonitrile

AcOH Acetic acid

AGC Automatic gain control

CAA Chloroacetamide

cDNA Complementary DNA

CID Collision-induced dissociation

CP Buffer compound pulldown buffer

CPTAC Clinical Proteomics Tumor Analysis Consortium

CV Coefficient of variation

DDA Data-dependent acquisition

DMEM Dulbecco's Modified Eagle Medium

DMSO Dimethyl sulfoxide

DNA Deoxyribonucleic acid

DTT Dithiothreitol

EC_{50} Half maximal effective concentration

EDTA Ethylenediaminetetraacetic acid

ESI Electrospray ionization

ETD Electron-transfer dissociation

EThcD electron-transfer/higher-energy collision dissociation

EtOH Ethanol

FA Formic acid

FBS Fetal bovine serum

FDR False discovery rate

FFPE Formalin-fixed, paraffin-embedded

FTMS Fourier transformation mass spectrometry

HCD higher-energy collision-induced dissociation

HPLC High-performance liquid chromatography

HPA Human protein atlas

HPP The Human Proteome project

hSAX Hydrophilic strong anion exchange

IAA Iodoacetamide

iBAQ Intensity-based absolute quantitation

IC_{50} Inhibitory dose resulting in a half-maximal inhibition

IHC Immunohistochemistry

ITMS Ion trap mass spectrometry

LC Liquid chromatography

LC-MS/MS Liquid chromatography-tandem mass spectrometry

LFQ Label-free quantitation

lm Linear model

lncRNA long non-coding RNA

LTQ Linear trap quadrupole

m/z Mass-to-charge ratio

MALDI Matrix-assisted laser desorption/ionization

MeOH Methanol

mRNA Messenger ribonucleic acid

RNA-Seq RNA Sequencing

MS Mass spectrometry or mass spectrometer

MS/MS Tandem mass spectrometry

MS^1 spectrum Precursor mass spectrum

MS^2 spectrum Fragment mass spectrum

NCBI National Center for Biotechnology Information

OLS Ordinary least squares regression

PBS Phosphate-buffered saline

PD Pulldown

PD/PD Pulldown of pulldown

ppm Parts per million

PSM Peptide-spectrum-match

RMA Reduced major axis regression

RP Reversed-phase

RPKM Reads per kilobase of exon model per million mapped reads

RT Retention time

SA Spectral angle

SAAV single amino acids variants

SDS Sodium dodecyl sulfate

SDS-PAGE Sodium dodecyl sulfate polyacrylamide gel electrophoresis

sORF Small open-reading frame

StageTips stop-and-go-extraction tips

TCEP tris(2-carboxyethyl)phosphine

TFA Trifluoroacetic acid

TIS Translation initiation site

Tris Tris(hydroxymethyl) aminomethane

UniProtKB UniProt knowledgebase

uORF Upstream open-reading frame

WES Whole-exome sequencing

References

1. Venter, J. C. *et al.* The sequence of the human genome. *Science* 291, 1304–1351 (2001).
2. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761 (2018).
3. Smith, L. M., Kelleher, N. L. & Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* 10, 186–187 (2013).
4. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* 14, 206–214 (2018).
5. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* 21, 255–261 (2003).
6. Siu, G. *et al.* The human T cell antigen receptor is encoded by variable, diversity, and joining gene segments that rearrange to generate a complete V gene. *Cell* 37, 393–401 (1984).
7. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U. S. A.* 106, 20216–20221 (2009).
8. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311 (2001).
9. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285 (2016).
10. Ponomarenko, E. A. *et al.* The Size of the Human Proteome: The Width and Depth. *Int. J. Anal. Chem.* 2016, 7436849 (2016).
11. Bennike, T., Birkelund, S., Stensballe, A. & Andersen, V. Biomarkers in inflammatory bowel diseases: current status and proteomics identification strategies. *World J. Gastroenterol.* 20, 3231–3244 (2014).
12. Legrain, P. *et al.* The human proteome project: current state and future direction. *Mol. Cell. Proteomics* 10, M111.009993 (2011).
13. Uhlén, M. *et al.* A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* 4, 1920–1932 (2005).
14. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587 (2014).
15. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* 509, 575–581 (2014).
16. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419 (2015).
17. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* 356, (2017).
18. Doll, S. *et al.* Region and cell-type resolved quantitative proteomic map of the human heart. *Nat. Commun.* 8, 1469 (2017).
19. Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* 8, 14271 (2017).
20. Elguoshy, A. *et al.* Why are they missing? : Bioinformatics characterization of missing human proteins. *J. Proteomics* 149, 7–14 (2016).
21. Chait, B. T. Mass Spectrometry: Bottom-Up or Top-Down? *Science* 314, 65–66 (2006).
22. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R., 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* 113, 2343–2394 (2013).
23. Wolters, D. A., Washburn, M. P. & Yates, J. R., 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 73, 5683–5690 (2001).
24. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* 11, M111.014050 (2012).
25. Steen, H. & Mann, M. The ABC’s (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* 5, 699–711 (2004).
26. Schmidt, A., Forne, I. & Imhof, A. Bioinformatic analysis of proteomics data. *BMC Syst. Biol.* 8 Suppl 2, S3 (2014).

27. Aitken, A. & Learmonth, M. P. Protein Determination by UV Absorption. in *The Protein Protocols Handbook* (ed. Walker, J. M.) 3–6 (Humana Press, 2009).

28. Isaacson, T. *et al.* Sample extraction techniques for enhanced proteomic analysis of plant tissues. *Nat. Protoc.* 1, 769–774 (2006).

29. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* 6, 359–362 (2009).

30. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* 68, 850–858 (1996).

31. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* 9, 1323–1329 (2010).

32. Tsiatsiani, L. & Heck, A. J. R. Proteomics beyond trypsin. *FEBS J.* 282, 2612–2626 (2015).

33. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protoc.* 11, 993–1006 (2016).

34. Glatter, T. *et al.* Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *J. Proteome Res.* 11, 5145–5156 (2012).

35. Ahn, J., Cao, M.-J., Yu, Y. Q. & Engen, J. R. Accessing the reproducibility and specificity of pepsin and other aspartic proteases. *Biochim. Biophys. Acta* 1834, 1222–1229 (2013).

36. Hunt, D. F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 255, 1261–1263 (1992).

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

38. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682 (1999).

39. Han, G. *et al.* Large-scale phosphoproteome analysis of human liver tissue by enrichment and fractionation of phosphopeptides with strong anion exchange chromatography. *Proteomics* 8, 1346–1361 (2008).

40. Nguyen, H. P. & Schug, K. A. The advantages of ESI-MS detection in conjunction with HILIC mode separations: Fundamentals and applications. *J. Sep. Sci.* 31, 1465–1480 (2008).

41. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics. *J. Proteome Res.* 13, 6176–6186 (2014).

42. Ritorto, M. S., Cook, K., Tyagi, K., Pedrioli, P. G. A. & Trost, M. Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide separation of complex proteomes. *J. Proteome Res.* 12, 2449–2457 (2013).

43. Ruprecht, B. *et al.* Hydrophilic Strong Anion Exchange (hSAX) Chromatography Enables Deep Fractionation of Tissue Proteomes. *Methods Mol. Biol.* 1550, 69–82 (2017).

44. Fournier, M. L., Gilmore, J. M., Martin-Brown, S. A. & Washburn, M. P. Multidimensional separations-based shotgun proteomics. *Chem. Rev.* 107, 3654–3686 (2007).

45. Hillenkamp, F., Karas, M., Beavis, R. C. & Chait, B. T. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* 63, 1193A–1203A (1991).

46. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246, 64–71 (1989).

47. Karas, M. & Krüger, R. Ion formation in MALDI: the cluster ionization mechanism. *Chem. Rev.* 103, 427–440 (2003).

48. Yarin, A. L., Koombhongse, S. & Reneker, D. H. Taylor cone and jetting from liquid droplets in electrospinning of nanofibers. *J. Appl. Phys.* 90, 4836–4846 (2001).

49. Kebarle, P. & Peschke, M. On the mechanisms by which the charged droplets produced by electrospray lead to gas phase ions. *Anal. Chim. Acta* 406, 11–35 (2000).

50. Iavarone, A. T. & Williams, E. R. Mechanism of charging and supercharging molecules in electrospray ionization. *J. Am. Chem. Soc.* 125, 2319–2327 (2003).

51. Hahne, H. *et al.* DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat. Methods* 10, 989–991 (2013).

52. Wiley, W. C. & McLaren, I. H. Time-of-Flight Mass Spectrometer with Improved Resolution. *Rev. Sci. Instrum.* 26, 1150–1157 (1955).

53. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* 10, M111.011015 (2011).

54. Douglas, D. J., Frank, A. J. & Mao, D. Linear ion traps in mass spectrometry. *Mass Spectrom. Rev.* 24, 1–29 (2005).

55. Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* 72, 1156–1162 (2000).

56. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* 40, 430–443 (2005).

57. Savaryn, J. P., Toby, T. K. & Kelleher, N. L. A researcher's guide to mass spectrometry-based proteomics. *Proteomics* 16, 2435–2443 (2016).

58. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* 28, 710–721 (2010).

59. Kohli, B. M., Eng, J. K., Nitsch, R. M. & Konietzko, U. An alternative sampling algorithm for use in liquid chromatography/tandem mass spectrometry experiments. *Rapid Commun. Mass Spectrom.* 19, 589–596 (2005).

60. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* 11, 601 (1984).

61. Wells, J. M. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* 402, 148–185 (2005).

62. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* 4, 709–712 (2007).

63. Jedrychowski, M. P. *et al.* Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol. Cell. Proteomics* 10, M111.009910 (2011).

64. Mikesh, L. M. *et al.* The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1764, 1811–1822 (2006).

65. Frese, C. K. *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal. Chem.* 84, 9668–9673 (2012).

66. Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *J. Am. Chem. Soc.* 120, 3265–3266 (1998).

67. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9528–9533 (2004).

68. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7, 655–667 (2007).

69. Craig, R., Cortens, J. C., Fenyo, D. & Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* 5, 1843–1849 (2006).

70. Seidler, J., Zinn, N., Boehm, M. E. & Lehmann, W. D. De novo sequencing of peptides by MS/MS. *Proteomics* 10, 634–649 (2010).

71. Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399 (1994).

72. Perkins, D. N., Pappin, D. J. C. & Creasy, D. M. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* (1999).

73. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* 10, 1794–1805 (2011).

74. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976–989 (1994).

75. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4, 207–214 (2007).

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

77. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical Statistical Model To Estimate

the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* 74, 5383–5392 (2002).

78. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* 7, 40–44 (2008).
79. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 73, 2092–2123 (2010).
80. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* 4, 1419–1440 (2005).
81. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75, 4646–4658 (2003).
82. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 4, 787 (2007).
83. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* 8, 2405–2417 (2009).
84. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* 404, 939–965 (2012).
85. Ong, S.-E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 1, 376–386 (2002).
86. Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75, 1895–1904 (2003).
87. Wiese, S., Reidegeld, K. A., Meyer, H. E. & Warscheid, B. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* 7, 340–350 (2007).
88. Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: ‘the good, the bad and the ugly’. *J. Proteome Res.* 8, 5347–5355 (2009).
89. Zhu, W., Smith, J. W. & Huang, C.-M. Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. Biotechnol.* 2010, 840518 (2010).
90. Higgs, R. E., Knierman, M. D., Gelfanova, V., Butler, J. P. & Hale, J. E. Comprehensive label-free method for the relative quantification of proteins from biological samples. *J. Proteome Res.* 4, 1442–1450 (2005).
91. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 389, 1017–1031 (2007).
92. Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* 5, 144–156 (2006).
93. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* 473, 337–342 (2011).
94. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 13, 2513–2526 (2014).
95. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 11, 2301 (2016).
96. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470 (1995).
97. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55, 641–658 (2009).
98. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19, 1720–1730 (1999).
99. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–3973 (2009).
100. Lundberg, E. *et al.* Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* 6, 450 (2010).
101. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* 7, 548 (2011).
102. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and

transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232 (2012).

103. Payne, S. H. The utility of protein and mRNA correlation. *Trends Biochem. Sci.* 40, 1–3 (2015).
104. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 165, 535–550 (2016).
105. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* 136, 731–745 (2009).
106. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44, 283–292 (1986).
107. Hall, M. N., Gabay, J., Débarbouillé, M. & Schwartz, M. A role for mRNA secondary structure in the control of translation initiation. *Nature* 295, 616–618 (1982).
108. Kozak, M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 12, 857–872 (1984).
109. Huez, I. *et al.* Two independent internal ribosome entry sites are involved in translation initiation of vascular endothelial growth factor mRNA. *Mol. Cell. Biol.* 18, 6178–6190 (1998).
110. Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* 349, 97–105 (2005).
111. Hinnebusch, A. G. Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol. Mol. Biol. Rev.* 75, 434–67, first page of table of contents (2011).
112. Brar, G. A. *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335, 552–557 (2012).
113. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* 352, 1413–1416 (2016).
114. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223 (2009).
115. Kaufman, R. J. Regulation of mRNA translation by protein folding in the endoplasmic reticulum. *Trends Biochem. Sci.* 29, 152–158 (2004).
116. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences* 107, 3645–3650 (2010).
117. Sabi, R. & Tuller, T. A comparative genomics study on the effect of individual amino acids on ribosome stalling. *BMC Genomics* 16 Suppl 10, S5 (2015).
118. Yu, C.-H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell* 59, 744–754 (2015).
119. McCaughan, K. K., Brown, C. M., Dalphin, M. E., Berry, M. J. & Tate, W. P. Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc. Natl. Acad. Sci. U. S. A.* 92, 5431–5435 (1995).
120. Tate, W. P. *et al.* The translational stop signal: codon with a context, or extended factor recognition element? *Biochimie* 78, 945–952 (1996).
121. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* 46, 582–592 (2018).
122. Preiss, T. & Hentze, M. W. Dual function of the messenger RNA cap structure in poly(A)-tail-promoted translation in yeast. *Nature* 392, 516–520 (1998).
123. Wang, X. *et al.* N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* 161, 1388–1399 (2015).
124. Pichon, X. *et al.* RNA binding protein/RNA element interactions and the control of translation. *Curr. Protein Pept. Sci.* 13, 294–304 (2012).
125. Iwakawa, H.-O. & Tomari, Y. The Functions of MicroRNAs: mRNA Decay and Translational Repression. *Trends Cell Biol.* 25, 651–665 (2015).
126. Cottrell, K. A. *Regulation of Gene Expression by RNA Binding Proteins and MicroRNAs*. (Washington University, 2017).
127. Rock, K. L. *et al.* Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell* 78, 761–771 (1994).

128. Toyama, B. H. & Hetzer, M. W. Protein homeostasis: live long, won't prosper. *Nat. Rev. Mol. Cell Biol.* 14, 55–61 (2013).

129. Hershko, A., Heller, H., Eytan, E. & Reiss, Y. The protein substrate binding site of the ubiquitin-protein ligase system. *J. Biol. Chem.* 261, 11992–11999 (1986).

130. Nash, P. *et al.* Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature* 414, 514–521 (2001).

131. Rechsteiner, M. & Rogers, S. W. PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* 21, 267–271 (1996).

132. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208 (2005).

133. Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* 6, 400 (2010).

134. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2, e270 (2014).

135. Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.* 11, e1005206 (2015).

136. Jaffe, J. D., Berg, H. C. & Church, G. M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 4, 59–77 (2004).

137. Li, J. *et al.* A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol. Cell. Proteomics* 10, M110.006536 (2011).

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

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

140. Komor, M. A. *et al.* Identification of differentially expressed splice variants by the proteogenomic pipeline Splicify. *Mol. Cell. Proteomics* (2017). doi:10.1074/mcp.TIR117.000056

141. Ruggles, K. V. *et al.* An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol. Cell. Proteomics* 15, 1060–1071 (2016).

142. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA* 8, (2017).

143. Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat. Commun.* 7, 11778 (2016).

144. Na, C. H. *et al.* Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. *Genome Res.* 28, 25–36 (2018).

145. Van Damme, P., Gawron, D., Van Criekinge, W. & Menschaert, G. N-terminal Proteomics and Ribosome Profiling Provide a Comprehensive View of the Alternative Translation Initiation Landscape in Mice and Men. *Mol. Cell. Proteomics* 13, 1245–1261 (2014).

146. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15, 193 (2014).

147. 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* 4, e08890 (2015).

148. Wang, X. *et al.* Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* 11, 1009–1017 (2012).

149. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 11, 1114–1125 (2014).

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

151. Reimand, J., Wagih, O. & Bader, G. D. The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* 3, 2651 (2013).

152. Creixell, P. *et al.* Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* 163, 202–217 (2015).

153. Menschaert, G. & Fenyö, D. Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrom. Rev.* 36, 584–599 (2017).

154. Ruggles, K. V. *et al.* Methods, Tools and Current Perspectives in Proteogenomics. *Mol. Cell. Proteomics* 16, 959–981 (2017).

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

156. Fermin, D. *et al.* Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* 7, R35 (2006).

157. Alfaro, J. A. *et al.* Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. *Genome Med.* 9, 62 (2017).

158. Sheynkman, G. M. *et al.* Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* 15, 703 (2014).

159. Park, H. *et al.* Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses. *Proteomics* 14, 2742–2749 (2014).

160. Wang, X. *et al.* Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *J. Proteome Res.* 11, 1009–1017 (2012).

161. Sheynkman, G. M., Shortreed, M. R., Frey, B. L. & Smith, L. M. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol. Cell. Proteomics* 12, 2341–2353 (2013).

162. Woo, S. *et al.* Proteogenomic database construction driven from large scale RNA-Seq data. *J. Proteome Res.* 13, 21–28 (2014).

163. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* 12, R16 (2011).

164. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15, 205–213 (2014).

165. Koch, A. *et al.* A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* 14, 2688–2698 (2014).

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

167. Wang, X. & Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* 29, 3235–3237 (2013).

168. Nagaraj, S. H. *et al.* PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization. *J. Proteome Res.* 14, 2255–2266 (2015).

169. Zhu, Y. *et al.* Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat. Commun.* 9, 903 (2018).

170. Krug, K. *et al.* Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* 12, 3420–3430 (2013).

171. Brosch, M., Saunders, G. I., Frankish, A. & Collins, M. O. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and ‘resurrected’ pseudogenes in the mouse genome. *Genome* (2011).

172. Branca, R. M. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* 11, 59 (2013).

173. Zhang, K. *et al.* A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics* 31, 3249–3253 (2015).

174. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* 2, 1896–1906 (2007).

175. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367–1372 (2008).

176. Uhlén, M. *et al.* Transcriptomics resources of human tissues and organs. *Mol. Syst. Biol.* 12, 862 (2016).

177. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287 (2012).

178. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26, 976–978 (2010).

179. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* 44, D1251–7 (2016).

180. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* 350, 1092–1096 (2015).

181. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* 163, 1515–1526 (2015).

182. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* 350, 1096–1101 (2015).

183. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183 (2004).

184. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082 (2018).

185. Duan, G., Li, X. & Köhn, M. The human DEPhOsphorylation database DEPOD: a 2015 update. *Nucleic Acids Res.* 43, D531–5 (2015).

186. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* 172, 650–665 (2018).

187. Culhane, A. C., Thioulouse, J., Perrière, G. & Higgins, D. G. MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics* 21, 2789–2790 (2005).

188. Médard, G. *et al.* Optimized chemical proteomics assay for kinase inhibitor profiling. *J. Proteome Res.* 14, 1574–1586 (2015).

189. Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* 14, 259–262 (2017).

190. Toprak, U. H. *et al.* Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Mol. Cell. Proteomics* 13, 2056–2071 (2014).

191. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* 13, 397–406 (2014).

192. Lund, P. K., Abadi, D. M. & Mathies, J. C. Lipid composition of normal human bone marrow as determined by column chromatography. *J. Lipid Res.* (1962).

193. Schmidt, T. *et al.* ProteomicsDB. *Nucleic Acids Res.* 46, D1271–D1281 (2018).

194. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* 45, D177–D182 (2017).

195. Geiger, T. *et al.* Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell. Proteomics* 12, 1709–1722 (2013).

196. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585 (2013).

197. Bhargava, P. & Schnellmann, R. G. Mitochondrial energetics in the kidney. *Nat. Rev. Nephrol.* 13, 629 (2017).

198. Gustafsson, Å. B. & Gottlieb, R. A. Heart mitochondria: gates of life and death. *Cardiovasc. Res.* 77, 334–343 (2008).

199. Williams, E. G. *et al.* Systems proteomics of liver mitochondria function. *Science* 352, aad0189 (2016).

200. Kostrzewa, M., Damian, M. S. & Müller, U. Superoxide dismutase 1: identification of a novel mutation in a case of familial amyotrophic lateral sclerosis. *Hum. Genet.* 98, 48–50 (1996).

201. Garcia, C. C. *et al.* Identification of a mutation in synapsin I, a synaptic vesicle protein, in a family with epilepsy. *J. Med. Genet.* 41, 183–186 (2004).

202. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45, D777–D783 (2017).

203. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* 16, 19–34 (2017).

204. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* 43, D1079–85 (2015).

205. Regard, J. B., Sato, I. T. & Coughlin, S. R. Anatomical profiling of G protein-coupled receptor expression. *Cell* 135, 561–571 (2008).

206. Hao, Y. & Tatonetti, N. P. Predicting G protein-coupled receptor downstream signaling by tissue expression. *Bioinformatics* 32, 3435–3443 (2016).

207. Wu, P., Nielsen, T. E. & Clausen, M. H. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* 36, 422–439 (2015).

208. Ghosh, A. & Giese, K. P. Calcium/calmodulin-dependent kinase II and Alzheimer's disease. *Mol. Brain* 8, 78 (2015).

209. Kong, H. Y. & Byun, J. Emerging roles of human prostatic Acid phosphatase. *Biomol. Ther.* 21, 10–20 (2013).

210. Kustatscher, G., Grabowski, P. & Rappsilber, J. Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* 13, 937 (2017).

211. Laurent, J. M. *et al.* Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* 10, 4209–4212 (2010).

212. Kosti, I., Jain, N., Aran, D., Butte, A. J. & Sirota, M. Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci. Rep.* 6, 24799 (2016).

213. Wilhelm, M. *et al.* Wilhelm et al. reply. *Nature* 547, E23 (2017).

214. Frejno, M. *et al.* Pharmacoproteomic characterisation of human colon and rectal cancer. *Mol. Syst. Biol.* 13, 951 (2017).

215. Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* 12, 883 (2016).

216. Bantscheff, M. *et al.* Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* 25, 1035 (2007).

217. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* 39, D301–8 (2011).

218. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57 (2009).

219. Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev.* 25, 1770–1782 (2011).

220. Mandel, C. R. *et al.* Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* 444, 953–956 (2006).

221. Takagaki, Y. & Manley, J. L. RNA recognition by the human polyadenylation factor CstF. *Mol. Cell. Biol.* 17, 3907–3914 (1997).

222. Mandel, C. R., Bai, Y. & Tong, L. Protein factors in pre-mRNA 3′-end processing. *Cell. Mol. Life Sci.* 65, 1099–1122 (2008).

223. Misra, A. & Green, M. R. From polyadenylation to splicing: Dual role for mRNA 3' end formation factors. *RNA Biol.* 13, 259–264 (2016).

224. Chen, C. Y. & Shyu, A. B. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem. Sci.* 20, 465–470 (1995).

225. Hudson, B. P., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat. Struct. Mol. Biol.* 11, 257–264 (2004).

226. Adachi, S. *et al.* ZFP36L1 and ZFP36L2 control LDLR mRNA stability via the ERK–RSK pathway. *Nucleic Acids Res.* 42, 10037–10049 (2014).

227. Aitken, C. E. & Lorsch, J. R. A mechanistic overview of translation initiation in eukaryotes. *Nat. Struct. Mol. Biol.* 19, 568–576 (2012).

228. Yi, X. *et al.* RNA processing and modification protein, carbon catabolite repression 4 (Ccr4), arrests the cell cycle through p21-dependent and p53-independent pathway. *J. Biol. Chem.* 287, 21045–21057 (2012).

229. Vikesaa, J. *et al.* RNA-binding IMPs promote cell adhesion and invadopodia formation. *EMBO J.* 25, 1456–1468 (2006).

230. Wächter, K., Köhn, M., Stöhr, N. & Hüttelmaier, S. Subcellular localization and RNP formation of

IGF2BPs (IGF2 mRNA-binding proteins) is modulated by distinct RNA-binding domains. *Biol. Chem.* 394, 1077–1090 (2013).

231. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476 (2008).
232. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 338, 1593–1599 (2012).
233. Ellis, J. D. *et al.* Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell* 46, 884–892 (2012).
234. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* 164, 805–817 (2016).
235. Ning, K. & Nesvizhskii, A. I. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics* 11, S14 (2010).
236. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515 (2010).
237. Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* 14, R70 (2013).
238. Ezkurdia, I. *et al.* Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* 14, 1880–1887 (2015).
239. Blakeley, P., Siepen, J. A., Lawless, C. & Hubbard, S. J. Investigating protein isoforms via proteomics: a feasibility study. *Proteomics* 10, 1127–1140 (2010).
240. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526, 68 (2015).
241. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65 (2012).
242. Chen, R. *et al.* Quantitative proteomics reveals that long non-coding RNA MALAT1 interacts with DBC1 to regulate p53 acetylation. *Nucleic Acids Res.* 45, 9947–9959 (2017).
243. Kearse, M. G. & Wilusz, J. E. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.* 31, 1717–1731 (2017).
244. Marx, H. *et al.* Annotation of the Domestic Pig Genome by Quantitative Proteogenomics. *J. Proteome Res.* 16, 2887–2898 (2017).
245. Bánfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* 22, 1646–1657 (2012).
246. Branca, R. M. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* 11, 59–62 (2014).
247. Gevaert, K. *et al.* Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* 21, 566–569 (2003).
248. Kleifeld, O. *et al.* Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat. Biotechnol.* 28, 281–288 (2010).
249. Ping, L. *et al.* Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's Disease. *Scientific Data* 5, 180036 (2018).
250. Vandenbrouck, Y. *et al.* Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. *J. Proteome Res.* 15, 3998–4019 (2016).
251. Wei, W. *et al.* Deep Coverage Proteomics Identifies More Low-Abundance Missing Proteins in Human Testis Tissue with Q-Exactive HF Mass Spectrometer. *J. Proteome Res.* 15, 3988–3997 (2016).
252. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665 (2015).
253. Matthews, D. E. An overview of phenylalanine and tyrosine kinetics in humans. *J. Nutr.* 137, 1549S–1555S; discussion 1573S–1575S (2007).
254. Anderson, L. & Seilhamer, J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537 (1997).
255. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA

levels? *Nature* 547, E19–E20 (2017).

256. Lackner, D. H. *et al.* A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol. Cell* 26, 145–155 (2007).
257. García-Martínez, J., González-Candelas, F. & Pérez-Ortín, J. E. Common gene expression strategies revealed by genome-wide analysis in yeast. *Genome Biol.* 8, R222 (2007).
258. Cheng, Z. *et al.* Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol. Syst. Biol.* 12, 855 (2016).
259. Jovanovic, M. *et al.* Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 347, 1259038 (2015).
260. Panda, A. C., Martindale, J. L. & Gorospe, M. Affinity Pulldown of Biotinylated RNA for Detection of Protein-RNA Complexes. *Bio Protoc* 6, (2016).
261. Galarneau, A. & Richard, S. Target RNA motif and target mRNAs of the Quaking STAR protein. *Nat. Struct. Mol. Biol.* 12, 691–698 (2005).
262. Marchese, D., de Groot, N. S., Lorenzo Gotor, N., Livi, C. M. & Tartaglia, G. G. Advances in the characterization of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA* 7, 793–810 (2016).
263. Shanmugam, A. K. & Nesvizhskii, A. I. Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics. *J. Proteome Res.* 14, 5169–5178 (2015).
264. Dimitrakopoulos, L. *et al.* Variant peptide detection utilizing mass spectrometry: laying the foundations for proteogenomic identification and validation. *Clin. Chem. Lab. Med.* 55, 1291–1304 (2017).
265. Lee, C.-Y. *et al.* Mining the human tissue proteome for protein citrullination. *Mol. Cell. Proteomics* (2018). doi:10.1074/mcp.RA118.000696
266. Mardinoglu, A. *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* 5, 3083 (2014).
267. Park, J. *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* eaar2131 (2018).
268. Specht, H. & Slavov, N. *Transformative opportunities for single-cell proteomics*. (PeerJ Preprints, 2018). doi:10.7287/peerj.preprints.26821v2
269. Spraggins, J. M. *et al.* Next-generation technologies for spatial proteomics: Integrating ultra-high speed MALDI-TOF and high mass resolution MALDI FTICR imaging mass spectrometry for protein analysis. *Proteomics* 16, 1678–1689 (2016).
270. Smith, L. M. & Kelleher, N. L. Proteoforms as the next proteomics currency. *Science* 359, 1106–1107 (2018).

Acknowledgements

Studying abroad for a PhD degree is a truly life-changing experience for me and this journey would not have been possible without numerous help, support and guidance that I received from many people. At the end of my thesis, I would like to express my sincere gratitude to all of them.

First, I would like to express my sincere thanks to my supervisor Prof. Bernhard Küster for providing me an excellent opportunity. Thank him for the continuous support and patient guidance of my doctoral studies and research. His immense knowledge, patience, and encouragement helped me a lot in all the time of research and the writing of manuscripts and this thesis. Thank him for providing such a great research atmosphere. Meanwhile, I would also like to thank Prof. Dr. Julien Gagneur and Prof. Dr. Dietmar Zehn for being part of my thesis committee.

I gratefully acknowledge the funding received towards my PhD from the China Scholarship Council.

I also would like to say many many thanks to Dr. Hannes Hanne. I greatly benefit from his invaluable guidance and significant assistance throughout my PhD time. He patiently taught me a lot about proteomics and mass spectrometry, especially during my 'childhood' in Küster lab. He helped my starting this work in the laboratory and provide tremendous suggestions, comments and feedbacks for the research, manuscripts and thesis. He guided me to the proteomic field and widen my knowledge in this field.

My special thanks go to all the people who make energetic contributions to complete this project. Thank the Human Protein Altas project colleagues for providing the tissue samples and RNA-Seq data. Thank Prof. Marily Theodoropoulou and Guillaume for providing the pituitary samples. Thank Basak Eraslan and Prof. Julien Gagneur for the sequence determinants analysis and collaboration in last 2 years. Thank Li-hua for the help in sample preparation. Thank Jana, Daniel, Svenja and Runsheng for MS measurements. Thank Thomas Wieland, Thomas Hopf, Mathias, Chen, Martin, and Tobias for the help in data analysis. Thank Karl, Steffi, Runsheng, and Maria for the help in RNA pulldown. Thank Andrea for the help in cell culture and Andy for the in-gel digestion.

I must thank all my colleagues for the tremendous help and friendly research environment during my PhD study. Whenever I need help or discussion, they are always there. I will never forget these four years with my excellent colleagues. In particular, I thank Gabi, Silvia and Steffi for their support in administrative issues and many others. Thank Huichao, Chen, Li-hua, Runsheng and Peng for taking care of me during my early days in Germany. Thank Heiner and Susan for teaching me how to take care of mass spectrometry.

Last but not least, I would like to thank my family. I sincerely thank my parents for unconditional love and support throughout my whole life. Many thanks to my brother for the constant encouragement. Without each of you, I'd be nowhere near the person I am and the person I'm still working on becoming. Words cannot express my appreciation. Love you.

Curriculum vitae

Personal information

Name	Dongxue Wang
Date and place of birth	22/10/1989, Liaoning, China
Address	Landshuter str.88, 85356 Freising, Germany
Phone	+49-017675983661; +49-8161-712260
E-mail	dx.wang@tum.de
Nationality	Chinese

Education

10.2014 - present	Technical University of Munich, Germany PhD student at Chair of Proteomics and Bioanalytics under supervision of Prof. Dr. Bernhard Küster Project: Proteogenomics of healthy human tissues
09.2011-06.2014	Shandong University, China Master of Science in Microbiology Thesis: Characterization of β-fructosidase from *A.oryzae* FS4
09.2007-06.2011	Shandong University, China Bachelor of Science in Biotechnology, Thesis: Single cell oil production from delignined corncob residue by *C. curvatus*
09.2004-06.2007	High school Lingyuan No. 1 Middle School, Lingyuan, Liaoning, China

Language	Chinese, native English, fluent

Conference attendance

Oral presentations:	ASMS, 2016, San Antonio, USA Title: Integrated analysis of human tissues with a multi-omics approach HUPO, 2017, Dublin, Ireland Title: Refining Human proteome: Integrated analysis of human tissues with a multi-omics approach
Poster presentations:	MaxQuant Summer School, 2015, Munich, Germany Title: Tissue based in-depth human proteome 9th European summer school – “Advanced proteomcis”, 2015, Brixen, Italy Title: Refining the human proteome: Tissue based in-depth human proteome Proteomic Forum, 2017, Potsdam, Germany Title: Proteogenomic analysis of health human tissues ASMS, 2018, San Diego, USA Title: A deep proteome and transcriptome abundance atlas of 29 healthy human tissues

List of publications

First-author:

[1] **A deep proteome and transcriptome abundance atlas of 29 healthy human tissues**, Wang D*, Eraslan B*, Wieland T, Hallström B, Hopf T, Zolg D, Zecha J, Asplund A, Li LH, Meng C, Frejno M, Schmidt T, Schnatbaum K, Wilhelm M, Ponten F, Uhlen M, Gagneur J, Hahne H, Küster B, submitted to Molecular System Biology (June 2018)

[2] **Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues**, Eraslan B*, Wang D*, Gusic M, Prokisch H, Hallström B, Uhlen M, Asplund A, Ponten F, Wieland T, Hopf T, Hahne H, Küster B, Gagneur J, submitted to Molecular System Biology (June 2018)

[3] **Purification, cloning, characterization, and N-glycosylation analysis of a novel β-fructosidase from *Aspergillus oryzae* FS4 synthesizing levan- and neolevan-type fructooligosaccharides.** Xu L*, Wang D*, Lu L, Jin L, Liu J, Song D, Guo Z, Xiao M. PLoS One. 2014 Dec 12;9(12): e114793. doi: 10.1371/journal.pone.0114793

Co- author:

[1] Mining the human tissue proteome for protein citrullination, Lee C.-Y, Wang D, Wilhelm, M, Zolg D, Schmidt T, Schnatbaum K, Reimer U, Pontén F, Uhlén M, Hahne H, Küster B, Mol Cell Proteomics. 2018 Jul;17(7):1378-1391. doi: 10.1074/mcp.RA118.000696.

[2] Hydrophilic Strong Anion Exchange (hSAX) Chromatography Enables Deep Fractionation of Tissue Proteomes, Ruprecht B, Wang D, Chiozzi RZ, Li LH, Hahne H, Küster B, Methods Mol Biol. 2017;1550:69-82. doi: 10.1007/978-1-4939-6747-6_7.

[3] Sophorolipid production from delignined corncob residue by Wickerhamiella domercqiae var. sophorolipid CGMCC 1576 and Cryptococcus curvatus ATCC 96219, Ma, X.-J., Li, H., Wang, D.-X., Song, X, Appl Microbiol Biotechnol. 2014 Jan;98(1):475-83. doi: 10.1007/s00253-013-4856-3.

*Contributed equally

Appendix

Appendix Figure 1. Abundance distribution of all proteins detected in 29 tissues.

Appendix Figure 2. Protein to mRNA abundance plots of 29 tissues.

Appendix Figure 3. The number of protein groups identified in MS data of 29 tissues by searching against Ensembl and RNA-Seq-based database.

Appendix Table 1. The list of tissues used in the main analyses.

Appendix Table 2. Sequence specific proteins identified by RNA pulldown.

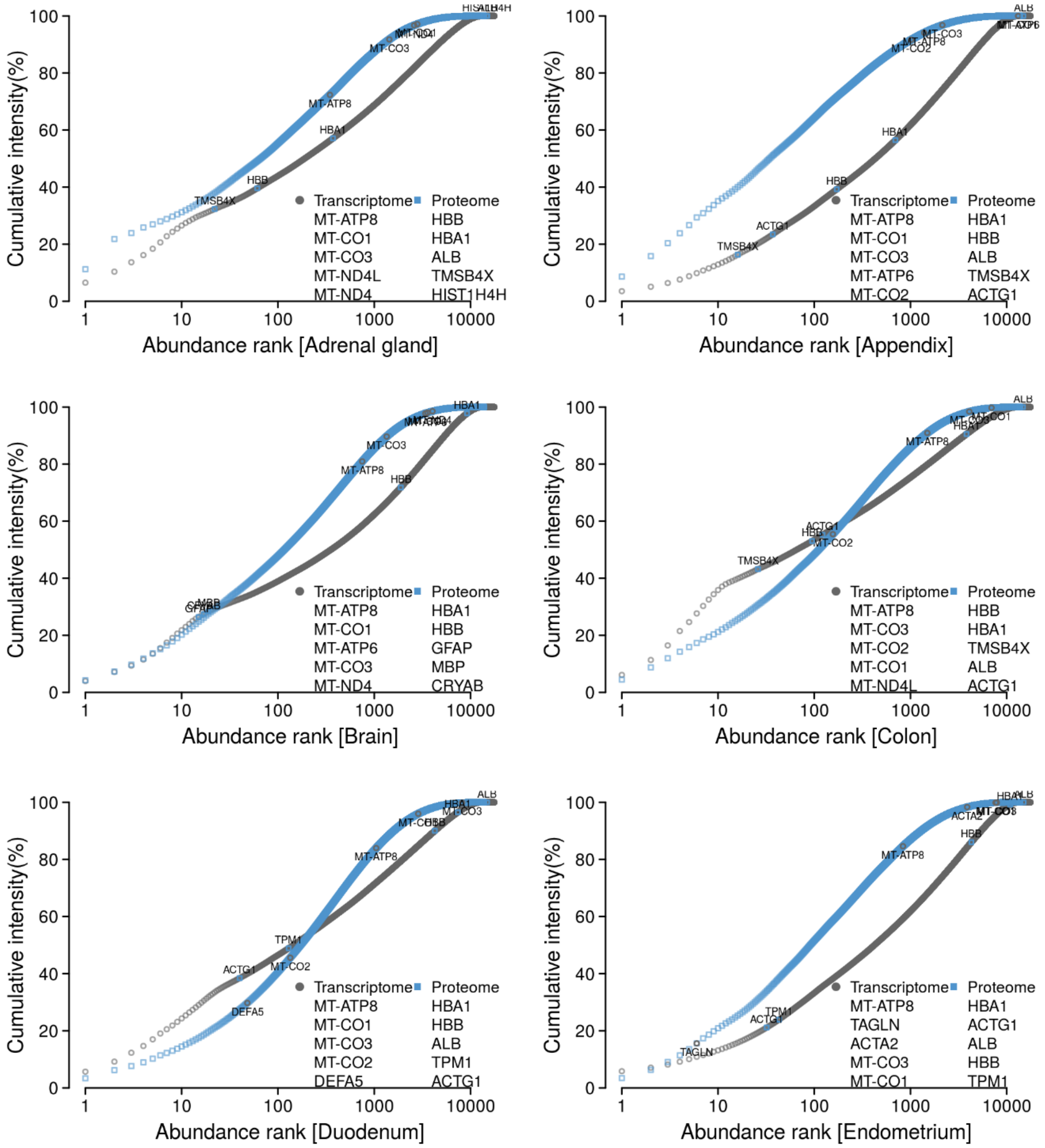
Appendix Table 3. List of variant peptides identified in the ultra-deep tonsil proteome**.**

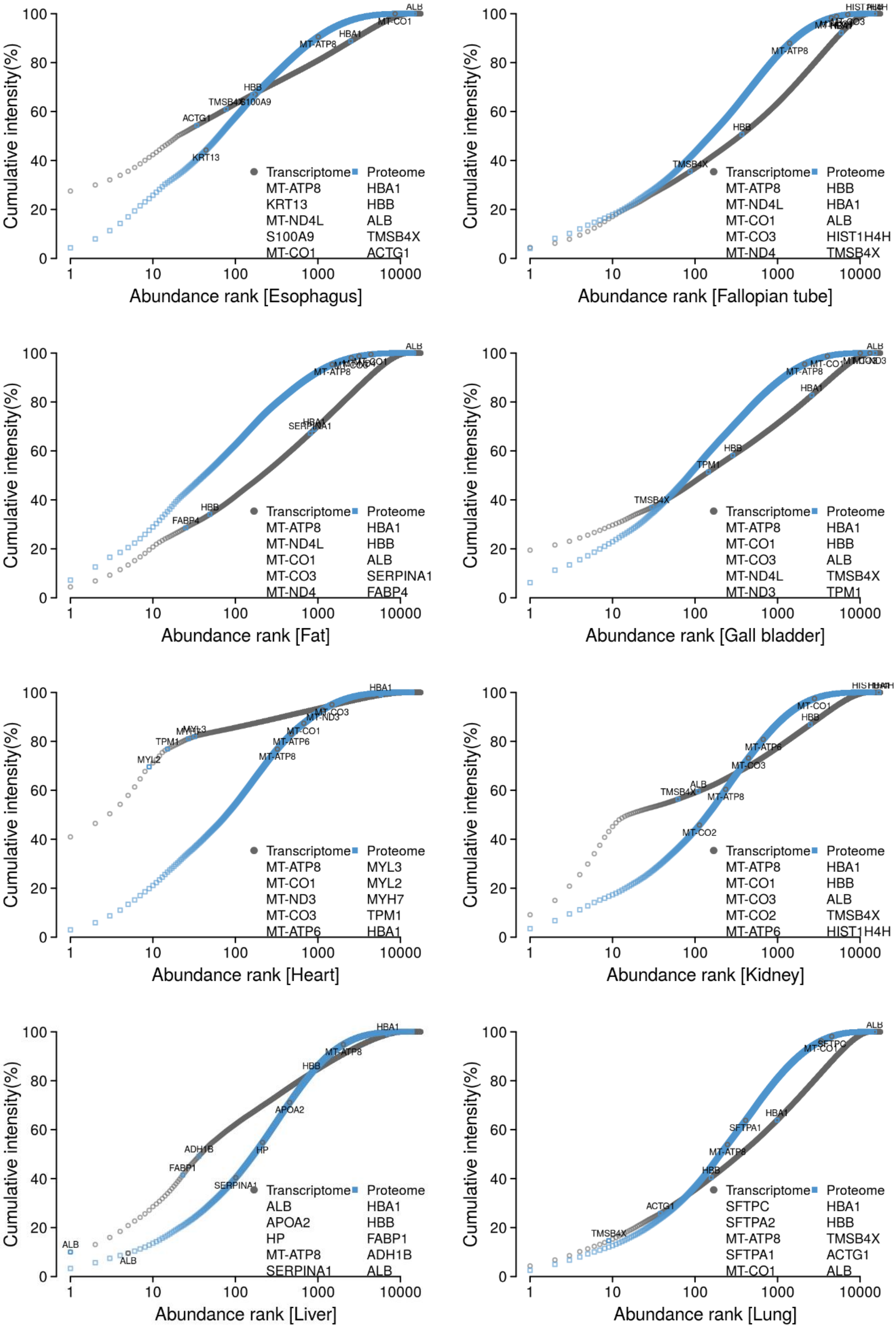
Appendix Table 4. List of peptides mapping to alternative translation initiation sties.

Appendix S1. Spectra of aTIS peptides confirmed by synthetic peptides

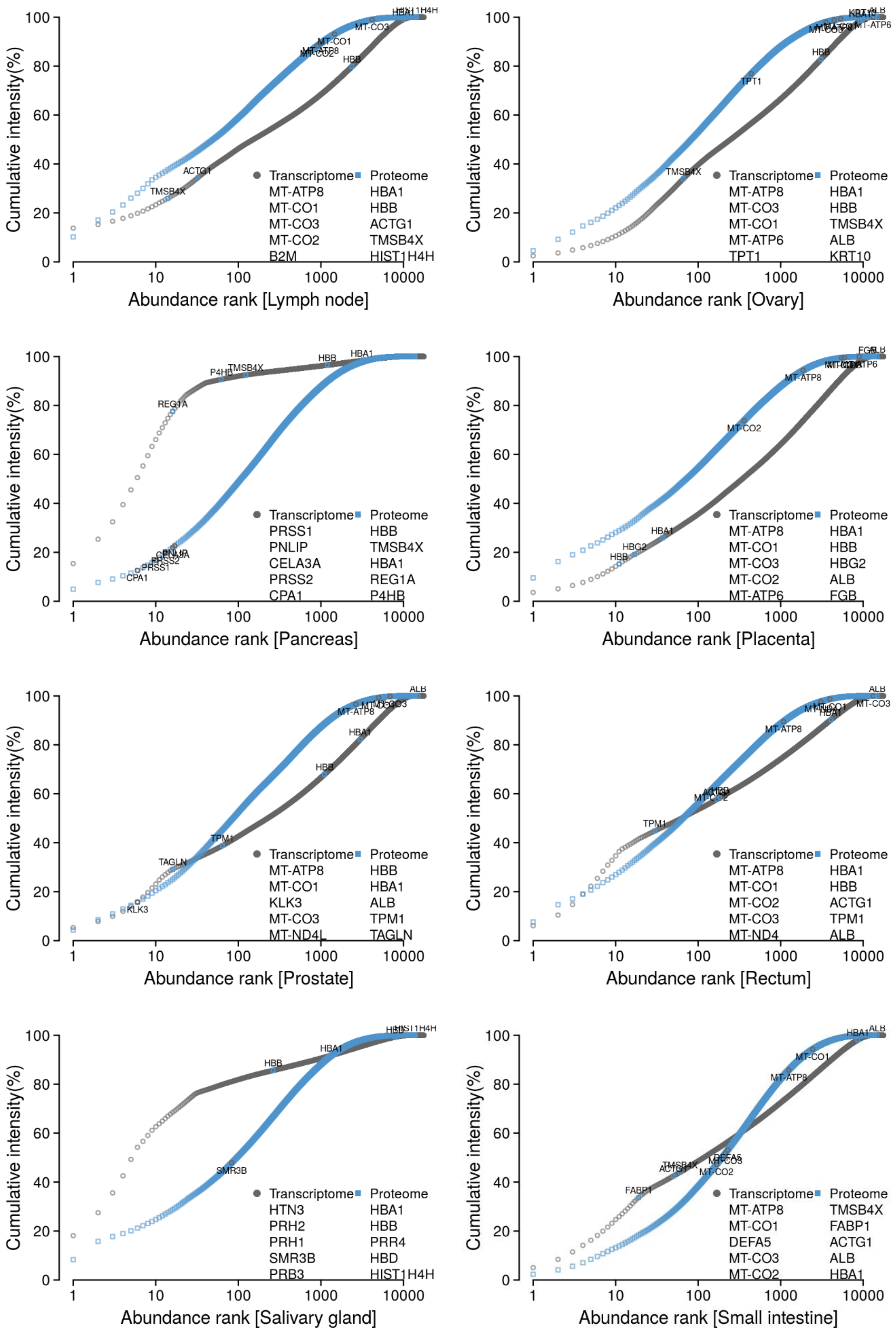
> **Appendix S1.1.** Spectra of aTIS peptides confirmed by spectral contrast angle analysis.
>
> **Appendix S1.2.** Spectra of aTIS peptides confirmed by manual inspection.

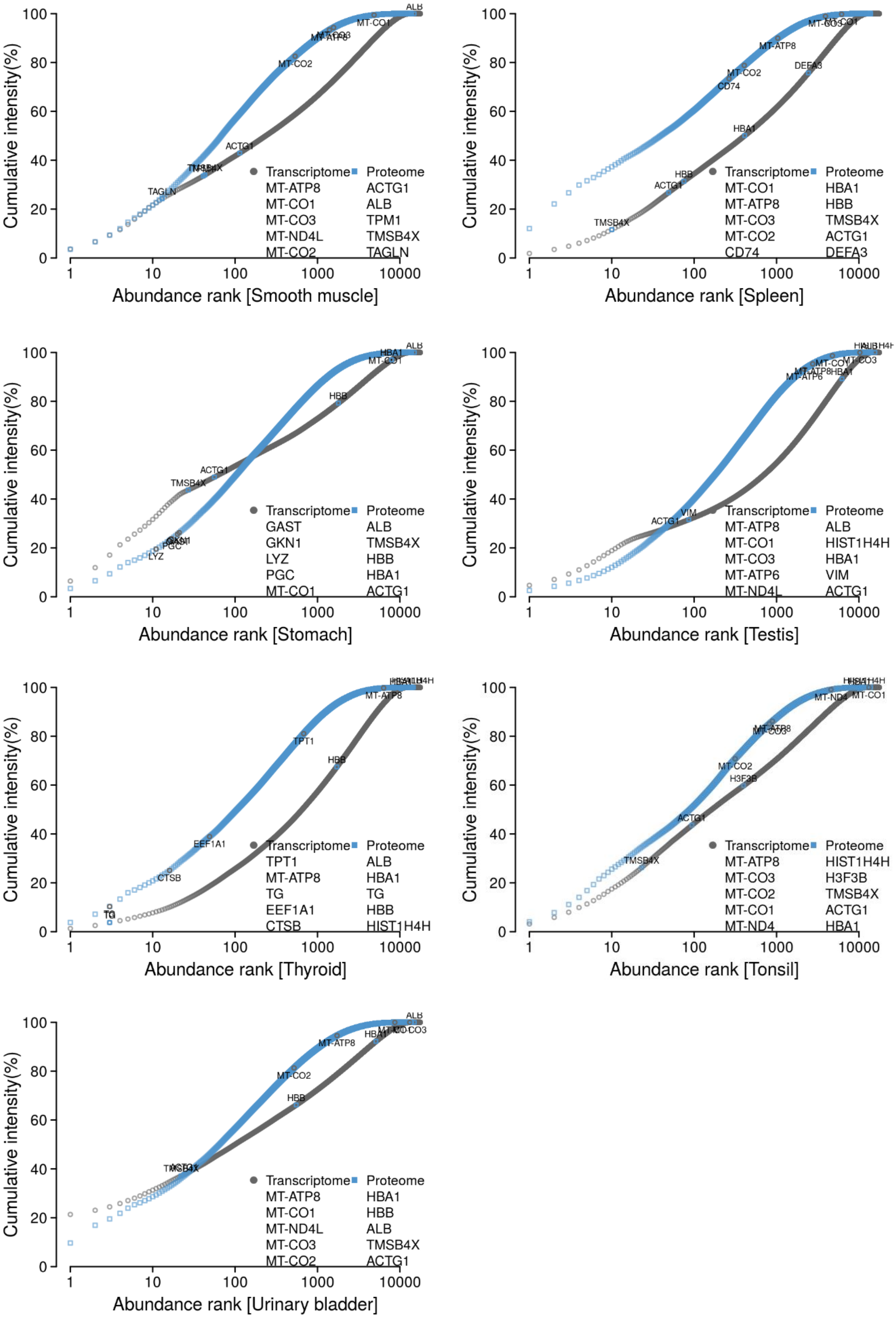
Appendix Figure 1. Abundance distribution of all proteins detected in 29 tissues. The grey dots / circle represent transcripts; the blue squares / hollow squares represent proteins. The gene names of top 5 abundant transcripts and proteins are shown in in the plot.



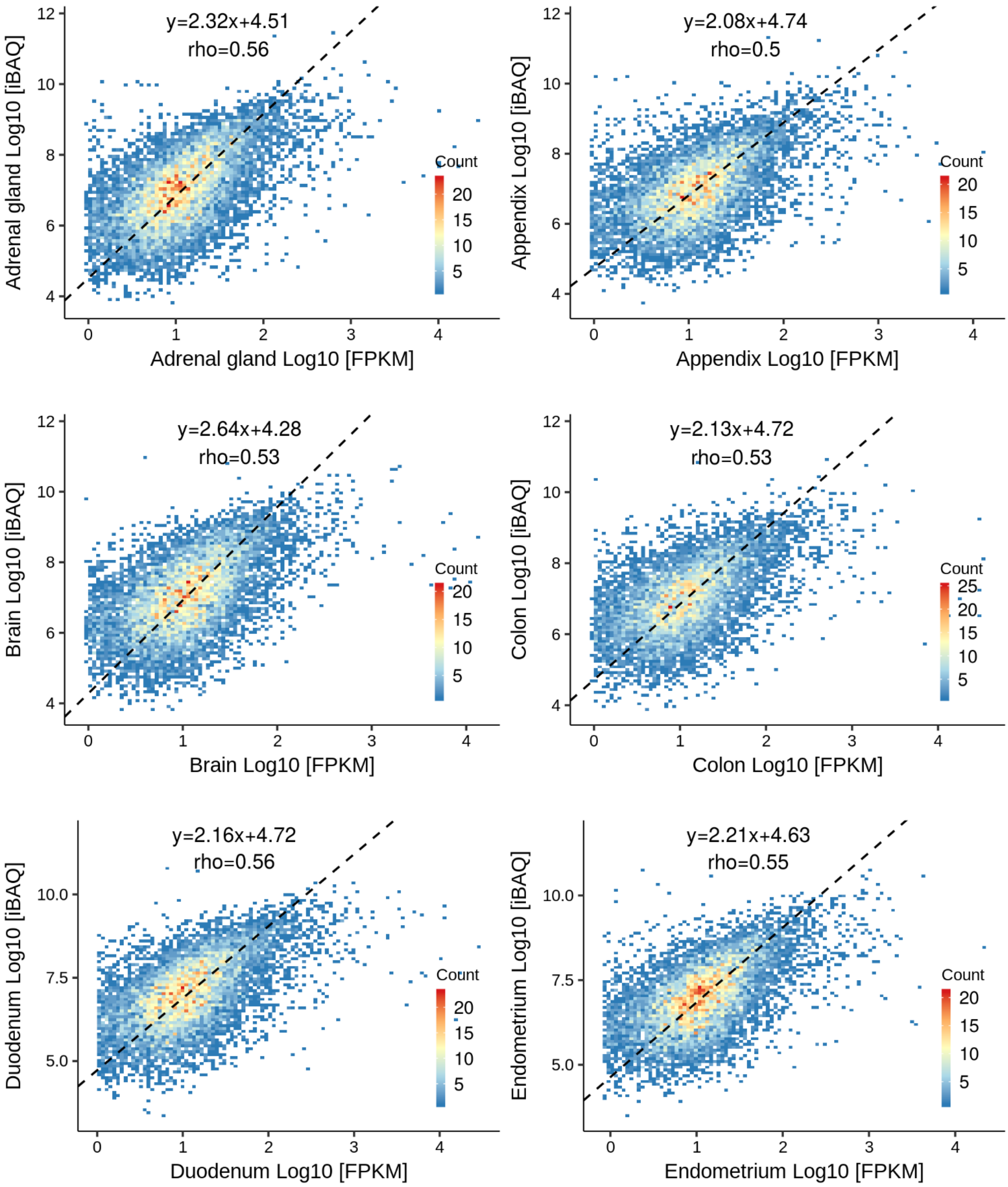


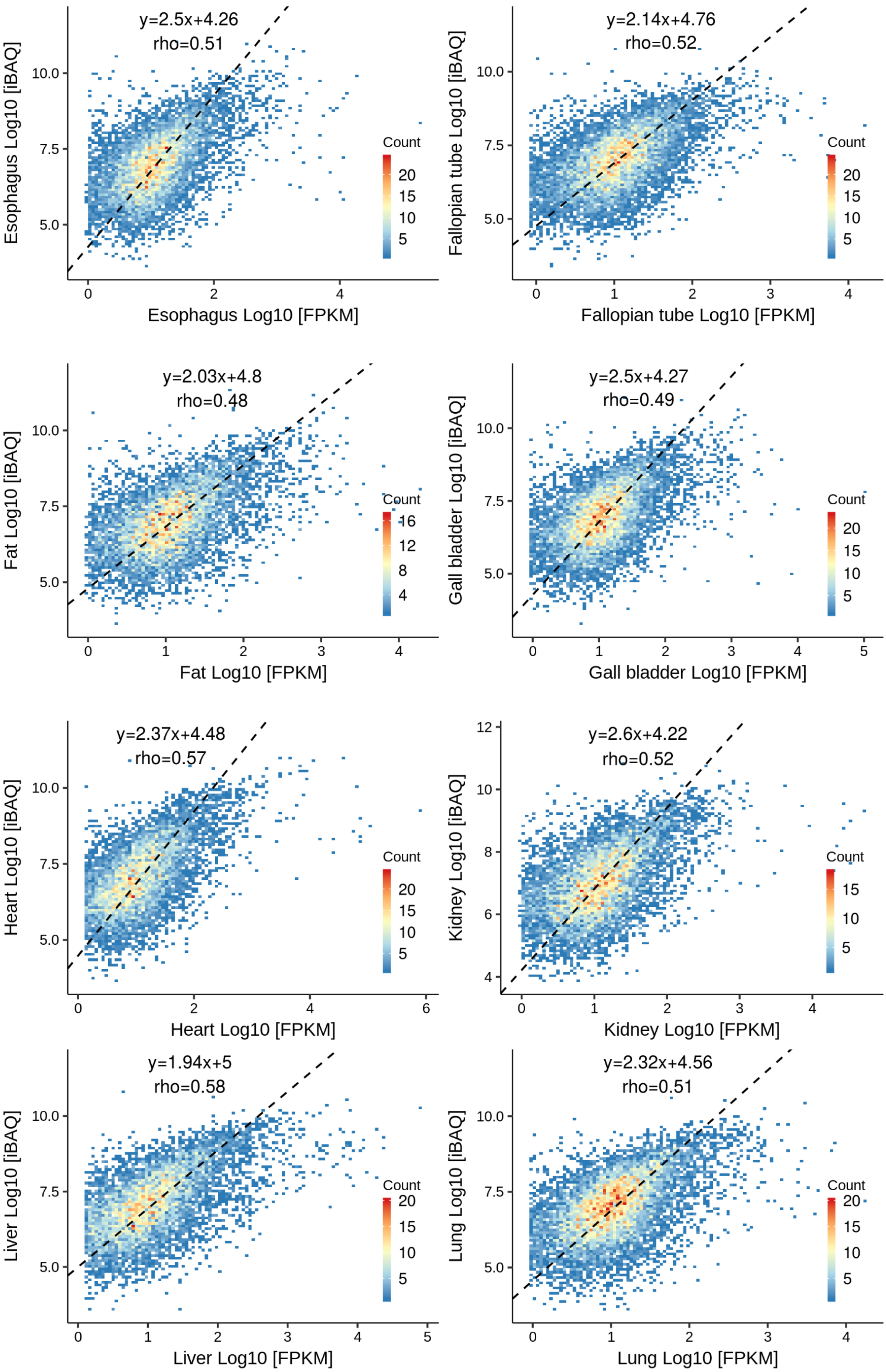
Cumulative intensity(%)
100
80
60
40
20
0
1
10
100
1000
10000
ALB
MT-CO1
HBA1
MT-ATP8
HBB
TMSB4X
S100A9
ACTG1
KRT13
● Transcriptome
■ Proteome
MT-ATP8
MT-ND4L
KRT13
S100A9
MT-CO1
HBA1
HBB
ALB
TMSB4X
ACTG1
Abundance rank [Esophagus]
HIST1H4H
MT-CO3
MT-ATP8
HBB
TMSB4X
MT-ND4L
MT-CO1
MT-ND4
HBA1
HIST1H4H
Abundance rank [Fallopian tube]
SERPINA1
FABP4
MT-ND4
MT-CO3
Abundance rank [Fat]
TPM1
MT-ND3
Abundance rank [Gall bladder]
MYL3
MYL2
MYH7
MT-ND3
MT-ATP6
Abundance rank [Heart]
MT-CO2
Abundance rank [Kidney]
ADH1B
FABP1
SERPINA1
HP
APOA2
Abundance rank [Liver]
SFTPC
SFTPA2
SFTPA1
Abundance rank [Lung]

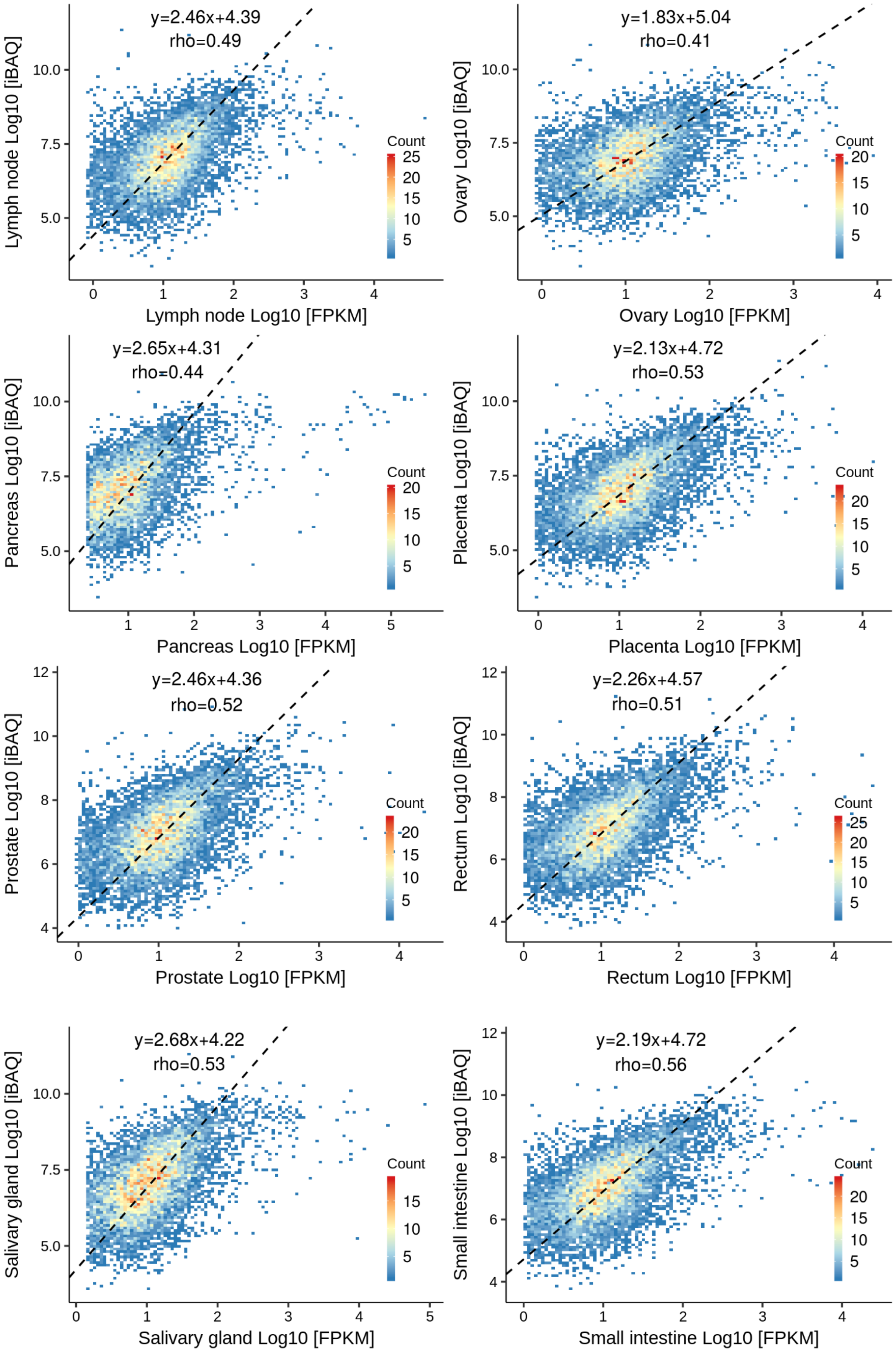

Cumulative intensity(%)
Abundance rank [Lymph node]
● Transcriptome ■ Proteome
MT-ATP8 HBA1
MT-CO1 HBB
MT-CO3 ACTG1
MT-CO2 TMSB4X
B2M HIST1H4H
Abundance rank [Ovary]
● Transcriptome ■ Proteome
MT-ATP8 HBA1
MT-CO3 HBB
MT-CO1 TMSB4X
MT-ATP6 ALB
TPT1 KRT10
Abundance rank [Pancreas]
● Transcriptome ■ Proteome
PRSS1 HBB
PNLIP TMSB4X
CELA3A HBA1
PRSS2 REG1A
CPA1 P4HB
Abundance rank [Placenta]
● Transcriptome ■ Proteome
MT-ATP8 HBA1
MT-CO1 HBB
MT-CO3 HBG2
MT-CO2 ALB
MT-ATP6 FGB
Abundance rank [Prostate]
● Transcriptome ■ Proteome
MT-ATP8 HBB
MT-CO1 HBA1
KLK3 ALB
MT-CO3 TPM1
MT-ND4L TAGLN
Abundance rank [Rectum]
● Transcriptome ■ Proteome
MT-ATP8 HBA1
MT-CO1 HBB
MT-CO2 ACTG1
MT-CO3 TPM1
MT-ND4 ALB
Abundance rank [Salivary gland]
● Transcriptome ■ Proteome
HTN3 HBA1
PRH2 HBB
PRH1 PRR4
SMR3B HBD
PRB3 HIST1H4H
Abundance rank [Small intestine]
● Transcriptome ■ Proteome
MT-ATP8 TMSB4X
MT-CO1 FABP1
DEFA5 ACTG1
MT-CO3 ALB
MT-CO2 HBA1
0
20
40
60
80
100
1
10
100
1000
10000

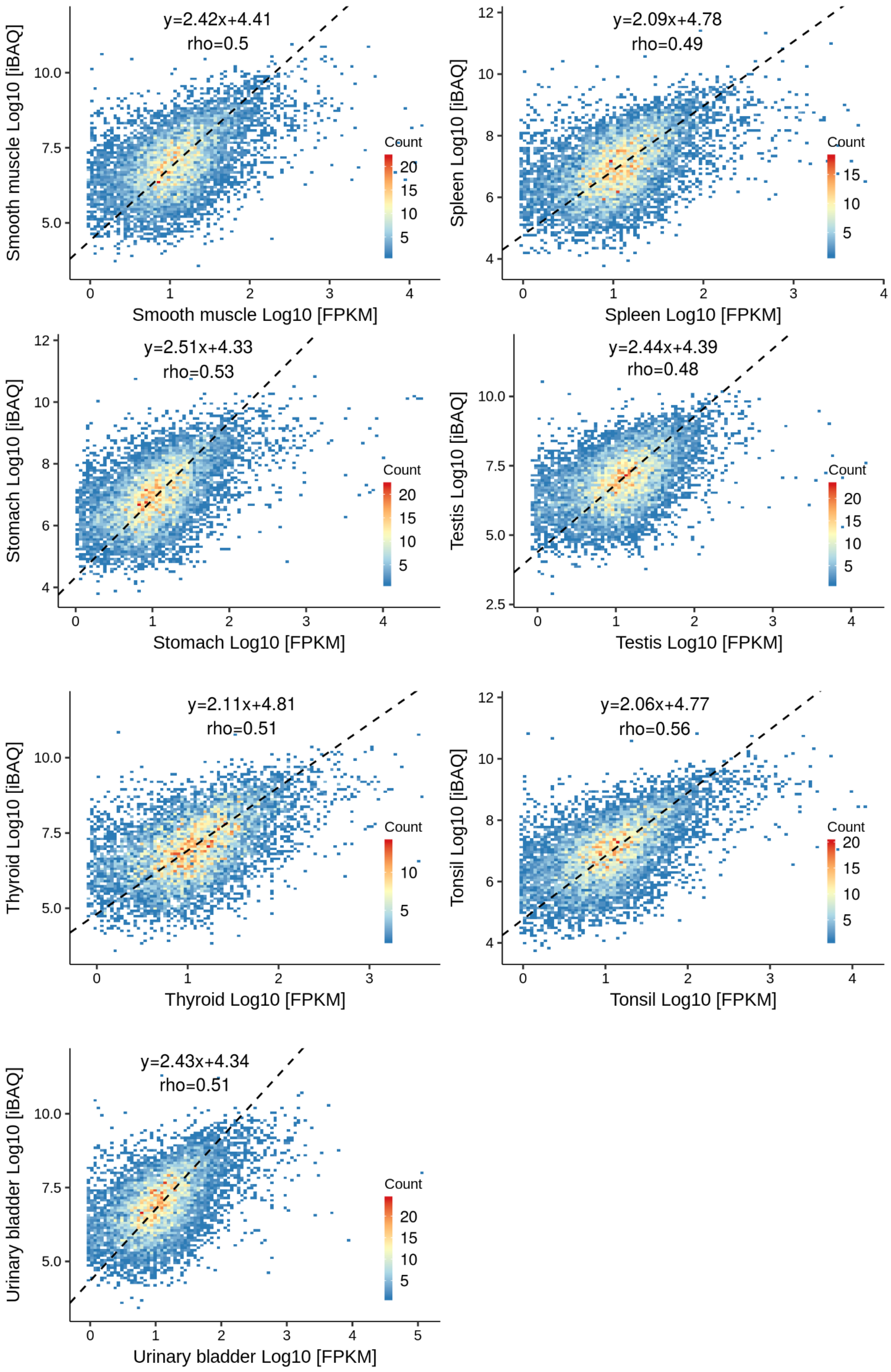

Cumulative intensity(%)
Abundance rank [Smooth muscle]
● Transcriptome ■ Proteome
MT-ATP8 ACTG1
MT-CO1 ALB
MT-CO3 TPM1
MT-ND4L TMSB4X
MT-CO2 TAGLN
ALB
MT-CO1
MT-CO3
MT-ATP8
MT-CO2
ACTG1
TMSB4X
TAGLN
Abundance rank [Spleen]
● Transcriptome ■ Proteome
MT-CO1 HBA1
MT-ATP8 HBB
MT-CO3 TMSB4X
MT-CO2 ACTG1
CD74 DEFA3
MT-CO3
MT-CO1
MT-ATP8
DEFA3
MT-CO2
CD74
HBA1
HBB
ACTG1
TMSB4X
Abundance rank [Stomach]
● Transcriptome ■ Proteome
GAST ALB
GKN1 TMSB4X
LYZ HBB
PGC HBA1
MT-CO1 ACTG1
ALB
HBA1
MT-CO1
HBB
ACTG1
TMSB4X
GKN1
PGC
LYZ
Abundance rank [Testis]
● Transcriptome ■ Proteome
MT-ATP8 ALB
MT-CO1 HIST1H4H
MT-CO3 HBA1
MT-ATP6 VIM
MT-ND4L ACTG1
MT-CO1
MT-CO3
MT-ATP8
HBA1
MT-ATP6
ACTG1
VIM
Abundance rank [Thyroid]
● Transcriptome ■ Proteome
TPT1 ALB
MT-ATP8 HBA1
TG TG
EEF1A1 HBB
CTSB HIST1H4H
MT-ATP8
TPT1
HBB
EEF1A1
CTSB
TG
Abundance rank [Tonsil]
● Transcriptome ■ Proteome
MT-ATP8 HIST1H4H
MT-CO3 H3F3B
MT-CO2 TMSB4X
MT-CO1 ACTG1
MT-ND4 HBA1
MT-ND4
MT-CO1
MT-ATP8
MT-CO3
MT-CO2
H3F3B
ACTG1
TMSB4X
Abundance rank [Urinary bladder]
● Transcriptome ■ Proteome
MT-ATP8 HBA1
MT-CO1 HBB
MT-ND4L ALB
MT-CO3 TMSB4X
MT-CO2 ACTG1
ALB
MT-CO3
HBA1
MT-ATP8
MT-CO2
HBB
TMSB4X
0
20
40
60
80
100
1
10
100
1000
10000

Appendix Figure 2. Protein to mRNA abundance plots of 29 tissues.

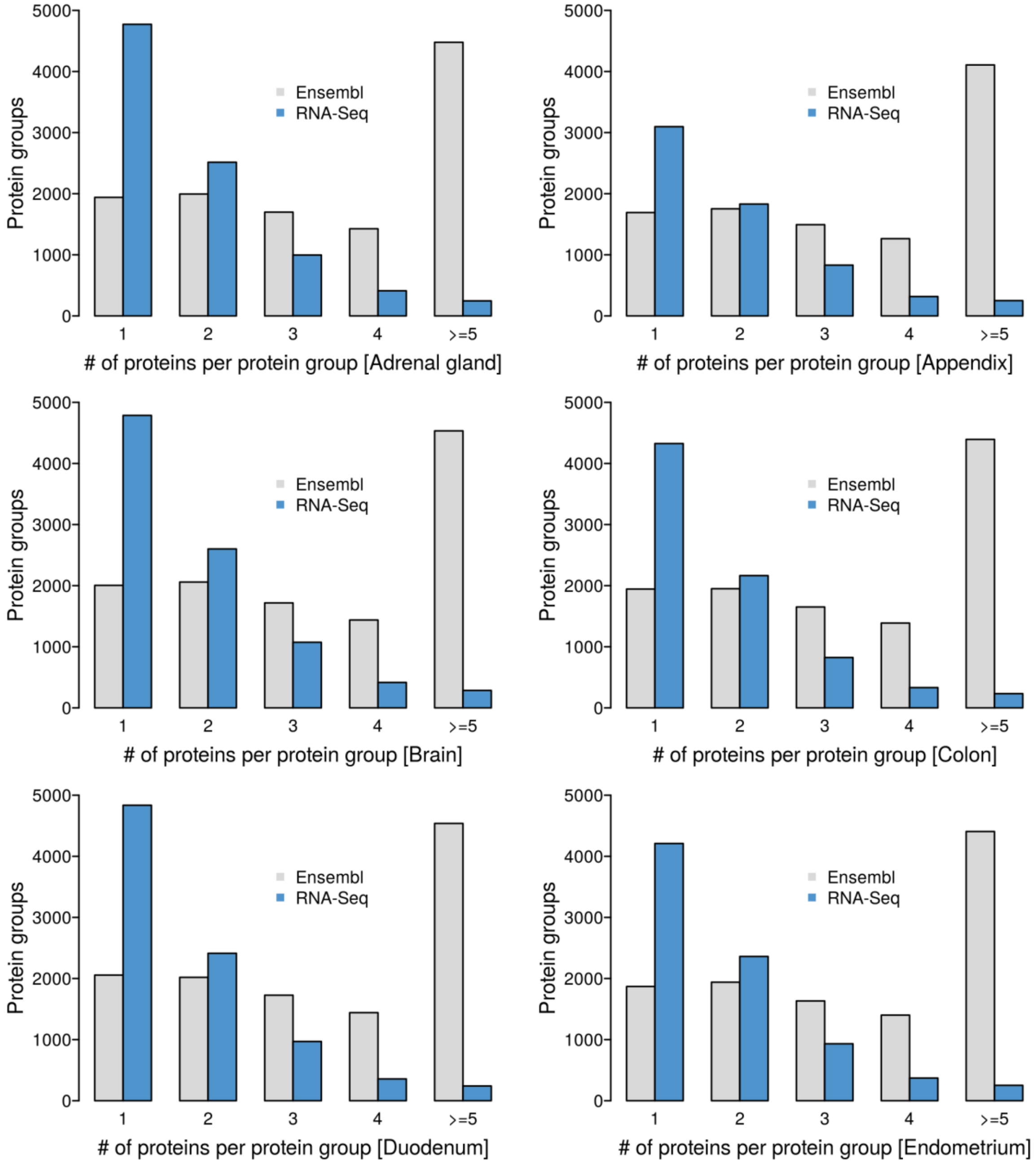


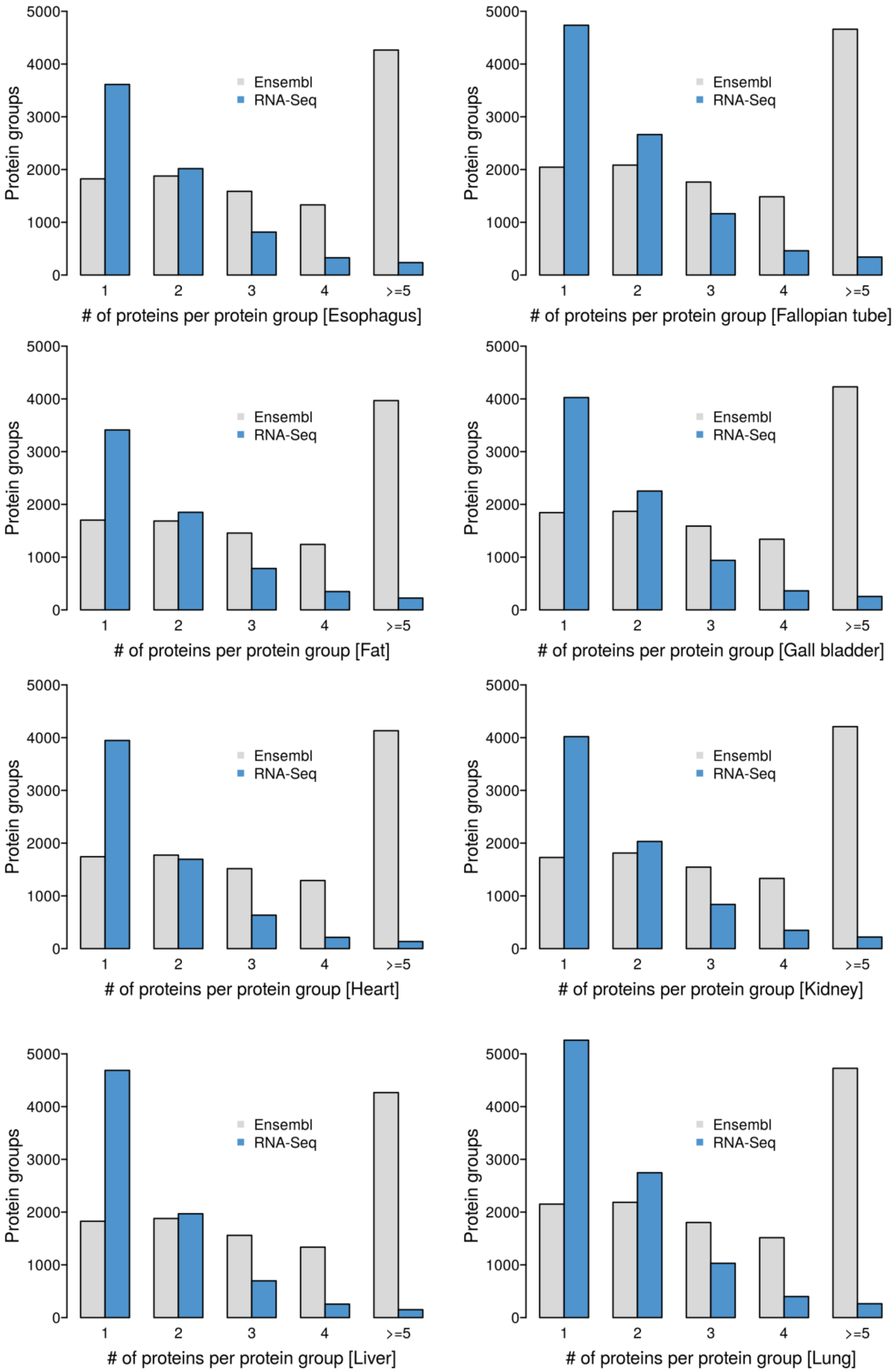

y=2.5x+4.26
rho=0.51
Esophagus Log10 [iBAQ]
10.0
7.5
5.0
0
2
4
Esophagus Log10 [FPKM]
Count
20
15
10
5
y=2.14x+4.76
rho=0.52
Fallopian tube Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
4
Fallopian tube Log10 [FPKM]
Count
20
15
10
5
y=2.03x+4.8
rho=0.48
Fat Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
4
Fat Log10 [FPKM]
Count
16
12
8
4
y=2.5x+4.27
rho=0.49
Gall bladder Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
4
5
Gall bladder Log10 [FPKM]
Count
20
15
10
5
y=2.37x+4.48
rho=0.57
Heart Log10 [iBAQ]
10.0
7.5
5.0
0
2
4
6
Heart Log10 [FPKM]
Count
20
15
10
5
y=2.6x+4.22
rho=0.52
Kidney Log10 [iBAQ]
12
10
8
6
4
0
1
2
3
4
Kidney Log10 [FPKM]
Count
15
10
5
y=1.94x+5
rho=0.58
Liver Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
4
5
Liver Log10 [FPKM]
Count
20
15
10
5
y=2.32x+4.56
rho=0.51
Lung Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
4
Lung Log10 [FPKM]
Count
20
15
10
5

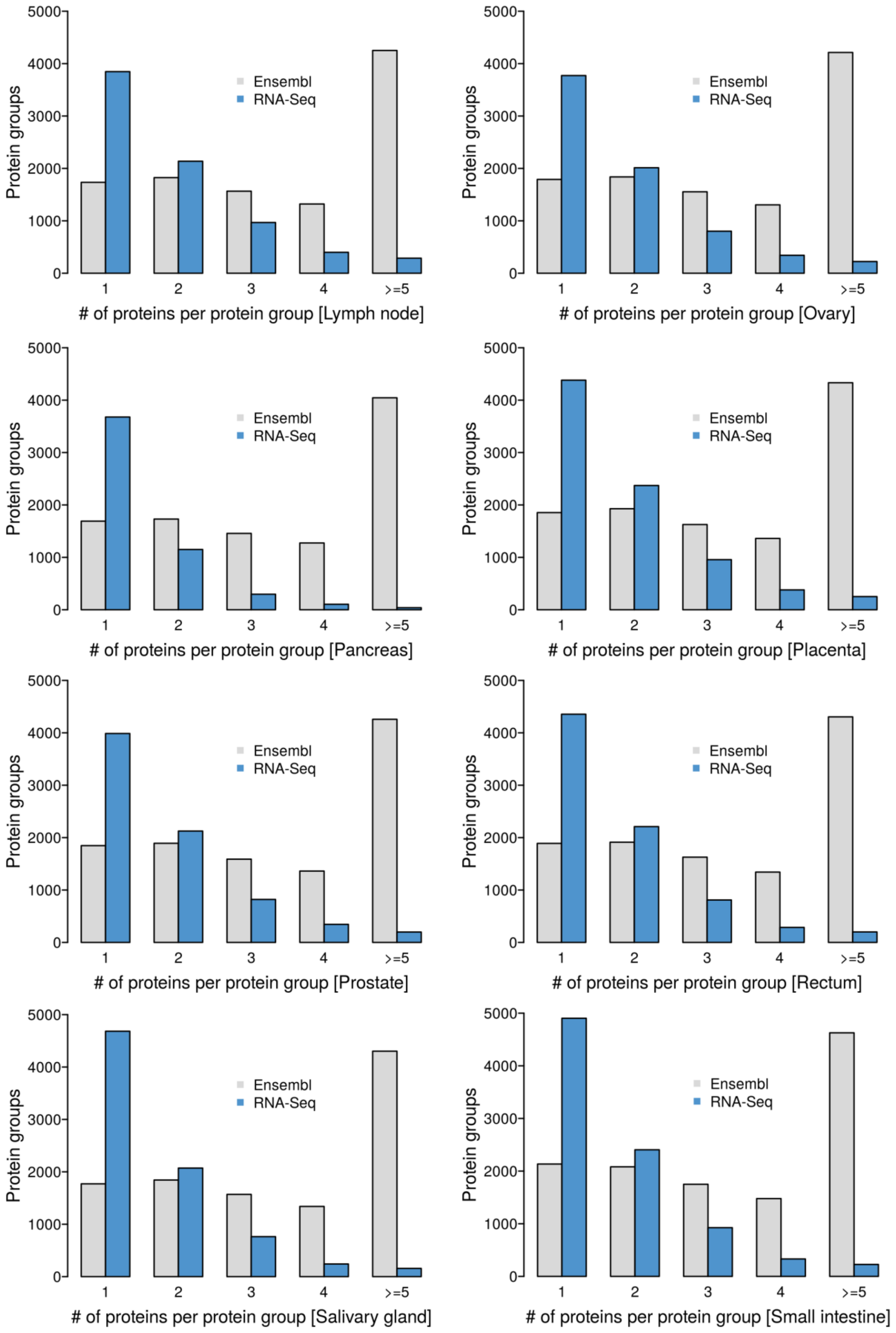

y=2.46x+4.39
rho=0.49
Lymph node Log10 [iBAQ]
Lymph node Log10 [FPKM]
Count
y=1.83x+5.04
rho=0.41
Ovary Log10 [iBAQ]
Ovary Log10 [FPKM]
Count
y=2.65x+4.31
rho=0.44
Pancreas Log10 [iBAQ]
Pancreas Log10 [FPKM]
Count
y=2.13x+4.72
rho=0.53
Placenta Log10 [iBAQ]
Placenta Log10 [FPKM]
Count
y=2.46x+4.36
rho=0.52
Prostate Log10 [iBAQ]
Prostate Log10 [FPKM]
Count
y=2.26x+4.57
rho=0.51
Rectum Log10 [iBAQ]
Rectum Log10 [FPKM]
Count
y=2.68x+4.22
rho=0.53
Salivary gland Log10 [iBAQ]
Salivary gland Log10 [FPKM]
Count
y=2.19x+4.72
rho=0.56
Small intestine Log10 [iBAQ]
Small intestine Log10 [FPKM]
Count

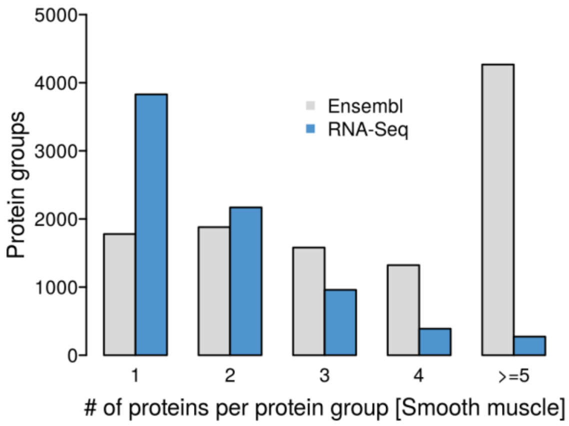

y=2.42x+4.41
rho=0.5
Smooth muscle Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
4
Smooth muscle Log10 [FPKM]
Count
20
15
10
5
y=2.09x+4.78
rho=0.49
Spleen Log10 [iBAQ]
12
10
8
6
4
0
1
2
3
4
Spleen Log10 [FPKM]
Count
15
10
5
y=2.51x+4.33
rho=0.53
Stomach Log10 [iBAQ]
12
10
8
6
4
0
1
2
3
4
Stomach Log10 [FPKM]
Count
20
15
10
5
y=2.44x+4.39
rho=0.48
Testis Log10 [iBAQ]
10.0
7.5
5.0
2.5
0
1
2
3
4
Testis Log10 [FPKM]
Count
20
15
10
5
y=2.11x+4.81
rho=0.51
Thyroid Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
Thyroid Log10 [FPKM]
Count
10
5
y=2.06x+4.77
rho=0.56
Tonsil Log10 [iBAQ]
12
10
8
6
4
0
1
2
3
4
Tonsil Log10 [FPKM]
Count
20
15
10
5
y=2.43x+4.34
rho=0.51
Urinary bladder Log10 [iBAQ]
10.0
7.5
5.0
0
1
2
3
4
5
Urinary bladder Log10 [FPKM]
Count
20
15
10
5

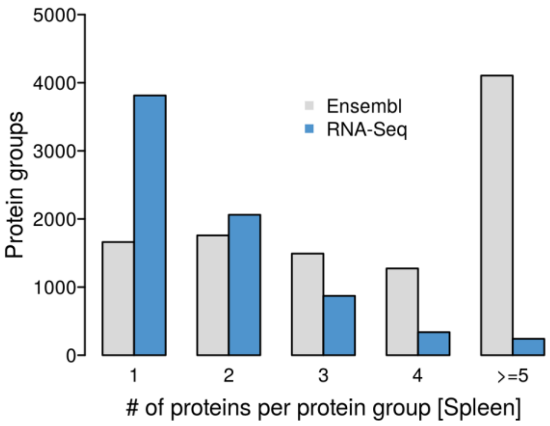
Appendix Figure 3. Searching the proteomic data of 29 tissues against their tissue-specific sequence databases constructed from RNA-Seq data drastically reduces the number of individual protein sequences in protein groups compared to searches against Ensembl.

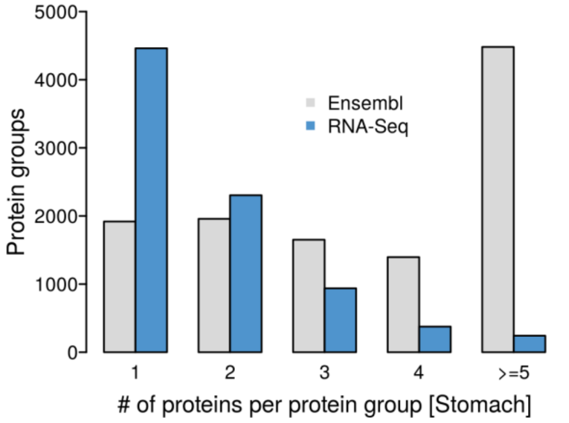


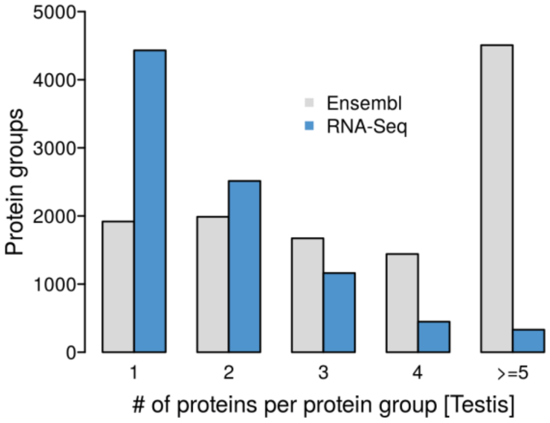

Protein groups
0
1000
2000
3000
4000
5000
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Esophagus]
# of proteins per protein group [Fallopian tube]
# of proteins per protein group [Fat]
# of proteins per protein group [Gall bladder]
# of proteins per protein group [Heart]
# of proteins per protein group [Kidney]
# of proteins per protein group [Liver]
# of proteins per protein group [Lung]

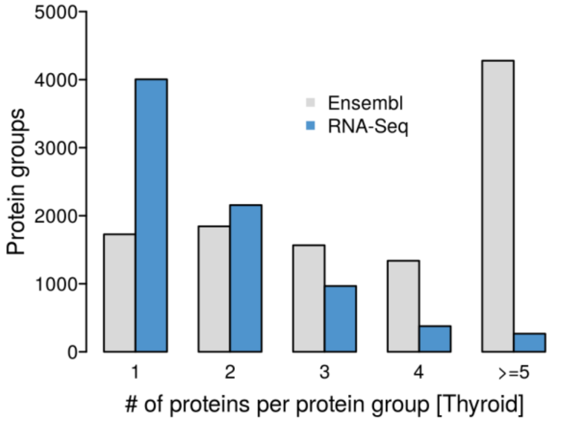

Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Lymph node]
# of proteins per protein group [Ovary]
# of proteins per protein group [Pancreas]
# of proteins per protein group [Placenta]
# of proteins per protein group [Prostate]
# of proteins per protein group [Rectum]
# of proteins per protein group [Salivary gland]
# of proteins per protein group [Small intestine]

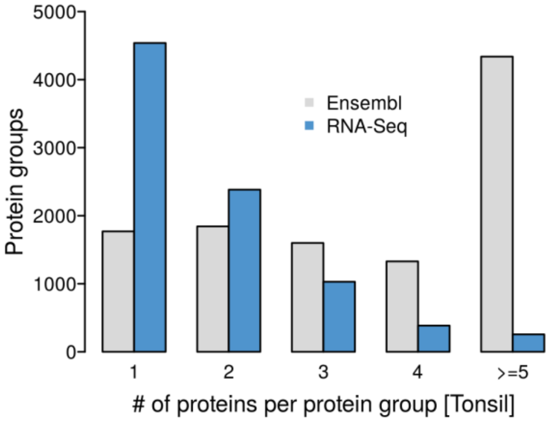

Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Smooth muscle]

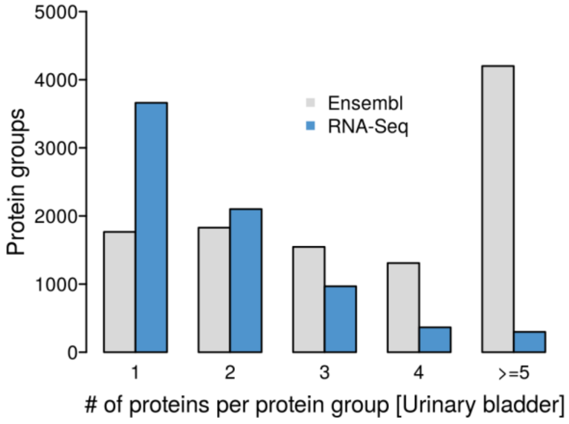

Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Spleen]


Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Stomach]


Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Testis]


Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Thyroid]


Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Tonsil]


Protein groups
5000
4000
3000
2000
1000
0
Ensembl
RNA-Seq
1
2
3
4
>=5
# of proteins per protein group [Urinary bladder]

Appendix Table 1. The list of tissues used in the main analyses

	Tissue	Tissue ID	MS Experiments	Proteases	fragmentation techniques	RNA-Seq Data ID	Studies
1	Adrenal Gland	V122	P015424	Trypsin	HCD	P282_105	Quantitative proteome; Isoforms; lncRNA; aTIS
2	Appendix	V154	P013677	Trypsin	HCD	P282_120	Quantitative proteome; Isoforms; lncRNA; aTIS
3	Bone Marrow	V364	P013674	Trypsin	HCD	-	Quantitative proteome; Isoforms; lncRNA; aTIS
4	Brain	V102	P013129	Trypsin	HCD	P262_159	Quantitative proteome; Isoforms; lncRNA; aTIS
5	Colon	V269	P013387	Trypsin	HCD	P973_107	Quantitative proteome; Isoforms; lncRNA; aTIS
6	Colon	V291	P010693	Trypsin	HCD	-	lncRNA; aTIS
7	Duodenum	V145	P013187	Trypsin	HCD	P282_126	Quantitative proteome; Isoforms; lncRNA; aTIS
8	Endometrium	V200	P013386	Trypsin	HCD	P414_119	Quantitative proteome; Isoforms; lncRNA; aTIS
9	Esophagus	V184	P013198	Trypsin	HCD	P414_107	Quantitative proteome; Isoforms; lncRNA; aTIS
10	Fallopian Tube	V277	P013559	Trypsin	HCD	P973_112	Quantitative proteome; Isoforms; lncRNA; aTIS
11	Fat	V315	P015159	Trypsin	HCD	P973_140	Quantitative proteome; Isoforms; lncRNA; aTIS
12	Gall Bladder	V179	P013197	Trypsin	HCD	P414_104	Quantitative proteome; Isoforms; lncRNA; aTIS
13	Heart	V191	P013201	Trypsin	HCD	P414_111	Quantitative proteome; Isoforms; lncRNA; aTIS
14	Kidney	V359	P013560	Trypsin	HCD	P1973_516	Quantitative proteome; Isoforms; lncRNA; aTIS
15	Liver	V358	P012502	Trypsin	HCD	P1973_515	Quantitative proteome; Isoforms; lncRNA; aTIS
16	Liver	V362	P013127	Trypsin	HCD	P1973_519	lncRNA; aTIS
17	Liver	V289	P010738	Trypsin	HCD	-	lncRNA; aTIS
18	Lung	V133	P013163	Trypsin	HCD	P282_114	Quantitative proteome; Isoforms; lncRNA; aTIS
19	Lung	V299	P010740	Trypsin	HCD	-	lncRNA; aTIS
20	Lymph node	V193	P015160	Trypsin	HCD	P414_113	Quantitative proteome; Isoforms; lncRNA; aTIS

21	Ovary	V233	P013679	Trypsin	HCD	P415_106	Quantitative proteome; Isoforms; lncRNA; aTIS
22	Pancreas	V229	P013678	Trypsin	HCD	P415_104	Quantitative proteome; Isoforms; lncRNA; aTIS
23	Pituitary	-	P018021	Trypsin	HCD	-	lncRNA; aTIS
24	Pituitary	-	P018020	Trypsin	HCD	-	lncRNA; aTIS
25	Placenta	V223	P013680	Trypsin	HCD	P415_102	Quantitative proteome; Isoforms; lncRNA; aTIS
26	Placenta	V262	P010695	Trypsin	HCD	-	lncRNA; aTIS
27	Prostate	V128	P013675	Trypsin	HCD	P282_109	Quantitative proteome; Isoforms; lncRNA; aTIS
28	Rectum	V321	P013681	Trypsin	HCD	P973_143	Quantitative proteome; Isoforms; lncRNA; aTIS
29	Salivary Gland	V240	P015161	Trypsin	HCD	P415_112	Quantitative proteome; Isoforms; lncRNA; aTIS
30	Small Intestine	V151	P013188	Trypsin	HCD	P282_136	Quantitative proteome; Isoforms; lncRNA; aTIS
31	Smooth Muscle	V266	P013562	Trypsin	HCD	P973_105	Quantitative proteome; Isoforms; lncRNA; aTIS
32	Spleen	V82	P013114	Trypsin	HCD	P262_150	Quantitative proteome; Isoforms; lncRNA; aTIS
33	Stomach	V90	P013128	Trypsin	HCD	P262_154	Quantitative proteome; Isoforms; lncRNA; aTIS
34	Stomach	V296	P010739	Trypsin	HCD	-	lncRNA; aTIS
35	Testis	V134	P013164	Trypsin	HCD	P282_115	Quantitative proteome; Isoforms; lncRNA; aTIS
36	Thyroid	V196	P013385	Trypsin	HCD	P414_115	Quantitative proteome; Isoforms; lncRNA; aTIS
37	Tonsil	V287	P018699	Chymotrypsin	HCD	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
38	Tonsil	V287	P018699	Chymotrypsin	CID	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
39	Tonsil	V287	P018444	LysN	HCD	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
40	Tonsil	V287	P018443	AspN	HCD	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
41	Tonsil	V287	P018442	ArgC	HCD	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
42	Tonsil	V287	P018441	GluC	HCD	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS

43	Tonsil	V287	P018440	LysC	HCD	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
44	Tonsil	V287	P010747	Trypsin	HCD	P973_118	Quantitative proteome; Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
45	Tonsil	V287	P010747	Trypsin	CID	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
46	Tonsil	V287	P010747	Trypsin	EThcD/ETD	P973_118	Ultra-deep tonsil proteome; Isoforms; Exomic variants; lncRNA; aTIS
47	Tonsil	V293	P010694	Trypsin	HCD	-	lncRNA; aTIS
48	Urinary Bladder	V176	P013196	Trypsin	HCD	P414_102	Quantitative proteome; Isoforms; lncRNA; aTIS

Appendix Table 2. Sequence specific proteins identified by RNA pulldown

	Sequence specific proteins			EC50 [log10, mol/L]							
				AAUAAA				CUGUCCU		UAUUUAU	
Gene name	AAUAAA	CUGUCCU	UAUUUAU	Exp_1	Exp_2	Exp_3	Random	Exp_1	Random	Exp_1	Random
CDKN2AIP	1	0	0	7.97	5.92	-	-	-	-	6.38	6.15
CSTF1	1	0	0	7.53	7.34	6.94	-	-	-	6.24	-
FIP1L1	1	0	0	7.51	7.08	6.97	6.31	-	-	5.86	-
CPSF1	1	0	0	7.51	7.00	7.06	6.33	-	-	5.77	-
CSTF3	1	0	0	7.50	-	6.69	-	-	-	6.22	-
CPSF3	1	0	0	7.48	7.17	7.07	6.22	-	-	6.22	-
P4HA1	1	0	0	7.46	7.12	6.97	-	-	-	-	-
CPSF2	1	0	0	7.46	7.12	6.92	6.26	-	-	5.82	-
NME2	1	0	0	7.18	5.04	-	-	-	-	-	-
SRSF10	1	0	0	5.87	6.01	-	-	-	-	-	-
IFIT5	1	0	0	5.84	6.05	5.88	-	-	-	-	-
ERI3	1	0	0	5.74	5.89	5.48	-	-	-	-	-
PABPN1	1	0	0	5.65	5.26	-	-	-	-	-	-
MKRN2	1	0	0	5.63	5.61	5.10	-	-	-	-	-
SYMPK	1	0	0	-	7.04	6.83	-	-	-	-	-
HMGB2	1	0	0	8.60	-	-	-	-	-	-	-
CXorf57	1	0	0	5.45	-	-	-	-	-	-	-
CPSF7	1	0	0	5.05	-	-	-	-	-	-	-
AASS	1	0	0	-	5.01	-	-	-	-	-	-
C1QBP	1	0	0	-	5.83	-	-	-	-	-	-
CSNK2B	1	0	0	-	5.45	-	-	-	-	-	-
GPALPP1	1	0	0	-	7.47	-	-	-	-	-	-
GPATCH11	1	0	0	-	6.91	-	-	-	-	-	-
HNRNPF	1	0	0	-	5.21	-	-	-	-	-	-
HNRNPR	1	0	0	-	5.55	-	-	-	-	-	-
HSP90AB2P	1	0	0	-	6.89	-	-	-	-	-	-
MCM4	1	0	0	-	7.30	-	-	-	-	-	-
P4HA2	1	0	0	-	7.41	-	-	-	-	-	-
P4HB	1	0	0	-	7.39	-	-	-	-	-	-
PARP12	1	0	0	-	4.95	-	-	-	-	-	-
PINX1	1	0	0	-	5.67	-	-	-	-	-	-
PTBP3	1	0	0	-	5.16	-	-	-	-	-	-
PUF60	1	0	0	-	6.05	-	-	-	-	-	-
RBM47	1	0	0	-	5.43	-	-	-	-	-	-
RPL7	1	0	0	-	5.16	-	-	-	-	-	-
RPS17	1	0	0	-	6.89	-	-	-	-	-	-
SF3B2	1	0	0	-	5.93	-	-	-	-	-	-
SF3B5	1	0	0	-	5.92	-	-	-	-	-	-

SF3B6	1	0	0	-	5.34	-	-	-	-	-	-
SRRM2	1	0	0	-	5.98	-	-	-	-	-	-
TIAL1	1	0	0	-	6.26	-	-	-	5.17	-	-
SRSF11	1	0	0	-	5.95	-	-	-	-	-	-
TIA1	1	0	0	-	5.68	-	-	-	-	-	-
TUBB2A	1	0	0	-	5.96	-	-	-	-	-	-
TUBB4B	1	0	0	-	5.85	-	-	-	-	-	-
U2AF1	1	0	0	-	5.84	-	-	-	-	-	-
U2AF2	1	0	0	-	5.99	-	-	-	-	-	-
ZGPAT	1	0	0	-	5.02	-	-	-	-	-	-
ZRANB2	1	0	0	-	5.72	-	-	-	-	-	-
MRPS25	0	0	1	-	-	-	-	-	-	7.76	-
ERAL1	0	0	1	-	-	-	-	-	-	7.56	-
MRPS2	0	0	1	-	-	-	-	-	-	7.54	-
MRPS9	0	0	1	-	-	-	-	-	-	7.53	-
MRPS26	0	0	1	-	-	-	-	-	-	7.46	-
MRPS23	0	0	1	-	-	-	-	-	-	7.40	-
PTCD3	0	0	1	-	-	-	-	-	-	7.30	-
DAP3	0	0	1	-	-	-	-	-	-	7.29	-
MRPS18B	0	0	1	-	-	-	-	-	-	7.29	-
MRPS28	0	0	1	-	-	-	-	-	-	7.25	-
MRPS34	0	0	1	-	-	-	-	-	-	7.22	-
ZFP36L2	0	0	1	5.74	-	-	-	-	-	7.21	-
ZFP36L1	0	0	1	-	-	-	-	-	-	7.13	-
MRPS27	0	0	1	-	-	-	-	-	-	7.08	-
MRPS7	0	0	1	-	-	-	-	-	-	7.07	-
MRPS22	0	0	1	-	-	-	-	-	-	7.07	-
MRPS31	0	0	1	-	-	-	-	-	-	7.03	-
ZCCHC11	0	0	1	-	-	-	-	-	-	6.97	-
CSTF2T	0	0	1	-	-	-	-	-	-	6.45	-
ZNFX1	0	0	1	-	-	-	-	-	-	6.36	-
MEX3D	0	0	1	-	-	-	-	-	-	6.15	-
MEX3C	0	0	1	-	-	-	-	-	-	6.10	-
ATIC	0	0	1	-	-	-	-	-	-	6.09	-
TTC37	0	0	1	-	-	-	-	-	-	6.02	-
SKIV2L	0	0	1	-	-	-	-	-	-	5.71	-
CNBP	0	0	1	-	-	-	-	-	-	5.65	-
RBM38	0	0	1	-	-	-	-	-	-	5.54	-
TRIM56	0	0	1	-	-	-	-	-	-	5.46	-
MYEF2	0	0	1	-	-	-	-	-	-	5.42	-
DHX37	0	0	1	-	-	-	-	-	-	5.33	-
PRPF31	0	0	1	-	-	-	-	-	-	5.23	-
HNRNPM	0	0	1	-	-	-	-	-	-	5.17	-
PRPF4	0	0	1	-	-	-	-	-	-	5.17	-
PRPF3	0	0	1	-	-	-	-	-	-	5.15	-
HNRNPA2B1	0	0	1	-	-	-	-	-	-	5.14	-

FASTKD2	0	0	1	-	-	-	-	-	-	5.10	-
LSM8	0	0	1	-	-	-	-	-	-	5.02	-
GRSF1	0	0	1	-	-	-	-	-	-	4.74	-
MRPL17	0	1	0	-	-	-	-	7.28	-	-	-
MRPL22	0	1	0	-	-	-	-	7.01	-	-	-
MRPS18A	0	1	0	-	-	-	-	6.99	-	-	-
MRPL53	0	1	0	-	-	-	-	6.99	-	-	-
MRPL43	0	1	0	-	-	-	-	6.98	-	-	-
MRPL14	0	1	0	-	-	-	-	6.98	-	-	-
RPS3A	0	1	0	-	-	-	-	6.98	-	-	-
MRPL49	0	1	0	-	-	-	-	6.96	-	-	-
MRPL12	0	1	0	-	-	-	-	6.94	-	-	-
MRPL11	0	1	0	-	-	-	-	6.93	-	-	-
MRPL38	0	1	0	-	-	-	-	6.93	-	-	-
RPS13	0	1	0	-	-	-	-	6.91	-	-	-
ICT1	0	1	0	-	-	-	-	6.90	-	-	-
MRPL19	0	1	0	-	-	-	-	6.87	-	-	-
MRPL21	0	1	0	-	-	-	-	6.82	-	-	-
MRPL48	0	1	0	-	-	-	-	6.75	-	-	-
MRPL13	0	1	0	-	-	-	-	6.71	-	-	-
RPS28	0	1	0	-	-	-	-	6.67	-	-	-
RPS10	0	1	0	-	-	-	-	6.65	-	-	-
MRPL50	0	1	0	-	-	-	-	6.61	-	-	-
MRPL37	0	1	0	-	-	-	-	6.61	-	-	-
MRPL39	0	1	0	-	-	-	-	6.61	-	-	-
MRPL1	0	1	0	-	-	-	-	6.59	-	-	-
ANGEL2	0	1	0	-	-	-	-	6.30	-	-	-
SRSF4	0	1	0	-	-	-	-	6.01	-	-	-
GNB2L1	0	1	0	-	-	-	-	5.53	-	-	-
HNRNPK	0	1	0	-	-	-	-	5.52	-	-	-
PCBP1	0	1	0	-	-	-	-	5.47	-	-	-
PCBP2	0	1	0	-	-	-	-	5.47	-	-	-
IGF2BP3	0	1	0	-	-	-	-	5.42	-	-	-
LRPPRC	0	0	0	-	5.42	5.66	-	5.64	5.53	-	-
RPS20	0	0	0	5.22	5.62	-	-	6.77	6.97	-	-
XRN2	0	0	0	6.41	6.05	5.76	5.87	6.83	6.96	6.39	-
DHX15	0	0	0	5.93	5.88	5.19	-	6.12	5.73	6.24	6.45
CELF2	0	0	0	5.90	-	-	5.54	5.57	5.79	6.29	5.87
DIS3L2	0	0	0	5.66	6.08	-	5.79	5.93	6.01	6.42	-
DHX9	0	0	0	5.37	5.41	5.39	-	6.26	5.94	5.73	6.01
ZC3HAV1	0	0	0	5.35	5.32	5.56	5.15	5.88	5.82	5.27	-
DDX21	0	0	0	-	-	-	-	6.94	7.50	-	-
DHX30	0	0	0	-	-	-	-	6.95	7.22	-	-
HNRNPU	0	0	0	-	-	-	-	6.42	7.23	-	-
IGF2BP2	0	0	0	-	-	-	-	5.53	5.27	-	5.58
RPS16	0	0	0	-	6.12	-	-	6.76	-	-	-

NCL	0	0	0	-	-	-	-	6.49	6.65	5.66	6.13
PABPC1	0	0	0	-	-	5.00	5.86	6.70	6.71	-	-
RBM4	0	0	0	-	-	5.38	6.45	6.43	-	-	-
SRSF6	0	0	0	-	-	-	-	5.94	-	-	5.35
PHAX	0	0	0	-	-	4.97	-	5.87	-	5.04	5.78
RPS12	0	0	0	-	-	-	-	6.41	7.07	-	-
RPS14	0	0	0	-	-	-	-	6.74	7.02	-	-
RPS18	0	0	0	-	-	-	-	7.05	7.27	-	-
RPS19	0	0	0	-	-	-	-	6.75	7.00	-	-
RPS25	0	0	0	-	-	-	-	7.14	7.24	-	-
RPS3	0	0	0	-	-	-	-	6.51	7.01	-	-
RPS5	0	0	0	-	-	-	-	6.52	6.88	-	-
RPSA	0	0	0	-	-	-	-	6.31	6.98	-	-
SRSF3	0	0	0	-	-	5.52	-	5.58	5.94	-	-
IGF2BP1	0	0	0	-	-	-	-	5.43	-	-	6.55
SRSF7	0	0	0	-	-	-	-	5.52	5.93	-	-
YTHDC2	0	0	0	-	-	-	-	5.65	7.18	-	-
SKIV2L2	0	0	0	-	5.54	-	6.08	5.46	5.69	6.21	6.29
ZCCHC8	0	0	0	-	5.61	-	6.19	5.51	5.85	6.21	6.32
SLIRP	0	0	0	-	5.80	-	-	5.82	5.53	-	-
SRSF9	0	0	0	-	-	-	-	5.34	-	5.10	5.39
TSN	0	0	0	7.53	6.94	-	-	-	-	7.12	6.88
WDR33	0	0	0	7.50	7.10	6.97	6.52	-	-	6.02	-
CPSF4	0	0	0	7.48	7.24	7.04	6.96	-	-	-	-
CSTF2	0	0	0	7.32	7.03	6.88	-	-	-	6.33	-
LSM1	0	0	0	6.97	6.48	-	-	-	-	6.40	-
LSM2	0	0	0	6.51	6.49	-	-	-	-	6.11	-
LSM4	0	0	0	6.48	6.19	-	-	-	-	6.22	-
PIP5K1A	0	0	0	5.99	5.43	5.52	5.39	-	-	-	-
MEX3A	0	0	0	5.91	5.75	-	-	-	-	6.00	-
SYNCRIP	0	0	0	5.70	6.05	5.45	-	-	6.99	-	-
HNRNPDL	0	0	0	5.36	6.00	-	5.52	-	-	5.46	-
HNRNPAB	0	0	0	5.17	5.87	-	5.19	-	-	5.51	-
RBMS3	0	0	0	-	5.87	5.98	-	-	-	6.56	5.91
XRN1	0	0	0	-	5.99	6.72	-	-	-	6.71	-
RBMS2	0	0	0	-	6.02	5.76	-	-	6.41	6.26	-
PATL1	0	0	0	6.58	6.31	6.09	6.02	-	6.37	6.39	6.23
LSM6	0	0	0	6.33	6.43	5.54	-	-	6.36	6.35	5.74
LIN28B	0	0	0	5.91	5.18	5.52	5.39	5.64	5.59	5.83	-
TOE1	0	0	0	5.87	5.97	5.19	-	-	6.89	5.85	5.98
TSNAX	0	0	0	-	-	7.52	-	-	-	-	6.96
RTCA	0	0	0	6.76	6.20	5.78	5.54	5.71	5.89	6.42	6.56
LSM3	0	0	0	6.57	6.32	-	6.04	-	-	6.16	5.88
SSB	0	0	0	6.36	6.28	5.13	5.61	5.58	5.48	6.52	6.35
NKRF	0	0	0	6.29	5.86	5.64	5.63	6.26	6.34	6.42	7.01
CELF1	0	0	0	6.18	-	-	5.51	5.17	5.73	6.26	5.71

MSI2	0	0	0	6.10	6.06	-	6.21	-	5.77	6.10	6.27
HNRNPUL1	0	0	0	6.04	6.03	5.48	5.54	6.82	6.45	6.19	6.40
DAZAP1	0	0	0	5.96	5.86	-	5.51	-	5.84	5.91	5.53
FUBP3	0	0	0	5.94	5.39	-	5.45	-	5.81	6.10	5.95
RBMS1	0	0	0	5.87	5.99	5.55	5.81	-	5.82	6.34	6.10
DHX36	0	0	0	5.81	5.56	5.08	5.37	5.30	5.33	6.00	6.30
KHSRP	0	0	0	5.65	-	-	5.45	-	5.85	5.76	-
MSI1	0	0	0	5.46	-	-	-	-	-	6.31	5.94
FUBP1	0	0	0	5.24	-	-	5.23	-	-	-	-
HNRNPD	0	0	0	5.19	-	-	5.50	-	5.61	6.18	-
CDC5L	0	0	0	-	-	-	-	-	7.37	-	-
CSDE1	0	0	0	-	-	-	6.52	-	5.75	-	-
CWF19L1	0	0	0	-	-	-	-	-	5.45	5.50	-
DDX5	0	0	0	-	-	-	-	-	5.72	-	-
DHX57	0	0	0	-	-	-	-	-	7.15	-	-
DNAJC9	0	0	0	-	-	-	-	-	7.23	-	-
GNL3	0	0	0	-	-	-	-	-	7.45	-	-
HARS2	0	0	0	-	-	-	-	-	-	-	5.59
HIST1H1C	0	0	0	-	-	-	-	-	7.34	-	-
HNRNPA0	0	0	0	-	-	-	5.04	-	-	5.31	-
HNRNPA1	0	0	0	-	-	-	-	-	-	5.53	5.56
HNRNPA3	0	0	0	-	-	-	-	-	-	-	5.51
HNRNPC	0	0	0	-	-	-	-	-	7.71	-	6.87
ILF2	0	0	0	-	-	-	-	-	7.05	-	-
ILF3	0	0	0	-	-	-	-	-	7.15	-	-
KRR1	0	0	0	-	-	-	-	-	6.84	-	-
MCM2	0	0	0	-	-	-	-	-	8.21	-	-
PDCD11	0	0	0	-	-	-	-	-	7.70	-	-
PHF5A	0	0	0	-	-	-	-	-	-	-	6.57
PLOD1	0	0	0	-	-	-	-	-	-	5.17	5.44
POLRMT	0	0	0	-	-	-	-	-	7.38	-	-
PPIE	0	0	0	-	-	-	-	-	5.94	6.52	-
PUM2	0	0	0	-	-	-	-	-	6.11	-	-
PURA	0	0	0	-	-	-	-	-	7.01	-	-
RBM11	0	0	0	-	-	-	-	-	-	-	6.84
RBM14	0	0	0	-	-	5.70	-	-	5.65	-	-
RBM42	0	0	0	-	-	-	-	-	6.96	-	-
RBM45	0	0	0	-	-	-	6.21	-	-	-	-
RPL17	0	0	0	-	-	-	-	-	7.14	-	-
RPL22	0	0	0	-	-	-	-	-	7.08	-	-
RPL23A	0	0	0	-	-	-	-	-	7.12	-	-
RPL27A	0	0	0	-	-	-	-	-	7.37	-	-
RPL3	0	0	0	-	-	-	-	-	7.44	-	-
RPS15A	0	0	0	-	-	-	-	-	7.03	-	-
RPS21	0	0	0	-	-	-	-	-	6.60	-	-
RPS29	0	0	0	-	-	-	-	-	6.91	-	-

RPS4X	0	0	0	-	-	-	-	-	7.28	-	-
RPS7	0	0	0	-	-	-	-	-	-	-	5.55
SETX	0	0	0	-	-	-	-	-	6.99	-	-
SF3B1	0	0	0	-	-	-	-	-	6.86	-	-
SNRPA1	0	0	0	-	-	-	-	-	6.14	-	-
SNRPD2	0	0	0	-	-	-	-	-	5.29	-	-
STAU1	0	0	0	-	-	-	-	-	7.20	-	-
SUPT16H	0	0	0	-	-	-	-	-	8.82	-	-
TARDBP	0	0	0	-	-	-	-	-	-	5.55	5.81
YBX3	0	0	0	-	-	-	-	-	7.39	-	-
ZCRB1	0	0	0	-	-	-	-	-	-	5.42	5.84
ELAVL2	0	0	0	-	5.01	-	-	-	-	5.42	-
ELAVL1	0	0	0	-	5.30	-	-	-	-	5.52	-
UPF1	0	0	0	-	5.31	-	-	-	5.55	-	-
SF3B3	0	0	0	-	5.31	-	-	-	6.78	-	-
PABPC4	0	0	0	-	5.37	-	-	-	6.93	4.87	-
RBM17	0	0	0	-	5.60	-	-	-	-	7.03	7.01
TFIP11	0	0	0	-	5.63	-	-	-	-	5.62	-
RBM7	0	0	0	-	5.71	-	6.26	-	5.50	6.51	6.33
MEX3B	0	0	0	-	5.78	-	-	-	-	6.14	-
XAB2	0	0	0	-	5.86	-	-	-	6.27	6.22	6.40

Appendix Table 3. List of variant peptides identified in the ultra-deep tonsil proteome

	Chromosome variant	Chromosome variant type	Protein	Protein variant	Modified Sequence	Protease cleave AA	Identification method	Exact blast hitted genes	Identified by MaxQuant	Manually confirmed	Spectral contrast angle confirmed	All confirmed
1	chr20:2309687_C>A	Homozygous	ENSP00000370867	T13K	_(ac)AALGVQSINWQKAFNR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
2	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
3	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
4	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
5	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
6	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
7	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
8	chr9:113373918_C>T	Heterozygous	ENSP00000238379	G146E	_LEGILEGLGLR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
9	chr12:120516687_C>T	Heterozygous	ENSP00000288532	A152T	_AQQNLSWEEITKEYQNEEDSLGGSR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
10	chr1:207580276_A>G	Heterozygous	ENSP00000356016	H1658R	_TPSHQDNFSPGQEVFYSCEPGYDLR_	Arg-C_HCD (R Cterm)	novel cleavage	CR1	Yes	Yes	Yes	Yes
11	chr11:237087_A>G	Heterozygous	ENSP00000333811	N13S	_MKDVPGFLQQSQSSGPGQPAVWHR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
12	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
13	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
14	chr12:121888906_T>C	Homozygous	ENSP00000261817	V17A	_QSGGSSQAGAVTVSDVQELM(ox)R_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
15	chr19:17219251_T>C	Homozygous	ENSP00000263897	L154S	_TGVAGSQPVSEKQSAAELDLVLQR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
16	chr9:113373918_C>T	Heterozygous	ENSP00000238379	G146E	_LEGILEGLGLR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
17	chr7:44107886_C>A	Heterozygous	ENSP00000223357	P273T	_VWPETPEEKAPAPAPEER_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
18	chr1:207580276_A>G	Heterozygous	ENSP00000356016	H1658R	_TPSHQDNFSPGQEVFYSCEPGYDLR_	Arg-C_HCD (R Cterm)	novel cleavage	CR1	Yes	Yes	Yes	Yes
19	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
20	chr3:13354079_C>A	Heterozygous	ENSP00000254508	R786L	_NPLLDLAAYDQEGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

21	chr2:219182118_G>A	Heterozygous	ENSP00000395249	R374H	_DLGEGEEGELAPPEDLLGHPQALSR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
22	chr1:179907853_C>G	Homozygous	ENSP00000271583	P277R	_SVLSSGYQKTPQEWAPQTAR_	Arg-C_HCD (R Cterm)	novel cleavage	TOR1AIP1	Yes	Yes	Yes	Yes
23	chr1:179907853_C>G	Homozygous	ENSP00000271583	P277R	_SVLSSGYQKTPQEWAPQTAR_	Arg-C_HCD (R Cterm)	novel cleavage	TOR1AIP1	Yes	Yes	-	Yes
24	chr1:145873487_G>A	Heterozygous	ENSP00000347802	P428S	_SQPEEQGPSQSPASETIR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
25	chr3:13354079_C>A	Heterozygous	ENSP00000254508	R786L	_NPLLDLAAYDQEGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
26	chr3:45035631_G>A	Heterozygous	ENSP00000296130	G106S	_GGTLSTPQTGSENDALYEYLR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
27	chr19:1068739_G>A	Heterozygous	ENSP00000316772	R139H	_FAEGLEKLKECVLHDDLLEAR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
28	chr19:1068739_G>A	Heterozygous	ENSP00000316772	R139H	_FAEGLEKLKECVLHDDLLEAR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
29	chr11:119081463_A>G	Heterozygous	ENSP00000479680	K345R	_DLHDQFQHQLR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
30	chr17:77498623_A>G	Homozygous	ENSP00000329161	M558V	_LNEGSSAM(ox)ANGVEEKEPEAPEM(ox)_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
31	chr15:74035800_G>T	Homozygous	ENSP00000268059	G780V	_AQTLGAVVPPGDSVR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
32	chr5:112209973_C>G	Homozygous	ENSP00000261486	S366T	_IAQTQPAESNTISR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
33	chr10:119670121_T>C	Heterozygous	ENSP00000358081	C151R	_GQVAAAAAAQPPASHGPER_	Arg-C_HCD (R Cterm)	novel cleavage	BAG3	Yes	Yes	Yes	Yes
34	chr1:179882939_T>C	Homozygous	ENSP00000271583	M146T	_LQQQHSEQPPLQPSPVTTR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
35	chr22:30292499_T>C	Heterozygous	ENSP00000215790	H468R	_GQLEKPPAPNQAMVVAAAGDACPPQR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
36	chr9:113373918_C>T	Heterozygous	ENSP00000238379	G146E	_RLEGILEGLGLR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
37	chr5:80437260_T>C	Homozygous	ENSP00000337159	I192T	_EQQNDTSSELQNR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
38	chr19:49451550_T>G	Homozygous	ENSP00000262265	M9L	_(ac)ANPKLLGLELSEAEAIGADSAR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
39	chr10:124484294_A>G	Homozygous	ENSP00000357832	Q94R	_LGFDISEQEVTAPAPAACQILKER_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
40	chr2:171554763_G>A	Heterozygous	ENSP00000319141	S266N	_GSMPAYSGNNMDKSDSELNNEVAAR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
41	chr8:144106920_G>A	Homozygous	ENSP00000318604	G236R	_GSGAEETSTMEEDR_	Arg-C_HCD (R Cterm)	novel cleavage	MAF1	Yes	Yes	Yes	Yes
42	chr8:144106920_G>A	Homozygous	ENSP00000318604	G236R	_GSGAEETSTM(ox)EEDR_	Arg-C_HCD (R Cterm)	novel cleavage	MAF1	Yes	Yes	-	Yes
43	chr17:77498623_A>G	Homozygous	ENSP00000329161	M558V	_LNEGSSAMANGVEEKEPEAPEM_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
44	chr17:77498623_A>G	Homozygous	ENSP00000329161	M558V	_LNEGSSAM(ox)ANGVEEKEPEAPEM(ox)_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
45	chr17:77498623_A>G	Homozygous	ENSP00000329161	M558V	_LNEGSSAM(ox)ANGVEEKEPEAPEM(ox)_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
46	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Arg-C_HCD (R Cterm)	novel cleavage	MYO1C	Yes	Yes	Yes	Yes
47	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Arg-C_HCD (R Cterm)	novel cleavage	MYO1C	Yes	Yes	Yes	Yes

48	chr10:119670121_T>C	Heterozygous	ENSP00000358081	C151R	_GQVAAAAAAQPPASHGPER_	Arg-C_HCD (R Cterm)	novel cleavage	BAG3	Yes	Yes	Yes	Yes
49	chr10:128119296_T>C	Heterozygous	ENSP00000357642	N104S	_YENESLQSGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
50	chr9:113373918_C>T	Heterozygous	ENSP00000238379	G146E	_RLEGILEGLGLR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
51	chr22:19895461_T>G	Heterozygous	ENSP00000383362	S269R	_TTGKEDTGTFDTVLWAIGR_	Arg-C_HCD (R Cterm)	novel cleavage	TXNRD2	Yes	Yes	Yes	Yes
52	chr22:19895461_T>G	Heterozygous	ENSP00000383362	S269R	_TTGKEDTGTFDTVLWAIGR_	Arg-C_HCD (R Cterm)	novel cleavage	TXNRD2	Yes	Yes	-	Yes
53	chr19:49451550_T>G	Homozygous	ENSP00000262265	M9L	_(ac)ANPKLLGLELSEAEAIGADSAR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
54	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
55	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
56	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
57	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
58	chr9:113373918_C>T	Heterozygous	ENSP00000238379	G146E	_RLEGILEGLGLR_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
59	chr14:24432070_G>T	Homozygous	ENSP00000251343	W270L	_EMDLGWKELPGEEAWER_	Arg-C_HCD (R Cterm)	direct sequence	-	-	Yes	-	Yes
60	chr14:24432070_G>T	Homozygous	ENSP00000251343	W270L	_EM(ox)DLGWKELPGEEAWER_	Arg-C_HCD (R Cterm)	direct sequence	-	Yes	Yes	-	Yes
61	chr17:41755893_T>A	Homozygous	ENSP00000311113	M697L	_DDLDATYRPMYSS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
62	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DEVKEQTFSGGTSQ_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
63	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_DAIWNLLHQAQEKFGK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
64	chr14:92810309_T>C	Heterozygous	ENSP00000163416	F350L	_DQSEGNSLQNQALQTLQERLHEA_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
65	chr17:41755893_T>A	Homozygous	ENSP00000311113	M697L	_DDLDATYRPMYSS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	-	Yes
66	chr17:41755893_T>A	Homozygous	ENSP00000311113	M697L	_DDLDATYRPMYSS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
67	chr1:186406173_A>G	Heterozygous	ENSP00000287859	D364G	_DYKFDGESAEEIR_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
68	chr16:28497395_T>C	Homozygous	ENSP00000416094	V776A	_DVVPHISAAGAGEALEGALGQGW_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
69	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DEVKEQTFSGGTSQ_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
70	chr1:114405659_A>G	Homozygous	ENSP00000351250	I840T	_DALASLENHVKTEPA_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
71	chr3:149198448_T>A	Homozygous	ENSP00000264613	E544D	_DPVCLAKMYYSAVDPTK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	-	Yes	-	Yes
72	chrX:47203135_G>A	Homozygous	ENSP00000338413	R447H	_DALECLPEDKEVLTEDKCLQHQNRY_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
73	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DEVKEQTFSGGTSQ_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
74	chr2:241253433_T>C	Heterozygous	ENSP00000312042	N418S	_DVSVAQEQIEGM(ox)VK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes

75	chr1:114405659_A>G	Homozygous	ENSP00000351250	I840T	_DALASLENHVKTEPA_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
76	chr3:149198448_T>A	Homozygous	ENSP00000264613	E544D	_DPVCLAKMYYSAVDPTK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	-	Yes	-	Yes
77	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_DTPEAGYFAVAVVKKSAS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
78	chr11:61008834_C>T	Homozygous	ENSP00000323280	A257V	_DAGVVCSEHQSWRLTGGA_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	-	Yes
79	chr19:34452021_G>T	Heterozygous	ENSP00000246548	Q304H	_DQQNEPHLGLK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
80	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_DTPEAGYFAVAVVKKSAS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
81	chr5:140668682_T>C	Homozygous	ENSP00000351100	C151R	_DDTGGIRLW_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
82	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_DAIWNLLHQAQEKFGK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	-	Yes
83	chr1:236537557_C>T	Heterozygous	ENSP00000238181	R36C	_DQLDPGTLIVICGHVPSDA_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
84	chr7:77618504_G>A	Heterozygous	ENSP00000248594	V322I	_DGVNEINTENMISSIEPEKQ_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	-	Yes
85	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_DTPEAGYFAVAVVKKSAS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
86	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_DTPEAGYFAVAVVKKSAS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
87	chr19:6468320_T>C	Homozygous	ENSP00000370889	M569V	_DSLSVGAKSAGSLRPSQSL_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	-	Yes
88	chr5:140668682_T>C	Homozygous	ENSP00000351100	C151R	_DDTGGIRLW_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
89	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_DAIWNLLHQAQEKFGK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	-	Yes
90	chr6:152219143_A>C	Homozygous	ENSP00000341887	F6897V	_DSLFVLHELGEQLKQQV_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	-	Yes
91	chr6:25604863_G>A	Homozygous	ENSP00000331983	G1202S	_DSSSPALSSVERS_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
92	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_DAIWNLLHQAQEKFGK_	Asp-N_HCD (D,B Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
93	chr10:133370622_G>A	Homozygous	ENSP00000357535	T75I	_KIFEEDPAVGAIVLTGGDKAF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
94	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_TAEFEEAQTSACLLQEELEKL_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	-	Yes
95	chr15:74035775_A>G	Homozygous	ENSP00000268059	S772G	_DMSSVVGAGEGRAQTL_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
96	chr13:46685919_T>C	Homozygous	ENSP00000308493	S234P	_KVLPQELVDLPLVKF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
97	chr1:89382706_G>A	Heterozygous	ENSP00000359485	E399K	_NKKGDFLLQNKESSVQY_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
98	chr12:10722889_T>C	Homozygous	ENSP00000228251	T75A	_AAAAGSEDAEKKVL_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	-	Yes
99	chr1:236558360_T>C	Heterozygous	ENSP00000355540	N1613S	_GAENPDPFVPVLSTAVKL_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	-	Yes
100	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_GIGEEEHDQEGRVIVAEF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
101	chr12:10722889_T>C	Homozygous	ENSP00000228251	T75A	_AAAAGSEDAEKKVL_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	-	Yes

102	chr12:8867893_A>G	Homozygous	ENSP00000299698	M1257V	_ATTAYVPSEEINLVVKSTENFQRTF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
103	chr7:134541075_A>G	Homozygous	ENSP00000352584	N313D	_RACNVLQSSHLEDYPFDAEY_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
104	chr12:8867893_A>G	Homozygous	ENSP00000299698	M1257V	_ATTAYVPSEEINLVVKSTENFQRTF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
105	chrX:47203135_G>A	Homozygous	ENSP00000338413	R447H	_DALECLPEDKEVLTEDKCLQHQNRY_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
106	chr7:134541075_A>G	Homozygous	ENSP00000352584	N313D	_RACNVLQSSHLEDYPFDAEY_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
107	chr17:41514660_G>C	Homozygous	ENSP00000254043	A421G	_AGIGIREASSGGGGSSSNF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	-	Yes	-	Yes
108	chr17:41514660_G>C	Homozygous	ENSP00000254043	A421G	_AGIGIREASSGGGGSSSNF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	-	Yes	-	Yes
109	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_KAVENINSTLGPALL_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
110	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_TAEFEEAQTSACLLQEELEKL_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
111	chr3:122540759_T>C	Heterozygous	ENSP00000353512	Y528C	_GFNVEEM(ox)CEAHAW_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	-	Yes
112	chr6:38682852_T>G	Heterozygous	ENSP00000362463	E111A	_ELTHNWGTEDDATQSY_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	-	Yes
113	chr6:38682852_T>G	Heterozygous	ENSP00000362463	E111A	_ELTHNWGTEDDATQSY_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
114	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_SHEGTFEAIQLDDEHIDSLSSF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
115	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_SHEGTFEAIQLDDEHIDSLSSF_	Chymotrypsin_HCD (F,L,M,W,Y, Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
116	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
117	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
118	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
119	chr12:52848656_A>G	Heterozygous	ENSP00000306261	L92P	_VTINQNPLTPLKIE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
120	chr19:35513204_A>G	Homozygous	ENSP00000342012	V91A	_AVGTGVRQVPGFGAADALGNRVGE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	-	Yes	-	Yes
121	chr21:29067014_C>T	Heterozygous	ENSP00000286788	V147I	_ILPNLICCSAKNLRDIDE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
122	chr11:1881538_G>A	Homozygous	ENSP00000308383	A100T	_GAQGTLDSGEPPQCRSPE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
123	chr11:1881538_G>A	Homozygous	ENSP00000308383	A100T	_GAQGTLDSGEPPQCRSPE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
124	chr21:29067014_C>T	Heterozygous	ENSP00000286788	V147I	_ILPNLICCSAKNLRDIDE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
125	chr11:1881538_G>A	Homozygous	ENSP00000308383	A100T	_GAQGTLDSGEPPQCRSPE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
126	chr11:1881538_G>A	Homozygous	ENSP00000308383	A100T	_GAQGTLDSGEPPQCRSPE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
127	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
128	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

129	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
130	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
131	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
132	chr21:46125854_G>A	Heterozygous	ENSP00000300527	R680H	_AIQLDDEHIDSLSSFKE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	-	Yes
133	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_NINSTLGPALLQKKLSVADQE_	V8-DE_HCD (B,D,E,Z Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
134	chr2:190296896_T>C	Homozygous	ENSP00000352706	T46A	_GCAGVITLNRPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
135	chr19:5957950_C>T	Heterozygous	ENSP00000034275	V16I	_PAIAPPIFVFQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
136	chr10:29495143_G>C	Heterozygous	ENSP00000348128	P1235A	_GLASPTAITPVASAICGK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
137	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
138	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
139	chr10:95263873_T>C	Heterozygous	ENSP00000360305	N175S	_TAASGVEASSRPLDHAQPPSSLVIDK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
140	chr10:95263873_T>C	Heterozygous	ENSP00000360305	N175S	_TAASGVEASSRPLDHAQPPSSLVIDK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
141	chr10:22550699_T>C	Heterozygous	ENSP00000326294	N111S	_IYIDDNSK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
142	chr8:27451068_A>C	Homozygous	ENSP00000332816	K838T	_SLDPMVYMNDTSPLTPEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
143	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
144	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_LTAEFEEAQTSACLLQEELEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
145	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_LTAEFEEAQTSACLLQEELEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
146	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_LTAEFEEAQTSACLLQEELEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
147	chr19:5957950_C>T	Heterozygous	ENSP00000034275	V16I	_PAIAPPIFVFQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
148	chr19:5957950_C>T	Heterozygous	ENSP00000034275	V16I	_PAIAPPIFVFQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
149	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
150	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
151	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
152	chr4:94657437_G>A	Homozygous	ENSP00000321746	S492N	_ILGEVINALK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
153	chr4:94657437_G>A	Homozygous	ENSP00000321746	S492N	_ILGEVINALK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
154	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GMGTVEGADQSNPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
155	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GM(ox)GTVEGADQSNPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes

156	chr8:141136609_T>C	Homozygous	ENSP00000428714	M68T	_EDSQTAGANCGTLGK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
157	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQM(ox)EAQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
158	chr19:10631494_A>G	Homozygous	ENSP00000336888	Q154R	_GVAEVLRDGDCPAVLIPSK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
159	chr10:95263873_T>C	Heterozygous	ENSP00000360305	N175S	_TAASGVEASSRPLDHAQPPSSLVIDK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
160	chr10:95263873_T>C	Heterozygous	ENSP00000360305	N175S	_TAASGVEASSRPLDHAQPPSSLVIDK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
161	chr4:67668674_C>T	Heterozygous	ENSP00000313454	A224T	_EIFISNITQTNPGIVTCLENHPHK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
162	chr11:67490352_C>A	Homozygous	ENSP00000279146	Q228K	_EQPGSPEWIQLDK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
163	chr11:67490352_C>A	Homozygous	ENSP00000279146	Q228K	_EQPGSPEWIQLDK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
164	chr2:20624738_C>T	Homozygous	ENSP00000305193	V260M	_LSPQDPSEDVSSM(ox)DPLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
165	chr2:68822807_G>T	Heterozygous	ENSP00000386241	R549S	_NSGEEEIDSLQSMVQELRK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
166	chr2:68822807_G>T	Heterozygous	ENSP00000386241	R549S	_NSGEEEIDSLQSMVQELRK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
167	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_SDNCEDTPEAGYFAVAVVKK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
168	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_SDNCEDTPEAGYFAVAVVKK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
169	chr17:67215926_C>T	Heterozygous	ENSP00000351524	V74M	_NYMQADEDCRHVLGEGLAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
170	chr22:46281710_T>C	Homozygous	ENSP00000370419	F243L	_DGLEFMQHSETLWK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
171	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
172	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_LTAEFEEAQTSACLLQEELEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
173	chr7:5987144_T>C	Homozygous	ENSP00000265849	K541E	_APETDDSFSDVDCHSNQEDTGCK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
174	chr3:150562658_C>G	Homozygous	ENSP00000273435	T92S	_NTVLATWQPYSTSK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
175	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
176	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GMGTVEGADQSNPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
177	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GM(ox)GTVEGADQSNPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
178	chr6:52264686_C>T	Heterozygous	ENSP00000229854	E777K	_EPFSSVEIQAALSK_	Lys-C/P_HCD (K Cterm)	novel cleavage	MCM3	Yes	Yes	-	Yes
179	chr8:141136609_T>C	Homozygous	ENSP00000428714	M68T	_EDSQTAGANCGTLGK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
180	chr8:141136609_T>C	Homozygous	ENSP00000428714	M68T	_EDSQTAGANCGTLGK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
181	chr12:132777676_C>T	Heterozygous	ENSP00000204726	D1238N	_LQAEANDLQIREGK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
182	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQMEAQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

183	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQM(ox)EAQVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
184	chr1:63631761_C>T	Heterozygous	ENSP00000360124	R239C	_ICIDAMHGVVGPYVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
185	chr7:95411704_G>C	Heterozygous	ENSP00000222572	A148G	_FEEGENSLLHLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
186	chr6:152219143_A>C	Homozygous	ENSP00000341887	F6897V	_DSLFVLHELGEQLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
187	chr6:152219143_A>C	Homozygous	ENSP00000341887	F6897V	_DSLFVLHELGEQLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
188	chr6:151353191_A>C	Homozygous	ENSP00000253332	E1600D	_EGEDPQASAQDETPITSAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
189	chr6:151353191_A>C	Homozygous	ENSP00000253332	E1600D	_EGEDPQASAQDETPITSAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
190	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
191	chr12:8852296_C>A	Homozygous	ENSP00000299698	D850E	_SHEYQLESWADSQTSSCLCADEAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
192	chr3:150562658_C>G	Homozygous	ENSP00000273435	T92S	_NTVLATWQPYSTSK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
193	chr6:52264686_C>T	Heterozygous	ENSP00000229854	E777K	_EPFSSVEIQAALSK_	Lys-C/P_HCD (K Cterm)	novel cleavage	MCM3	Yes	Yes	-	Yes
194	chr7:95411704_G>C	Heterozygous	ENSP00000222572	A148G	_FEEGENSLLHLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
195	chr10:87745954_C>T	Heterozygous	ENSP00000354436	S610F	_VLTDYYRFLEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
196	chr21:39499884_T>C	Heterozygous	ENSP00000332513	V188A	_EGSEDAGNLPEAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
197	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
198	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
199	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
200	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
201	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
202	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
203	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
204	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
205	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
206	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
207	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
208	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
209	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes

210	chr9:69251072_C>T	Heterozygous	ENSP00000366453	S1010F	_EESYDFFK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
211	chr5:76627482_A>G	Homozygous	ENSP00000274364	K532E	_VLWLDEIQQAVDEANVDEDRAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
212	chr5:76627482_A>G	Homozygous	ENSP00000274364	K532E	_VLWLDEIQQAVDEANVDEDRAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
213	chr5:76627482_A>G	Homozygous	ENSP00000274364	K532E	_VLWLDEIQQAVDEANVDEDRAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
214	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DSRPSQAAGDNQGDEVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
215	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DSRPSQAAGDNQGDEVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
216	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DSRPSQAAGDNQGDEVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
217	chr12:27647813_G>T	Homozygous	ENSP00000228425	V148L	_LNATEEMLQQELLSRTSLETQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
218	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
219	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
220	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
221	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
222	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
223	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	-	Yes	-	Yes
224	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	-	Yes	-	Yes
225	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_SDNCEDTPEAGYFAVAVVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
226	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_SDNCEDTPEAGYFAVAVVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
227	chr4:94252628_T>C	Homozygous	ENSP00000346217	V301A	_VFAEDQDMQYASQSEVPNGK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
228	chr4:94252628_T>C	Homozygous	ENSP00000346217	V301A	_VFAEDQDMQYASQSEVPNGK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
229	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
230	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
231	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
232	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
233	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
234	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
235	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
236	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

237	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
238	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
239	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
240	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
241	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
242	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
243	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
244	chr2:203251967_C>T	Homozygous	ENSP00000348380	S97L	_TLDPFETMLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
245	chr2:203251967_C>T	Homozygous	ENSP00000348380	S97L	_TLDPFETMLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
246	chrX:123900661_A>C	Homozygous	ENSP00000347858	Q423P	_DSMPDESSQTSLQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
247	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DSRPSQAAGDNQGDEVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
248	chr12:106239791_C>T	Heterozygous	ENSP00000367265	A348T	_EAADTERLTLQALTEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
249	chr19:11447460_G>A	Heterozygous	ENSP00000465461	A291T	_YRSEALPTDLPTPSAPDLTEPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
250	chr2:37228889_C>T	Homozygous	ENSP00000234170	V102I	_ASLIEEDEPAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
251	chr2:37228889_C>T	Homozygous	ENSP00000234170	V102I	_ASLIEEDEPAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
252	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
253	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
254	chr5:160408651_A>G	Homozygous	ENSP00000297151	M229T	_HQWGEEEPNSQTEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
255	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	-	Yes	-	Yes
256	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_SDNCEDTPEAGYFAVAVVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
257	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_SDNCEDTPEAGYFAVAVVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
258	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
259	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_VNDDIIVNWVNETLREAEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
260	chr7:857859_G>A	Heterozygous	ENSP00000374225	E393K	_TISAVGEQLLPTVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
261	chr7:857859_G>A	Heterozygous	ENSP00000374225	E393K	_TISAVGEQLLPTVK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
262	chr21:29067014_C>T	Heterozygous	ENSP00000286788	V147I	_AHEILPNLICCSAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
263	chr21:29067014_C>T	Heterozygous	ENSP00000286788	V147I	_AHEILPNLICCSAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

264	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
265	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
266	chr12:25089661_C>G	Homozygous	ENSP00000346442	L141V	_SVNVRQSENTSANEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
267	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
268	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
269	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
270	chr9:94606867_C>T	Homozygous	ENSP00000364475	R218K	_DFDPAVTEYIQRK_	Lys-C/P_HCD (K Cterm)	novel cleavage	FBP1	Yes	Yes	Yes	Yes
271	chr9:94606867_C>T	Homozygous	ENSP00000364475	R218K	_DFDPAVTEYIQRK_	Lys-C/P_HCD (K Cterm)	novel cleavage	FBP1	Yes	Yes	Yes	Yes
272	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
273	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
274	chr5:160408651_A>G	Homozygous	ENSP00000297151	M229T	_HQWGEEEPNSQTEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
275	chr20:17619712_G>A	Heterozygous	ENSP00000246043	S1199L	_GELESSDQVREHTLHLEAELEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
276	chr8:120285917_G>C	Heterozygous	ENSP00000247781	V303L	_TLTYFNYDQSGDFQTLTFEGPEIRK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
277	chr9:111597344_C>A	Homozygous	ENSP00000311572	A27S	_TSELPPLK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
278	chr21:29067014_C>T	Heterozygous	ENSP00000286788	V147I	_AHEILPNLICCSAK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
279	chr3:149198448_T>A	Homozygous	ENSP00000264613	E544D	_MYYSAVDPTK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	-	Yes	-	Yes
280	chr3:149198448_T>A	Homozygous	ENSP00000264613	E544D	_MYYSAVDPTK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	-	Yes	-	Yes
281	chr13:45967538_T>A	Homozygous	ENSP00000242848	E1429D	_DLGSVQGFEDTNK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
282	chr9:94606867_C>T	Homozygous	ENSP00000364475	R218K	_DFDPAVTEYIQRK_	Lys-C/P_HCD (K Cterm)	novel cleavage	FBP1	Yes	Yes	Yes	Yes
283	chr2:171976274_T>C	Homozygous	ENSP00000264108	V314A	_LMQGFNEDMAIEAQQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
284	chr2:171976274_T>C	Homozygous	ENSP00000264108	V314A	_LMQGFNEDMAIEAQQK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	-	Yes
285	chr6:167022440_C>T	Heterozygous	ENSP00000230248	T184I	_ANDEANQSDISVSLSEPK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
286	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
287	chr20:17619712_G>A	Heterozygous	ENSP00000246043	S1199L	_GELESSDQVREHTLHLEAELEK_	Lys-C/P_HCD (K Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
288	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_KSDNCEDTPEAGYFAVAVVK_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
289	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_KSDNCEDTPEAGYFAVAVVK_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
290	chr12:8867893_A>G	Homozygous	ENSP00000299698	M1257V	_KYATTAYVPSEEINLVV_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes

291	chr8:27451068_A>C	Homozygous	ENSP00000332816	K838T	_KSLDPMVYMNDTSPLTPE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
292	chr2:68822807_G>T	Heterozygous	ENSP00000386241	R549S	_KNSGEEEIDSLQSMVQELR_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
293	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_KSDNCEDTPEAGYFAVAVVK_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
294	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_KSDNCEDTPEAGYFAVAVVK_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
295	chr2:68822807_G>T	Heterozygous	ENSP00000386241	R549S	_KNSGEEEIDSLQSM(ox)VQELR_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
296	chr2:171554763_G>A	Heterozygous	ENSP00000319141	S266N	_KSDSELNNEVAAR_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
297	chr19:1077986_A>G	Heterozygous	ENSP00000316772	S439G	_KAEEEQAGSAPGAGGTAT_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
298	chr6:151353191_A>C	Homozygous	ENSP00000253332	E1600D	_KEGEDPQASAQDETPITSA_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
299	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
300	chr5:76627482_A>G	Homozygous	ENSP00000274364	K532E	_KVLWLDEIQQAVDEANVDEDRA_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
301	chr19:1077986_A>G	Heterozygous	ENSP00000316772	S439G	_KAEEEQAGSAPGAGGTAT_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
302	chr19:1077986_A>G	Heterozygous	ENSP00000316772	S439G	_KAEEEQAGSAPGAGGTAT_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
303	chr7:128754552_C>T	Heterozygous	ENSP00000408838	A82V	_KIDVDKDGFVTEGEL_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
304	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
305	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
306	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
307	chr5:76627482_A>G	Homozygous	ENSP00000274364	K532E	_KVLWLDEIQQAVDEANVDEDRA_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
308	chr2:27201045_G>A	Heterozygous	ENSP00000310208	L573F	_KLLSLFPLSCQ_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
309	chr1:63631761_C>T	Heterozygous	ENSP00000360124	R239C	_KICIDAMHGVVGPYVK_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
310	chr19:1077986_A>G	Heterozygous	ENSP00000316772	S439G	_KAEEEQAGSAPGAGGTAT_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
311	chr19:1077986_A>G	Heterozygous	ENSP00000316772	S439G	_KAEEEQAGSAPGAGGTAT_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
312	chr19:1077986_A>G	Heterozygous	ENSP00000316772	S439G	_KAEEEQAGSAPGAGGTAT_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
313	chr7:128754552_C>T	Heterozygous	ENSP00000408838	A82V	_KIDVDKDGFVTEGEL_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
314	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_KEDAIWNLLHQAQE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
315	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_KEDAIWNLLHQAQE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
316	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
317	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes

318	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
319	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
320	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
321	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
322	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_KEDAIWNLLHQAQE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
323	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
324	chr13:46134156_T>C	Homozygous	ENSP00000315757	K533E	_KVNDDIIVNWVNETLREAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
325	chr17:34961863_T>G	Homozygous	ENSP00000354518	H99Q	_KEASQGSSASSAPQSV_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
326	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_KAVENINSTLGPALLQ_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
327	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_KAVENINSTLGPALLQ_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
328	chr9:94606867_C>T	Homozygous	ENSP00000364475	R218K	_KDFDPAVTEYIQR_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
329	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_KGFGTDEQAIIDCLGSCSN_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
330	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_KGFGTDEQAIIDCLGSCSN_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
331	chr16:21179532_A>C	Heterozygous	ENSP00000233047	E154D	_KSADFEGLYQE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
332	chr11:62525410_G>T	Homozygous	ENSP00000367263	Q3003K	_KVDIDVPDVGVQGPDWHL_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
333	chr3:133766289_A>G	Homozygous	ENSP00000385834	I448V	_KSDNCEDTPEAGYFAVAVVK_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
334	chrX:77682471_C>G	Homozygous	ENSP00000362441	E929Q	_KEQSFTSLEVR_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
335	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_KAVENINSTLGPALLQ_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
336	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_KGFGTDEQAIIDCLGSCSN_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
337	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_KGFGTDEQAIIDCLGSCSN_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
338	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_KGFGTDEQAIIDCLGSCSN_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	-	Yes
339	chr2:37228889_C>T	Homozygous	ENSP00000234170	V102I	_KASLIEEDEPAE_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
340	chr11:62525410_G>T	Homozygous	ENSP00000367263	Q3003K	_KVDIDVPDVGVQGPDWHL_	Lys-N_HCD (K Nterm)	direct sequence	-	Yes	Yes	Yes	Yes
341	chr4:48842653_A>G	Heterozygous	ENSP00000264312	T53A	_SVPLAAASM(ox)LITQGLISK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
342	chr19:35513269_C>A	Heterozygous	ENSP00000342012	E69D	_VSDALGQGTR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
343	chr9:115084301_T>C	Heterozygous	ENSP00000265131	Q680R	_VPGDQTSTIIR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
344	chr17:45241412_C>G	Homozygous	ENSP00000327442	P173A	_FSESTAM(ox)GASR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes

345	chr1:152985411_G>A	Homozygous	ENSP00000357751	V61I	_IPEPCQPK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
346	chr1:152985411_G>A	Homozygous	ENSP00000357751	V61I	_IPEPCQPK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
347	chr12:132777676_C>T	Heterozygous	ENSP00000204726	D1238N	_LQAEANDLQIR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
348	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
349	chr17:5368468_G>C	Heterozygous	ENSP00000339569	M628I	_DVQEQIAVLM(ox)QSR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
350	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
351	chr17:45241412_C>G	Homozygous	ENSP00000327442	P173A	_FSESTAM(ox)GASR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
352	chr2:203289829_C>T	Homozygous	ENSP00000348380	L346F	_LTPVSAQFQDIEGK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
353	chr2:203289829_C>T	Homozygous	ENSP00000348380	L346F	_LTPVSAQFQDIEGK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
354	chr12:6563433_C>T	Heterozygous	ENSP00000313272	R257Q	_DFGAQQEEGR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
355	chr12:52772156_C>T	Heterozygous	ENSP00000330101	A359T	_TQYEEIAQR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
356	chr3:122699903_G>A	Heterozygous	ENSP00000418194	R450K	_SEDVQSIEVQVK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
357	chr12:105152394_G>A	Homozygous	ENSP00000328062	V901I	_KLGITPEGQSYLDQFR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
358	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
359	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQM(ox)EAQVK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
360	chr12:105152394_G>A	Homozygous	ENSP00000328062	V901I	_LGITPEGQSYLDQFR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
361	chr8:143926863_T>C	Heterozygous	ENSP00000323856	H1459R	_YSELTTLTSQYIK_	Trypsin/P_CID (K,R Cterm)	novel cleavage	PLEC	Yes	Yes	-	Yes
362	chr3:45035631_G>A	Heterozygous	ENSP00000296130	G106S	_GGTLSTPQTGSENDALYEYLR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
363	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
364	chr3:45035631_G>A	Heterozygous	ENSP00000296130	G106S	_GGTLSTPQTGSENDALYEYLR_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
365	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_CID (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
366	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_EThcD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
367	chr9:115084301_T>C	Heterozygous	ENSP00000265131	Q680R	_VPGDQTSTIIR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
368	chr17:80332480_A>G	Homozygous	ENSP00000425956	D1380G	_IM(ox)CTVDHQGQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
369	chr4:67514497_G>A	Homozygous	ENSP00000273853	L341F	_TISPAESTALFQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
370	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
371	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes

372	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
373	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
374	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	MYO1C	Yes	Yes	Yes	Yes
375	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	MYO1C	Yes	Yes	Yes	Yes
376	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	MYO1C	Yes	Yes	Yes	Yes
377	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	MYO1C	Yes	Yes	Yes	Yes
378	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
379	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
380	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
381	chr11:237087_A>G	Heterozygous	ENSP00000333811	N13S	_DVPGFLQQSQSSGPGQPAVWHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
382	chr1:153003465_C>G	Homozygous	ENSP00000295367	L149V	_VPVPGYTKVPEPCPSTVTPGPAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
383	chr1:153003465_C>G	Homozygous	ENSP00000295367	L149V	_VPVPGYTKVPEPCPSTVTPGPAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
384	chr1:153003465_C>G	Homozygous	ENSP00000295367	L149V	_VPVPGYTKVPEPCPSTVTPGPAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
385	chr1:153003465_C>G	Homozygous	ENSP00000295367	L149V	_VPVPGYTKVPEPCPSTVTPGPAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
386	chr1:152985411_G>A	Homozygous	ENSP00000357751	V61I	_IPEPCQPKVPEPCPSTVTPAPAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
387	chr3:158649111_G>A	Heterozygous	ENSP00000264263	V215I	_GIIDLIEER_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
388	chr3:158649111_G>A	Heterozygous	ENSP00000264263	V215I	_GIIDLIEER_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
389	chr9:113373918_C>T	Heterozygous	ENSP00000238379	G146E	_LEGILEGLGLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
390	chr1:152911235_G>C	Heterozygous	ENSP00000357753	V480L	_QLELPEQQEGQLK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
391	chr11:74342502_C>T	Homozygous	ENSP00000298198	V531I	_DITTGYDSSQPNKK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
392	chr1:223766378_A>C	Heterozygous	ENSP00000295006	K568Q	_QDIQSDGFSIETCK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
393	chr20:17624595_A>T	Homozygous	ENSP00000246043	L1043H	_EKLHSLTQAKEESEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
394	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
395	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
396	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
397	chr18:31542836_G>A	Heterozygous	ENSP00000261590	R773K	_DMAGAQAAAVALNEEFLK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
398	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes

399	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
400	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
401	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
402	chr19:11447460_G>A	Heterozygous	ENSP00000465461	A291T	_SEALPTDLPTPSAPDLTEPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
403	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	-	Yes
404	chr1:153003465_C>G	Homozygous	ENSP00000295367	L149V	_VPVPGYTKVPEPCPSTVTPGPAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
405	chr19:4844778_G>C	Heterozygous	ENSP00000221957	Q284E	_AQEALLQLSQALSLMETVKEGVDQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
406	chr19:4844778_G>C	Heterozygous	ENSP00000221957	Q284E	_AQEALLQLSQALSLMETVKEGVDQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
407	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
408	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
409	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
410	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
411	chr15:72346551_T>C	Homozygous	ENSP00000268097	I436V	_DFYVVEPLAFEGTPEQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
412	chr3:149198448_T>A	Homozygous	ENSP00000264613	E544D	_MYYSAVDPTKDIFTGLIGPM(ox)K_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
413	chr3:149198448_T>A	Homozygous	ENSP00000264613	E544D	_MYYSAVDPTKDIFTGLIGPM(ox)K_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
414	chr3:129257608_C>G	Heterozygous	ENSP00000325002	L240V	_QVEEEDGSRDSPLFDFIESCLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
415	chr2:219182118_G>A	Heterozygous	ENSP00000395249	R374H	_DLGEGEEGELAPPEDLLGHPQALSR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
416	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
417	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
418	chr11:121621073_G>A	Homozygous	ENSP00000260197	V1967I	_WESPYDSPDQDLLYAIAVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
419	chr1:152985411_G>A	Homozygous	ENSP00000357751	V61I	_IPEPCQPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
420	chr15:85580757_G>A	Homozygous	ENSP00000354718	V897M	_GNTDSSLQSMGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
421	chr7:857859_G>A	Heterozygous	ENSP00000374225	E393K	_TISAVGEQLLPTVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
422	chr3:46459778_C>T	Heterozygous	ENSP00000231751	A29T	_SVQWCTVSQPEATK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
423	chr21:29067014_C>T	Heterozygous	ENSP00000286788	V147I	_AHEILPNLICCSAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
424	chr11:67490352_C>A	Homozygous	ENSP00000279146	Q228K	_QITPLLLNYCQCK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	AIP	Yes	Yes	Yes	Yes
425	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

426	chr19:10631494_A>G	Homozygous	ENSP00000336888	Q154R	_DGDCPAVLIPSKPLAR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	SLC44A2	Yes	Yes	Yes	Yes
427	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
428	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
429	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
430	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
431	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
432	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
433	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
434	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
435	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
436	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
437	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
438	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
439	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
440	chr1:201386394_T>C	Heterozygous	ENSP00000356282	K337E	_NLPSLAEQGASDPPTVASR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
441	chr12:121888906_T>C	Homozygous	ENSP00000261817	V17A	_QSGGSSQAGAVTVSDVQELM(ox)R_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
442	chr12:121888906_T>C	Homozygous	ENSP00000261817	V17A	_QSGGSSQAGAVTVSDVQELM(ox)R_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
443	chr12:121888906_T>C	Homozygous	ENSP00000261817	V17A	_QSGGSSQAGAVTVSDVQELM(ox)R_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
444	chr12:121888906_T>C	Homozygous	ENSP00000261817	V17A	_QSGGSSQAGAVTVSDVQELM(ox)R_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
445	chr1:179882939_T>C	Homozygous	ENSP00000271583	M146T	_LQQQHSEQPPLQPSPVTTR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
446	chr1:153003465_C>G	Homozygous	ENSP00000295367	L149V	_VPVPGYTKVPEPCPSTVTPGPAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
447	chr3:158649111_G>A	Heterozygous	ENSP00000264263	V215I	_GIIDLIEER_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
448	chr2:203251967_C>T	Homozygous	ENSP00000348380	S97L	_TLDPFETMLK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
449	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GM(ox)GTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
450	chrX:123900661_A>C	Homozygous	ENSP00000347858	Q423P	_DSMPDESSQTSLQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
451	chr1:223766378_A>C	Heterozygous	ENSP00000295006	K568Q	_QDIQSDGFSIETCK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
452	chr1:223766378_A>C	Heterozygous	ENSP00000295006	K568Q	_QDIQSDGFSIETCK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

453	chr19:1068739_G>A	Heterozygous	ENSP00000316772	R139H	_LKECVLHDDLLEAR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
454	chr12:105152394_G>A	Homozygous	ENSP00000328062	V901I	_LGITPEGQSYLDQFR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
455	chr12:105152394_G>A	Homozygous	ENSP00000328062	V901I	_LGITPEGQSYLDQFR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
456	chr19:49451550_T>G	Homozygous	ENSP00000262265	M9L	_LLGLELSEAEAIGADSAR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
457	chr2:171976274_T>C	Homozygous	ENSP00000264108	V314A	_LMQGFNEDMAIEAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
458	chr2:171976274_T>C	Homozygous	ENSP00000264108	V314A	_LMQGFNEDMAIEAQQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
459	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
460	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
461	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
462	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
463	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
464	chr2:68822807_G>T	Heterozygous	ENSP00000386241	R549S	_NSGEEEIDSLQSM(ox)VQELR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
465	chr1:89381852_C>T	Homozygous	ENSP00000359485	L344F	_FPTDTLQELLDVHAACER_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
466	chr2:68822807_G>T	Heterozygous	ENSP00000386241	R549S	_NSGEEEIDSLQSMVQELRK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
467	chr17:56901433_G>A	Homozygous	ENSP00000323889	P358L	_LQELTPSSGDPGEHDPASTHK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
468	chr17:56901433_G>A	Homozygous	ENSP00000323889	P358L	_LQELTPSSGDPGEHDPASTHK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
469	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
470	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
471	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
472	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
473	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
474	chr5:76627482_A>G	Homozygous	ENSP00000274364	K532E	_VLWLDEIQQAVDEANVDEDRAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
475	chr3:129257608_C>G	Heterozygous	ENSP00000325002	L240V	_QVEEEDGSRDSPLFDFIESCLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
476	chr3:129257608_C>G	Heterozygous	ENSP00000325002	L240V	_QVEEEDGSRDSPLFDFIESCLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
477	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
478	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
479	chr20:2309687_C>A	Homozygous	ENSP00000370867	T13K	_(ac)AALGVQSINWQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

480	chr2:10577680_C>T	Homozygous	ENSP00000263837	D585N	_NGTLSVSNTTVGSK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
481	chr7:92083384_A>G	Homozygous	ENSP00000348573	N2792S	_ISSSSQTPQILVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
482	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
483	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
484	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
485	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
486	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
487	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
488	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
489	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
490	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
491	chr4:48842653_A>G	Heterozygous	ENSP00000264312	T53A	_SVPLAAASM(ox)LITQGLISK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
492	chr4:48842653_A>G	Heterozygous	ENSP00000264312	T53A	_SVPLAAASM(ox)LITQGLISK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
493	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPKTVAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
494	chr10:133420037_A>G	Homozygous	ENSP00000323047	I293V	_VLTGTGNVNVIQPNYPAAAR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
495	chr13:24455080_C>T	Homozygous	ENSP00000371419	A899T	_QITLHALSLVGEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
496	chr5:139478334_T>C	Homozygous	ENSP00000331288	H232R	_FLDKLPQQTGDR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
497	chr3:150562658_C>G	Homozygous	ENSP00000273435	T92S	_NTVLATWQPYSTSK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
498	chr17:4953081_A>G	Homozygous	ENSP00000324105	N71S	_AVENINSTLGPALLQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
499	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
500	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
501	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
502	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
503	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
504	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
505	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
506	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

507	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
508	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
509	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
510	chr1:201386394_T>C	Heterozygous	ENSP00000356282	K337E	_NLPSLAEQGASDPPTVASR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
511	chr2:238081747_G>A	Homozygous	ENSP00000254663	A183T	_VSGQTEVDDILAAVRPTTR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
512	chr2:238081747_G>A	Homozygous	ENSP00000254663	A183T	_VSGQTEVDDILAAVRPTTR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
513	chr6:25604863_G>A	Homozygous	ENSP00000331983	G1202S	_KLGNDAVSQDSSSPALSSVER_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
514	chr1:225419442_C>T	Homozygous	ENSP00000272163	S154N	_FNLSQESSYIATQYSLRPR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
515	chr1:225419442_C>T	Homozygous	ENSP00000272163	S154N	_FNLSQESSYIATQYSLRPR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
516	chr1:225419442_C>T	Homozygous	ENSP00000272163	S154N	_FNLSQESSYIATQYSLRPR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
517	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
518	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
519	chr20:17619712_G>A	Heterozygous	ENSP00000246043	S1199L	_EHTLHLEAELEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
520	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
521	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
522	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
523	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
524	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
525	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
526	chr12:6601981_C>A	Homozygous	ENSP00000349508	E139D	_KEEEEEDDDDDDSKEPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
527	chr3:129257608_C>G	Heterozygous	ENSP00000325002	L240V	_QVEEEDGSRDSPLFDFIESCLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
528	chr3:129257608_C>G	Heterozygous	ENSP00000325002	L240V	_QVEEEDGSRDSPLFDFIESCLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
529	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
530	chr5:149368339_G>C	Heterozygous	ENSP00000274569	E390D	_KQPQDAAVWR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
531	chr20:2309687_C>A	Homozygous	ENSP00000370867	T13K	_(ac)AALGVQSINWQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
532	chr20:2309687_C>A	Homozygous	ENSP00000370867	T13K	_(ac)AALGVQSINWQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
533	chr19:1094005_G>A	Homozygous	ENSP00000215587	S44F	_AQFGDKPSEGRPR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

534	chr1:153032377_C>T	Homozygous	ENSP00000306461	T11I	_QPCIPPPQLQQQQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
535	chr4:48842653_A>G	Heterozygous	ENSP00000264312	T53A	_SVPLAAASM(ox)LITQGLISK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
536	chr4:48842653_A>G	Heterozygous	ENSP00000264312	T53A	_SVPLAAASM(ox)LITQGLISK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
537	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPKTVAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
538	chr1:159016687_G>C	Homozygous	ENSP00000295809	S179T	_TSLSAPPNTSSTENPKTVAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
539	chr10:133420037_A>G	Homozygous	ENSP00000323047	I293V	_VLTGTGNVNVIQPNYPAAAR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
540	chr19:17219251_T>C	Homozygous	ENSP00000263897	L154S	_QSAAELDLVLQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
541	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GMGTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
542	chr11:61125134_A>G	Heterozygous	ENSP00000342681	H461R	_LSAYPALEGALHR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	CD5	Yes	Yes	-	Yes
543	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GM(ox)GTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
544	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GM(ox)GTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
545	chr8:141136609_T>C	Homozygous	ENSP00000428714	M68T	_EDSQTAGANCGTLGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
546	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQQMEAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
547	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQQM(ox)EAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
548	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQQM(ox)EAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
549	chr21:29067014_C>T	Heterozygous	ENSP00000286788	V147I	_AHEILPNLICCSAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
550	chr7:44964464_T>C	Heterozygous	ENSP00000258787	Q861R	_DKDGFGAVLFSSHVR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	MYO1G	Yes	Yes	Yes	Yes
551	chr12:6451307_A>G	Homozygous	ENSP00000266557	H233R	_SNKGESPVEPAEPCR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
552	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
553	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
554	chr1:154601491_T>C	Homozygous	ENSP00000357456	K89R	_NTNSVPETAPAAIPETR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
555	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
556	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
557	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
558	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
559	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
560	chr1:145873487_G>A	Heterozygous	ENSP00000347802	P428S	_SQPEEQGPSQSPASETIR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

561	chr5:32400161_A>G	Heterozygous	ENSP00000265069	I520T	_LQSTGNKAEDTKGTECVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
562	chr6:25604863_G>A	Homozygous	ENSP00000331983	G1202S	_LGNDAVSQDSSSPALSSVER_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
563	chr2:233274722_A>G	Heterozygous	ENSP00000318259	T137A	_SVSSFPVPQDNVDAHPGSGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
564	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	Yes	Yes
565	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	-	Yes
566	chr1:225419442_C>T	Homozygous	ENSP00000272163	S154N	_FNLSQESSYIATQYSLRPR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
567	chr1:225419442_C>T	Homozygous	ENSP00000272163	S154N	_FNLSQESSYIATQYSLRPR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
568	chr1:159032587_C>A	Homozygous	ENSP00000295809	R409S	_LPQEQSQLPNPSEASTTFPESHLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
569	chr1:159032587_C>A	Homozygous	ENSP00000295809	R409S	_LPQEQSQLPNPSEASTTFPESHLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
570	chr3:191369951_A>T	Heterozygous	ENSP00000376249	L121F	_LLQEKEFQEEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
571	chr1:209777415_C>G	Homozygous	ENSP00000355990	Q109E	_FLENEHQELQAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
572	chr16:70514394_G>A	Homozygous	ENSP00000315775	T162I	_SEDYEQAAAHIHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
573	chr11:67490352_C>A	Homozygous	ENSP00000279146	Q228K	_EQPGSPEWIQLDK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
574	chr11:67490352_C>A	Homozygous	ENSP00000279146	Q228K	_EQPGSPEWIQLDK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
575	chr3:13354079_C>A	Heterozygous	ENSP00000254508	R786L	_NPLLDLAAYDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
576	chr3:13354079_C>A	Heterozygous	ENSP00000254508	R786L	_NPLLDLAAYDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
577	chr1:63631761_C>T	Heterozygous	ENSP00000360124	R239C	_ICIDAMHGVVGPYVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
578	chr10:80166946_G>A	Heterozygous	ENSP00000265447	R197C	_GFGTDEQAIIDCLGSCSNK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
579	chr1:89381852_C>T	Homozygous	ENSP00000359485	L344F	_FPTDTLQELLDVHAACER_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
580	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_SVNGKEDAIWNLLHQAQEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
581	chr9:133332630_G>A	Heterozygous	ENSP00000361092	T175M	_KAEEAMEAQEVVEATPEGACTEPR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
582	chr19:4844778_G>C	Heterozygous	ENSP00000221957	Q284E	_AQEALLQLSQALSLMETVKEGVDQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
583	chr19:4844778_G>C	Heterozygous	ENSP00000221957	Q284E	_AQEALLQLSQALSLMETVKEGVDQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
584	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
585	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
586	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
587	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes

588	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
589	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
590	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
591	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
592	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
593	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
594	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
595	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
596	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
597	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
598	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
599	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
600	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
601	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
602	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
603	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
604	chr8:143926863_T>C	Heterozygous	ENSP00000323856	H1459R	_YSELTTLTSQYIK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	PLEC	Yes	Yes	Yes	Yes
605	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
606	chr14:20456995_T>G	Heterozygous	ENSP00000216714	D148E	_VSYGIGEEEHDQEGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
607	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
608	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
609	chr5:149368339_G>C	Heterozygous	ENSP00000274569	E390D	_KQPQDAAVWR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
610	chr20:2309687_C>A	Homozygous	ENSP00000370867	T13K	_(ac)AALGVQSINWQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
611	chr3:49722606_T>C	Homozygous	ENSP00000309092	Q184R	_INAGM(ox)YILSPAVLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
612	chr10:133420037_A>G	Homozygous	ENSP00000323047	I293V	_VLTGTGNVNVIQPNYPAAAR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
613	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GMGTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
614	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GM(ox)GTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes

615	chr13:46685919_T>C	Homozygous	ENSP00000308493	S234P	_VLPQELVDLPLVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
616	chr8:141136609_T>C	Homozygous	ENSP00000428714	M68T	_EDSQTAGANCGTLGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
617	chr2:203289829_C>T	Homozygous	ENSP00000348380	L346F	_LTPVSAQFQDIEGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
618	chr11:74342502_C>T	Homozygous	ENSP00000298198	V531I	_DITTGYDSSQPNKK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
619	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQMEAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
620	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQM(ox)EAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
621	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQM(ox)EAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
622	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
623	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
624	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
625	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
626	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
627	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
628	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_M(ox)QYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
629	chr12:132808278_G>A	Heterozygous	ENSP00000204726	P264L	_TEDSNAGNSGGNVLAPDSTK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
630	chr14:24440156_T>C	Homozygous	ENSP00000382327	Q270R	_RTLATGYQYSFPELGAALK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	SDR39U1	Yes	Yes	Yes	Yes
631	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	Yes	Yes
632	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	Yes	Yes
633	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	Yes	Yes
634	chr1:159032587_C>A	Homozygous	ENSP00000295809	R409S	_LPQEQSQLPNPSEASTTFPESHLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
635	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
636	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
637	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
638	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
639	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
640	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
641	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes

642	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
643	chr14:105469487_G>A	Homozygous	ENSP00000389425	A152T	_GTYLGLANHGQTR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
644	chr3:49722606_T>C	Homozygous	ENSP00000309092	Q184R	_INAGMYILSPAVLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
645	chr3:49722606_T>C	Homozygous	ENSP00000309092	Q184R	_INAGM(ox)YILSPAVLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
646	chr3:49722606_T>C	Homozygous	ENSP00000309092	Q184R	_INAGM(ox)YILSPAVLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
647	chr3:49722606_T>C	Homozygous	ENSP00000309092	Q184R	_INAGM(ox)YILSPAVLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
648	chr2:75696287_T>C	Homozygous	ENSP00000318690	N249S	_DIDLSCGSGSSK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
649	chr11:64344456_T>C	Homozygous	ENSP00000349238	W639R	_EAPQGELVPEAR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
650	chr5:76652825_A>G	Heterozygous	ENSP00000274364	I724V	_QTFIDNTDSVVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
651	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GMGTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
652	chr11:119081463_A>G	Heterozygous	ENSP00000479680	K345R	_DLHDQFQHQLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
653	chr16:3671772_G>C	Heterozygous	ENSP00000246957	D395E	_GVVDSEEIPLNLSR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
654	chr5:80437260_T>C	Homozygous	ENSP00000337159	I192T	_EQQNDTSSELQNR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
655	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQQM(ox)EAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
656	chr1:36286832_C>T	Homozygous	ENSP00000346634	A201V	_DSRPSQAAGDNQGDEVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
657	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
658	chr19:35527026_G>C	Homozygous	ENSP00000430242	A419G	_EGGQFGHDIHHTAGQAGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
659	chr12:132808278_G>A	Heterozygous	ENSP00000204726	P264L	_TEDSNAGNSGGNVLAPDSTK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
660	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	-	Yes
661	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	Yes	Yes
662	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_RDAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
663	chr11:67490352_C>A	Homozygous	ENSP00000279146	Q228K	_EQPGSPEWIQLDK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
664	chr8:143926863_T>C	Heterozygous	ENSP00000323856	H1459R	_YSELTTLTSQYIK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	PLEC	Yes	Yes	Yes	Yes
665	chr8:143926863_T>C	Heterozygous	ENSP00000323856	H1459R	_YSELTTLTSQYIK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	PLEC	Yes	Yes	-	Yes
666	chr10:133370622_G>A	Homozygous	ENSP00000357535	T75I	_IFEEDPAVGAIVLTGGDK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
667	chr2:68822807_G>T	Heterozygous	ENSP00000386241	R549S	_NSGEEEIDSLQSMVQELR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
668	chr3:45035631_G>A	Heterozygous	ENSP00000296130	G106S	_GGTLSTPQTGSENDALYEYLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes

669	chr3:45035631_G>A	Heterozygous	ENSP00000296130	G106S	_GGTLSTPQTGSENDALYEYLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
670	chr17:77498623_A>G	Homozygous	ENSP00000329161	M558V	_LNEGSSAMANGVEEKEPEAPEM(ox)_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
671	chr11:9748015_C>G	Heterozygous	ENSP00000315630	Q505E	_VLKEQALQEAMEQLEELELERK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
672	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_QKLTAEFEEAQTSACLLQEELEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
673	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_QKLTAEFEEAQTSACLLQEELEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
674	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_LTAEFEEAQTSACLLQEELEKLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
675	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_LTAEFEEAQTSACLLQEELEKLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
676	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
677	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
678	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
679	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
680	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
681	chr3:49722606_T>C	Homozygous	ENSP00000309092	Q184R	_INAGMYILSPAVLR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
682	chr1:179882939_T>C	Homozygous	ENSP00000271583	M146T	_LQQQHSEQPPLQPSPVTTR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
683	chr1:179882939_T>C	Homozygous	ENSP00000271583	M146T	_LQQQHSEQPPLQPSPVTTR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
684	chr1:89382706_G>A	Heterozygous	ENSP00000359485	E399K	_ESSVQYCQAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	GBP6	Yes	Yes	-	Yes
685	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GMGTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
686	chr16:3671772_G>C	Heterozygous	ENSP00000246957	D395E	_GVVDSEEIPLNLSR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
687	chr16:3671772_G>C	Heterozygous	ENSP00000246957	D395E	_GVVDSEEIPLNLSR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
688	chr11:74342502_C>T	Homozygous	ENSP00000298198	V531I	_DITTGYDSSQPNKK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
689	chr18:57573273_C>T	Heterozygous	ENSP00000262093	R96Q	_LFLDQDLMTLPIQNK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
690	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
691	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
692	chr19:35527026_G>C	Homozygous	ENSP00000430242	A419G	_EGGQFGHDIHHTAGQAGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
693	chr19:35527026_G>C	Homozygous	ENSP00000430242	A419G	_EGGQFGHDIHHTAGQAGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
694	chr19:11447460_G>A	Heterozygous	ENSP00000465461	A291T	_SEALPTDLPTPSAPDLTEPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
695	chr12:121888906_T>C	Homozygous	ENSP00000261817	V17A	_QSGGSSQAGAVTVSDVQELM(ox)R_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes

696	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	-	Yes
697	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	-	Yes
698	chr1:205771540_G>C	Heterozygous	ENSP00000235932	Q104E	_WKEDLDSK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
699	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
700	chr3:46450505_C>T	Heterozygous	ENSP00000231751	R291H	_EDAIWNLLHQAQEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
701	chr20:17615510_C>A	Heterozygous	ENSP00000246043	R1324L	_QKLTAEFEEAQTSACLLQEELEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
702	chr1:26557020_A>C	Heterozygous	ENSP00000363281	K324T	_EITPPFKPAVAQPDDTFYFDTEFTSR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
703	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
704	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
705	chr9:115084301_T>C	Heterozygous	ENSP00000265131	Q680R	_VPGDQTSTIIR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
706	chr9:115084301_T>C	Heterozygous	ENSP00000265131	Q680R	_VPGDQTSTIIR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
707	chr17:80381694_C>G	Heterozygous	ENSP00000425956	L4698V	_VLQEQHQLSSR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
708	chr19:35513204_A>G	Homozygous	ENSP00000342012	V91A	_QVPGFGAADALGNR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
709	chr19:35513204_A>G	Homozygous	ENSP00000342012	V91A	_QVPGFGAADALGNR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
710	chr15:74035800_G>T	Homozygous	ENSP00000268059	G780V	_AQTLGAVVPPGDSVR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
711	chr4:67514497_G>A	Homozygous	ENSP00000273853	L341F	_TISPAESTALFQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
712	chr4:67514497_G>A	Homozygous	ENSP00000273853	L341F	_TISPAESTALFQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
713	chr4:67514497_G>A	Homozygous	ENSP00000273853	L341F	_TISPAESTALFQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
714	chr2:88992663_A>G	Heterozygous	ENSP00000419211	F31L	_ITQSPSSLSASTGDR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
715	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	MYO1C	Yes	Yes	Yes	Yes
716	chr17:1470224_T>C	Homozygous	ENSP00000352834	Q826R	_NVLDTSWPTPPPALR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	MYO1C	Yes	Yes	-	Yes
717	chr10:119670121_T>C	Heterozygous	ENSP00000358081	C151R	_GQVAAAAAAQPPASHGPER_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	BAG3	Yes	Yes	Yes	Yes
718	chr22:22322826_C>A	Heterozygous	ENSP00000374825	D70E	_LLIYENNKRPSGIPDR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
719	chr5:150396792_G>C	Heterozygous	ENSP00000325223	G1354A	_GMGTVEGADQSNPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
720	chr13:75849112_T>C	Heterozygous	ENSP00000317802	M1162T	_YLDQIGNTTSSQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
721	chr20:17624595_A>T	Homozygous	ENSP00000246043	L1043H	_LHSLTQAKEESEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
722	chr3:15218485_G>A	Heterozygous	ENSP00000253693	A128T	_TSYETTDKVLQNK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes

723	chr1:152911235_G>C	Heterozygous	ENSP00000357753	V480L	_QLELPEQQEGQLK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
724	chrX:123900661_A>C	Homozygous	ENSP00000347858	Q423P	_DSMPDESSQTSLQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
725	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQM(ox)EAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
726	chr1:89384183_A>T	Homozygous	ENSP00000359485	D520V	_LQEQQQM(ox)EAQVK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
727	chr18:57573273_C>T	Heterozygous	ENSP00000262093	R96Q	_LFLDQDLMTLPIQNK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
728	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
729	chr2:215370366_C>T	Homozygous	ENSP00000323534	V2230I	_GATYNIIVEALKDQQR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
730	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
731	chr2:85394936_T>C	Homozygous	ENSP00000263867	H335R	_MQYAPNTQVEILPQGR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
732	chr19:35527026_G>C	Homozygous	ENSP00000430242	A419G	_EGGQFGHDIHHTAGQAGK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
733	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
734	chr1:89056963_G>C	Homozygous	ENSP00000359504	T349S	_VQLPTESLQELLDLHR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
735	chr18:63979835_G>A	Heterozygous	ENSP00000331368	R68Q	_DGDIHQGFQSLLSEVNR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
736	chr19:11447460_G>A	Heterozygous	ENSP00000465461	A291T	_SEALPTDLPTPSAPDLTEPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
737	chr19:11447460_G>A	Heterozygous	ENSP00000465461	A291T	_SEALPTDLPTPSAPDLTEPK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
738	chr1:39448691_G>C	Homozygous	ENSP00000289893	S5172T	_TLLPEDTQKLDNFLGEVR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
739	chr10:119423182_C>T	Heterozygous	ENSP00000376609	A119V	_SPVFIVQVGQDLVSQTEEK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
740	chr12:8863977_A>G	Homozygous	ENSP00000299698	H1229R	_NAYGGFSSTQDTVVALQALAK_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	A2ML1	Yes	Yes	Yes	Yes
741	chr12:132808278_G>A	Heterozygous	ENSP00000204726	P264L	_SLADYRTEDSNAGNSGGNVLAPDSTK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
742	chr12:132808278_G>A	Heterozygous	ENSP00000204726	P264L	_SLADYRTEDSNAGNSGGNVLAPDSTK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
743	chr22:42810944_A>C	Heterozygous	ENSP00000263245	S355R	_YFDEPVELR_	Trypsin/P_HCD (K,R Cterm)	novel cleavage	ARFGAP3	Yes	Yes	-	Yes
744	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
745	chr17:41518389_T>C	Homozygous	ENSP00000254043	T147A	_QTPASPECDYSQYFK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
746	chr15:72346551_T>C	Homozygous	ENSP00000268097	I436V	_DFYVVEPLAFEGTPEQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	-	Yes
747	chr3:149198448_T>A	Homozygous	ENSP00000264613	E544D	_MYYSAVDPTKDIFTGLIGPM(ox)K_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
748	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
749	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes

750	chr17:41502942_T>C	Homozygous	ENSP00000246635	T298A	_DAEEWFHAK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	-	Yes	-	Yes
751	chr5:76627482_A>G	Homozygous	ENSP00000274364	K532E	_VLWLDEIQQAVDEANVDEDR_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
752	chr15:72346551_T>C	Homozygous	ENSP00000268097	I436V	_ISYGPDWKDFYVVEPLAFEGTPEQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes
753	chr15:72346551_T>C	Homozygous	ENSP00000268097	I436V	_ISYGPDWKDFYVVEPLAFEGTPEQK_	Trypsin/P_HCD (K,R Cterm)	direct sequence	-	Yes	Yes	Yes	Yes

Appendix Table 4. List of peptides mapping to alternative translation initiation sties

	Modified sequence	Ensembl gene_id	Gene name	Start codon	Identified by MaxQuant	Spectral contrast angle confirmed	Manually confirmed	All confirmed	Annotation	External information
1	_(ac)AAAAGDM(ox)DNAGKER_	ENSG00000125814	NAPB	GCG	Yes	Yes	Yes	Yes	N-terminal extension	-
2	_(ac)AATEPELLDDQEAKR_	ENSG00000168259	DNAJC7	AUG	Yes	Yes	Yes	Yes	Novel N-terminus	-
3	_(ac)AESAFSFKK_	ENSG00000198612	COPS8	AUG	Yes	-	Yes	Yes	Novel N-terminus	-
4	_(ac)AESAFSFKKLL_	ENSG00000198612	COPS8	AUG	Yes	Yes	-	Yes	Novel N-terminus	-
5	_(ac)AQIQQGGPDEKEK_	ENSG00000113391	FAM172A	AUG	Yes	Yes	Yes	Yes	Novel N-terminus	-
6	_(ac)AQIQQGGPDEKEKTTALK_	ENSG00000113391	FAM172A	AUG	Yes	-	Yes	Yes	Novel N-terminus	-
7	_(ac)AQSNMFTVADVLSQDELRK_	ENSG00000046651	OFD1	AUG	Yes	-	Yes	Yes	Novel N-terminus	-
8	_(ac)ATTQISKDELDELK_	ENSG00000102024	PLS3	AUG	Yes	Yes	Yes	Yes	Novel N-terminus	-
9	_(ac)ATTQISKDELDELKEAFA_	ENSG00000102024	PLS3	AUG	-	-	Yes	Yes	Novel N-terminus	-
10	_(ac)ATTQISKDELDELKEAFAK_	ENSG00000102024	PLS3	AUG	Yes	Yes	Yes	Yes	Novel N-terminus	-
11	_(ac)LAKVDATEESDLAQQYGVR_	ENSG00000185624	P4HB	UUG	-	-	Yes	Yes	Novel N-terminus	-
12	_(ac)LAKVDATEESDLAQQYGVRGYPTIK_	ENSG00000185624	P4HB	UUG	-	-	Yes	Yes	Novel N-terminus	-
13	_(ac)LASQGDSISSQLGPIHPPPR_	ENSG00000088247	KHSRP	CUG	-	-	Yes	Yes	Novel N-terminus	-
14	_(ac)LEAENNLAAYR_	ENSG00000131095	GFAP	CUG	-	-	Yes	Yes	Novel N-terminus	-
15	_(ac)LEAENNLAAYRQEADEATLAR_	ENSG00000131095	GFAP	CUG	-	-	Yes	Yes	Novel N-terminus	-
16	_(ac)LEVIKPFCVILPEIQKPER_	ENSG00000058262	SEC61A1	CUG	-	-	Yes	Yes	Novel N-terminus	-
17	_(ac)LVIGQNGILSTPAVSCIIRK_	ENSG00000079739	PGM1	UUG	-	-	Yes	Yes	Novel N-terminus	-
18	_(ac)LVIQDKNKYNTPK_	ENSG00000122406	RPL5	UUG	-	-	Yes	Yes	Novel N-terminus	-
19	_(ac)LVLGDLHIPHR_	ENSG00000111237	VPS29	UUG	-	-	Yes	Yes	Novel N-terminus	-
20	_(ac)M(ox)DFNVKKLAADAGTFLSR_	ENSG00000097033	SH3GLB1	AUG	Yes	-	Yes	Yes	Novel N-terminus	-
21	_(ac)M(ox)EKEFEQIDK_	ENSG00000196396	PTPN1	AUG	-	-	Yes	Yes	Novel N-terminus	-
22	_(ac)M(ox)HTALSDLYLEHLLQK_	ENSG00000130830	MPP1	AUG	Yes	-	Yes	Yes	Novel N-terminus	Kim_novel_N-termini

23	_(ac)M(ox)HTALSDLYLEHLLQKR_	ENSG00000130830	MPP1	AUG	Yes	Yes	-	Yes	Novel N-terminus	-
24	_(ac)M(ox)HYSNISASADQALDR_	ENSG00000079308	TNS1	AUG	-	-	Yes	Yes	Novel N-terminus	-
25	_(ac)MKAENSHNAGQVDTR_	ENSG00000131503	ANKHD1	AUG	-	-	Yes	Yes	Novel N-terminus	-
26	_(ac)MQFEMHDNVK_	ENSG00000105427	CNFN	AUG	Yes	-	Yes	Yes	N-terminal extension	-
27	_(ac)MQFEMHDNVKG_	ENSG00000105427	CNFN	AUG	Yes	Yes	Yes	Yes	N-terminal extension	-
28	_(ac)VAKPVVEMDGDEM(ox)TR_	ENSG00000182054	IDH2	GUG	-	-	Yes	Yes	Novel N-terminus	-
29	_(ac)VALNQEVM(ox)APEATK_	ENSG00000140459	CYP11A1	GUG	-	-	Yes	Yes	Novel N-terminus	-
30	_(ac)VALSPAGVQNLVK_	ENSG00000112992	NNT	GUG	-	-	Yes	Yes	Novel N-terminus	-
31	_(ac)VDSSQAGAHPAWVNTR_	ENSG00000100577	GSTZ1	AUG	Yes	-	Yes	Yes	N-terminal extension	-
32	_(ac)VEADIPGHGQEVLIR_	ENSG00000198125	MB	GUG	-	-	Yes	Yes	Novel N-terminus	-
33	_(ac)VEQATKPSFESGR_	ENSG00000147274	RBMX	GUG	-	-	Yes	Yes	Novel N-terminus	-
34	_(ac)VGEEEHVYSFPNK_	ENSG00000089159	PXN	GUG	-	-	Yes	Yes	Novel N-terminus	-
35	_(ac)VVFEGNHYFYSPYPTK_	ENSG00000163902	RPN1	GUG	-	-	Yes	Yes	Novel N-terminus	-
36	_(ac)VVNSETPVVVDFHAQWCGPCK_	ENSG00000100348	TXN2	GUG	-	-	Yes	Yes	Novel N-terminus	-
37	_AAAAAAVAVAAAAPHSAAK_	ENSG00000080493	SLC4A4	GUG	Yes	Yes	Yes	Yes	N-terminal extension	-
38	_AAAADGERPGPGPLLVGCGR_	ENSG00000129245	FXR2	GUG	Yes	Yes	Yes	Yes	N-terminal extension	Kim_gene_ extension
39	_AAPAAPAPGEAAAEALPAAAGAR_	ENSG00000139044	B4GALNT3	UCG	Yes	-	Yes	Yes	Novel coding	-
40	_AAPAGGGAEAGPGGGPGGAGGAAAK_	ENSG00000054793	ATP9A	GCG	Yes	Yes	-	Yes	N-terminal extension	-
41	_AASDREPPEVR_	ENSG00000115138	POMC	CUG	Yes	Yes	Yes	Yes	Novel coding	-
42	_AAVGPARLPPPAAPPPAPPPGR_	ENSG00000129245	FXR2	GUG	Yes	Yes	-	Yes	N-terminal extension	-
43	_AAVGPARLPPPAAPPPAPPPGRR_	ENSG00000129245	FXR2	GUG	-	-	Yes	Yes	N-terminal extension	-
44	_AGGAADM(ox)TDNIPLQPVRQK_	ENSG00000054793	ATP9A	GCG CCG	Yes	Yes	Yes	Yes	N-terminal extension	-
45	_AGGAADMTDNIPLQPVR_	ENSG00000054793	ATP9A	GCG CCG	Yes	Yes	Yes	Yes	N-terminal extension	-
46	_AGGEGPAGGAPGACGGGAPGGAR_	ENSG00000061938	TNK2	GUG	Yes	-	Yes	Yes	uORF	-
47	_AGSAALVAEPGAR_	ENSG00000100422	CERK	CUG GCG	Yes	Yes	Yes	Yes	Novel coding	-
48	_AGSSGSSLMEEK_	ENSG00000138030	KHK	CUG	Yes	-	Yes	Yes	N-terminal extension	-
49	_AGYKQQQDFYVR_	ENSG00000008710	PKD1	CUG	-	-	Yes	Yes	Novel coding	-
50	_ALPADSLGTQAQGLEPR_	ENSG00000128294	TPST2	GUG CUG	Yes	-	Yes	Yes	N-terminal extension uORF	Kim_gene_extension

				AUG						
51	_ANADPGPTGGTAPDSPR_	ENSG00000105443	CYTH2	GUG	Yes	Yes	Yes	Yes	N-terminal extension	-
52	_APPELAPGPPAAADAR_	ENSG00000170892 ENSG00000278622 ENSG00000274672 ENSG00000278712 ENSG00000278605 ENSG00000274796 ENSG00000274129 ENSG00000273896 ENSG00000275165 ENSG00000274078	TSEN34	AUG UUG GUG	Yes	Yes	Yes	Yes	Novel coding	-
53	_AVAAAAAAAPDPGGR_	ENSG00000077147	TM9SF3	CCG GCG	Yes	Yes	Yes	Yes	Novel coding	-
54	_AVPAGADQEDHADGR_	ENSG00000175550	DRAP1	CUG	Yes	Yes	Yes	Yes	Novel coding	-
55	_AVPAGADQEDHADGRR_	ENSG00000175550	DRAP1	CUG	Yes	Yes	Yes	Yes	Novel coding	-
56	_DAEQEEEVQR_	ENSG00000175550	DRAP1	CUG	Yes	Yes	Yes	Yes	Novel coding	Branca_different _frame
57	_DVDLDVLVEALKPVGTFK_	ENSG00000169515	CCDC8	CUG GUG	Yes	Yes	Yes	Yes	N-terminal extension	-
58	_EAEAGGGGGACAVGLGR_	ENSG00000214753 ENSG00000234857	HNRNPUL2 HNRNPUL2-BSCL2	CUG	Yes	Yes	Yes	Yes	Novel coding	Kim_novel_coding_regions
59	_EAGAGAEAAAGSAR_	ENSG00000182446	NPLOC4	CUG	Yes	Yes	Yes	Yes	N-terminal extension	-
60	_EANEDKTMFANLK_	ENSG00000117620	SLC35A3	CUG UUG GUG	Yes	-	Yes	Yes	N-terminal extension	-
61	_EEEEEAAPPPPPPPPPR_	ENSG00000187555	USP7	GCG	Yes	Yes	Yes	Yes	N-terminal extension	-
62	_EVVSSQVDDLTSHNEHLCK_	ENSG00000135951	TSGA10	UUG	Yes	Yes	Yes	Yes	Novel coding uORF	-
63	_FAGEEAAGGGGGGGGGDGGEAAESDR_	ENSG00000214753 ENSG00000234857	HNRNPUL2 HNRNPUL2-BSCL2	CUG GCG	Yes	Yes	Yes	Yes	Novel coding	Kim_novel_coding_regions
64	_GAAATAADPPGGLGLSPGPGR_	ENSG00000119729	RHOQ	GUG	Yes	Yes	-	Yes	N-terminal extension	-
65	_GAVDPGVLSVPRPAPPGPPAADGR_	ENSG00000099364	FBXL19	GUG CUG	Yes	Yes	-	Yes	N-terminal extension	-
66	_GSGADPEAGFAQPPTR_	ENSG00000189060	H1F0	CUG	Yes	Yes	Yes	Yes	N-terminal extension	-
67	_GSLGGGAM(ox)VGQLSEGAIAAIMQK_	ENSG00000132383	RPA1	GUG CUG	-	-	Yes	Yes	N-terminal extension	Kim_gene_extension
68	_HELNM(ox)AHNAGAAAAAGTHSAK_	ENSG00000049618	ARID1B	CUG GUG AUG	Yes	Yes	Yes	Yes	N-terminal extension	-
69	_IDLICQQNNIIVLEDTIK_	ENSG00000135951	TSGA10	UUG GUG	Yes	Yes	Yes	Yes	Novel coding uORF	-
70	_IDLICQQNNIIVLEDTIKR_	ENSG00000135951	TSGA10	UUG GUG	Yes	-	Yes	Yes	Novel coding uORF	-

71	_KEEDSTVGMLQIGEDVDYLLIPR_	ENSG00000169515	CCDC8	CUG GUG UUG	Yes	Yes	Yes	Yes	N-terminal extension	-
72	_LATVEETVVR_	ENSG00000105568	PPP2R1A	CUG	Yes	-	Yes	Yes	N-terminal extension	-
73	_LDDGSALGR_	ENSG00000284194	SCO2	CUG	Yes	Yes	-	Yes	Novel coding	-
74	_LDLPLVLR_	ENSG00000134419	RPS15A	CUG	-	-	Yes	Yes	uORF	-
75	_LDMESDDAR_	ENSG00000173457	PPP1R14B	CUG	Yes	Yes	-	Yes	N-terminal extension	-
76	_LDVTPLSLGIETAGGVM(ox)TVLIK_	ENSG00000109971	HSPA8	UUG	Yes	Yes	-	Yes	N-terminal extension	-
77	_LDVTPLSLGIETAGGVM(ox)TVLIKR_	ENSG00000109971	HSPA8	UUG	Yes	Yes	-	Yes	N-terminal extension	-
78	_LELEVELR_	ENSG00000173214	MFSD4B	CUG	Yes	-	Yes	Yes	N-terminal extension	-
79	_LGCTELAVATATGAGGAAGQR_	ENSG00000117620	SLC35A3	CUG UUG	Yes	Yes	Yes	Yes	Novel coding uORF	-
80	_LGLLLEELRR_	ENSG00000150527	CTAGE5	UUG	Yes	Yes	-	Yes	uORF	-
81	_LGM(ox)FLSFPTTK_	ENSG00000206172 ENSG00000188536	HBA1 HBA2	CUG	Yes	Yes	-	Yes	N-terminal extension	-
82	_LGNWM(ox)SPADFQR_	ENSG00000100889	PCK2	CUG	Yes	Yes	-	Yes	N-terminal extension	-
83	_LLEPAGAHLSSGSSEPLVEPGR_	ENSG00000119383	PTPA	AUG	Yes	-	Yes	Yes	N-terminal extension	Kim_gene_extension
84	_LLTPGACSSEVPSAVPSR_	ENSG00000067225	PKM	AUG	Yes	Yes	Yes	Yes	N-terminal extension Novel coding	-
85	_LPDDEEPPNMASESGK_	ENSG00000126522	ASL	CUG GCG	Yes	Yes	Yes	Yes	N-terminal extension	-
86	_LSAPQPGPDILQAPAR_	ENSG00000170892 ENSG00000278622 ENSG00000274672 ENSG00000278712 ENSG00000278605 ENSG00000274796 ENSG00000274129 ENSG00000273896 ENSG00000275165 ENSG00000274078	TSEN34	AUG UUG GUG	Yes	Yes	Yes	Yes	Novel coding	-
87	_LVEPDAGAGVAVM(ox)K_	ENSG00000167969	ECI1	CUG	Yes	Yes	-	Yes	N-terminal extension	-
88	_LVFHSFGGGTGSGFTSLLM(ox)ER_	ENSG00000127824	TUBA4A	CUG	Yes	Yes	-	Yes	N-terminal extension	-
89	_LVMAQPGPASQPDVSLQQR_	ENSG00000029725	RABEP1	CUG	Yes	Yes	Yes	Yes	N-terminal extension	-
90	_M(ox)EAASGGYAEVGR_	ENSG00000131503	ANKHD1	AUG	Yes	Yes	-	Yes	N-terminal extension	-
91	_MEAATLFLK_	ENSG00000107201	DDX58	AUG	-	-	Yes	Yes	uORF	-
92	_NFVVDSANKELEEAK_	ENSG00000135951	TSGA10	UUG GUG	Yes	Yes	Yes	Yes	Novel coding uORF	-

93	_QDLSSASLPGAVAALSPLR_	ENSG00000128050	PAICS	AUG	Yes	Yes	-	Yes	N-terminal extension	-
94	_RFAGEEAAGGGGGGGDGGEAAESDR_	ENSG00000214753 ENSG00000234857	HNRNPUL2 HNRNPUL2-BSCL2	CUG GCG	Yes	-	Yes	Yes	Novel coding	Kim_novel_coding_regions
95	_RM(ox)PHEELPSLQRPR_	ENSG00000145703	IQGAP2	CCG CUG	Yes	Yes	Yes	Yes	N-terminal extension	Kim_gene_extension
96	_SAVVPAAGLGTTGGFALDPTPR_	ENSG00000134955	SLC37A2	AUG	Yes	Yes	Yes	Yes	Novel coding	-
97	_SDAMDTESQYSGYSYK_	ENSG00000162738	VANGL2	GUG CUG	Yes	Yes	Yes	Yes	N-terminal extension	-
98	_SGPAPARPGPDSDSAAGAAR_	ENSG00000054598	FOXC1	CUG	Yes	Yes	-	Yes	Novel coding	-
99	_SGPPGNGGPGEGEGGEAR_	ENSG00000107829	FBXW4	AUG GUG	Yes	Yes	Yes	Yes	N-terminal extension	-
100	_SISEENM(ox)LANSASVR_	ENSG00000157851	DPYSL5	CUG ACG	Yes	Yes	Yes	Yes	N-terminal extension	Kim_gene_extension
101	_SQQAPEGGSCQPGK_	ENSG00000112759	SLC29A1	CUG AUG	Yes	Yes	Yes	Yes	N-terminal extension	-
102	_SSDSSLASPGAALQTGPVVR_	ENSG00000189060	H1F0	CUG	Yes	Yes	Yes	Yes	N-terminal extension	-
103	_TLEADKDHYKSEAQHLR_	ENSG00000135951	TSGA10	UUG GUG	Yes	-	Yes	Yes	N-terminal extension	-
104	_TLQVSPPGGGPEVAFR_	ENSG00000136883	KIF12	AUG CUG GUG	Yes	-	Yes	Yes	uORF	-
105	_TM(ox)AGAPTVSLPELR_	ENSG00000215845	TSTD1	AUG	Yes	Yes	Yes	Yes	N-terminal extension	-
106	_TMESGSTAASEEAR_	ENSG00000108946	PRKAR1A	GCG GUG UCG CUG	Yes	-	Yes	Yes	N-terminal extension	Kim_gene_extension
107	_TPGAM(ox)GGDLVLGLGALR_	ENSG00000104783	KCNN4	GUG	Yes	-	Yes	Yes	N-terminal extension Novel coding	-
108	_TQPASPATVGDFQGEGAHLPSGAR_	ENSG00000173809	TDRD12	GUG CUG	-	-	Yes	Yes	N-terminal extension	-
109	_TRAEYESEAEGVM(ox)AGQAF_	ENSG00000115541	HSPE1	CUG	Yes	-	Yes	Yes	N-terminal extension	-
110	_VAAAAEEAAAAGPR_	ENSG00000130725	UBE2M	UUG GUG	Yes	Yes	Yes	Yes	N-terminal extension	Branca_N-terminal_ extension
111	_VDLEPGTM(ox)DSVR_	ENSG00000196230 ENSG00000183311 ENSG00000232575 ENSG00000232421 ENSG00000235067 ENSG00000224156 ENSG00000229684 ENSG00000227739 ENSG00000258947 ENSG00000176014	TUBB TUBB3 TUBB6 TUBB4A	GUG	Yes	Yes	-	Yes	N-terminal extension	-

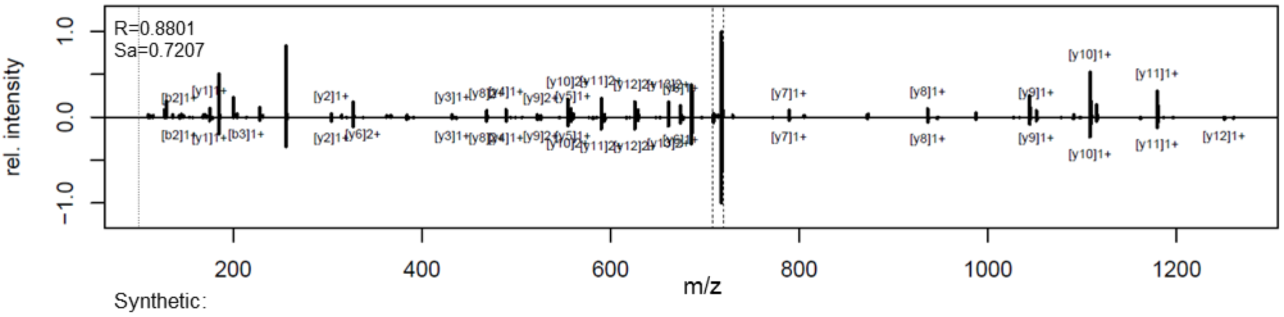
		ENSG00000104833								
112	_VEEAEAMAETEER_	ENSG00000091527	CDV3	CCG UCG	Yes	Yes	Yes	Yes	N-terminal extension	Kim_gene_extension
113	_VEIVDSVEAYATM(ox)LR_	ENSG00000079739	PGM1	GUG	Yes	Yes	-	Yes	N-terminal extension	-
114	_VPPPSWSEQCECAR_	ENSG00000082458	DLG3	CUG GUG	Yes	Yes	Yes	Yes	N-terminal extension	-
115	_VQTFTGGVVSAHTGAPE_	ENSG00000080603	SRCAP	GUG	Yes	-	Yes	Yes	N-terminal extension	-
116	_VSGGGDGHHQPPAAVPAGER_	ENSG00000148841	ITPRIP	CUG	Yes	-	Yes	Yes	Novel coding	-
117	_VVTVPAYFNDSQR_	ENSG00000109971	HSPA8	GUG	Yes	Yes	-	Yes	N-terminal extension	-

Appendix S1. Spectra of aTIS peptides confirmed by synthetic peptides

Appendix S1.1. Spectra of aTIS peptides confirmed by spectral contrast angle analysis

(ac)AAAAGDM(ox)DNAGKER

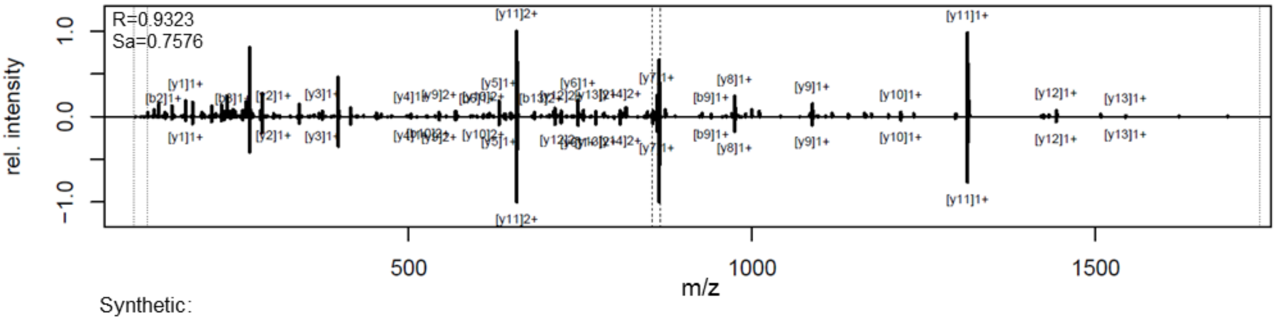
Sample:
01697_C02_P018020_S00_N11_R2 FTMS + c NSI d Full ms2 717.82@hcd25.00 [100.00−1485.00] 8780 Score 108.61



Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 717.8181@hcd28.00
[100.00−1446.00] 7412 147.39

(ac)AATEPELLDDQEAKR

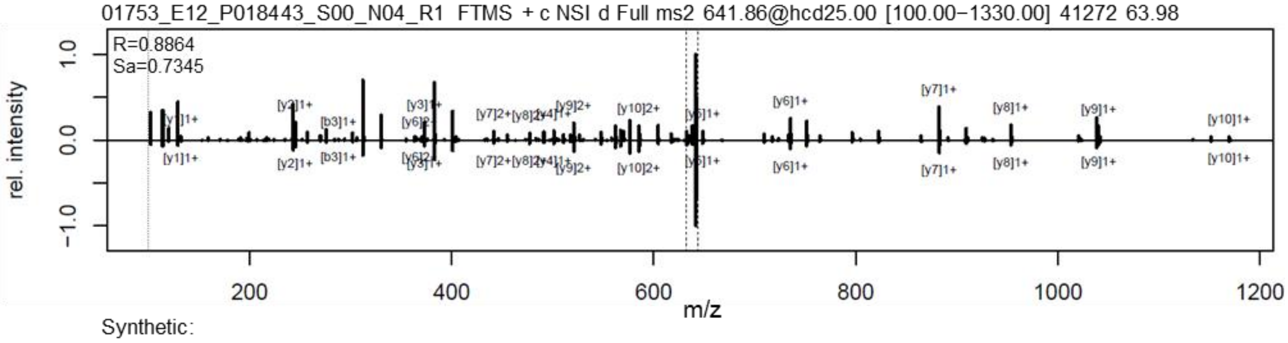
Sample:
01279_A03_P013163_B00_N17_R1 FTMS + c NSI d Full ms2 864.93@hcd25.00 [119.00−1785.00] 22786 87.5



Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 864.4267@hcd28.00
[100.00−1739.00] 31457 202.4

(ac)AESAFSFKKLL

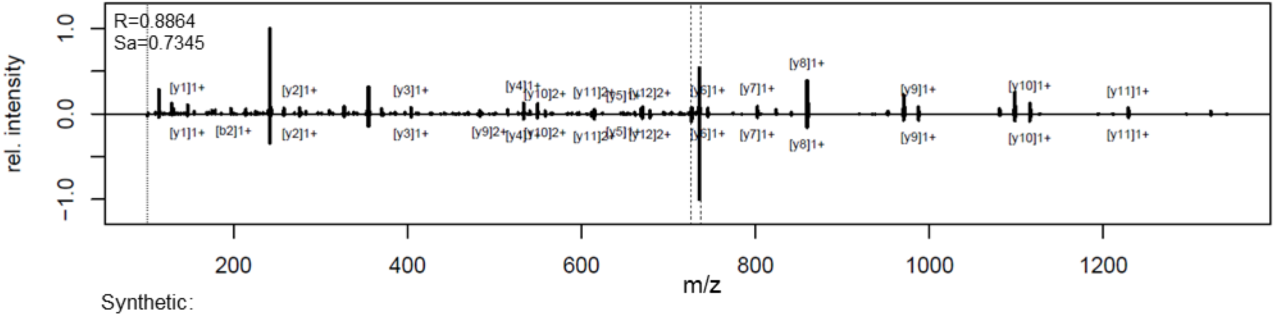
Sample:
01753_E12_P018443_S00_N04_R1 FTMS + c NSI d Full ms2 641.86@hcd25.00 [100.00−1330.00] 41272 63.98



Synthetic:
02383a_BE1−TUM_ac_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 641.8573@hcd28.00
[100.00−1294.00] 49924 170.89

(ac)AQIQQGGPDEKEK

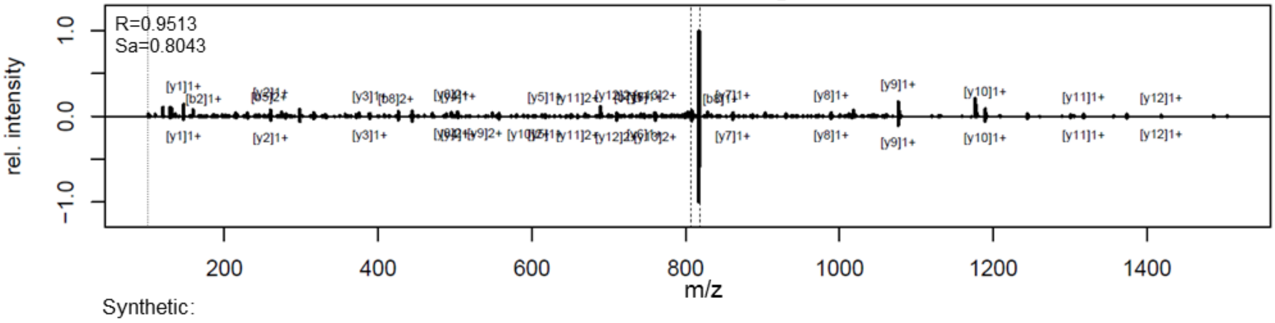
Sample:
01347_C02_P013678_S00_N11_R2 FTMS + c NSI d Full ms2 735.37@hcd25.00 [101.33−1520.00] 7791 81.92



Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01-DDA−1h−R1 FTMS + c NSI d Full ms2 735.3661@hcd28.00
[100.00−1481.00] 12573 217.68

(ac)ATTQISKDELDELK

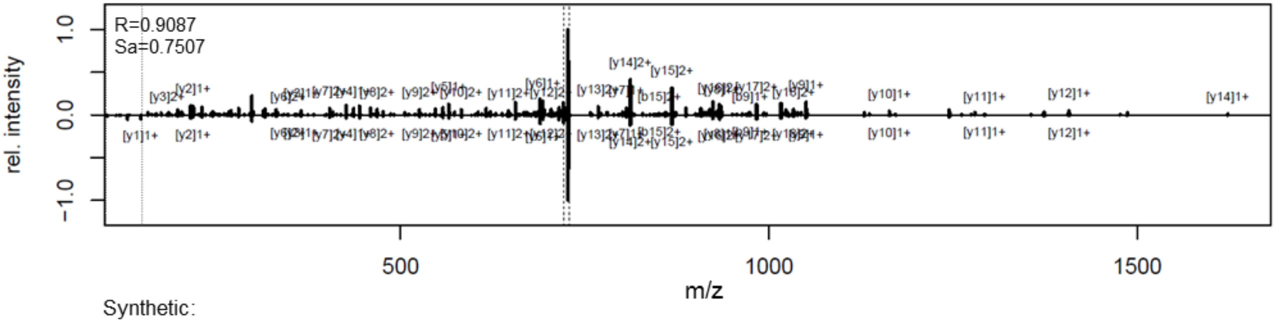
Sample:
01752_G02_P018440_S00_N15_R1 FTMS + c NSI d Full ms2 816.92@hcd25.00 [100.00−1685.00] 27885 137



Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01-DDA−1h−R1 FTMS + c NSI d Full ms2 816.9193@hcd28.00
[100.00−1644.00] 35674 231.71

(ac)ATTQISKDELDELKEAFAK

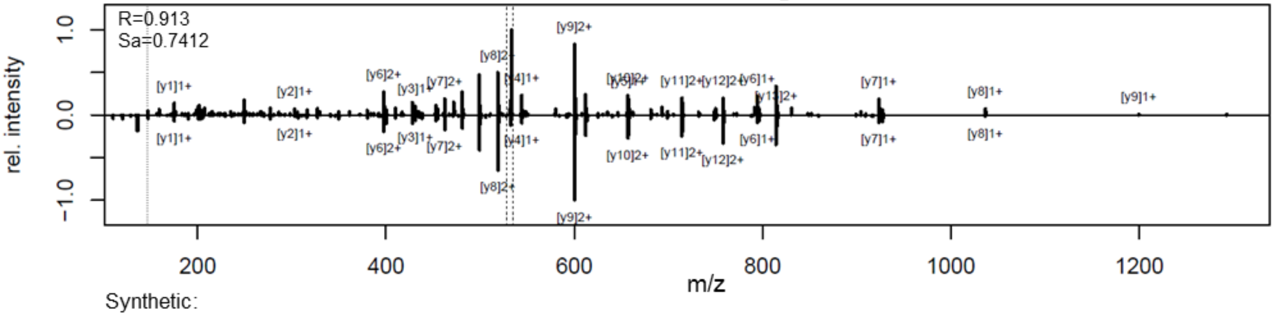
Sample:
01088_D03_P010740_S00_N20_R1 FTMS + c NSI d Full ms2 727.38@hcd25.00 [149.67−2245.00] 41200 164.04



Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01-DDA−1h−R1 FTMS + c NSI d Full ms2 727.3781@hcd28.00
[100.00−2000.00] 48506 292.75

(ac)M(ox)HTALSDLYLEHLLQKR

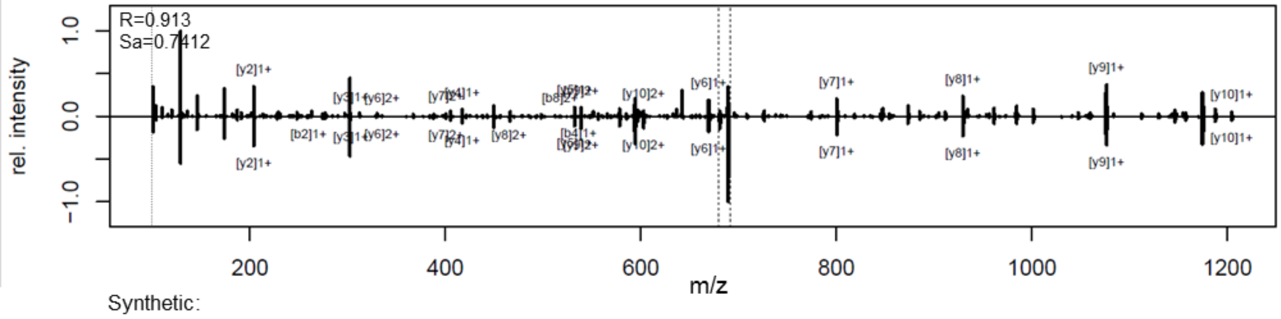
Sample:
01077_C03_P010695_S00_N19_R1 FTMS + c NSI d Full ms2 532.53@hcd25.00 [146.33−2195.00] 42810 68.58



Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 532.5332@hcd28.00
[100.00−2000.00] 45838 136.81

(ac)MQFEMHDNVKG

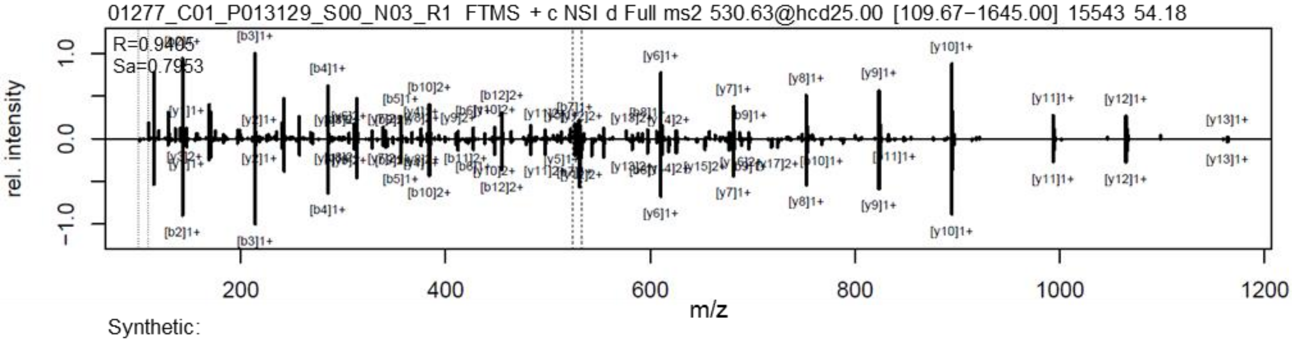
Sample:
01754_F02_P018444_S00_N14_R1 FTMS + c NSI d Full ms2 689.30@hcd25.00 [100.00−1425.00] 25851 56.63



Synthetic:
02383a_BE1−TUM_ac_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 689.3009@hcd28.00
[100.00−1389.00] 31910 173.37

AAAAAAVAVAAAAPHSAAK

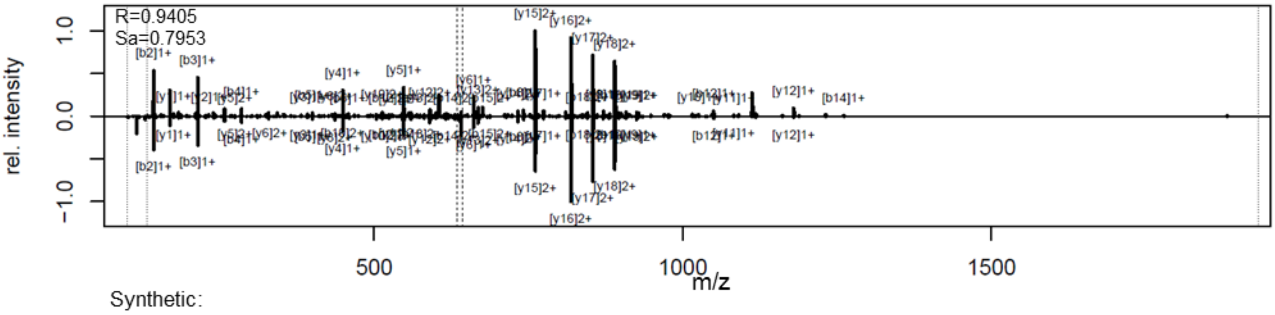
Sample:
01277_C01_P013129_S00_N03_R1 FTMS + c NSI d Full ms2 530.63@hcd25.00 [109.67−1645.00] 15543 54.18



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 530.6304@hcd28.00
[100.00−1602.00] 26755 109.72

AAAADGERPGPGPLLVGCGR

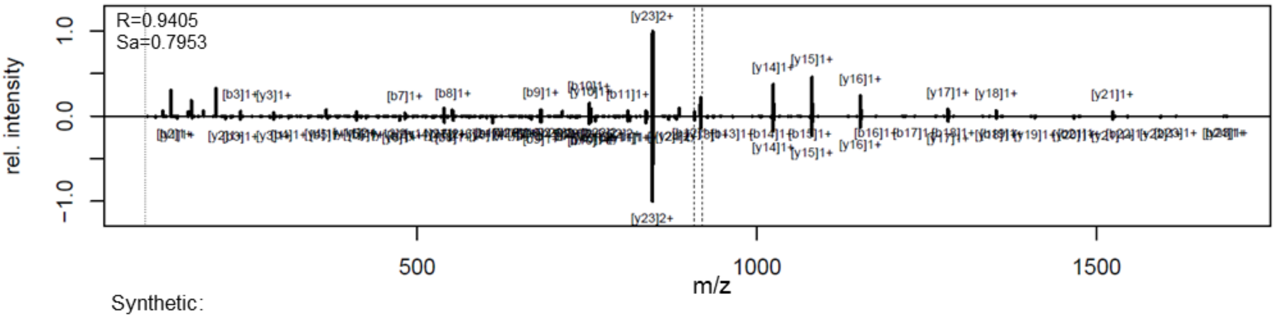
Sample:
01348_F01_P013679_S00_N06_R2 FTMS + c NSI d Full ms2 641.33@hcd25.00 [132.33−1985.00] 18931 125.33



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 640.9941@hcd28.00 [100.00−1933.00] 30212 148.26

AAPAGGGAEAGPGGGPGGAGGAAAK

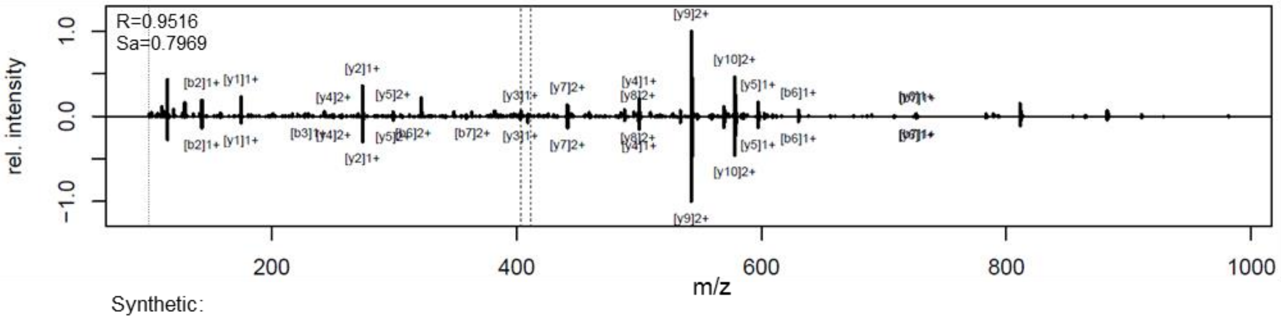
Sample:
01697_E01_P018020_S00_N05_R2 FTMS + c NSI d Full ms2 917.45@hcd25.00 [100.00−1890.00] 12718 55.16



Synthetic:
02383a_BC4−TUM_unmod_proteogenom_4_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 917.4459@hcd28.00 [100.00−1845.00] 11082 281.16

AASDREPPEVR

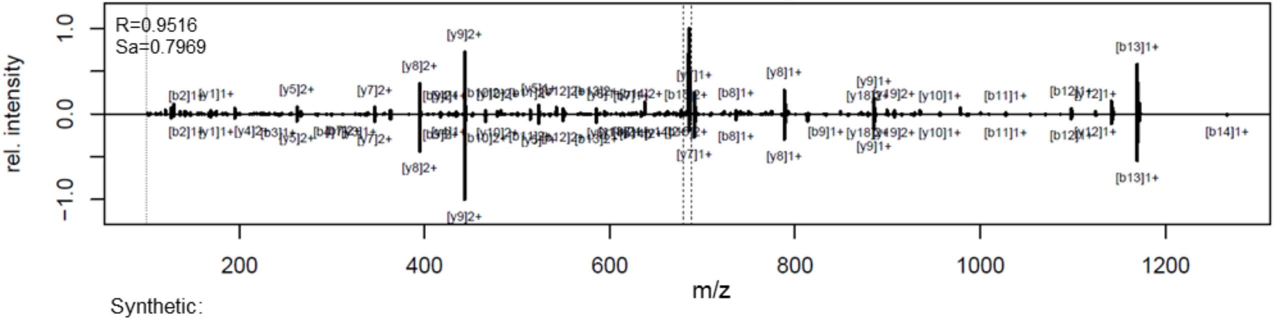
Sample:
01697_B02_P018020_S00_N10_R2 FTMS + c NSI d Full ms2 409.54@hcd25.00 [100.00−1275.00] 10835 62.18



Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 409.5422@hcd28.00 [100.00−1239.00] 8648 169.82

AAVGPARLPPPAAPPPAPPPGR

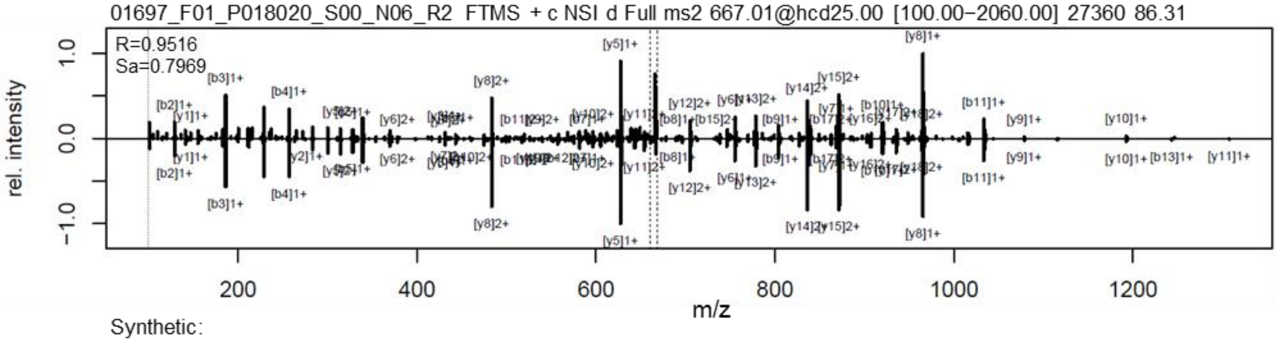
Sample:
01698_A01_P018021_S00_N01_R1 FTMS + c NSI d Full ms2 685.73@hcd25.00 [100.00−2120.00] 31219 51.72



Synthetic:
02383a_BC3-TUM_unmod_proteogenom_3_01_01-DDA−1h−R1 FTMS + c NSI d Full ms2 685.7285@hcd28.00 [100.00−2000.00] 30730 100.02

AGGAADM(ox)TDNIPLQPVRQK

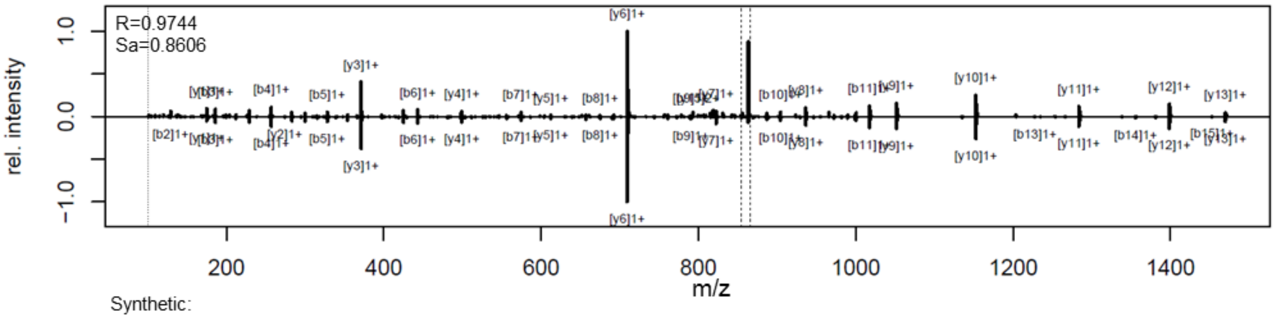
Sample:
01697_F01_P018020_S00_N06_R2 FTMS + c NSI d Full ms2 667.01@hcd25.00 [100.00−2060.00] 27360 86.31



Synthetic:
02383a_BC3-TUM_unmod_proteogenom_3_01_01-DDA−1h−R1 FTMS + c NSI d Full ms2 667.0085@hcd28.00 [100.00−2000.00] 26628 241.08

AGGAADMTDNIPLQPVR

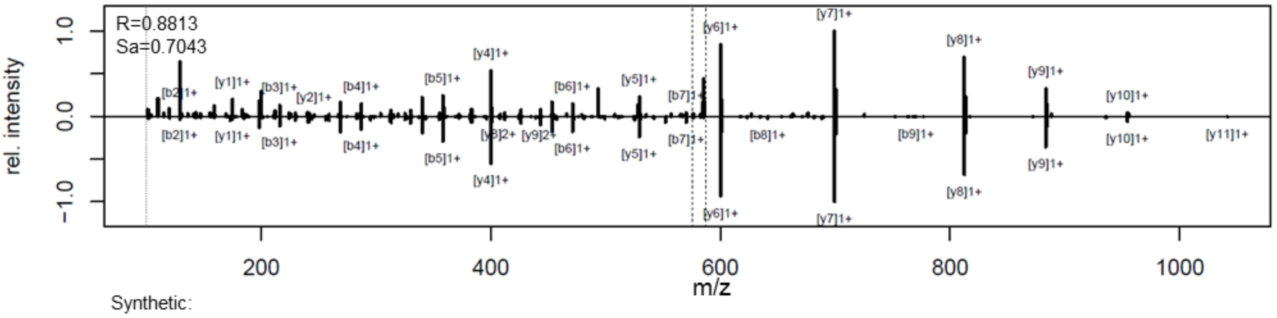
Sample:
01698_G01_P018021_S00_N07_R1 FTMS + c NSI d Full ms2 863.94@hcd25.00 [100.00−1780.00] 40898 112.65



Synthetic:
02383a_BC3-TUM_unmod_proteogenom_3_01_01-DDA−1h−R1 FTMS + c NSI d Full ms2 863.4351@hcd28.00 [100.00−1737.00] 39475 264.92

AGSAALVAEPGAR

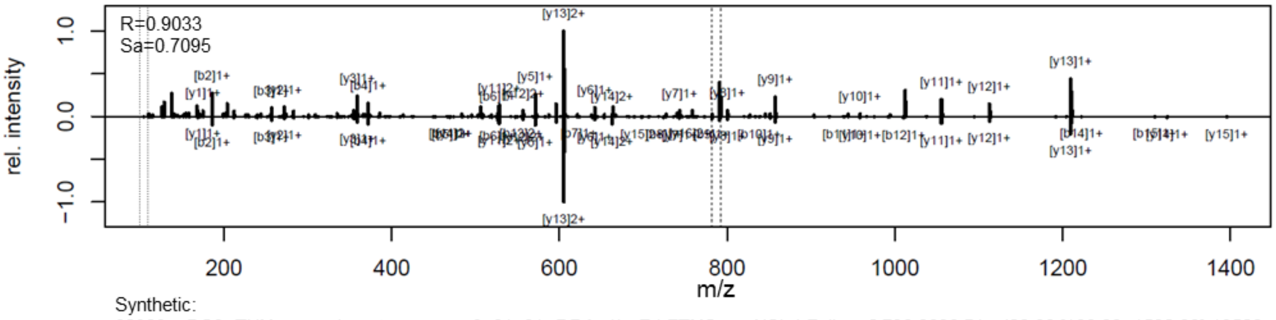
Sample:
01698_F01_P018021_S00_N06_R1 FTMS + c NSI d Full ms2 585.32@hcd25.00 [100.00−1215.00] 22524 57.33



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 585.3168@hcd28.00 [100.00−1181.00] 23075 152.4

ANADPGPTGGTAPDSPR

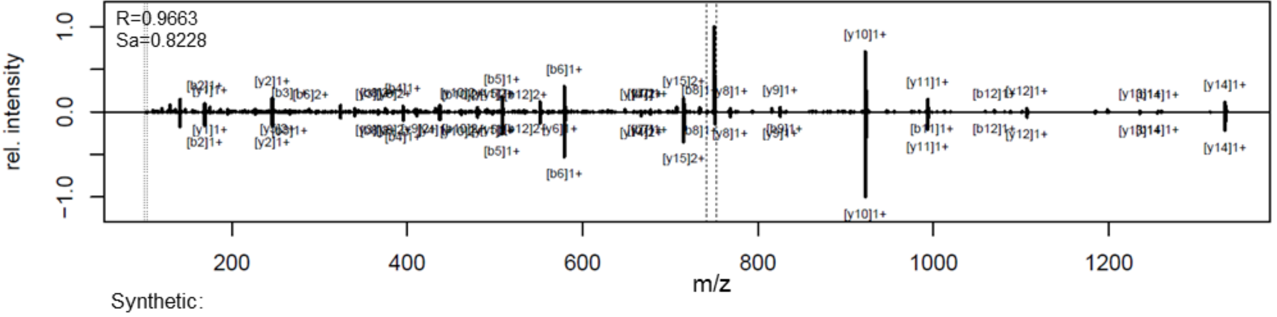
Sample:
01697_A02_P018020_S00_N09_R2 FTMS + c NSI d Full ms2 790.87@hcd25.00 [100.00−1635.00] 13478 62.47



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 790.8690@hcd28.00 [100.00−1592.00] 13528 219.24

APPELAPGPPAAADAR

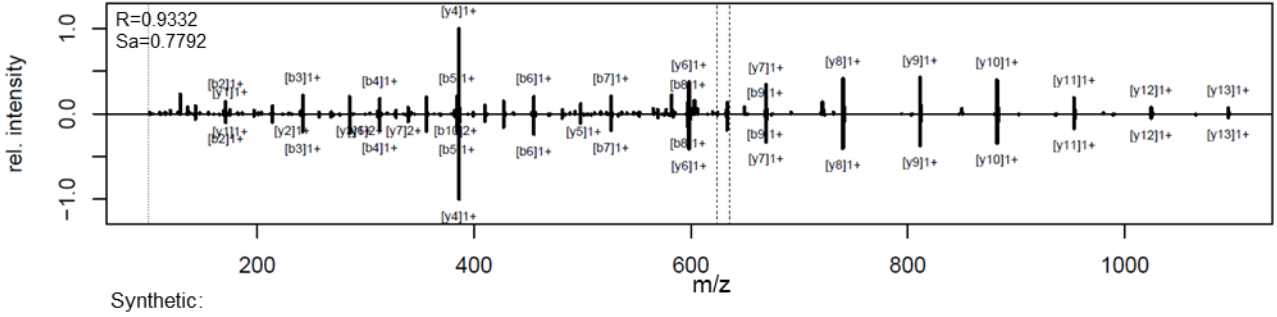
Sample:
01284_A02_P013188_B00_N09_R1 FTMS + c NSI d Full ms2 750.89@hcd25.00 [103.33−1550.00] 18542 70.11



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 750.8944@hcd28.00
[100.00−1512.00] 28704 176.42

AVAAAAAAAPDPGGR

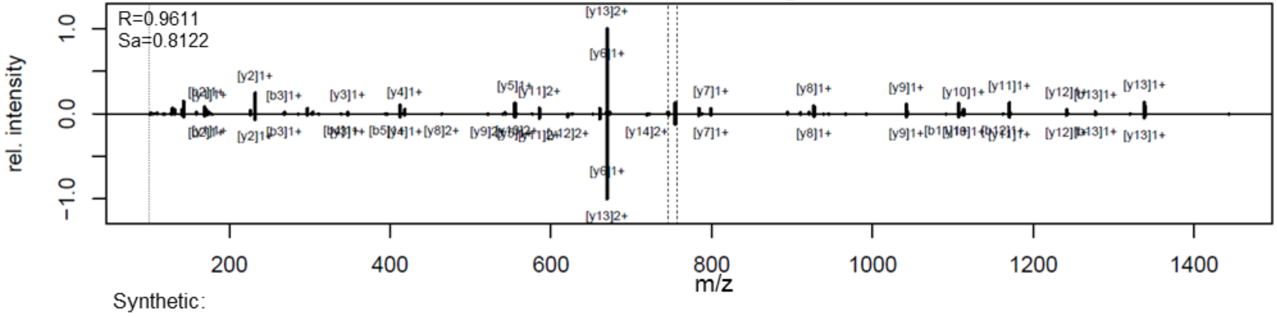
Sample:
01753_F01_P018442_S00_N06_R1 FTMS + c NSI d Full ms2 633.33@hcd25.00 [100.00−1310.00] 13477 65.84



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 633.3334@hcd28.00
[100.00−1277.00] 17717 100.11

AVPAGADQEDHADGR

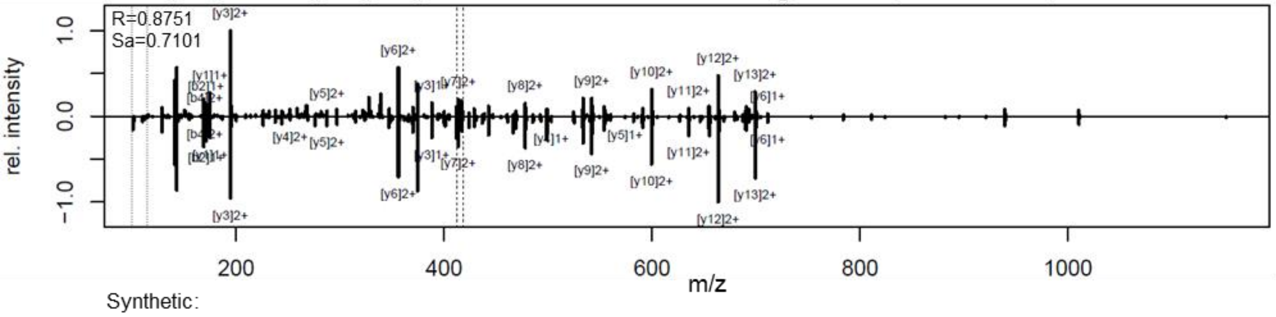
Sample:
01753_B03_P018442_S00_N18_R1 FTMS + c NSI d Full ms2 754.84@hcd25.00 [100.00−1560.00] 7527 89.49



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 754.8401@hcd28.00
[100.00−1520.00] 8029 118.21

AVPAGADQEDHADGRR

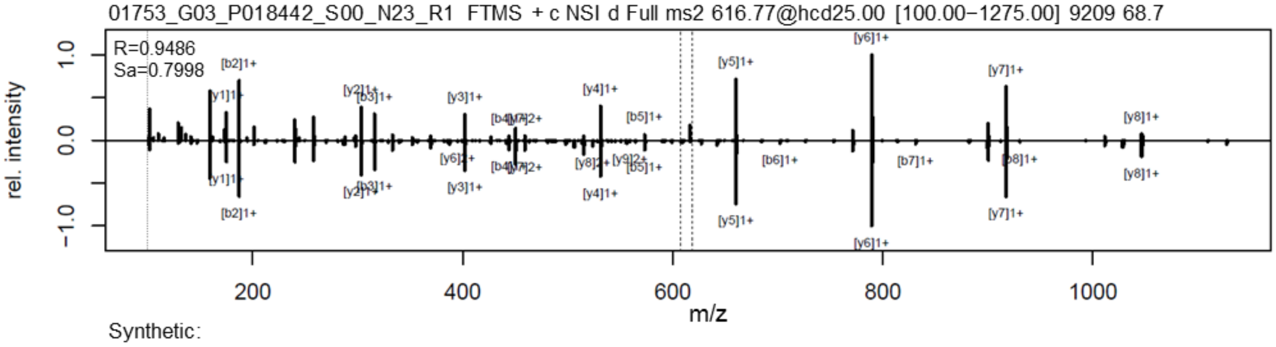
Sample:
01277_F02_P013129_S00_N14_R1 FTMS + c NSI d Full ms2 416.95@hcd25.00 [114.67−1720.00] 3293 61.09



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 416.9488@hcd28.00
[100.00−1678.00] 6445 130.72

DAEQEEEVQR

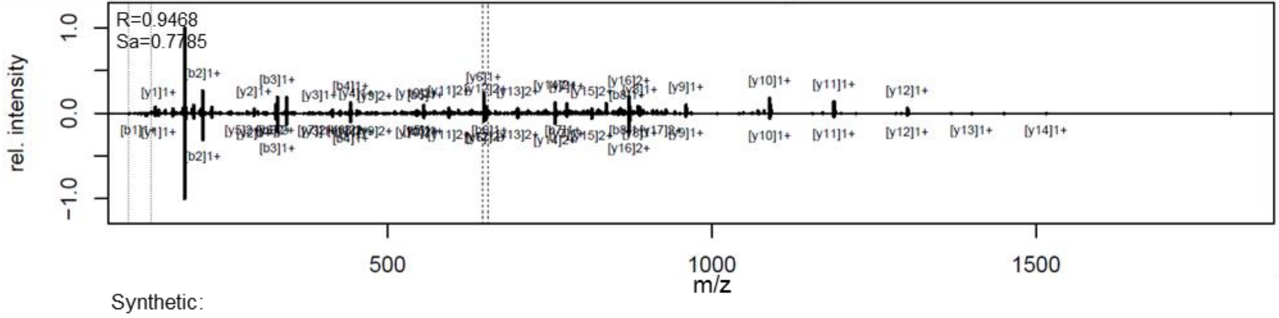
Sample:
01753_G03_P018442_S00_N23_R1 FTMS + c NSI d Full ms2 616.77@hcd25.00 [100.00−1275.00] 9209 68.7



Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 616.7730@hcd28.00
[100.00−1244.00] 8383 257.4

DVDLDVLVEALKPVGTFK

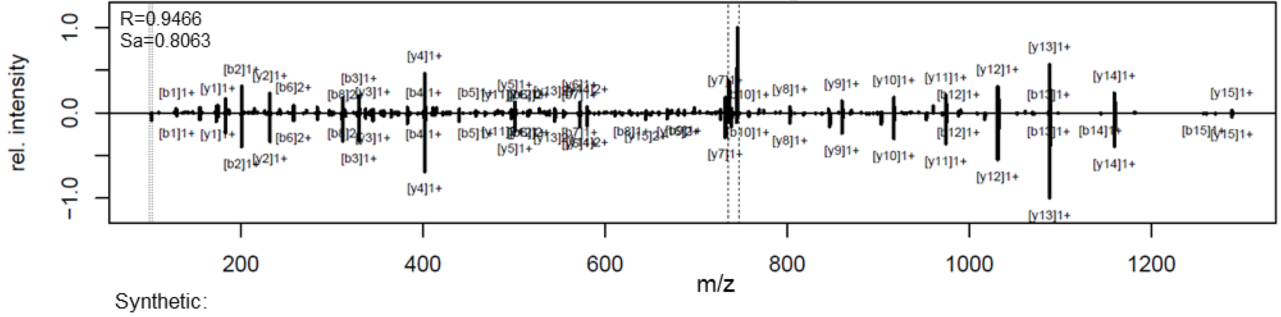
Sample:
01320_H02_P013559_S00_N16_R1 FTMS + c NSI d Full ms2 653.70@hcd25.00 [134.67−2020.00] 51987 91.44



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 653.7014@hcd28.00
[100.00−1972.00] 64622 161.86

EAEAGGGGGACAVGLGR

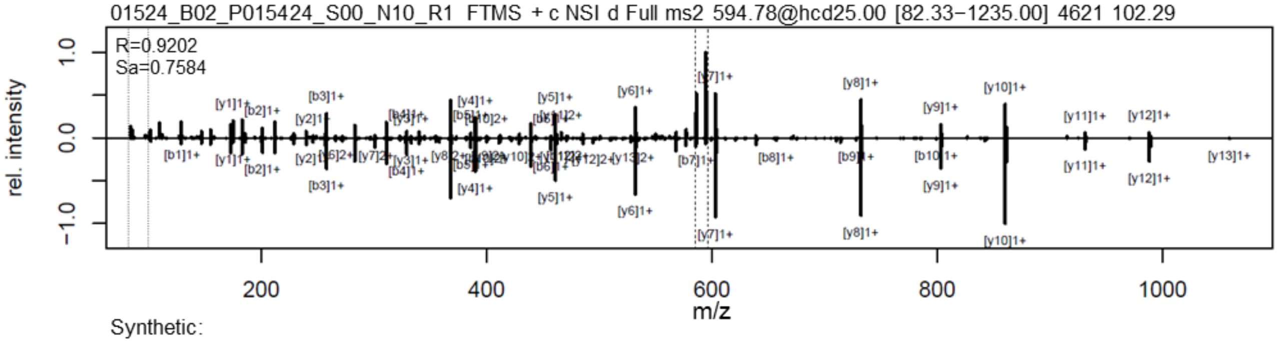
Sample:
01076_C02_P010694_S00_N11_R1 FTMS + c NSI d Full ms2 744.85@hcd25.00 [102.67−1540.00] 13028 98.26



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 744.8469@hcd28.00
[100.00−1500.00] 21638 260.96

EAGAGAEAAAGSAR

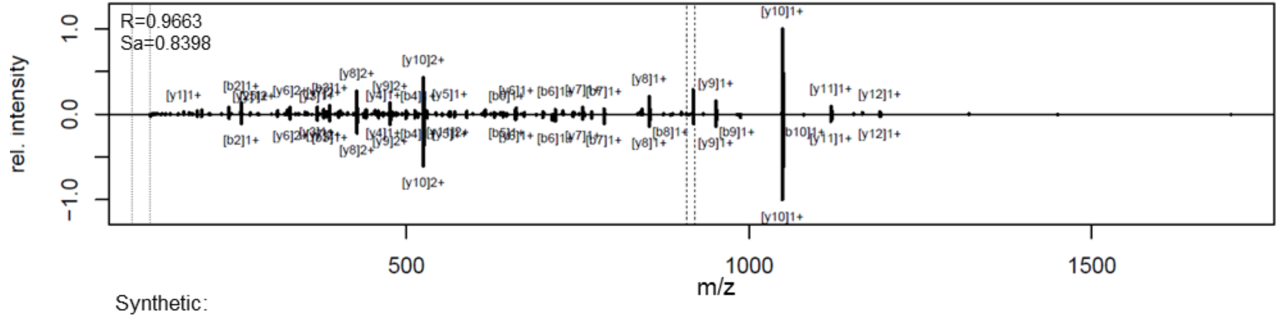
Sample:
01524_B02_P015424_S00_N10_R1 FTMS + c NSI d Full ms2 594.78@hcd25.00 [82.33−1235.00] 4621 102.29



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 594.7842@hcd28.00
[100.00−1200.00] 8489 195.94

EEEEEAAPPPPPPPPPR

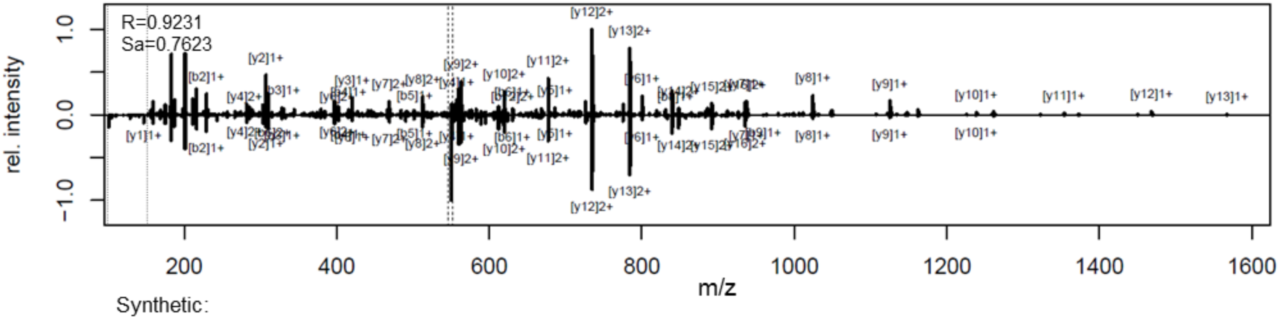
Sample:
01343_C03_P013674_S00_N19_R2 FTMS + c NSI d Full ms2 918.45@hcd25.00 [126.33−1895.00] 12840 55.92



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 918.9474@hcd28.00
[100.00−1848.00] 22367 101.2

EVVSSQVDDLTSHNEHLCK

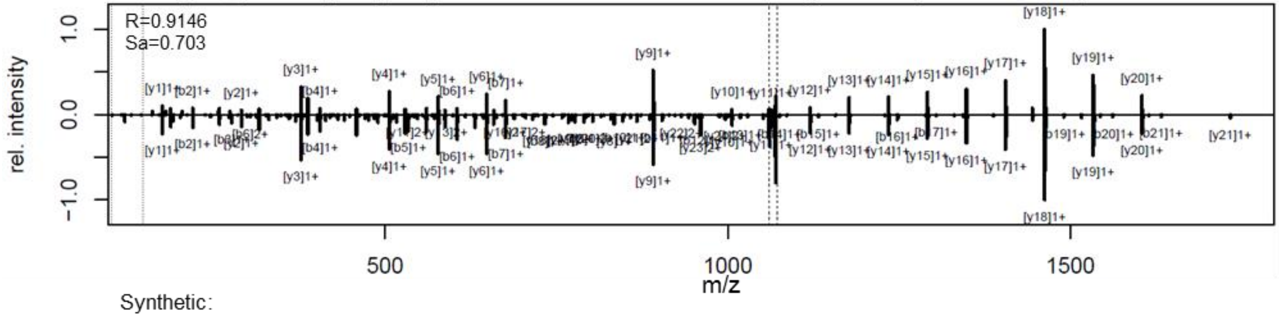
Sample:
01320_B03_P013559_S00_N18_R1 FTMS + c NSI d Full ms2 550.26@hcd25.00 [151.00−2265.00] 17779 123.82



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 550.2620@hcd28.00
[100.00−2000.00] 26586 190.55

FAGEEAAGGGGGGGGDGGEAAESDR

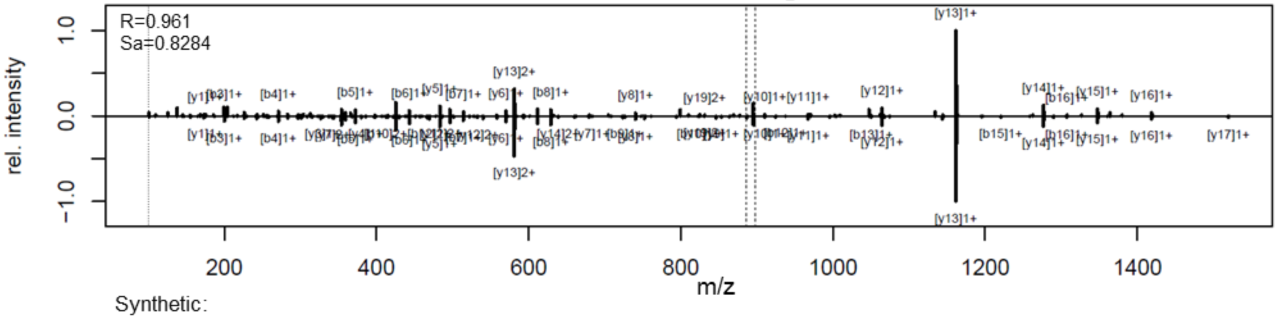
Sample:
01284_H03_P013188_B00_N24_R1 FTMS + c NSI d Full ms2 1069.94@hcd25.00 [146.67−2200.00] 9200 206.13



Synthetic:
02383a_BC4−TUM_unmod_proteogenom_4_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1069.9390@hcd28.00 [100.00−2000.00] 14920 341.65

GAAATAADPPGGLGLSPGPGR

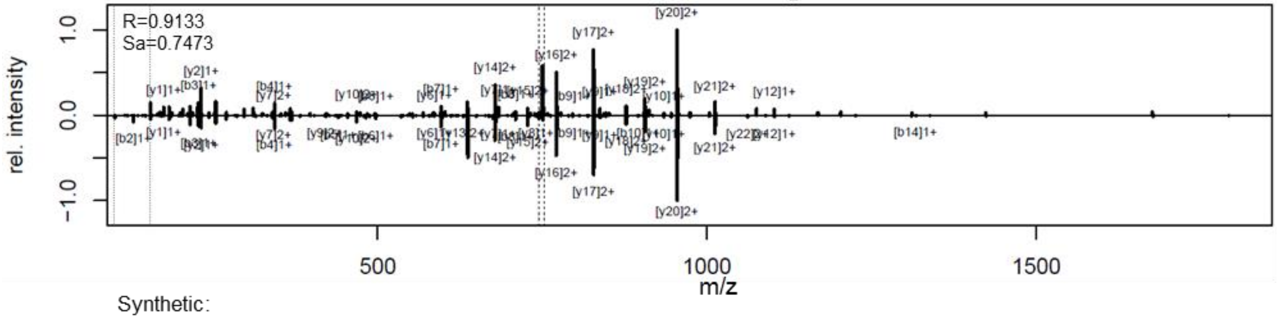
Sample:
01753_F01_P018442_S00_N06_R1 FTMS + c NSI d Full ms2 895.46@hcd25.00 [100.00−1845.00] 24048 72.47



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 895.4653@hcd28.00 [100.00−1801.00] 33019 168.99

GAVDPGVLSVPRPAPPGPPAADGR

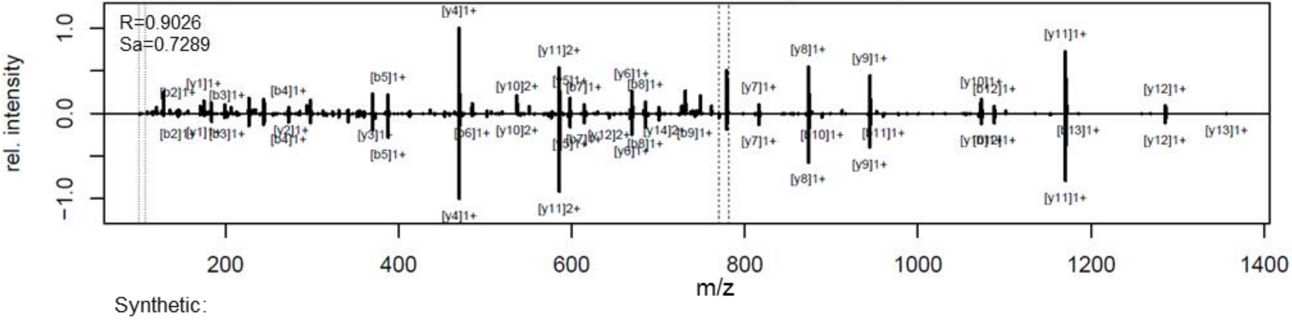
Sample:
01284_F01_P013188_B00_N06_R1 FTMS + c NSI d Full ms2 750.74@hcd25.00 [154.33−2315.00] 31097 70.11



Synthetic:
02383a_BC4−TUM_unmod_proteogenom_4_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 751.0729@hcd28.00 [100.00−2000.00] 34536 174.7

GSGADPEAGFAQPPTR

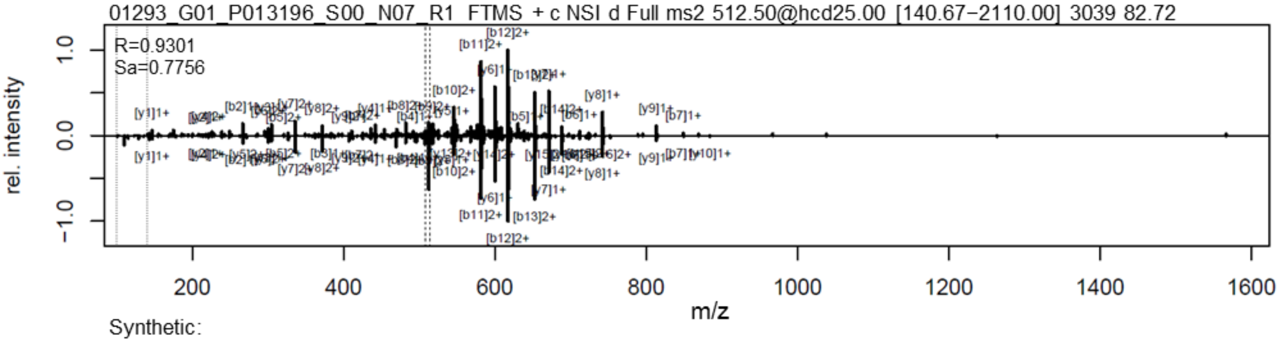
Sample:
01323_C02_P013562_S00_N11_R1 FTMS + c NSI d Full ms2 779.87@hcd25.00 [107.33−1610.00] 17035 61.63



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 779.3688@hcd28.00
[100.00−1569.00] 25710 235.35

HELNM(ox)AHNAGAAAAAGTHSAK

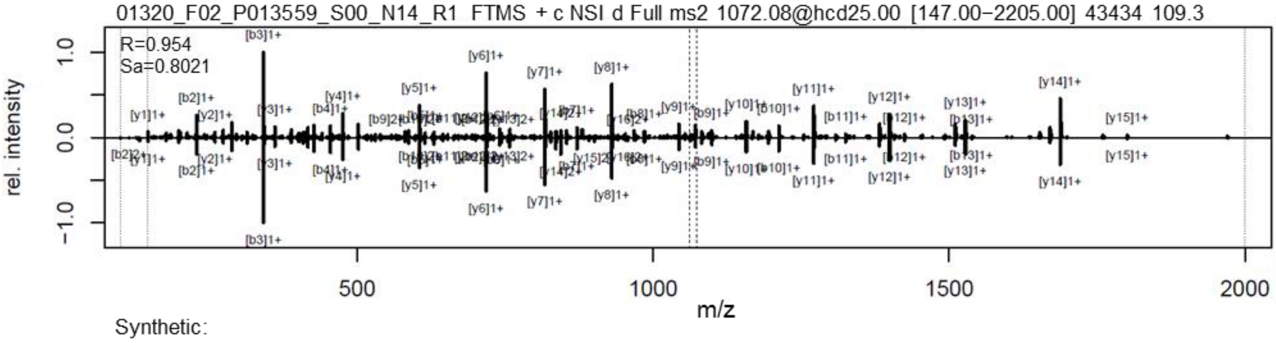
Sample:
01293_G01_P013196_S00_N07_R1 FTMS + c NSI d Full ms2 512.50@hcd25.00 [140.67−2110.00] 3039 82.72



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 512.4957@hcd28.00
[100.00−2000.00] 7650 108.08

IDLICQQNNIIVLEDTIK

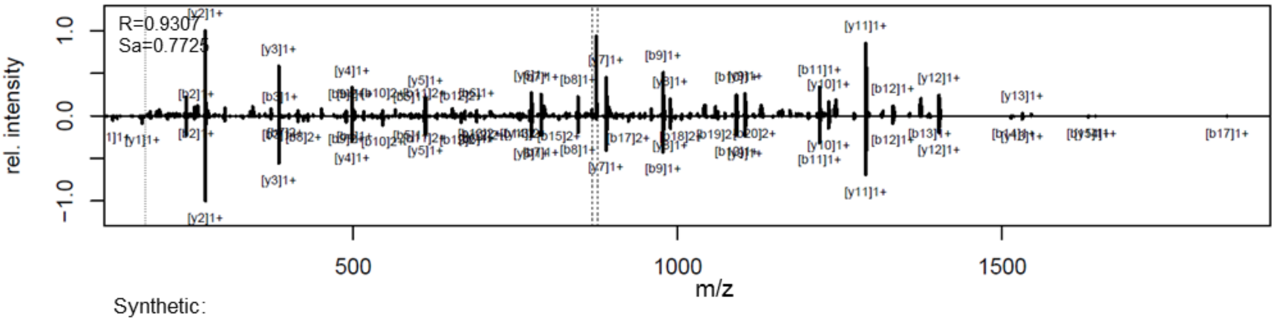
Sample:
01320_F02_P013559_S00_N14_R1 FTMS + c NSI d Full ms2 1072.08@hcd25.00 [147.00−2205.00] 43434 109.3



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1071.5784@hcd28.00
[100.00−2000.00] 56581 258.38

KEEDSTVGMLQIGEDVDYLLIPR

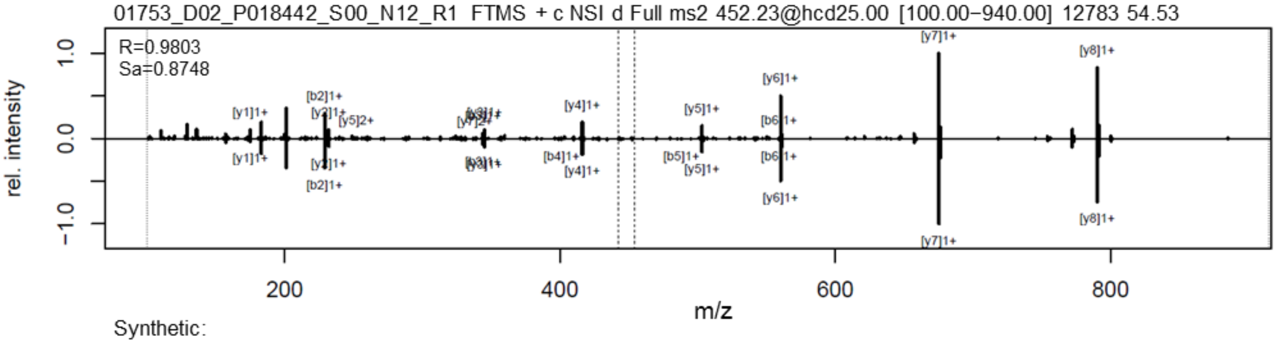
Sample:
01280_H03_P013164_S00_N24_R1 FTMS + c NSI d Full ms2 874.78@hcd25.00 [179.67−2695.00] 52165 71.29



Synthetic:
02383a_BC4−TUM_unmod_proteogenom_4_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 874.4459@hcd28.00 [100.00−2000.00] 60379 228.29

LDDGSALGR

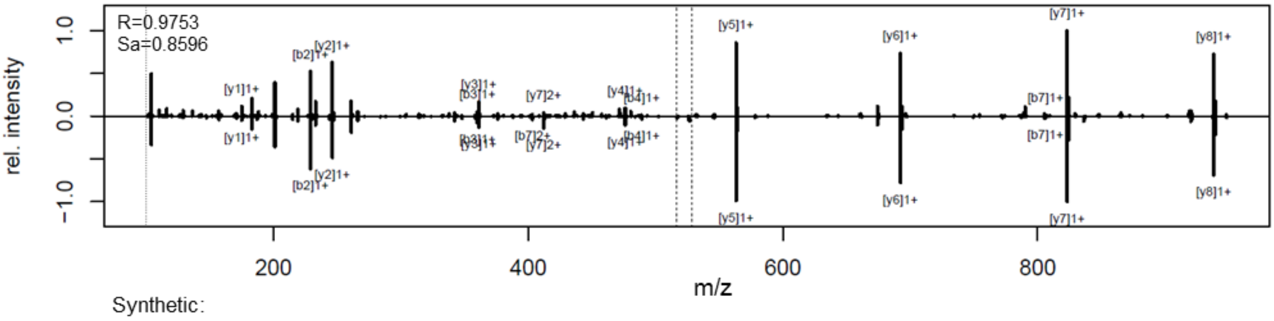
Sample:
01753_D02_P018442_S00_N12_R1 FTMS + c NSI d Full ms2 452.23@hcd25.00 [100.00−940.00] 12783 54.53



Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 452.2298@hcd28.00 [100.00−915.00] 16766 141.02

LDMESDDAR

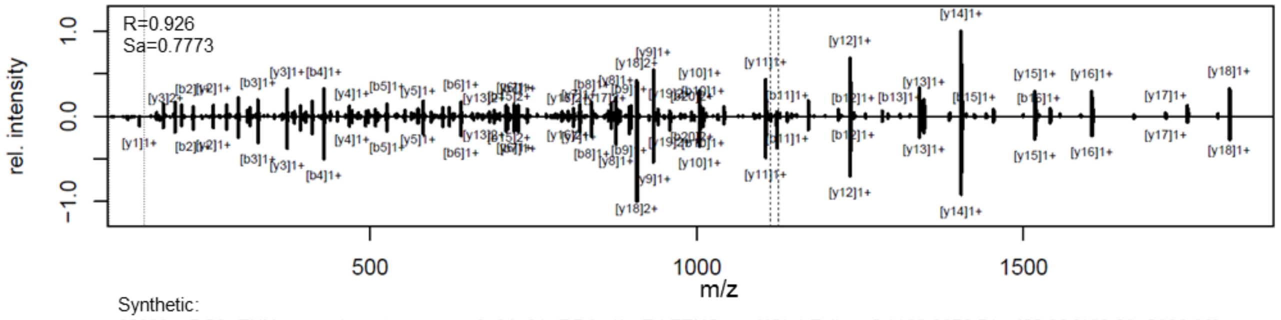
Sample:
01753_E03_P018442_S00_N21_R1 FTMS + c NSI d Full ms2 526.22@hcd25.00 [100.00−1095.00] 13298 67.02



Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 526.2215@hcd28.00 [100.00−1063.00] 15688 109.66

LDVTPLSLGIETAGGVM(ox)TVLIK

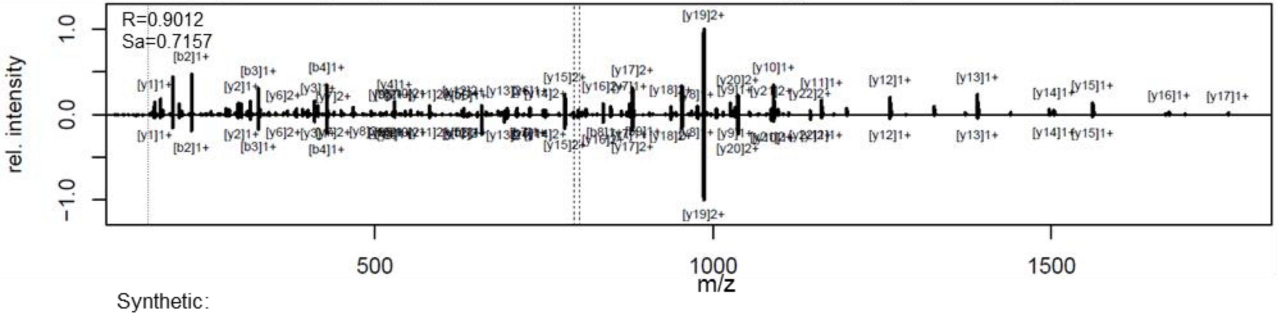
Sample:
01270_B02_P013114_S00_N10_R1 FTMS + c NSI d Full ms2 1122.64@hcd25.00 [154.00−2310.00] 50640 132.98



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1122.6378@hcd28.00 [100.00−2000.00] 64488 183.24

LDVTPLSLGIETAGGVM(ox)TVLIKR

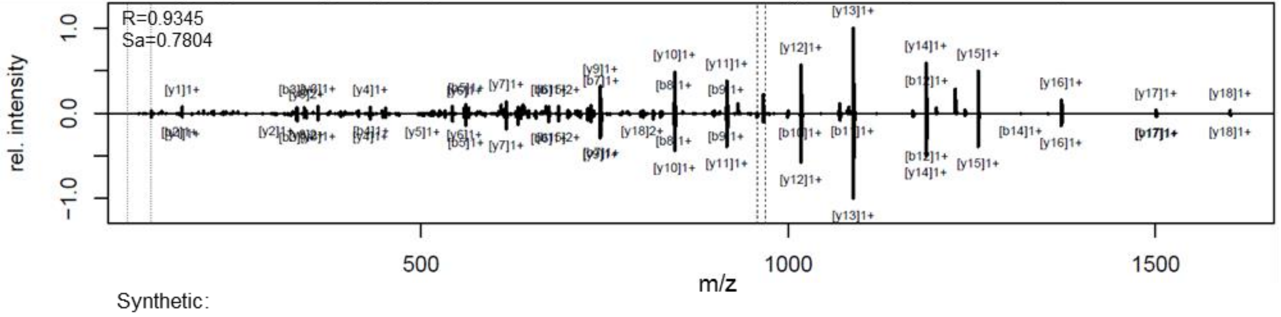
Sample:
01284_G01_P013188_B00_N07_R1 FTMS + c NSI d Full ms2 800.79@hcd25.00 [164.67−2470.00] 53810 189



Synthetic:
02383a_BC4−TUM_unmod_proteogenom_4_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 800.7929@hcd28.00 [100.00−2000.00] 58477 161.2

LGCTELAVATATGAGGAAGQR

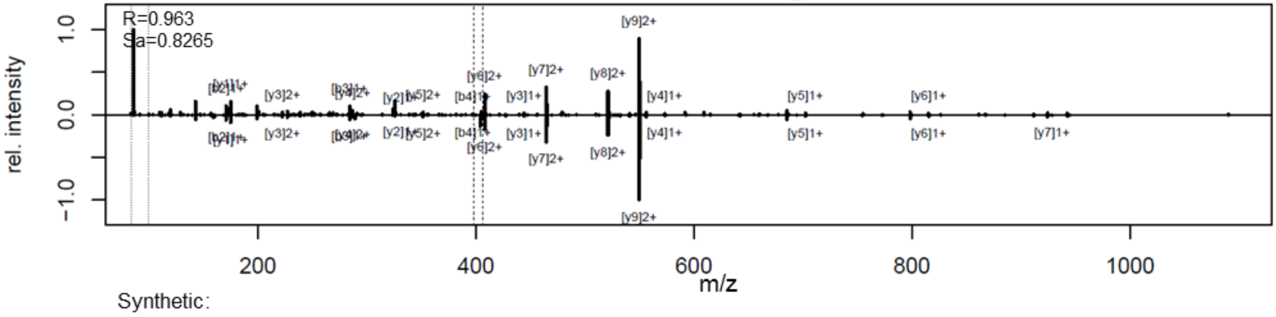
Sample:
01226_F01_P012502_S00_N06_R1 FTMS + c NSI d Full ms2 966.49@hcd25.00 [132.67−1990.00] 26180 95.06



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 966.9875@hcd28.00 [100.00−1944.00] 34338 173.21

LGLLLEELRR

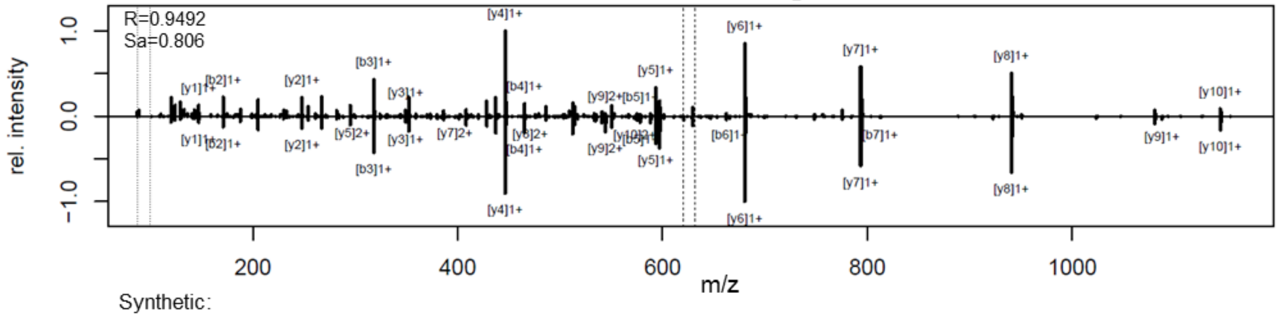
Sample:
01283_A02_P013187_S00_N09_R1 FTMS + c NSI d Full ms2 404.59@hcd25.00 [84.00−1260.00] 31201 54.01



Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 404.5871@hcd28.00 [100.00−1224.00] 52146 150.35

LGM(ox)FLSFPTTK

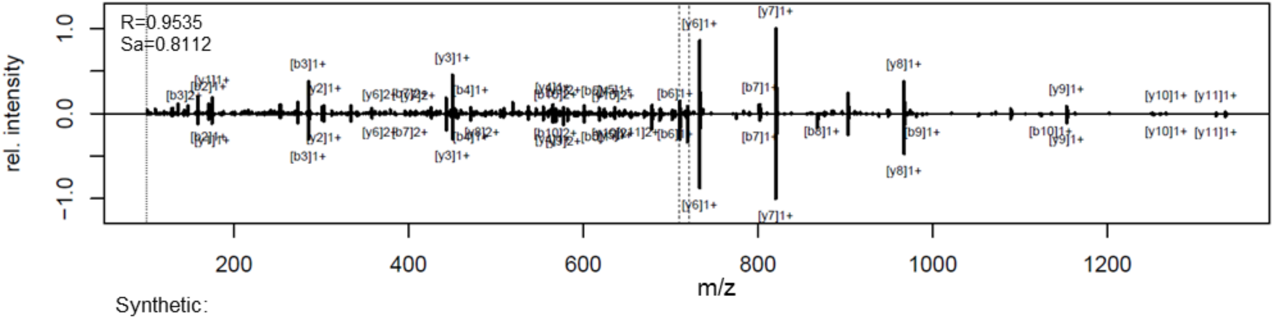
Sample:
01294_D01_P013197_S00_N04_R1 FTMS + c NSI d Full ms2 629.83@hcd25.00 [87.00−1305.00] 37394 62.01



Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 629.3305@hcd28.00 [100.00−1269.00] 50380 174.4

LGNWM(ox)SPADFQR

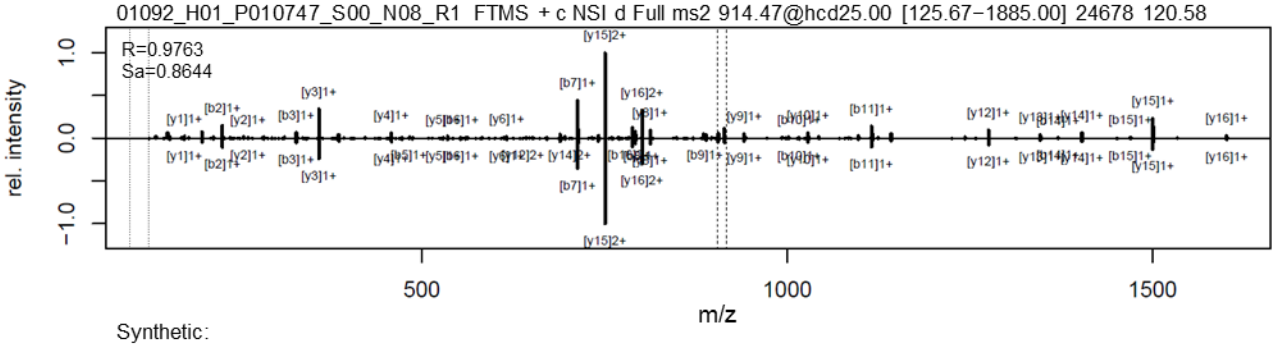
Sample:
01226_A02_P012502_S00_N09_R1 FTMS + c NSI d Full ms2 719.33@hcd25.00 [99.00−1485.00] 25928 70.37



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 719.3323@hcd28.00 [100.00−1449.00] 38515 185.37

LLTPGACSSEVPSAVPSR

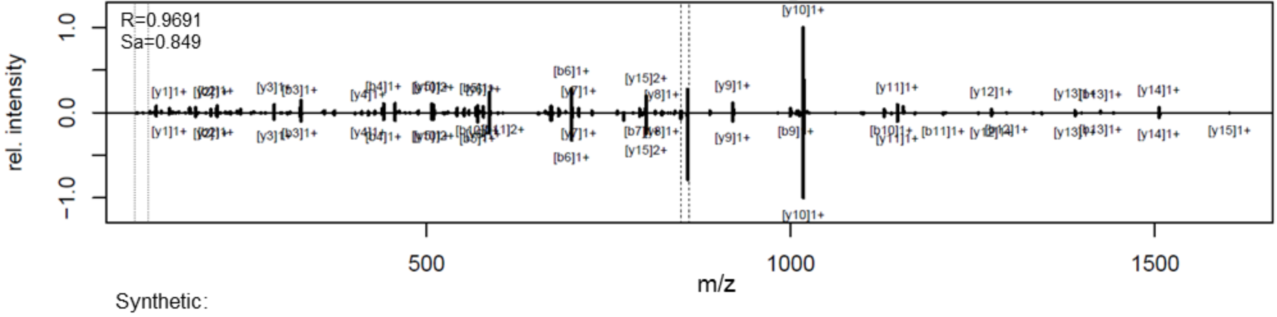
Sample:
01092_H01_P010747_S00_N08_R1 FTMS + c NSI d Full ms2 914.47@hcd25.00 [125.67−1885.00] 24678 120.58



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 914.4696@hcd28.00
[100.00−1839.00] 33851 161.2

LPDDEEPPNMASESGK

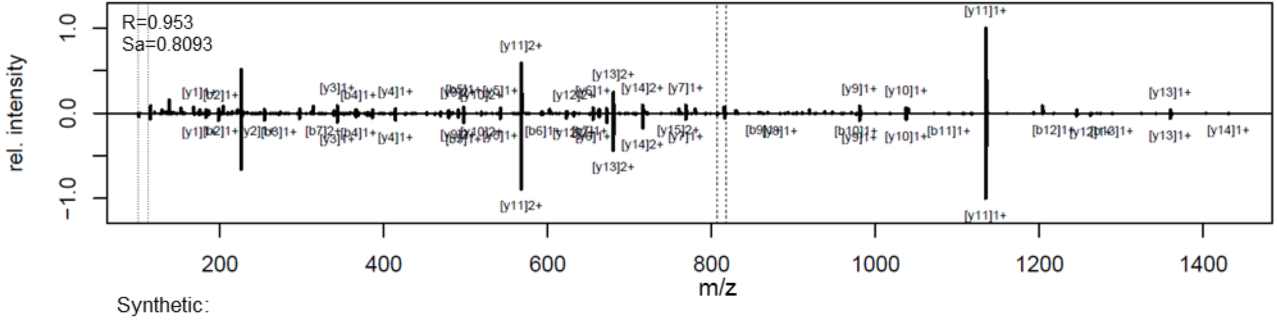
Sample:
01226_E03_P012502_S00_N21_R1 FTMS + c NSI d Full ms2 858.37@hcd25.00 [118.00−1770.00] 16150 101.06



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 858.3750@hcd28.00
[100.00−1727.00] 24709 181.72

LSAPQPGPDILQAPAR

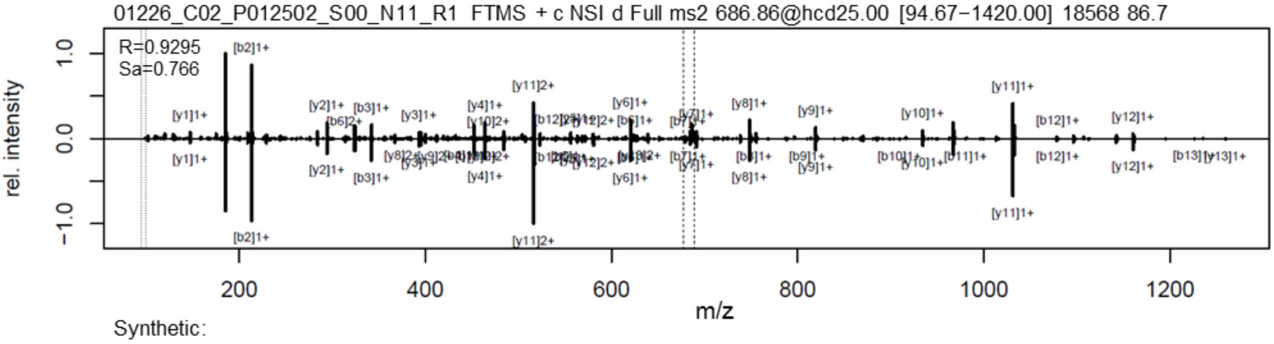
Sample:
01283_F01_P013187_S00_N06_R1 FTMS + c NSI d Full ms2 815.95@hcd25.00 [112.33−1685.00] 26533 78.08



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 815.9498@hcd28.00
[100.00−1642.00] 39211 182.09

LVEPDAGAGVAVM(ox)K

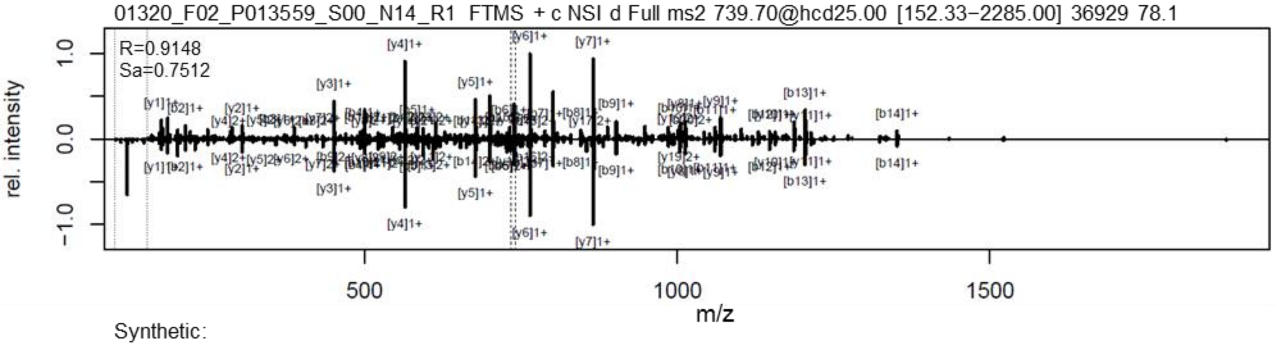
Sample:
01226_C02_P012502_S00_N11_R1 FTMS + c NSI d Full ms2 686.86@hcd25.00 [94.67−1420.00] 18568 86.7



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 686.8623@hcd28.00 [100.00−1384.00] 27497 108.58

LVFHSFGGGTGSGFTSLLM(ox)ER

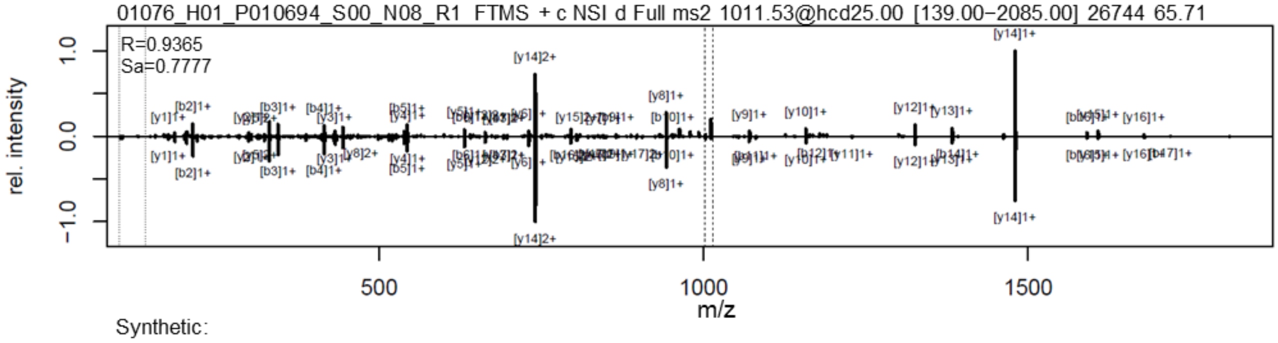
Sample:
01320_F02_P013559_S00_N14_R1 FTMS + c NSI d Full ms2 739.70@hcd25.00 [152.33−2285.00] 36929 78.1



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 739.7003@hcd28.00 [100.00−2000.00] 51239 131.02

LVMAQPGPASQPDVSLQQR

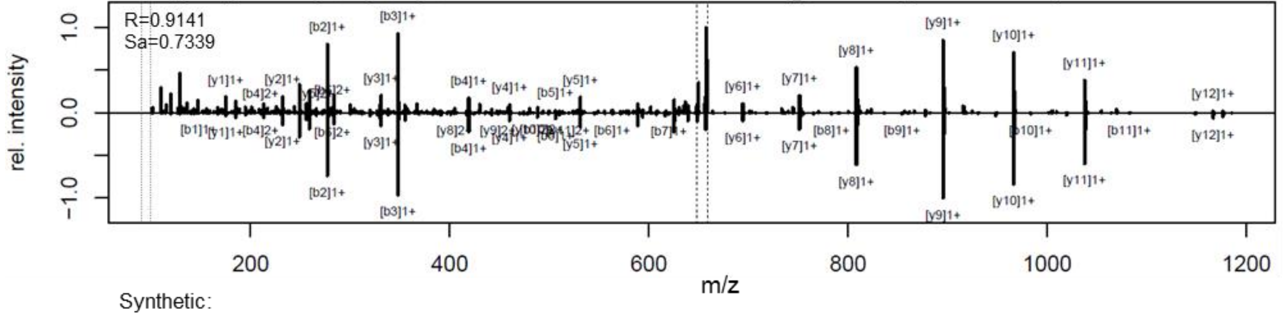
Sample:
01076_H01_P010694_S00_N08_R1 FTMS + c NSI d Full ms2 1011.53@hcd25.00 [139.00−2085.00] 26744 65.71



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1012.0276@hcd28.00 [100.00−2000.00] 34316 203.56

M(ox)EAASGGYAEVGR

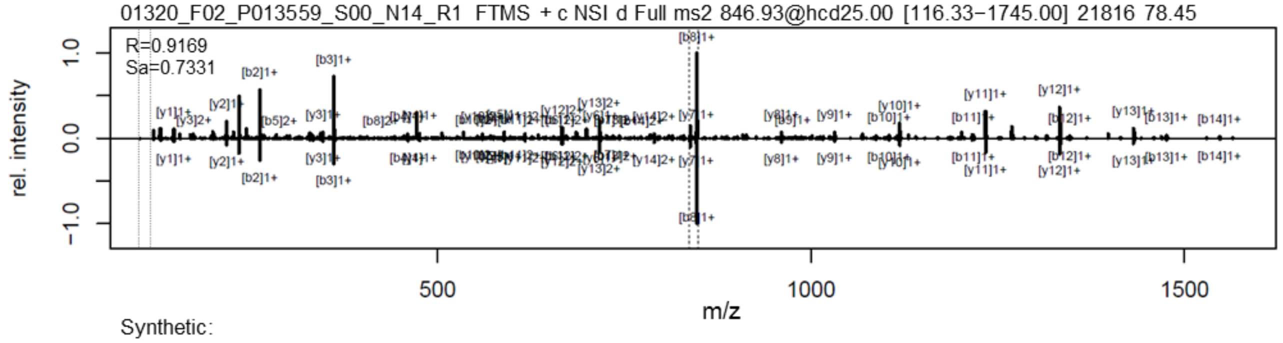
Sample:
01308_E02_P013387_B00_N13_R1 FTMS + c NSI d Full ms2 657.29@hcd25.00 [90.67−1360.00] 8387 58.19



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 657.2933@hcd28.00 [100.00−1325.00] 16741 215.51

NFVVDSANKELEEAK

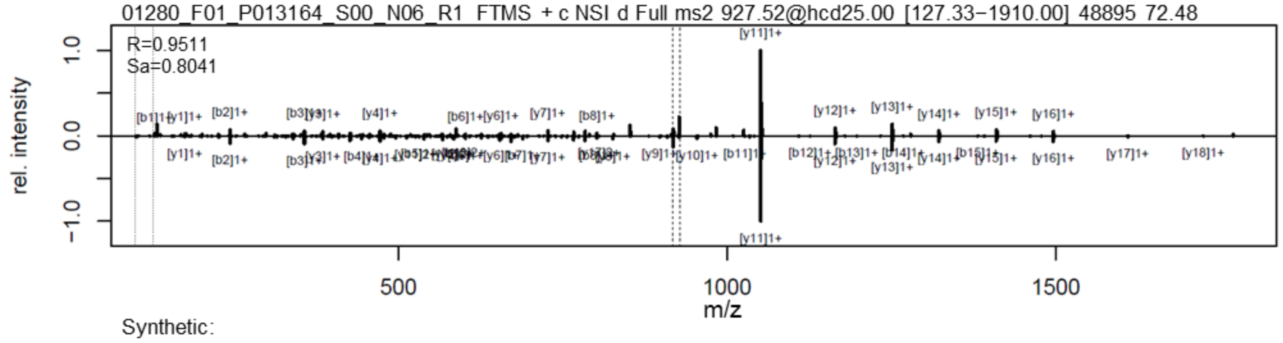
Sample:
01320_F02_P013559_S00_N14_R1 FTMS + c NSI d Full ms2 846.93@hcd25.00 [116.33−1745.00] 21816 78.45



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 657.2933@hcd28.00 [100.00−1325.00] 16741 215.51

QDLSSASLPGAVAALSPLR

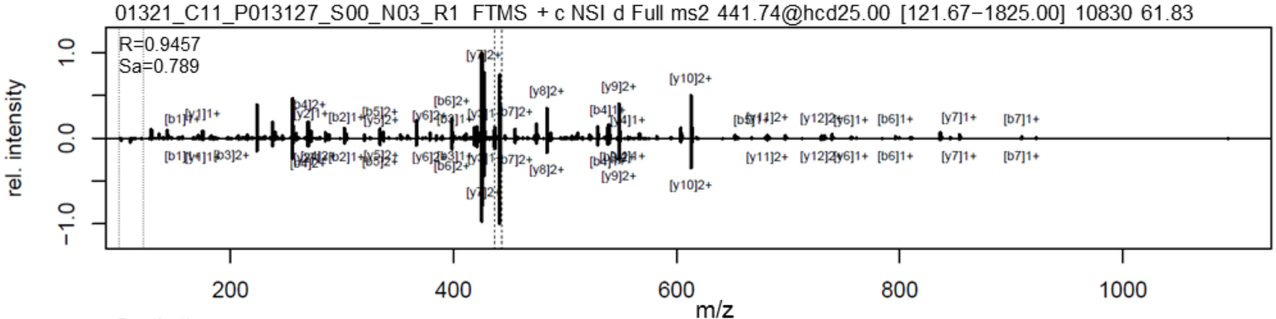
Sample:
01280_F01_P013164_S00_N06_R1 FTMS + c NSI d Full ms2 927.52@hcd25.00 [127.33−1910.00] 48895 72.48



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 927.0103@hcd28.00 [100.00−1865.00] 56118 247

RM(ox)PHEELPSLQRPR

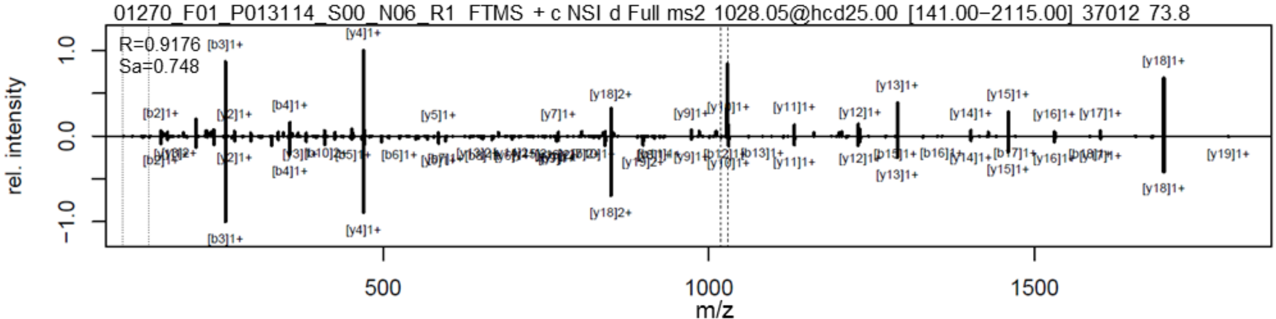
Sample:
01321_C11_P013127_S00_N03_R1 FTMS + c NSI d Full ms2 441.74@hcd25.00 [121.67−1825.00] 10830 61.83



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 441.2349@hcd28.00
[100.00−1775.00] 19876 104.94

SAVVPAAGLGTTGGFALDPTPR

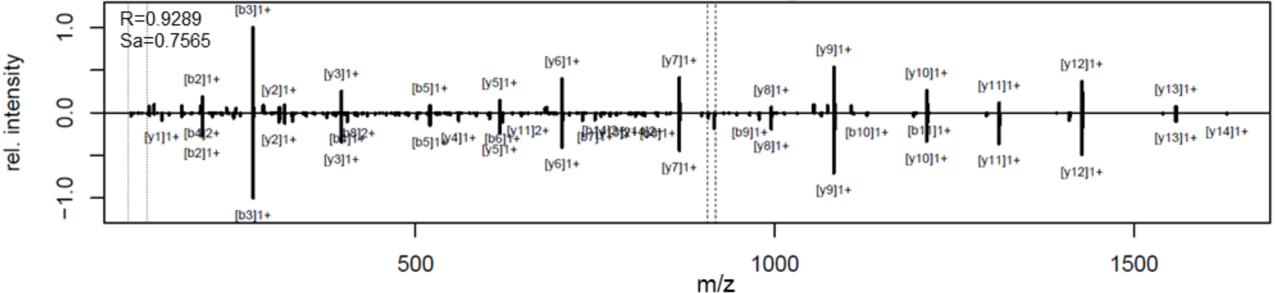
Sample:
01270_F01_P013114_S00_N06_R1 FTMS + c NSI d Full ms2 1028.05@hcd25.00 [141.00−2115.00] 37012 73.8



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1028.0464@hcd28.00
[100.00−2000.00] 48482 222.92

SDAMDTESQYSGYSYK

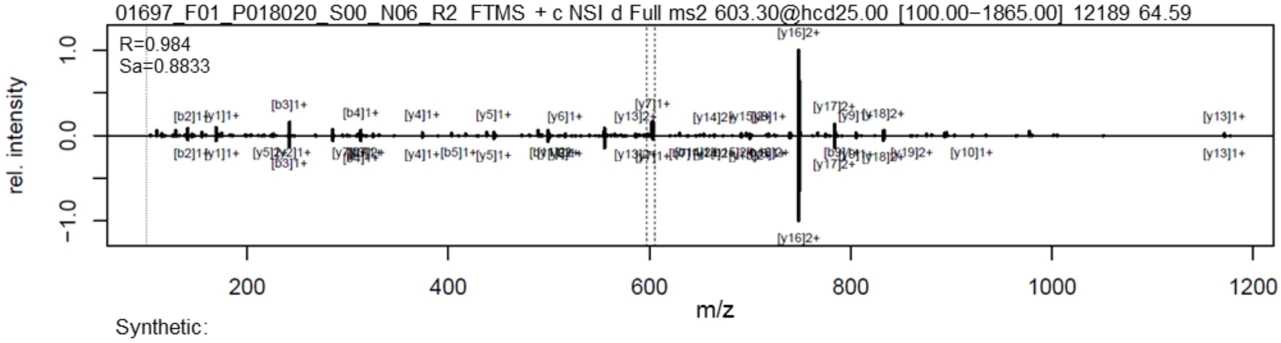
Sample:
01344_C04_P013675_S00_N27_R1 FTMS + c NSI d Full ms2 916.88@hcd25.00 [126.00−1890.00] 12612 108.37



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 916.3720@hcd28.00
[100.00−1843.00] 27132 276.1

SGPAPARPGPDSDSAAGAAR

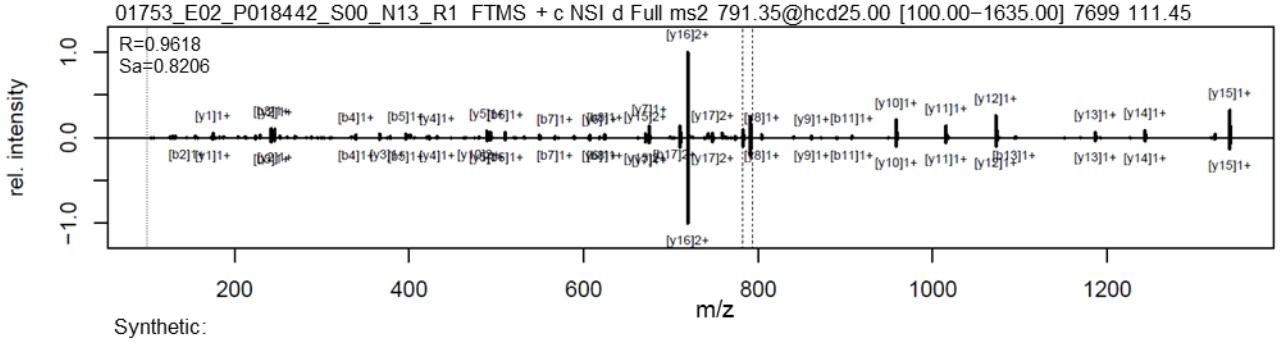
Sample:
01697_F01_P018020_S00_N06_R2 FTMS + c NSI d Full ms2 603.30@hcd25.00 [100.00−1865.00] 12189 64.59



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 603.2944@hcd28.00
[100.00−1820.00] 11417 103.4

SGPPGNGGPGEGEGGEAR

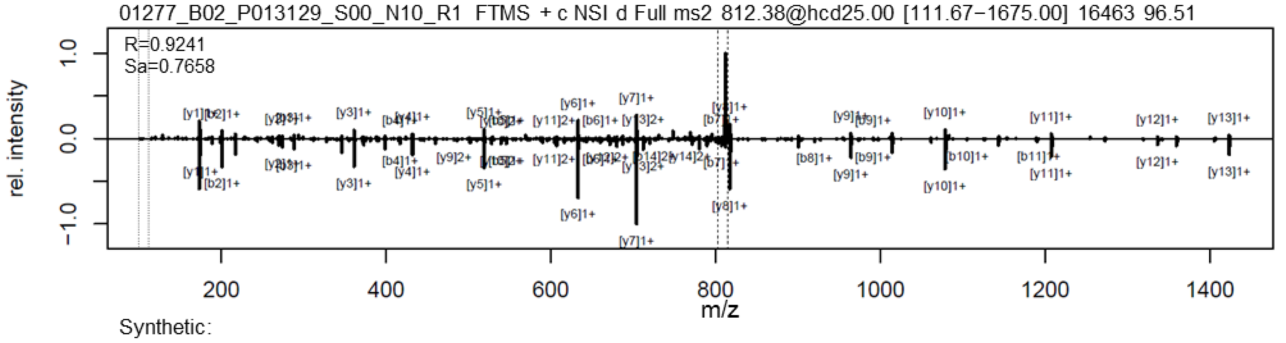
Sample:
01753_E02_P018442_S00_N13_R1 FTMS + c NSI d Full ms2 791.35@hcd25.00 [100.00−1635.00] 7699 111.45



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 791.3486@hcd28.00
[100.00−1593.00] 8810 229.93

SISEENM(ox)LANSASVR

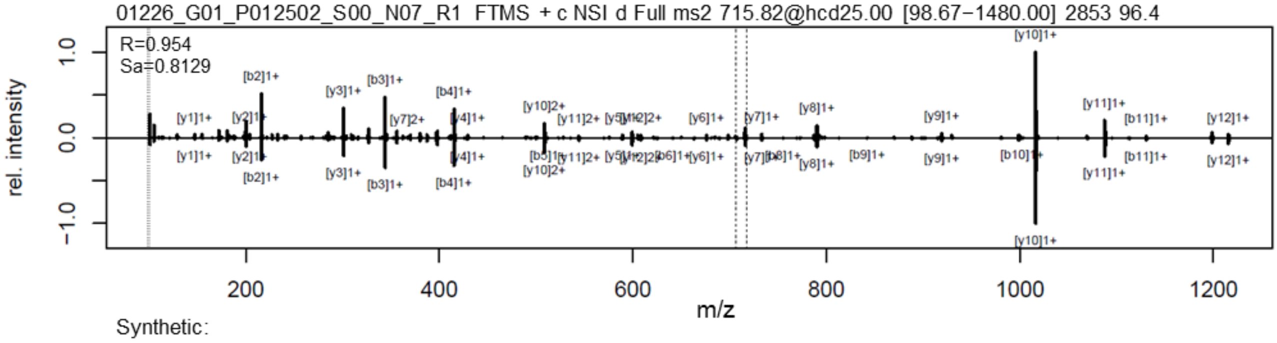
Sample:
01277_B02_P013129_S00_N10_R1 FTMS + c NSI d Full ms2 812.38@hcd25.00 [111.67−1675.00] 16463 96.51



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 812.3860@hcd28.00
[100.00−1635.00] 26687 231.13

SQQAPEGGSCQPGK

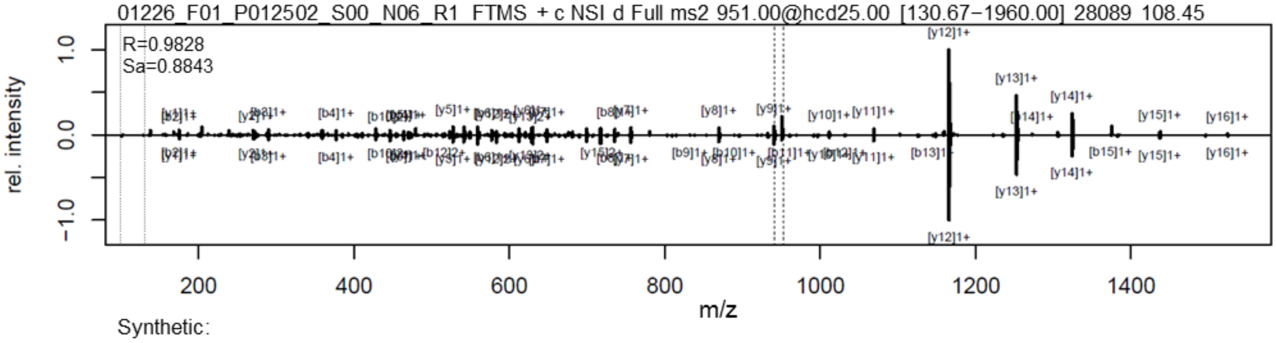
Sample:
01226_G01_P012502_S00_N07_R1 FTMS + c NSI d Full ms2 715.82@hcd25.00 [98.67−1480.00] 2853 96.4



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 715.8200@hcd28.00 [100.00−1442.00] 4793 209.84

SSDSSLASPGAALQTGPVVR

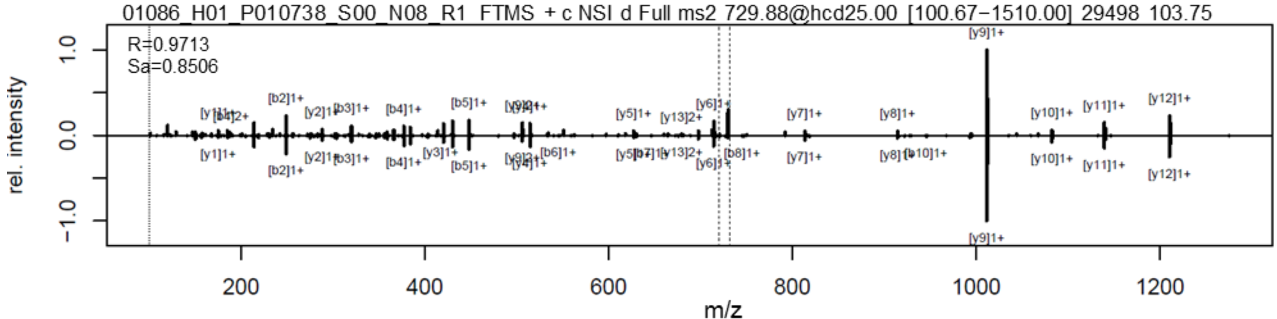
Sample:
01226_F01_P012502_S00_N06_R1 FTMS + c NSI d Full ms2 951.00@hcd25.00 [130.67−1960.00] 28089 108.45



Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 950.9911@hcd28.00 [100.00−1912.00] 36556 195.96

TM(ox)AGAPTVSLPELR

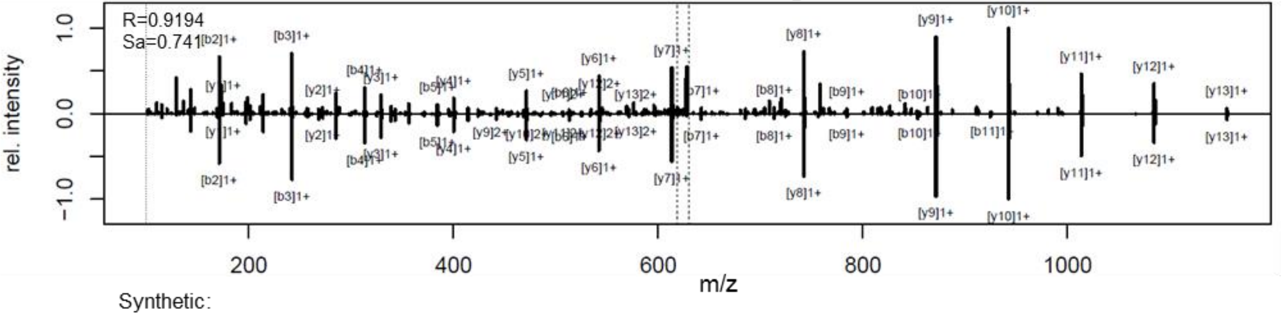
Sample:
01086_H01_P010738_S00_N08_R1 FTMS + c NSI d Full ms2 729.88@hcd25.00 [100.67−1510.00] 29498 103.75



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 729.8849@hcd28.00 [100.00−1470.00] 44529 132.51

VAAAAEEAAAAGPR

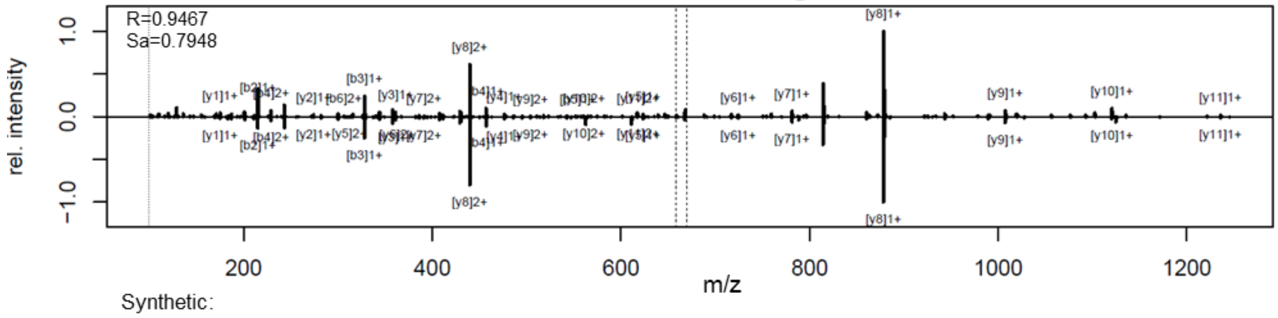
Sample:
01697_B02_P018020_S00_N10_R2 FTMS + c NSI d Full ms2 627.83@hcd25.00 [100.00−1300.00] 18473 65.31



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 627.8256@hcd28.00
[100.00−1266.00] 18545 164.64

VDLEPGTM(ox)DSVR

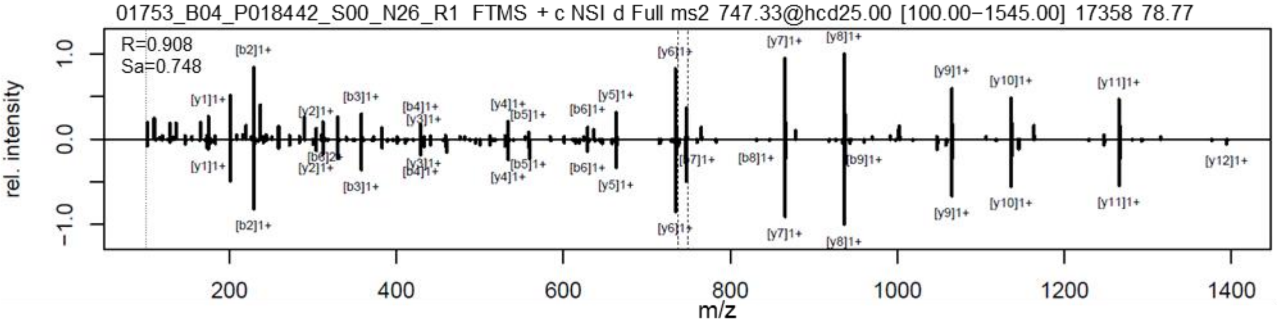
Sample:
01698_G02_P018021_S00_N15_R1 FTMS + c NSI d Full ms2 667.82@hcd25.00 [100.00−1380.00] 31551 76.93



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 667.8167@hcd28.00
[100.00−1346.00] 26018 113.24

VEEAEAMAETEER

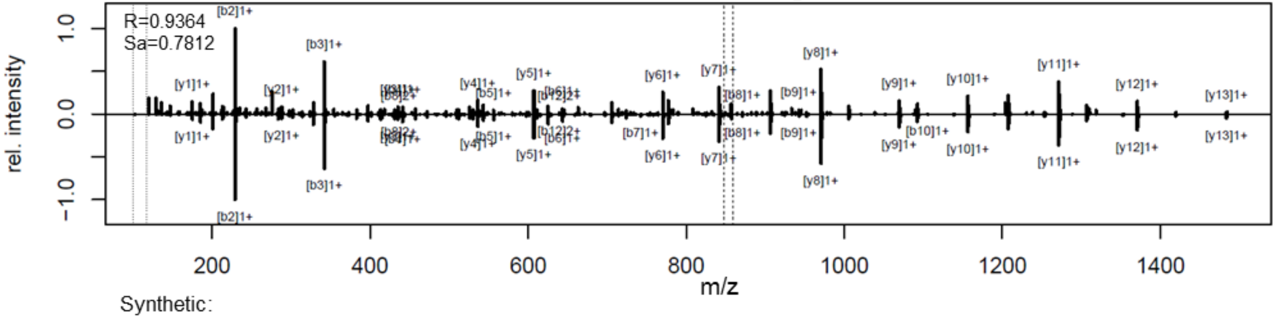
Sample:
01753_B04_P018442_S00_N26_R1 FTMS + c NSI d Full ms2 747.33@hcd25.00 [100.00−1545.00] 17358 78.77



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 747.3232@hcd28.00
[100.00−1505.00] 23921 139.3

VEIVDSVEAYATM(ox)LR

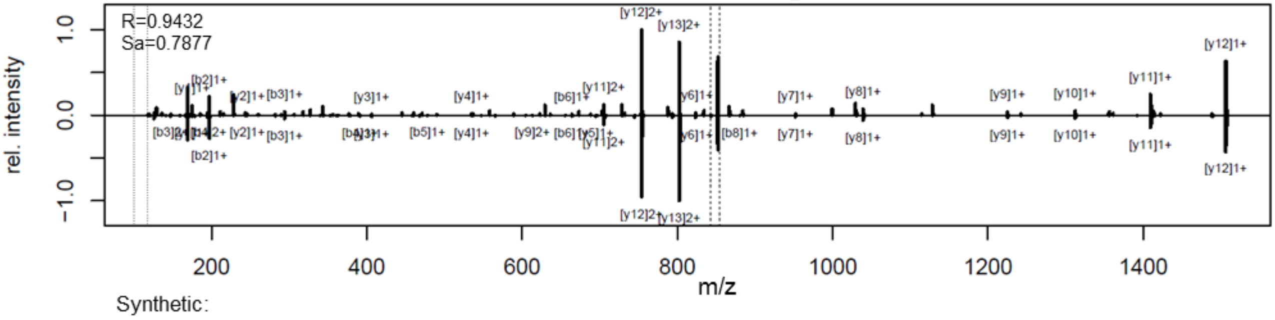
Sample:
01226_B03_P012502_S00_N18_R1 FTMS + c NSI d Full ms2 856.43@hcd25.00 [117.67−1765.00] 42907 100.47



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 856.9391@hcd28.00
[100.00−1724.00] 56670 205.3

VPPPSWSEQCECAR

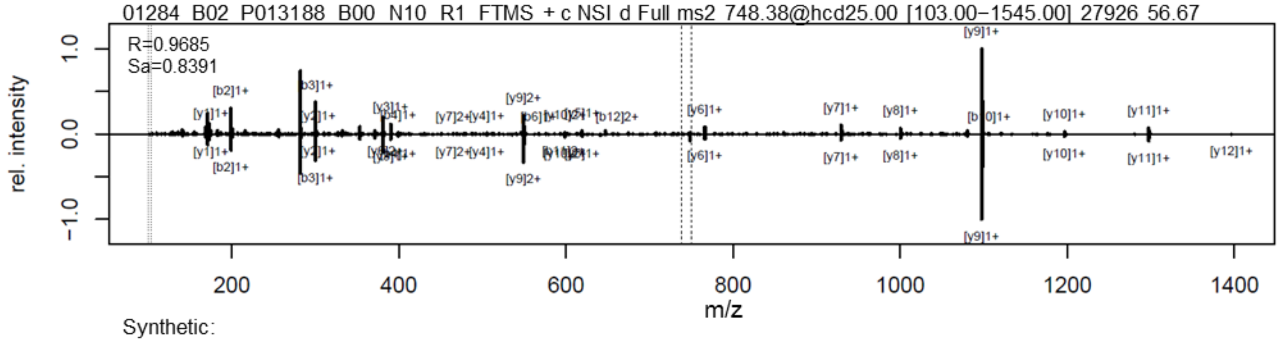
Sample:
01277_D02_P013129_S00_N12_R1 FTMS + c NSI d Full ms2 851.87@hcd25.00 [117.00−1755.00] 14176 56.59



Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 852.3707@hcd28.00
[100.00−1715.00] 25842 139.3

VVTVPAYFNDSQR

Sample:
01284_B02_P013188_B00_N10_R1 FTMS + c NSI d Full ms2 748.38@hcd25.00 [103.00−1545.00] 27926 56.67

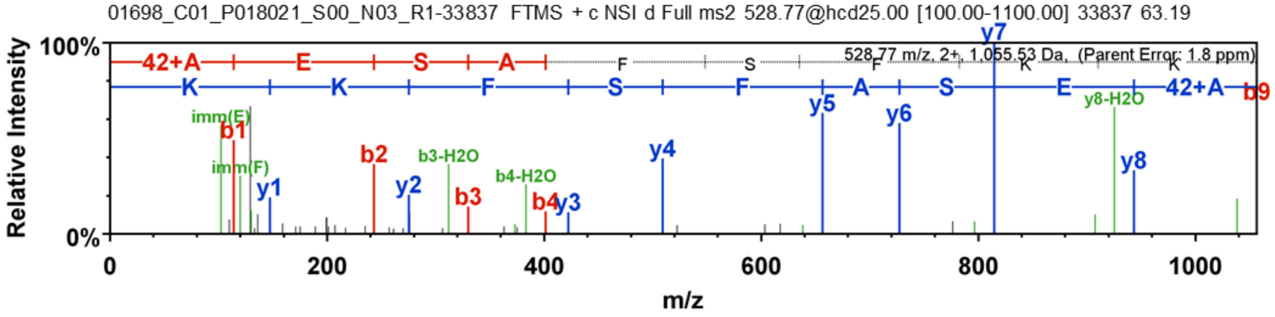


Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 748.3811@hcd28.00
[100.00−1507.00] 40071 139.78

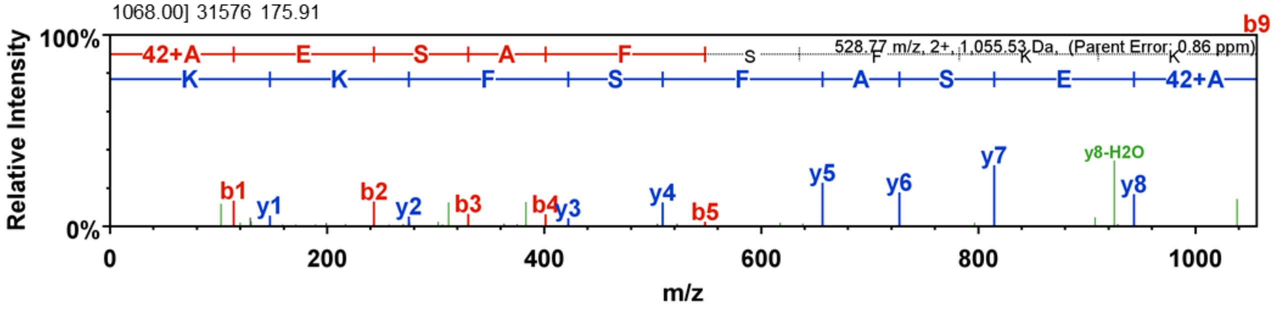
Appendix S1.2. Spectra of aTIS peptides confirmed by manual inspection. These spectra were visualized in Scaffold 4.

(ac)AESAFSFKK

Sample:
01698_C01_P018021_S00_N03_R1-33837 FTMS + c NSI d Full ms2 528.77@hcd25.00 [100.00-1100.00] 33837 63.19

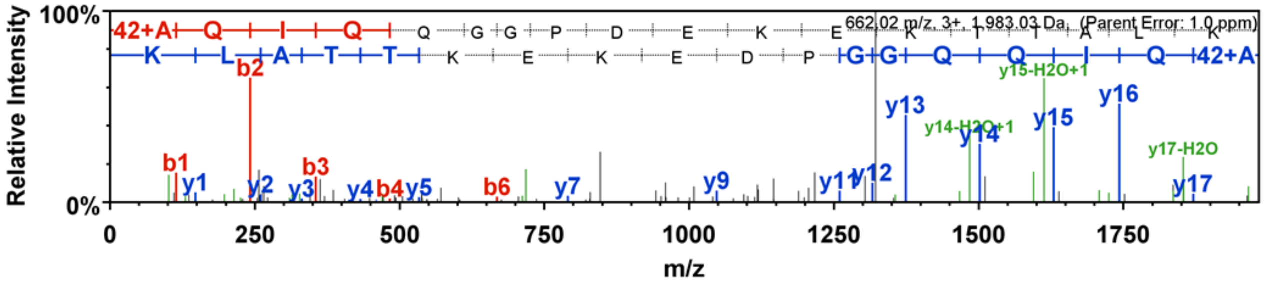


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 528.7720@hcd28.00 [100.00-1068.00] 31576 175.91

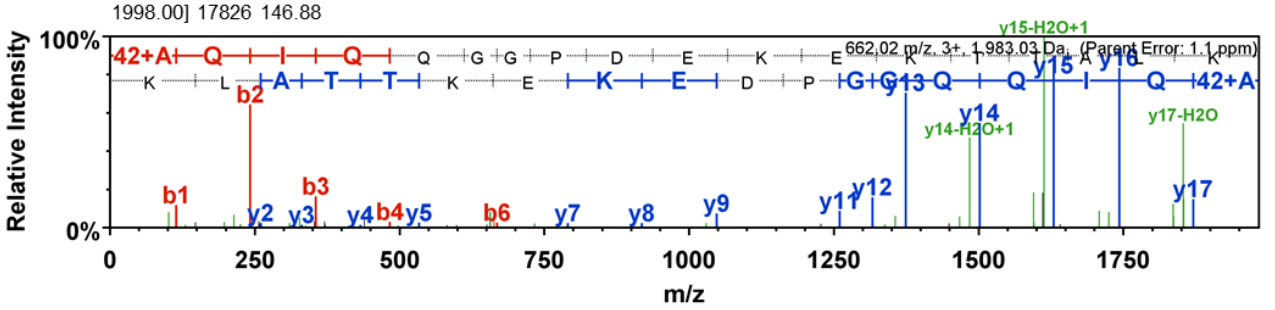


(ac)AQIQQGGPDEKEKTTALK

Sample:
01753_A12_P018443_S00_N08_R1-14630 FTMS + c NSI d Full ms2 662.02@hcd25.00 [100.00-2045.00] 14630 52.17

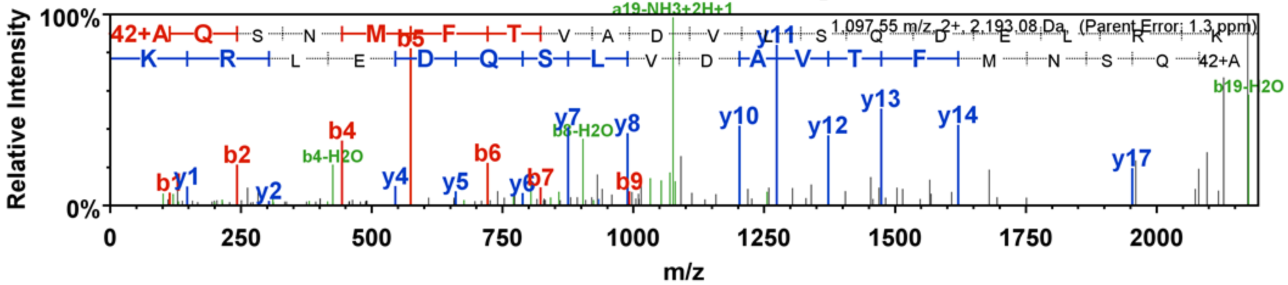


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 662.3511@hcd28.00 [100.00-1998.00] 17826 146.88

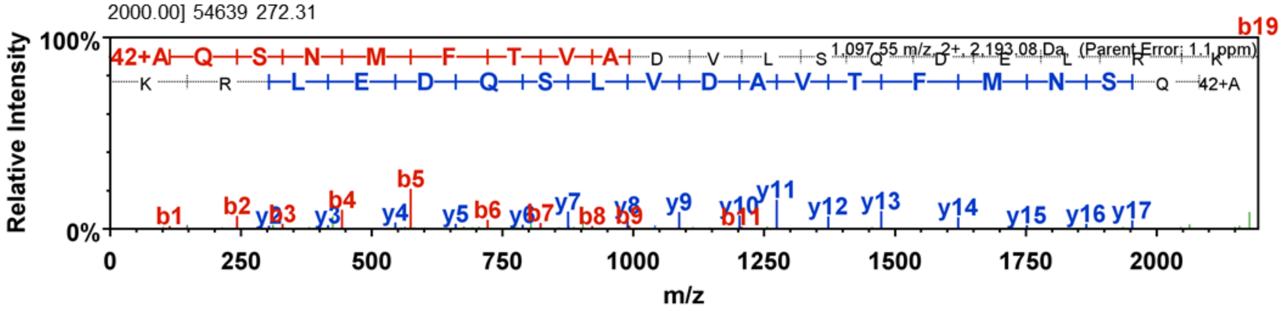


(ac)AQSNMFTVADVLSQDELRK

Sample:
01752_C02_P018440_S00_N11_R1-44868 FTMS + c NSI d Full ms2 1097.55@hcd25.00 [100.00-2260.00] 44868 71.9

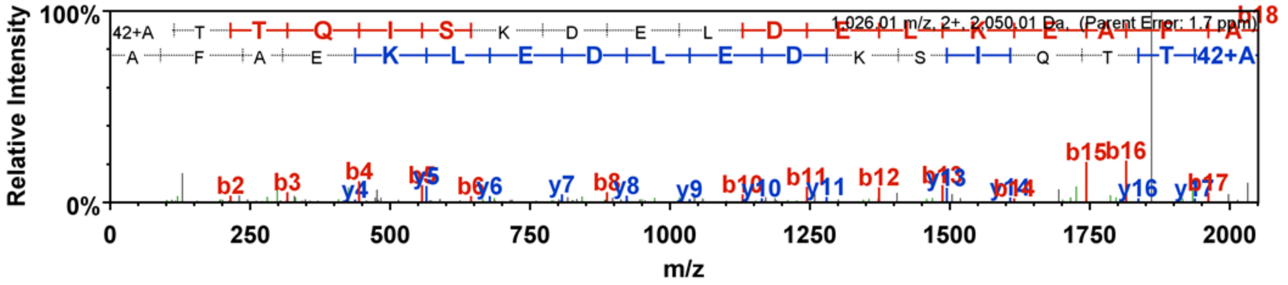


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1098.0470@hcd28.00 [100.00-2000.00] 54639 272.31

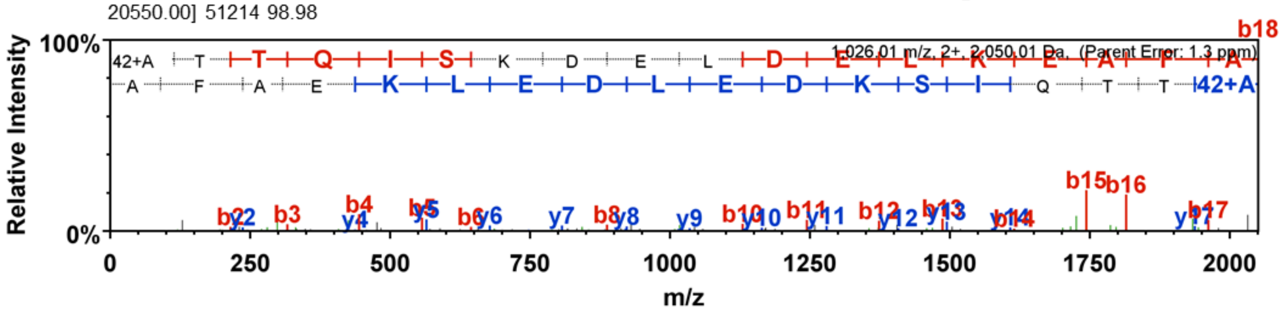


(ac) ATTQISKDELDELKEAFA

Sample:
01754_C03_P018444_S00_N19_R1-44868 FTMS + c NSI d Full ms2 1026.01@hcd25.00 [100.00-2055.00] 41657 90

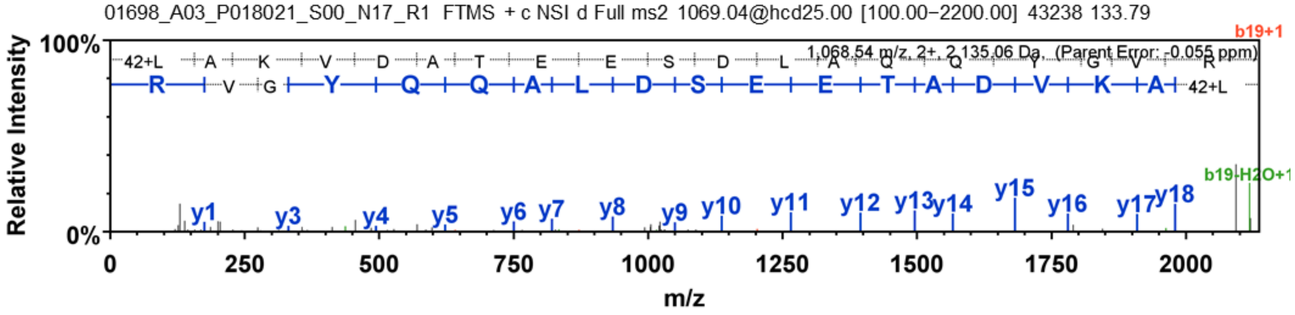


Synthetic:
02383a_BE1-TUM_ac_proteogenom_2_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 1026.01@hcd28.00 [100.00-20550.00] 51214 98.98

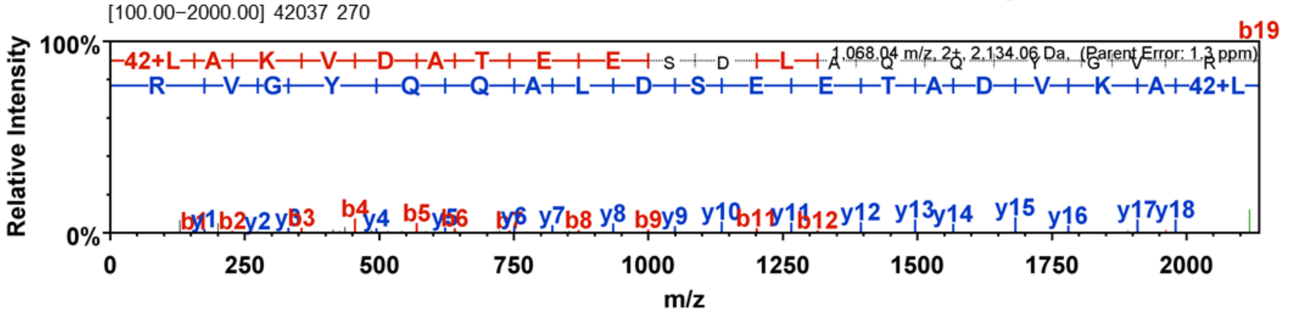


(ac)LAKVDATEESDLAQQYGVR

Sample:
01698_A03_P018021_S00_N17_R1 FTMS + c NSI d Full ms2 1069.04@hcd25.00 [100.00−2200.00] 43238 133.79

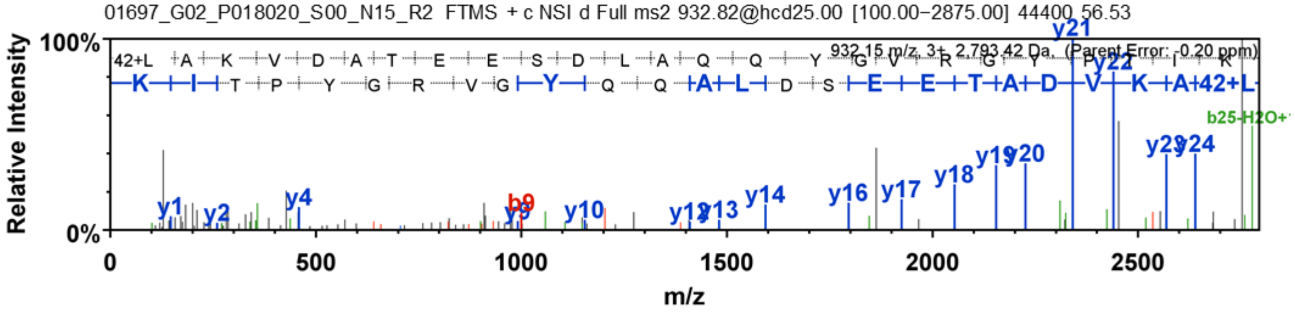


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1068.5370@hcd28.00
[100.00−2000.00] 42037 270

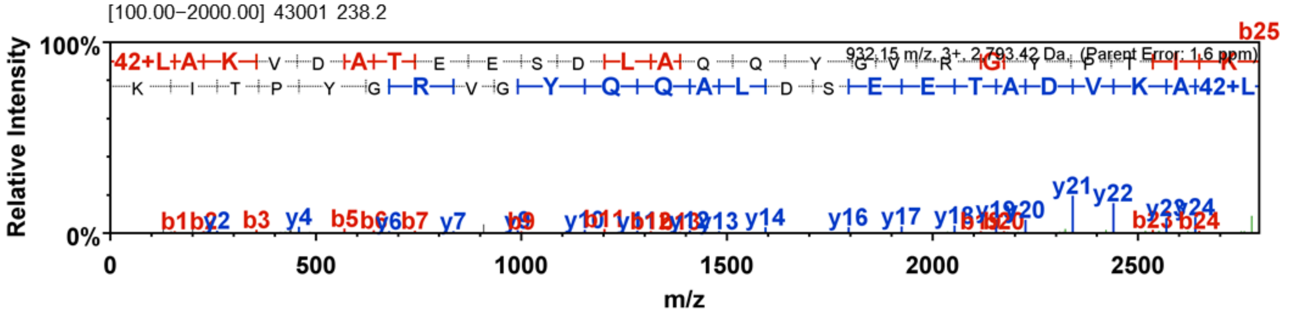


(ac)LAKVDATEESDLAQQYGVRGYPTIK

Sample:
01697_G02_P018020_S00_N15_R2 FTMS + c NSI d Full ms2 932.82@hcd25.00 [100.00−2875.00] 44400 56.53

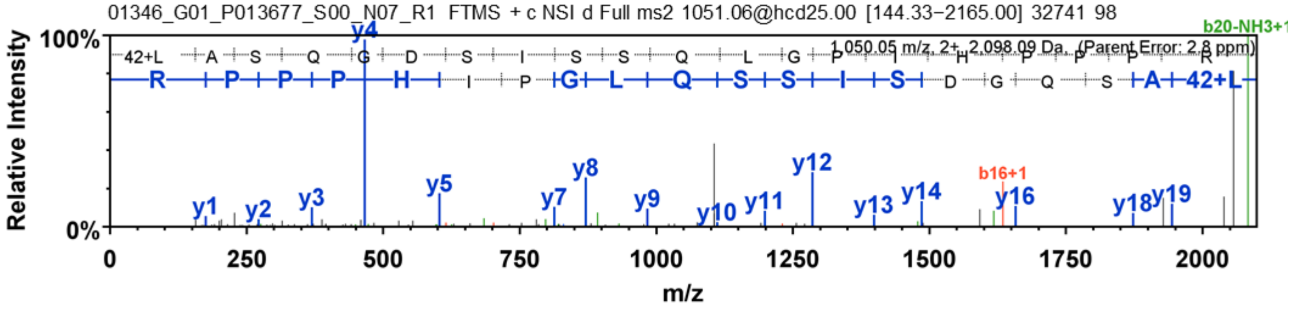


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 932.4819@hcd28.00
[100.00−2000.00] 43001 238.2

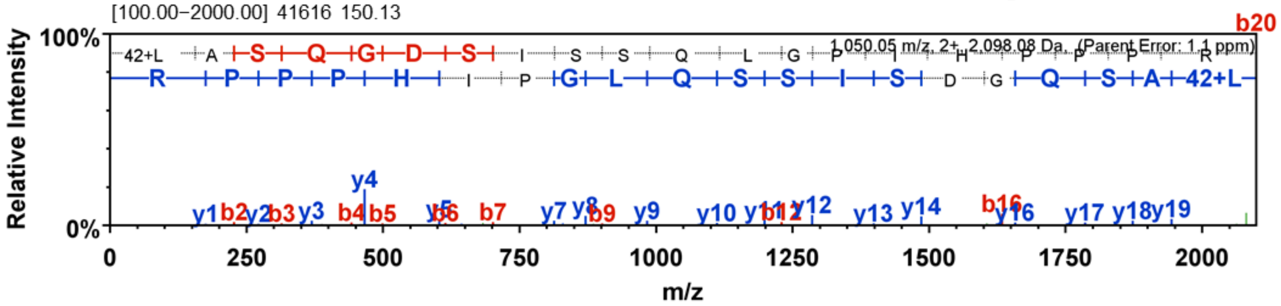


(ac)LASQGDSISSQLGPIHPPPR

Sample:
01346_G01_P013677_S00_N07_R1 FTMS + c NSI d Full ms2 1051.06@hcd25.00 [144.33–2165.00] 32741 98

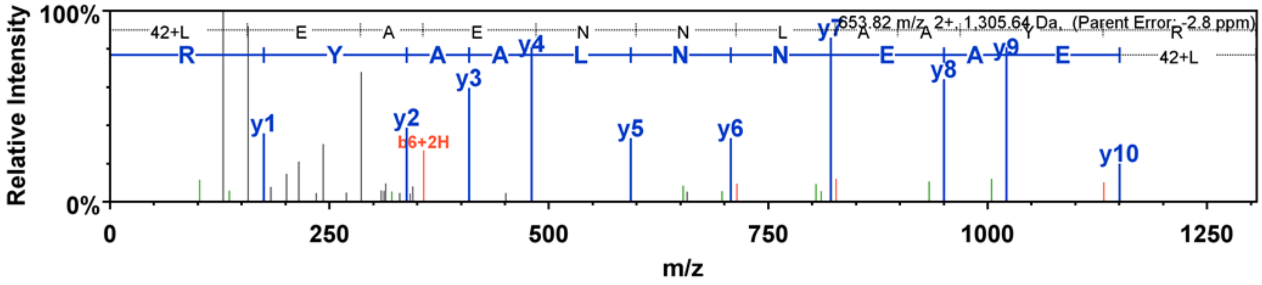


Synthetic:
02383a_BE1–TUM_ac_proteogenom_1_01_01–DDA–1h–R1 FTMS + c NSI d Full ms2 1050.5500@hcd28.00
[100.00–2000.00] 41616 150.13

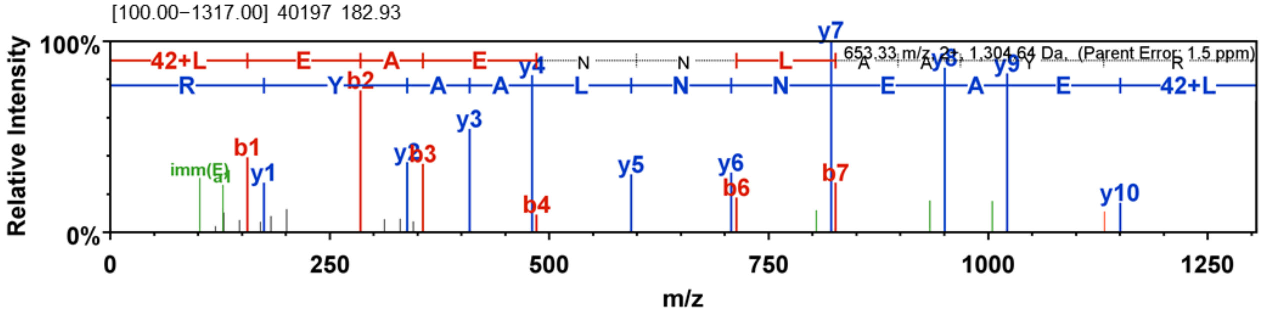


(ac)LEAENNLAAYR

Sample:
01277_H02_P013129_S00_N16_R1 FTMS + c NSI d Full ms2 653.82@hcd25.00 [90.33–1355.00] 25903 80.48

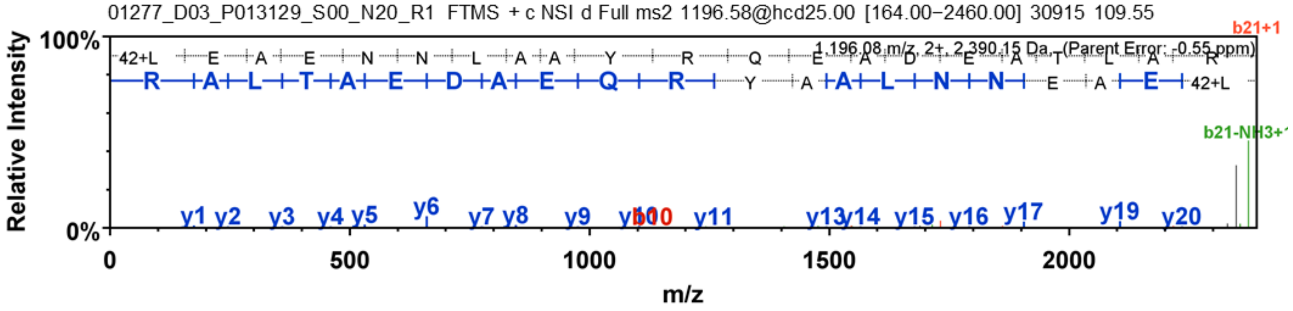


Synthetic:
02383a_BE1–TUM_ac_proteogenom_1_01_01–DDA–1h–R1 FTMS + c NSI d Full ms2 653.3260@hcd28.00
[100.00–1317.00] 40197 182.93

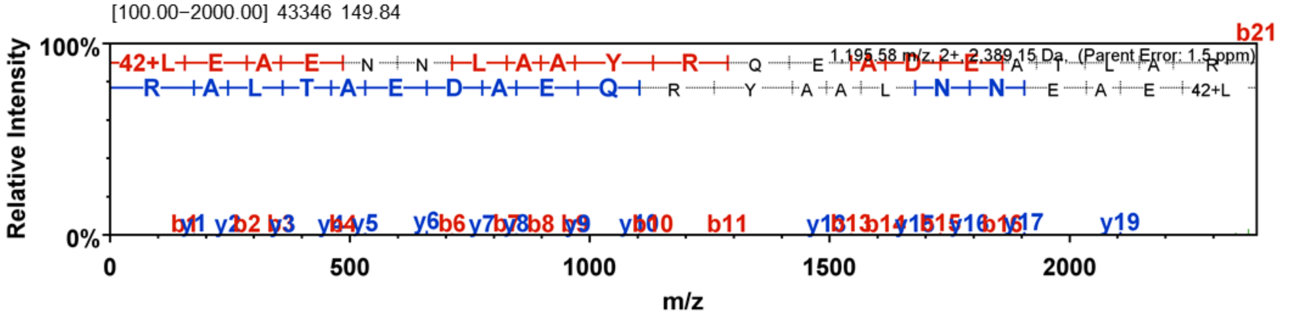


(ac)LEAENNLAAYRQEADEATLAR

Sample:
01277_D03_P013129_S00_N20_R1 FTMS + c NSI d Full ms2 1196.58@hcd25.00 [164.00-2460.00] 30915 109.55

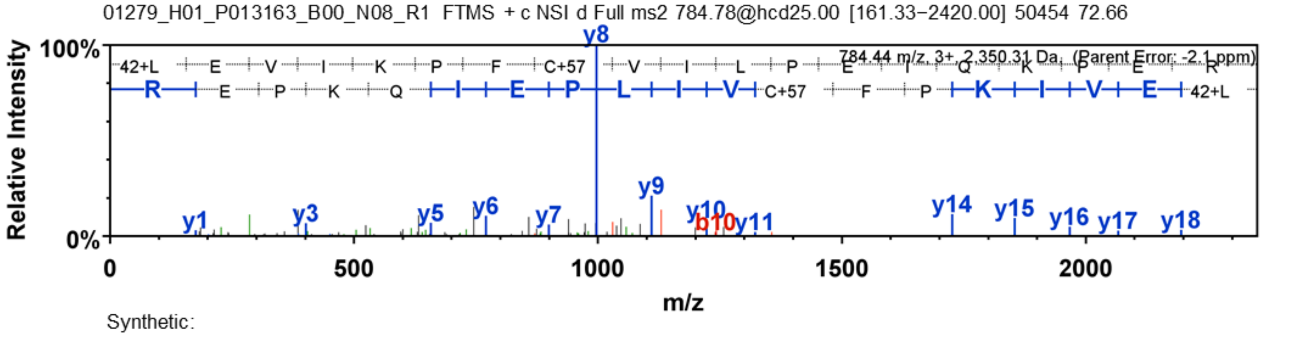


Synthetic:
02383a_BE1-TUM_ac_proteogenom_1_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 1196.0847@hcd28.00
[100.00-2000.00] 43346 149.84

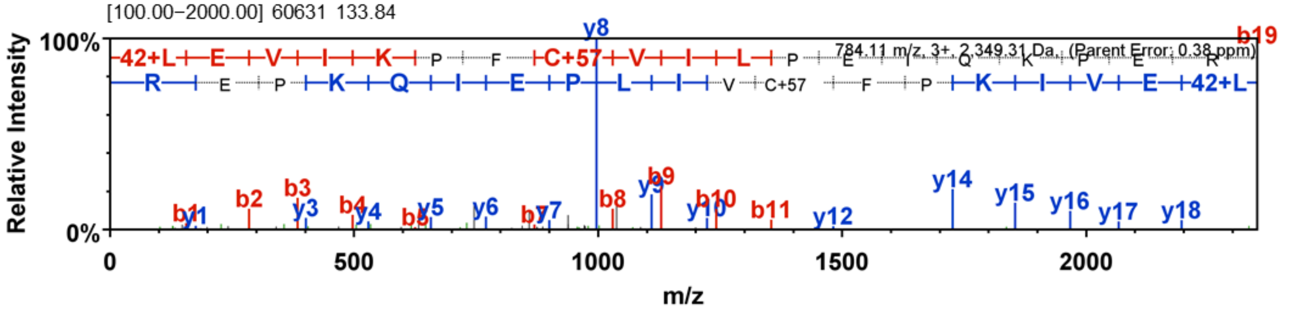


(ac)LEVIKPFCVILPEIQKPER

Sample:
01279_H01_P013163_B00_N08_R1 FTMS + c NSI d Full ms2 784.78@hcd25.00 [161.33-2420.00] 50454 72.66

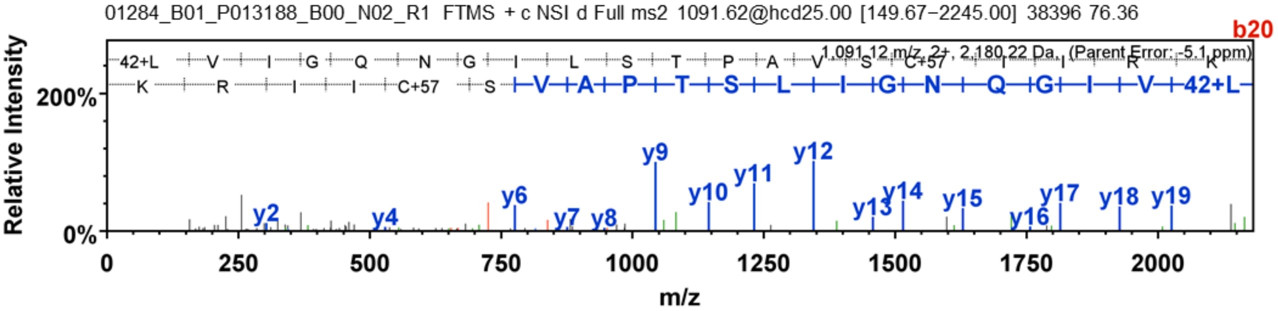


Synthetic:
02383a_BE1-TUM_ac_proteogenom_1_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 784.4474@hcd28.00
[100.00-2000.00] 60631 133.84

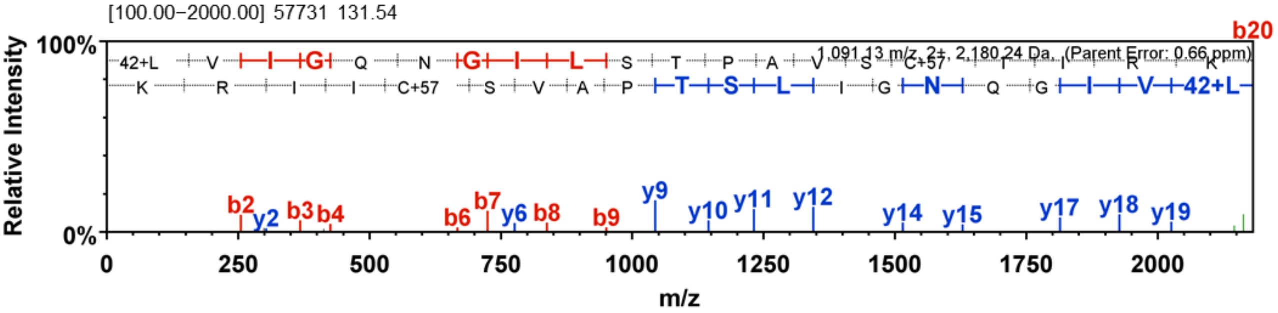


(ac)LVIGQNGILSTPAVSCIIRK

Sample:
01284_B01_P013188_B00_N02_R1 FTMS + c NSI d Full ms2 1091.62@hcd25.00 [149.67−2245.00] 38396 76.36

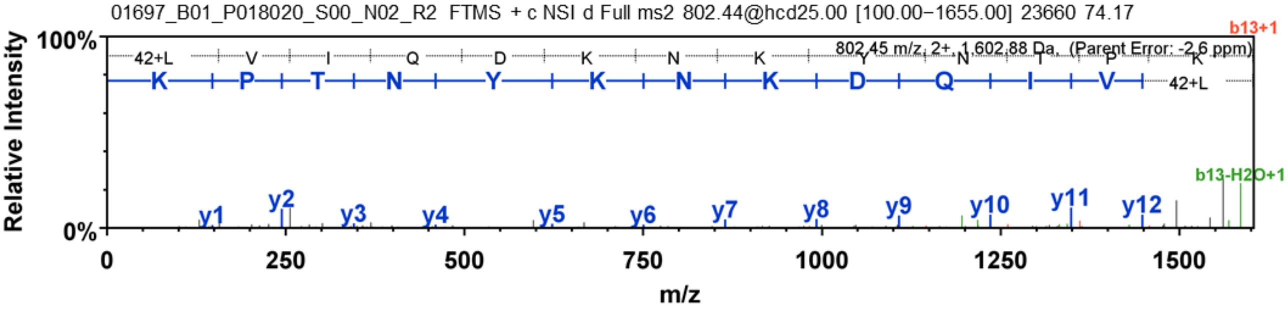


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1091.6279@hcd28.00
[100.00−2000.00] 57731 131.54

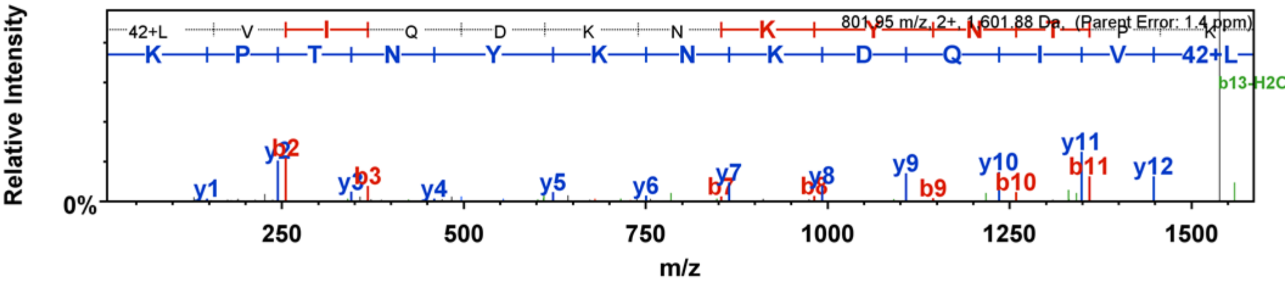


(ac)LVIQDKNKYNTPK

Sample:
01697_B01_P018020_S00_N02_R2 FTMS + c NSI d Full ms2 802.44@hcd25.00 [100.00−1655.00] 23660 74.17

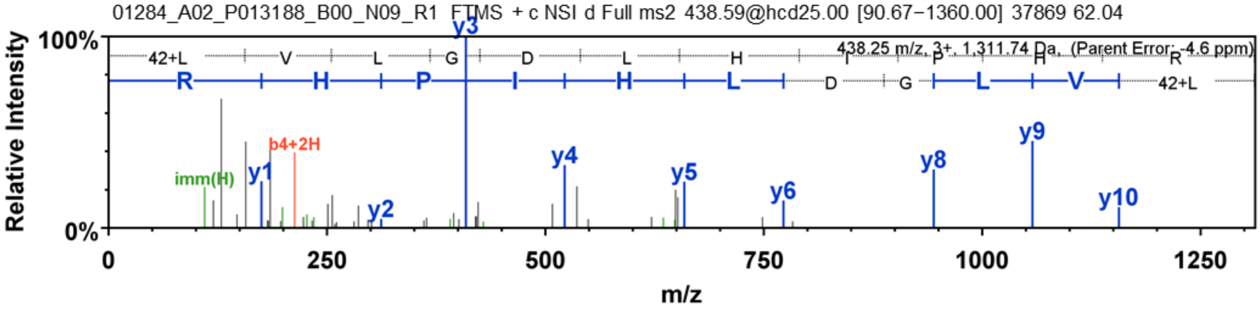


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 801.9467@hcd28.00
[100.00−1614.00] 26742 169.04

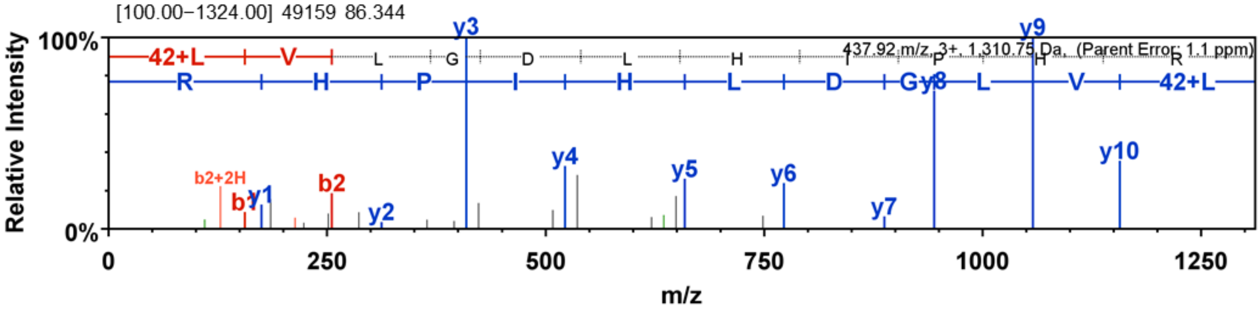


(ac)LVLGDLHIPHR

Sample:
01284_A02_P013188_B00_N09_R1 FTMS + c NSI d Full ms2 438.59@hcd25.00 [90.67−1360.00] 37869 62.04

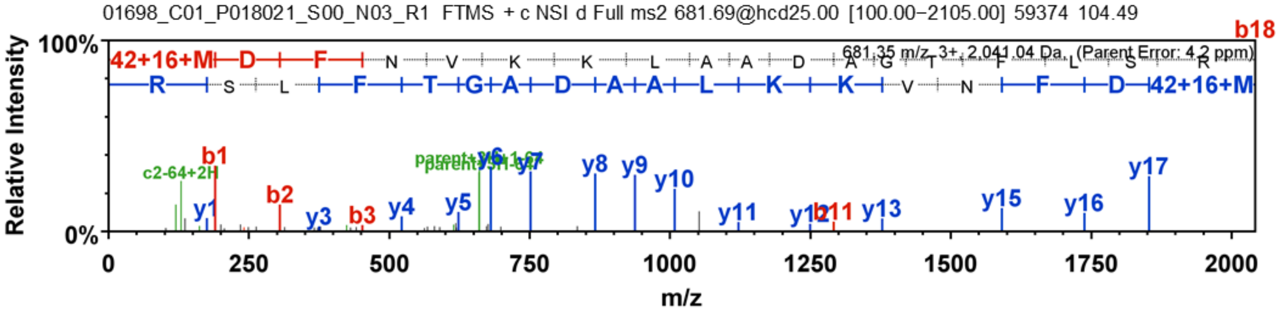


Synthetic:
02383a_BE1-TUM_ac_proteogenom_1_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 437.9230@hcd28.00 [100.00−1324.00] 49159 86.344

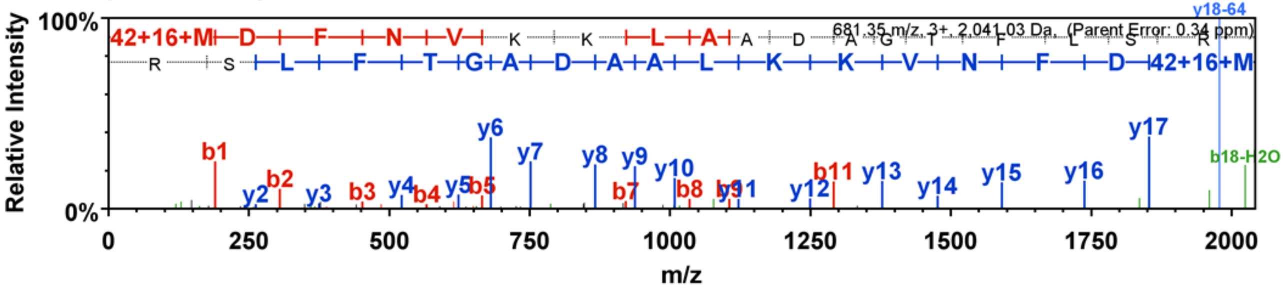


(ac)M(ox)DFNVKKLAADAGTFLSR

Sample:
01698_C01_P018021_S00_N03_R1 FTMS + c NSI d Full ms2 681.69@hcd25.00 [100.00−2105.00] 59374 104.49

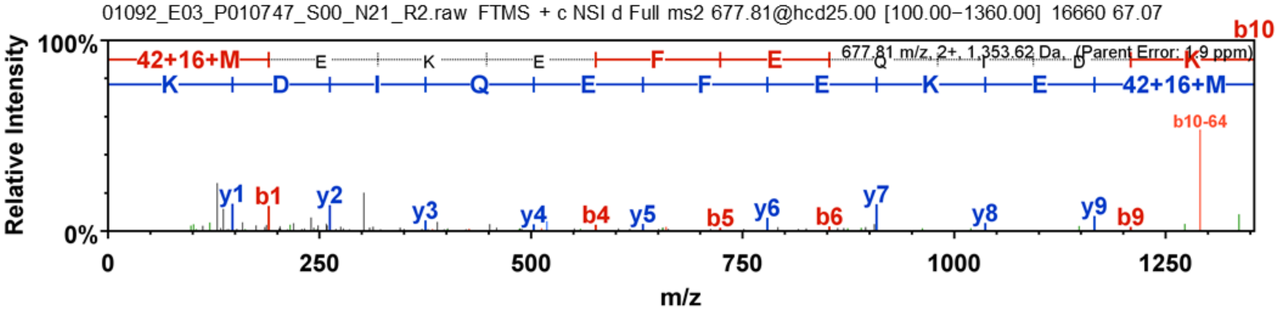


Synthetic:
02383a_BE1-TUM_ac_proteogenom_1_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 681.6862@hcd28.00 [100.00−2000.00] 57959 157.85

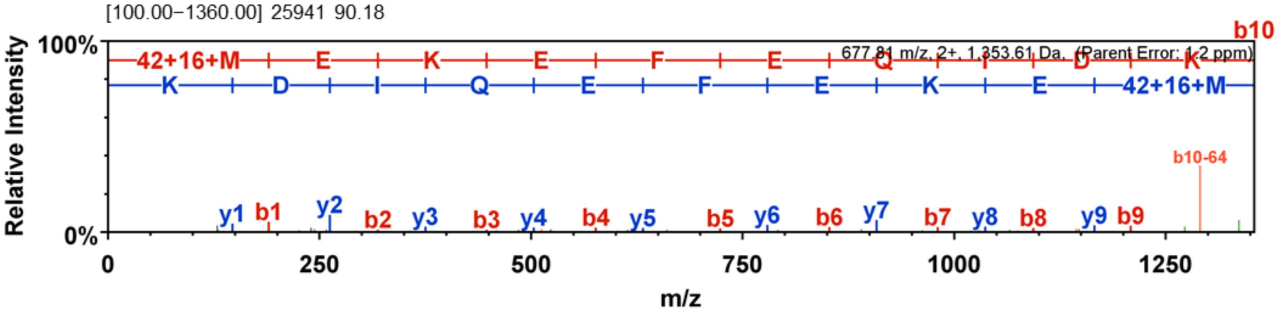


(ac)M(ox)EKEFEQIDK

Sample:
01092_E03_P010747_S00_N21_R2.raw FTMS + c NSI d Full ms2 677.81@hcd25.00 [100.00−1360.00] 16660 67.07

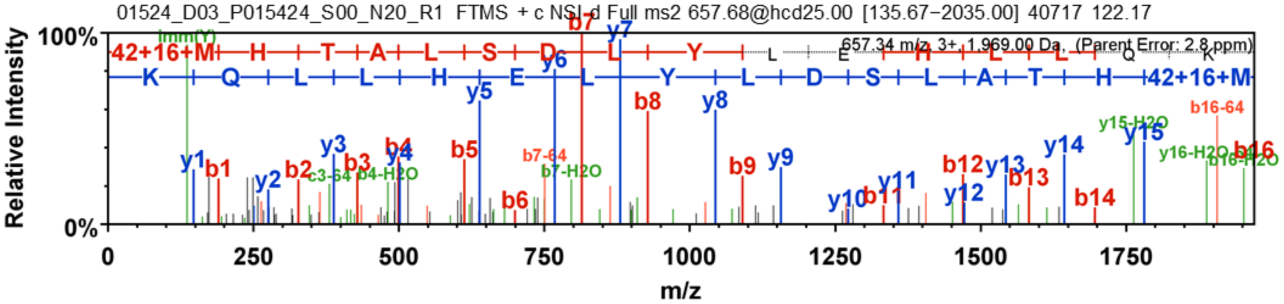


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 677.81@hcd28.00 [100.00−1360.00] 25941 90.18

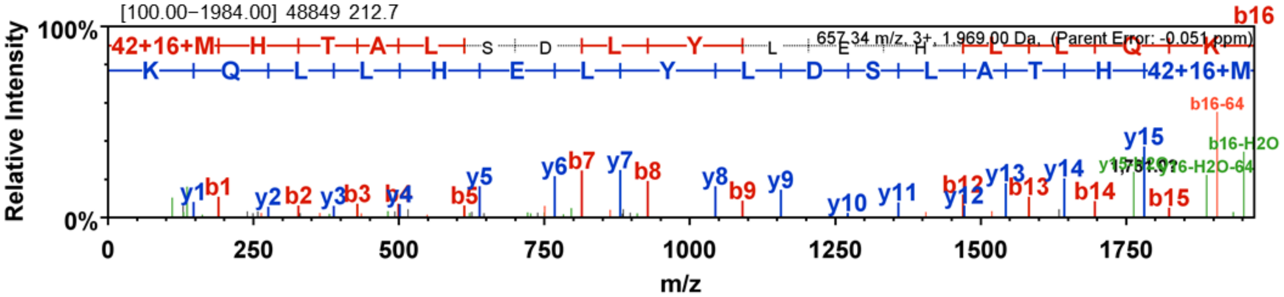


(ac)M(ox)HTALSDLYLEHLLQK

Sample:
01524_D03_P015424_S00_N20_R1 FTMS + c NSI d Full ms2 657.68@hcd25.00 [135.67−2035.00] 40717 122.17

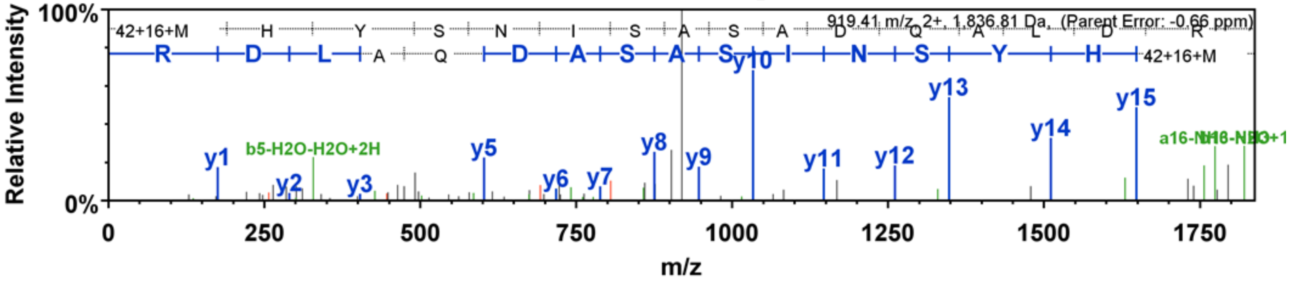


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 657.6732@hcd28.00 [100.00−1984.00] 48849 212.7

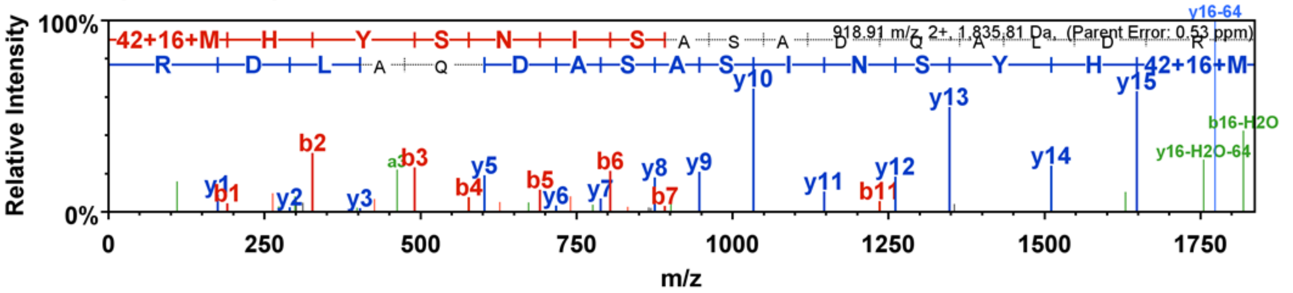


(ac)M(ox)HYSNISASADQALDR

Sample:
01307_E02_P013386_S00_N13_R1 FTMS + c NSI d Full ms2 919.91@hcd25.00 [126.33−1895.00] 20308 78.33

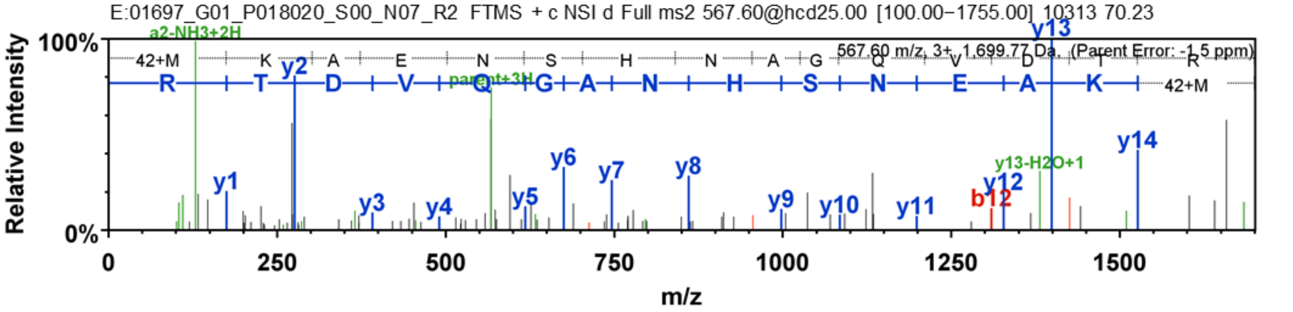


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 918.9131@hcd28.00 [100.00−1848.00] 29645 208.59

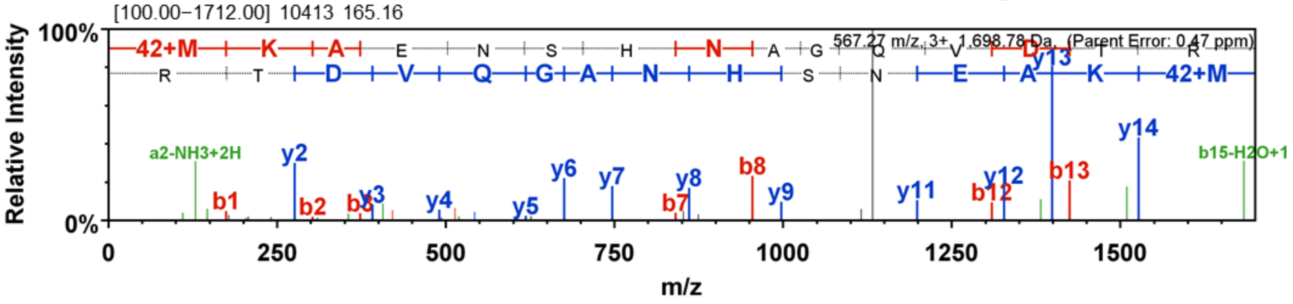


(ac)MKAENSHNAGQVDTR

Sample:
E:01697_G01_P018020_S00_N07_R2 FTMS + c NSI d Full ms2 567.60@hcd25.00 [100.00−1755.00] 10313 70.23

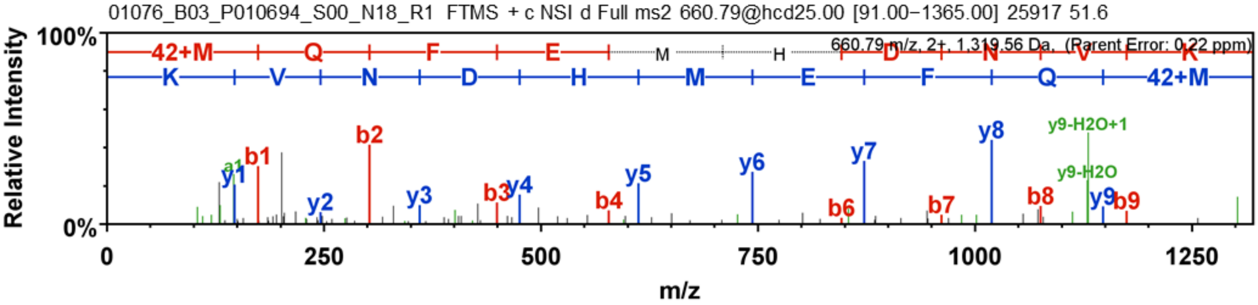


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 567.2661@hcd28.00 [100.00−1712.00] 10413 165.16

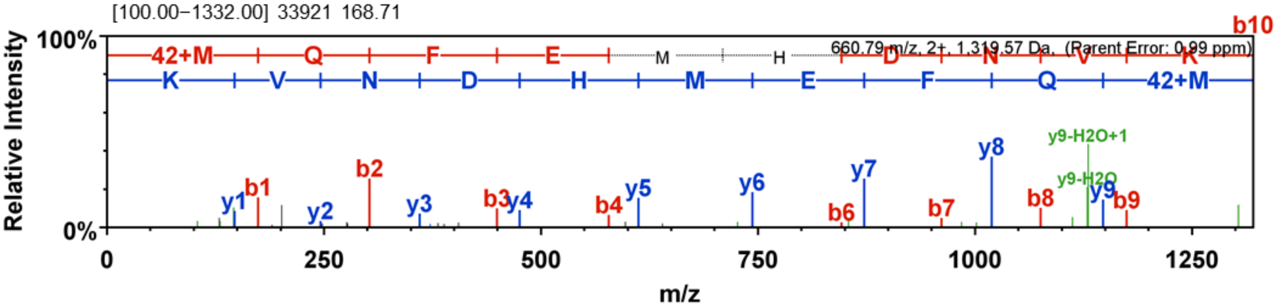


(ac)MQFEMHDNVK

Sample:
01076_B03_P010694_S00_N18_R1 FTMS + c NSI d Full ms2 660.79@hcd25.00 [91.00−1365.00] 25917 51.6

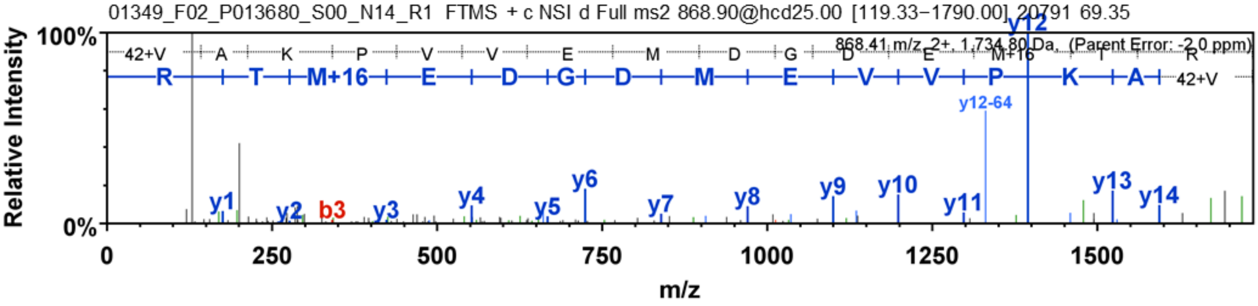


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 660.7897@hcd28.00
[100.00−1332.00] 33921 168.71

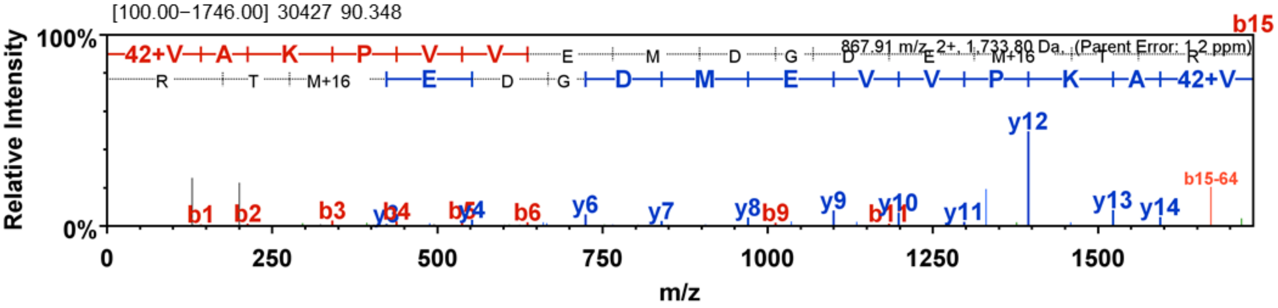


(ac)VAKPVVEMDGDEM(ox)TR

Sample:
01349_F02_P013680_S00_N14_R1 FTMS + c NSI d Full ms2 868.90@hcd25.00 [119.33−1790.00] 20791 69.35

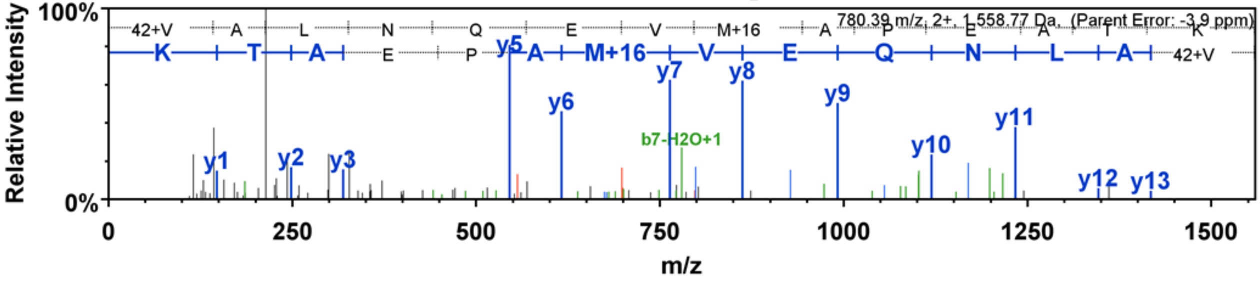


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 867.9063@hcd28.00
[100.00−1746.00] 30427 90.348

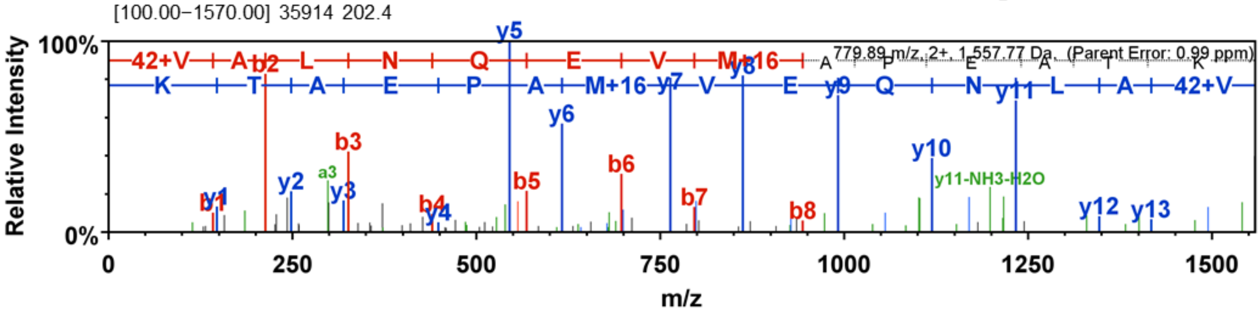


(ac)VALNQEVM(ox)APEATK

Sample:
01524_B02_P015424_S00_N10_R1 FTMS + c NSI d Full ms2 780.39@hcd25.00 [107.33−1610.00] 27444 77.01

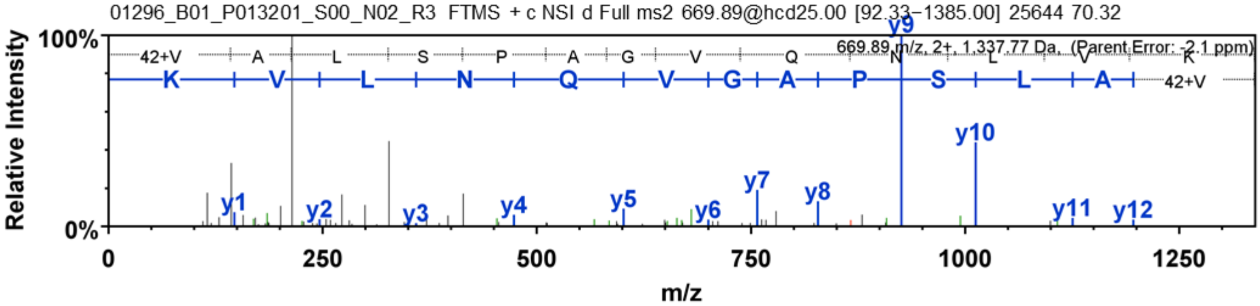


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 779.8934@hcd28.00
[100.00−1570.00] 35914 202.4

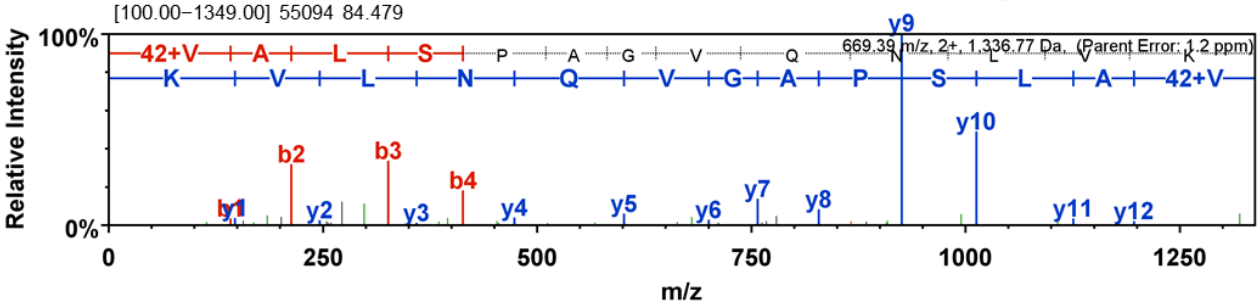


(ac)VALSPAGVQNLVK

Sample:
01296_B01_P013201_S00_N02_R3 FTMS + c NSI d Full ms2 669.89@hcd25.00 [92.33−1385.00] 25644 70.32

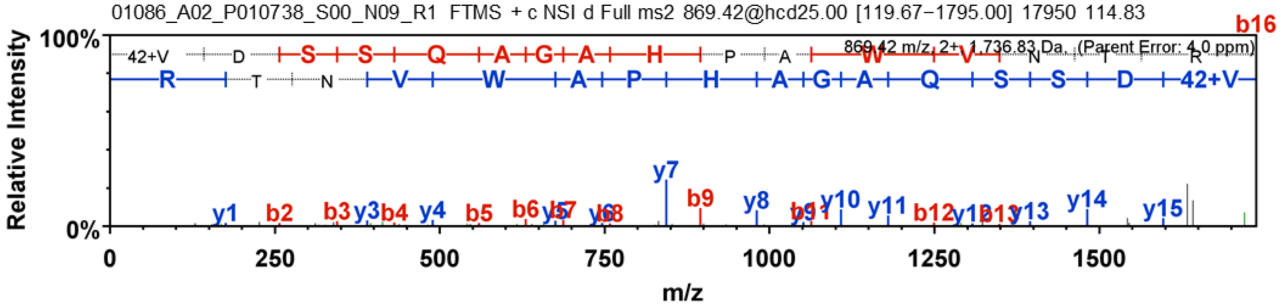


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 669.3935@hcd28.00
[100.00−1349.00] 55094 84.479

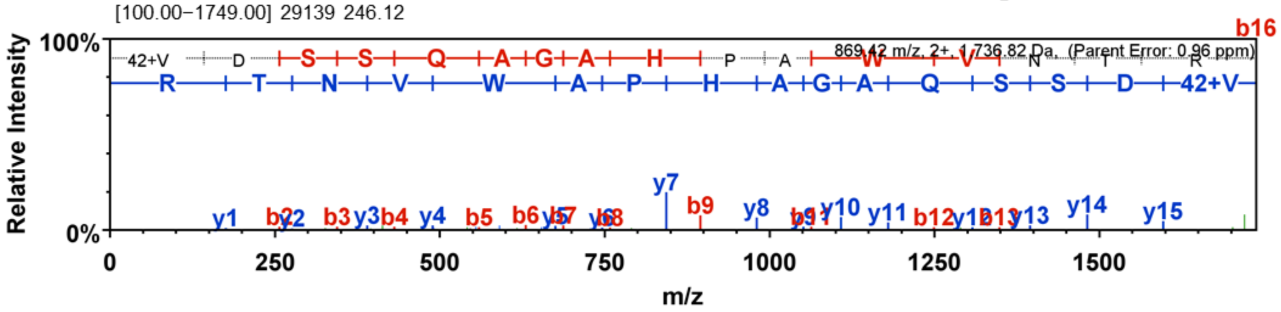


(ac)VDSSQAGAHPAWVNTR

Sample:
01086_A02_P010738_S00_N09_R1 FTMS + c NSI d Full ms2 869.42@hcd25.00 [119.67−1795.00] 17950 114.83

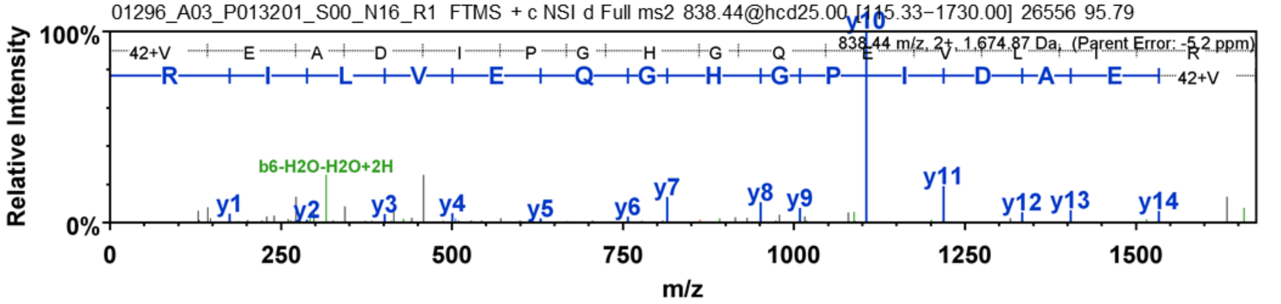


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 869.4197@hcd28.00 [100.00−1749.00] 29139 246.12

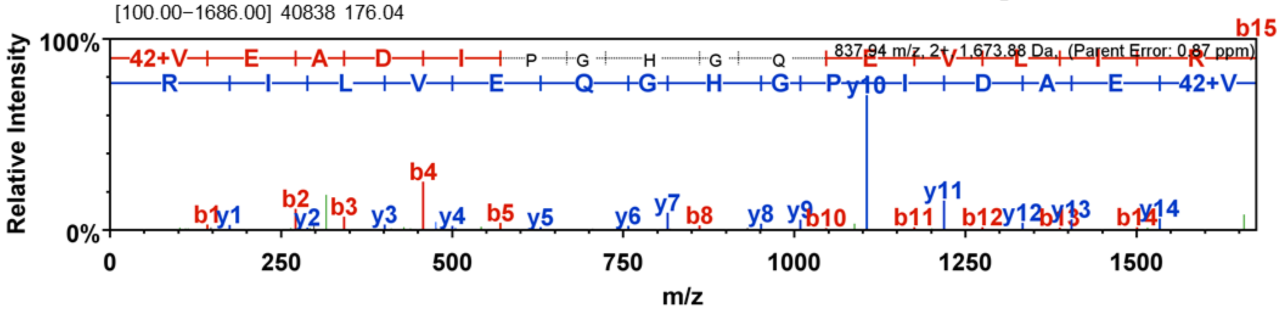


(ac)VEADIPGHGQEVLIR

Sample:
01296_A03_P013201_S00_N16_R1 FTMS + c NSI d Full ms2 838.44@hcd25.00 [115.33−1730.00] 26556 95.79

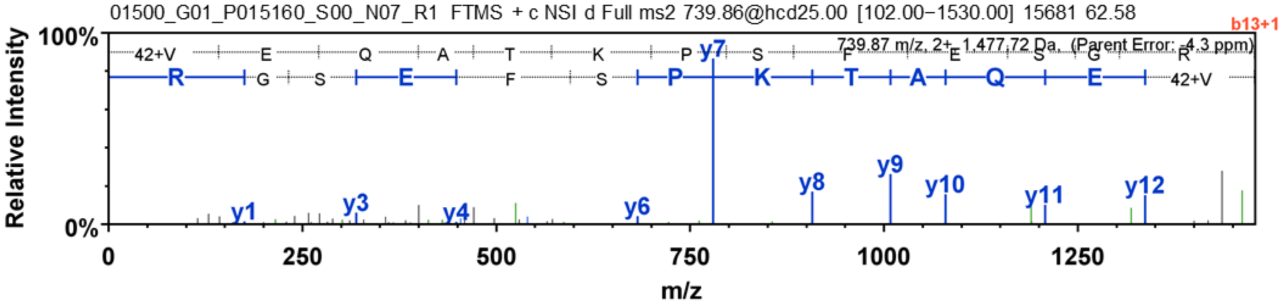


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 837.9448@hcd28.00 [100.00−1686.00] 40838 176.04

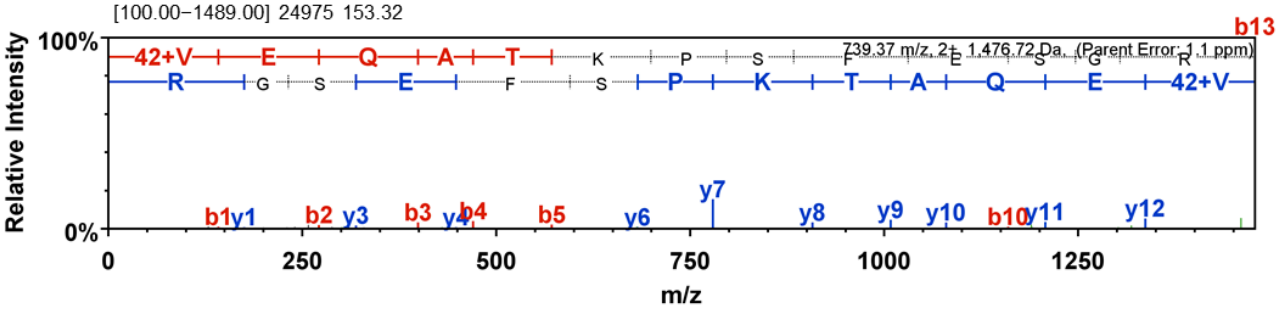


(ac)VEQATKPSFESGR

Sample:
01500_G01_P015160_S00_N07_R1 FTMS + c NSI d Full ms2 739.86@hcd25.00 [102.00−1530.00] 15681 62.58

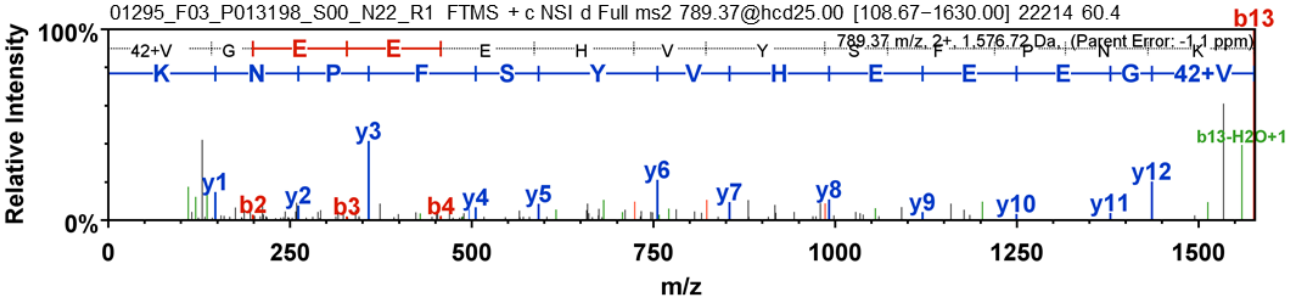


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 739.3691@hcd28.00
[100.00−1489.00] 24975 153.32

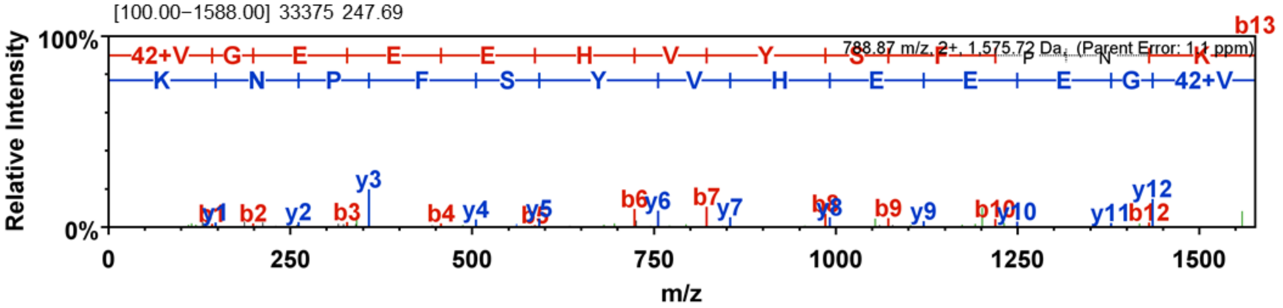


(ac)VGEEEHVYSFPNK

Sample:
01295_F03_P013198_S00_N22_R1 FTMS + c NSI d Full ms2 789.37@hcd25.00 [108.67−1630.00] 22214 60.4

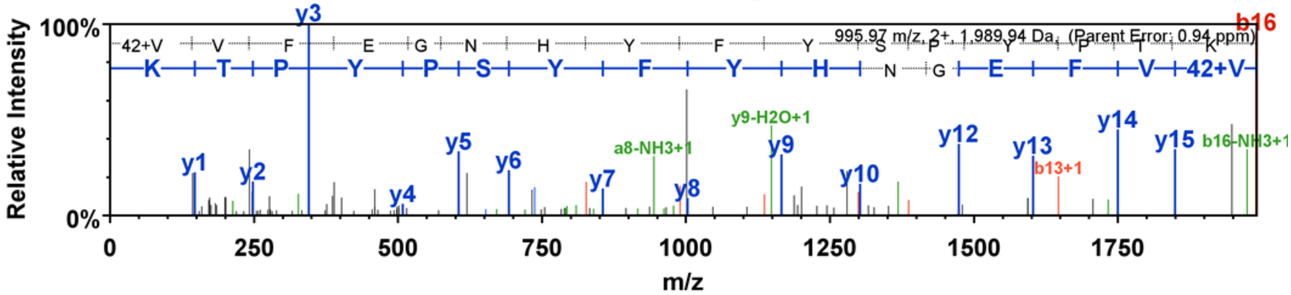


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 788.8680@hcd28.00
[100.00−1588.00] 33375 247.69

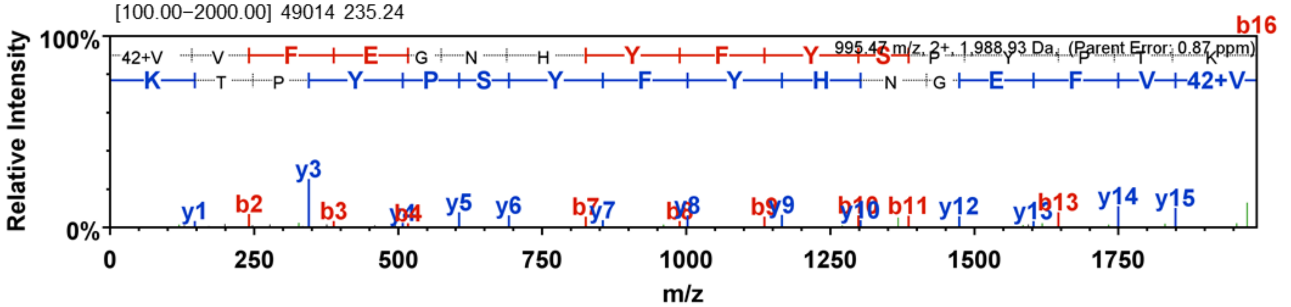


(ac)VVFEGNHYFYSPYPTK

Sample:
01226_C04_P012502_S00_N27_R1 FTMS + c NSI d Full ms2 995.98@hcd25.00 [136.67−2050.00] 35974 86

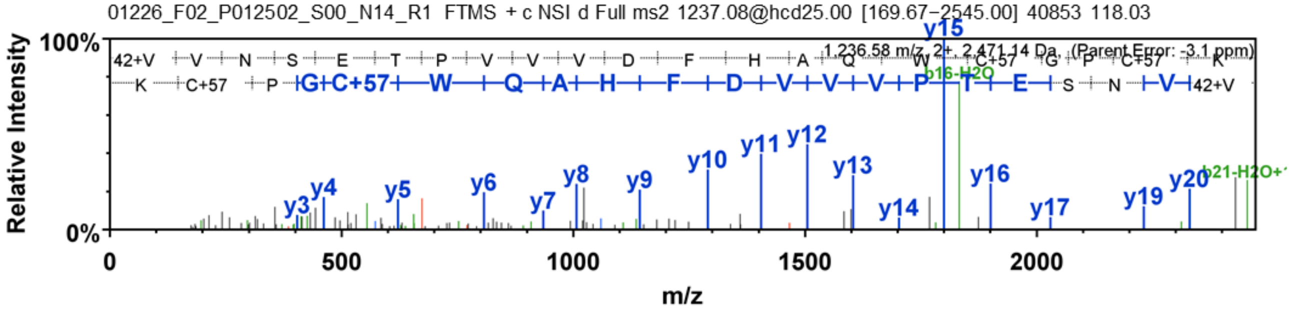


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 995.9744@hcd28.00 [100.00−2000.00] 49014 235.24

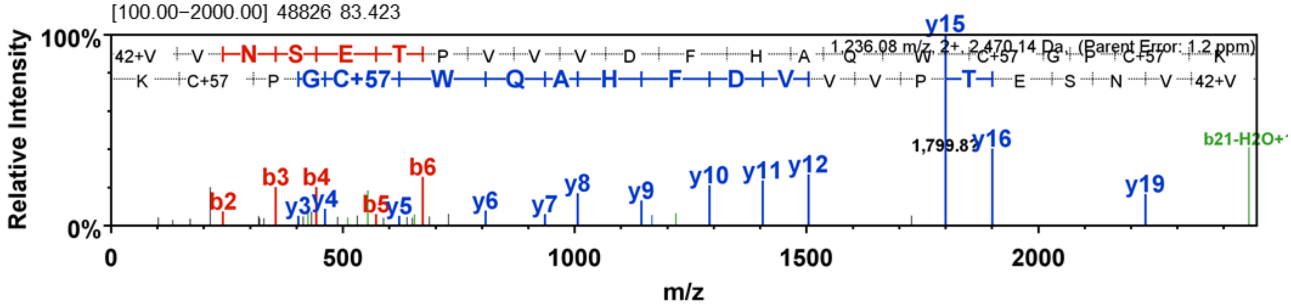


(ac)VVNSETPVVVDFHAQWCGPCK

Sample:
01226_F02_P012502_S00_N14_R1 FTMS + c NSI d Full ms2 1237.08@hcd25.00 [169.67−2545.00] 40853 118.03

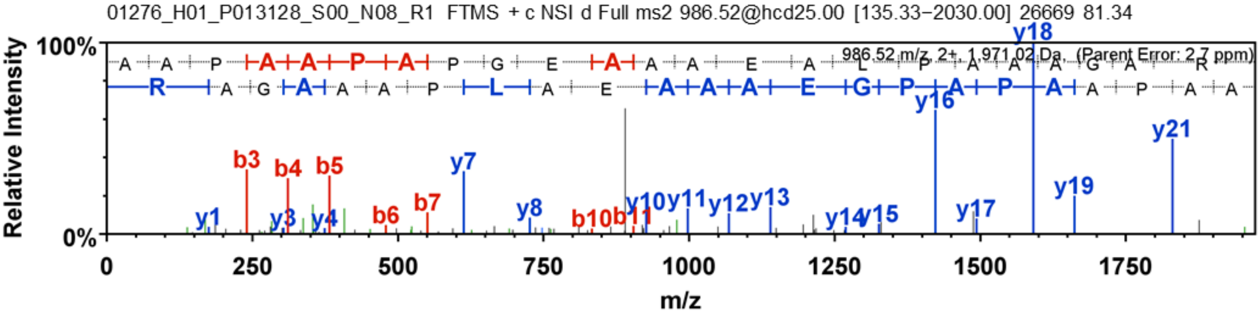


Synthetic:
02383a_BE1−TUM_ac_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 1236.5804@hcd28.00 [100.00−2000.00] 48826 83.423

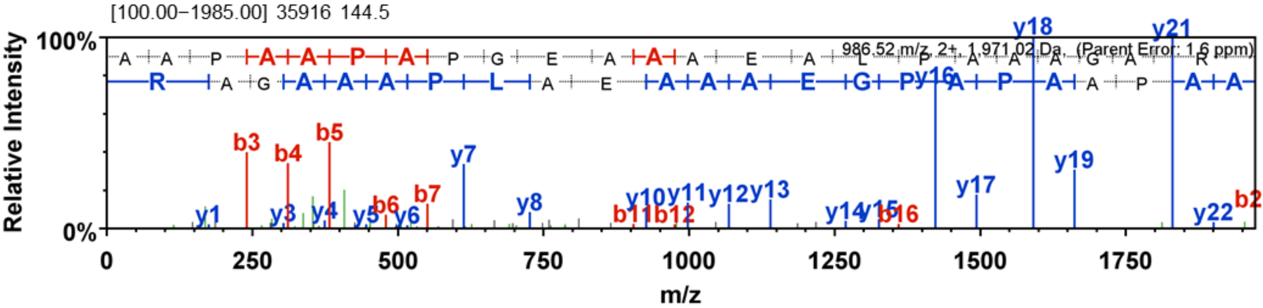


AAPAAPAPGEAAAEALPAAAGAR

Sample:
01276_H01_P013128_S00_N08_R1 FTMS + c NSI d Full ms2 986.52@hcd25.00 [135.33−2030.00] 26669 81.34

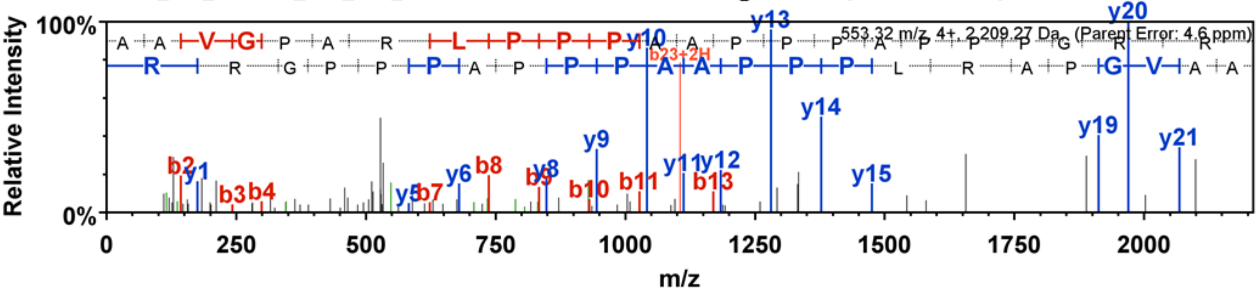


Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 987.0201@hcd28.00
[100.00−1985.00] 35916 144.5

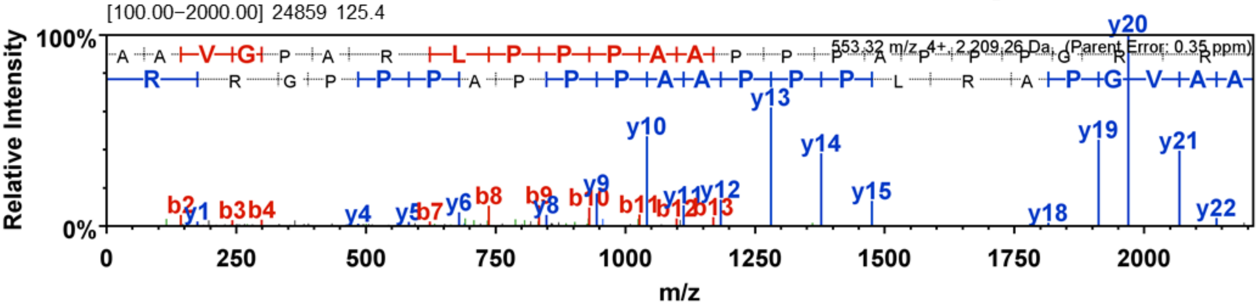


AAVGPARLPPPAAPPPAPPPGRR

Sample:
01698_F01_P018021_S00_N06_R1 FTMS + c NSI d Full ms2 553.33@hcd25.00 [100.00−2280.00] 26869 118.03

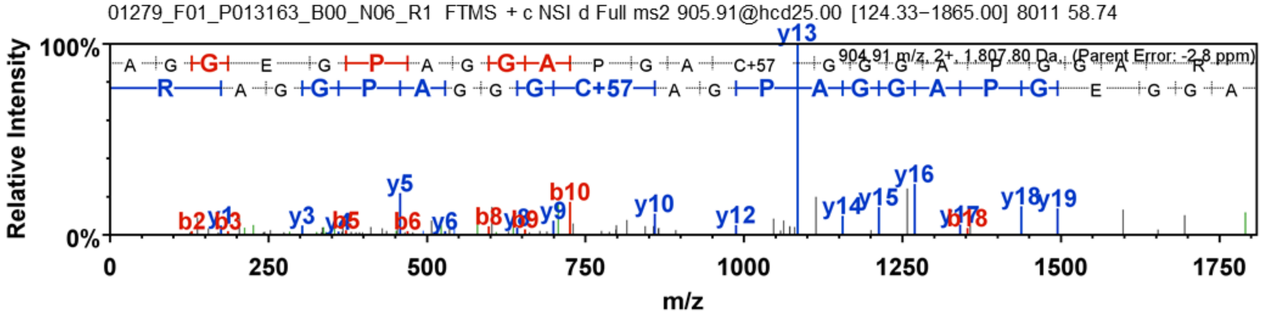


Synthetic:
02383a_BC4-TUM_unmod_proteogenom_4_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 553.57@hcd28.00
[100.00−2000.00] 24859 125.4

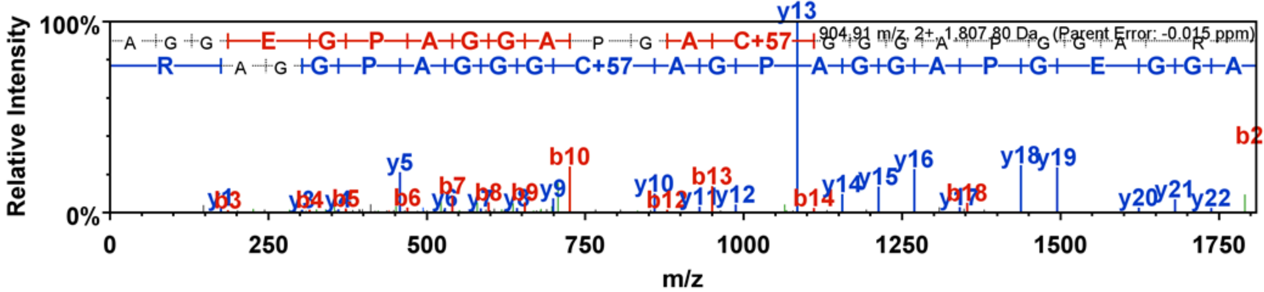


AGGEGPAGGAPGACGGGAPGGAR

Sample:
01279_F01_P013163_B00_N06_R1 FTMS + c NSI d Full ms2 905.91@hcd25.00 [124.33−1865.00] 8011 58.74

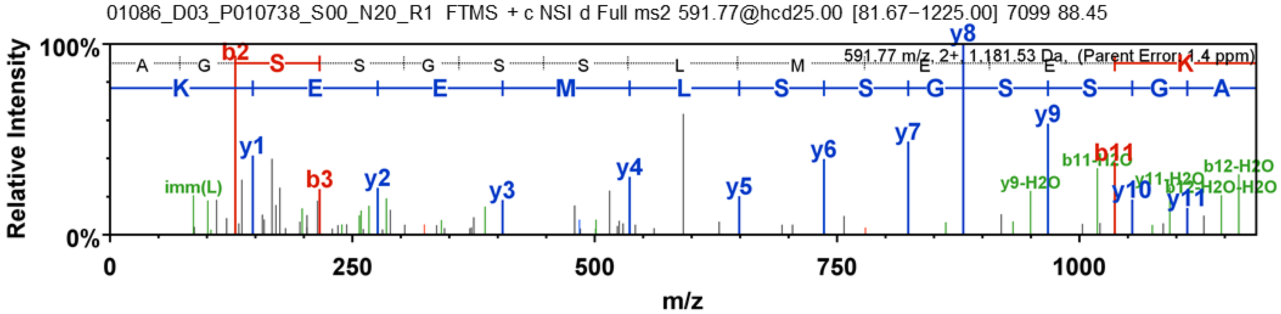


Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 905.4098@hcd28.00 [100.00−1821.00] 13695 183.18

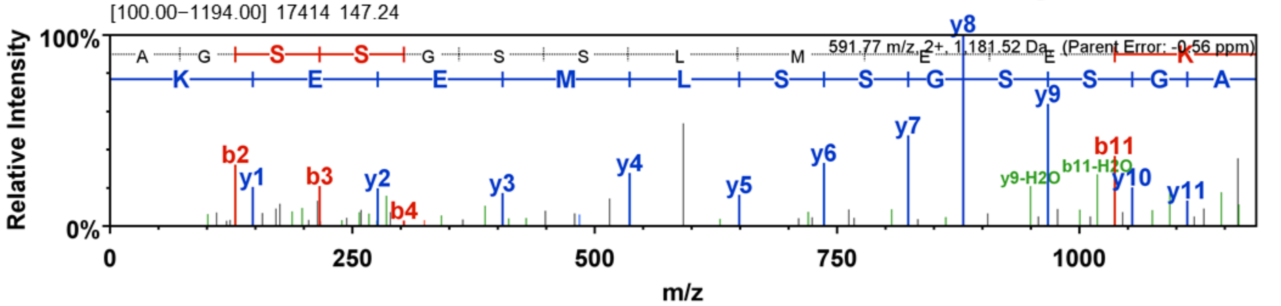


AGSSGSSLMEEK

Sample:
01086_D03_P010738_S00_N20_R1 FTMS + c NSI d Full ms2 591.77@hcd25.00 [81.67−1225.00] 7099 88.45

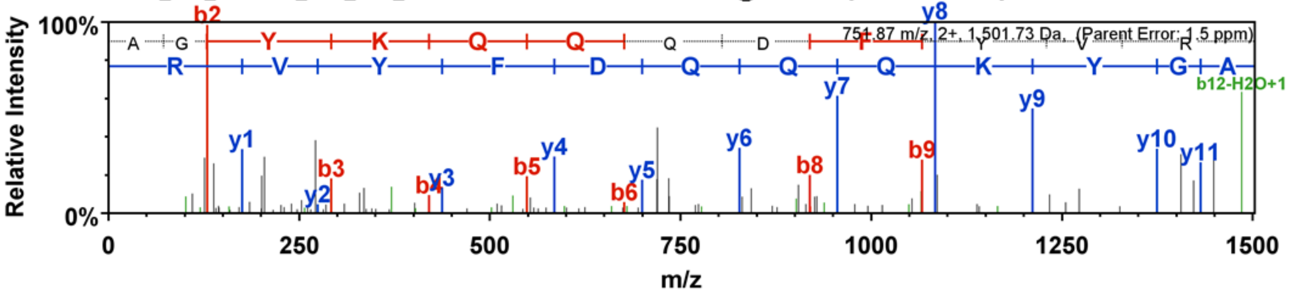


Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 591.7686@hcd28.00 [100.00−1194.00] 17414 147.24

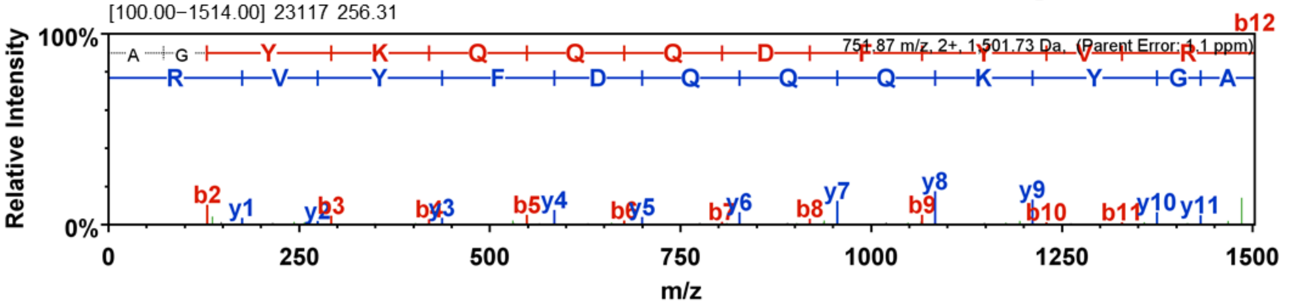


AGYKQQQDFYVR

Sample:
01697_F01_P018020_S00_N06_R2 FTMS + c NSI d Full ms2 751.88@hcd25.00 [100.00−1555.00] 23075 64.48

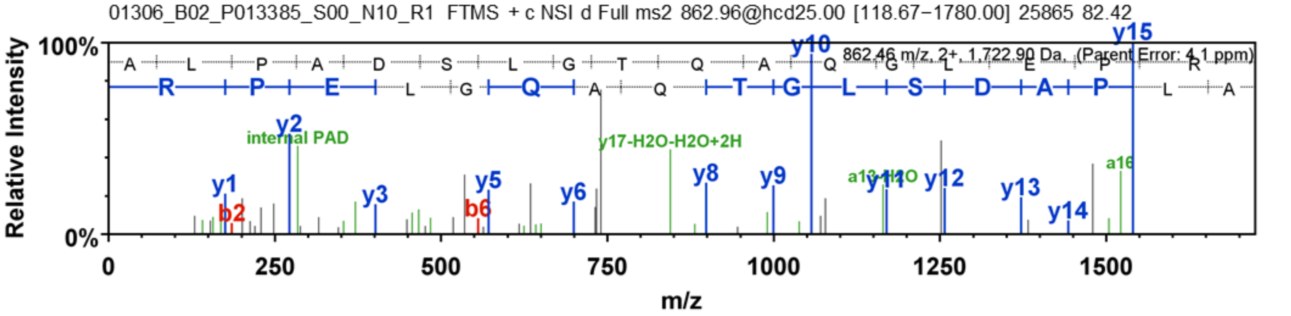


Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 751.8738@hcd28.00 [100.00−1514.00] 23117 256.31

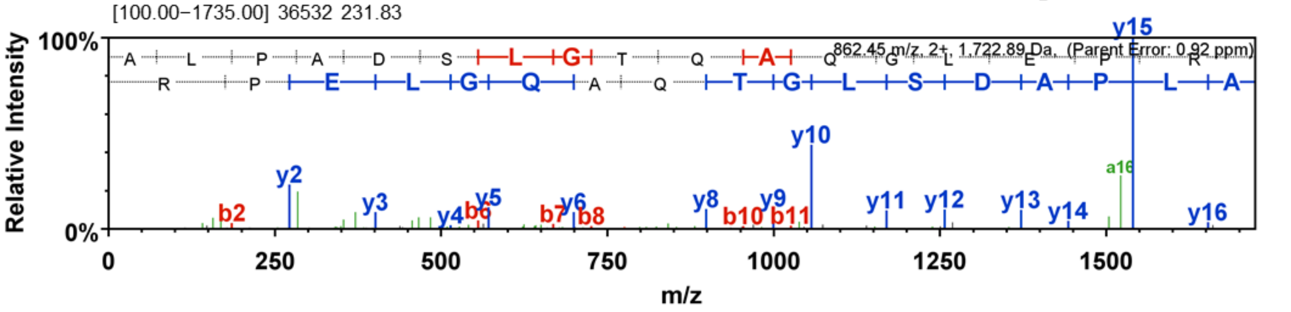


ALPADSLGTQAQGLEPR

Sample:
01306_B02_P013385_S00_N10_R1 FTMS + c NSI d Full ms2 862.96@hcd25.00 [118.67−1780.00] 25865 82.42

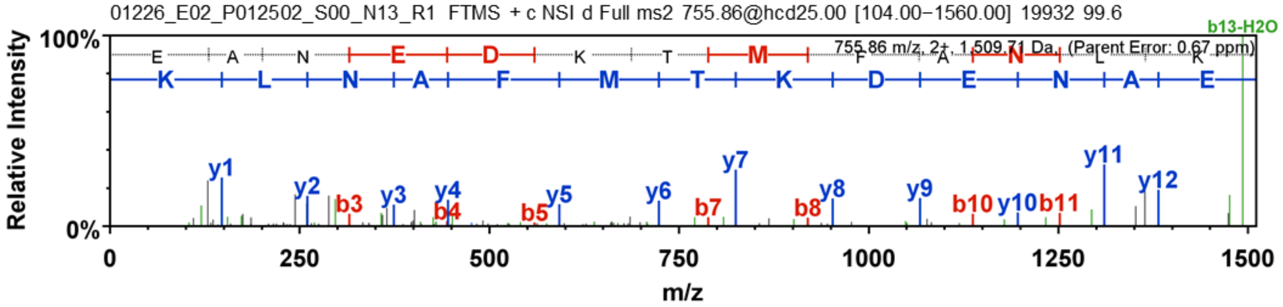


Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 862.4528@hcd28.00 [100.00−1735.00] 36532 231.83

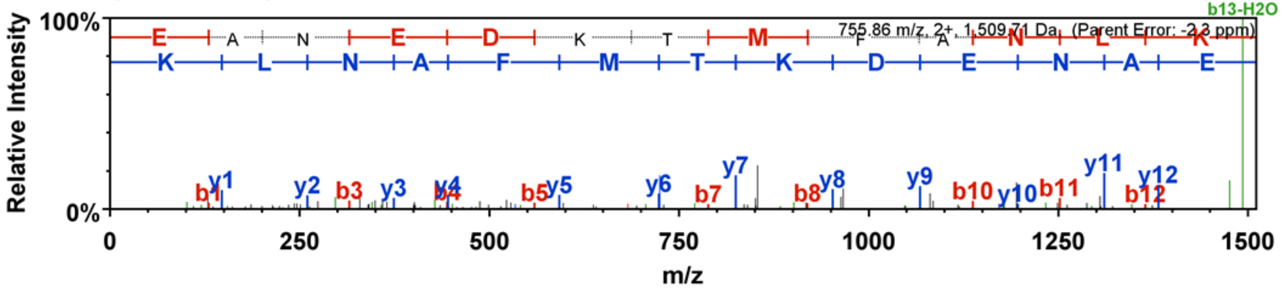


EANEDKTMFANLK

Sample:
01226_E02_P012502_S00_N13_R1 FTMS + c NSI d Full ms2 755.86@hcd25.00 [104.00−1560.00] 19932 99.6

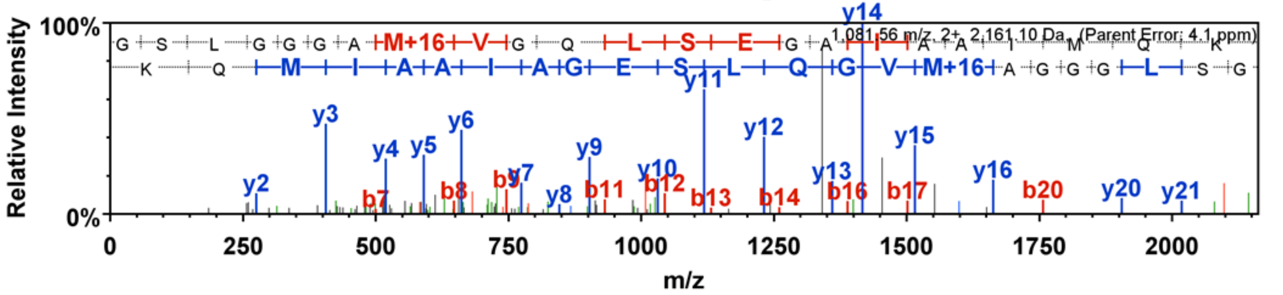


Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 755.8623@hcd28.00
[100.00−1522.00] 31848 221.63

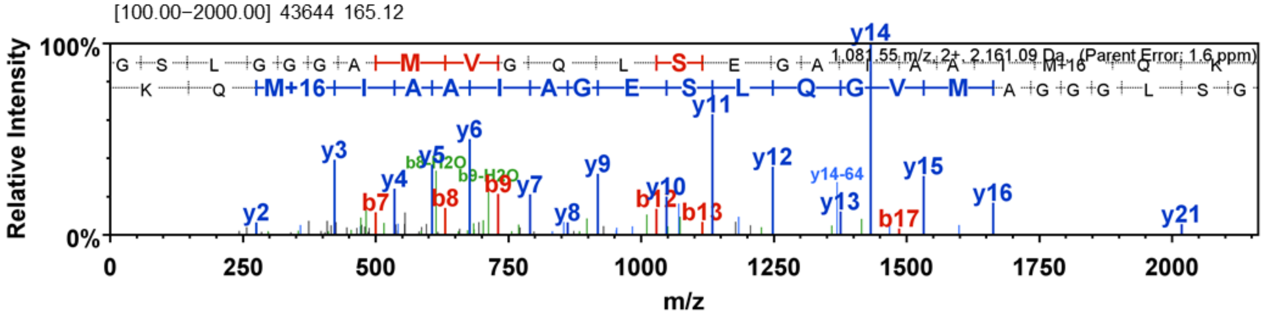


GSLGGGAM(ox)VGQLSEGAIAAIMQK

Sample:
01280_G01_P013164_S00_N07_R1 FTMS + c NSI d Full ms2 1082.05@hcd25.00 [148.33−2225.00] 42633 139.22

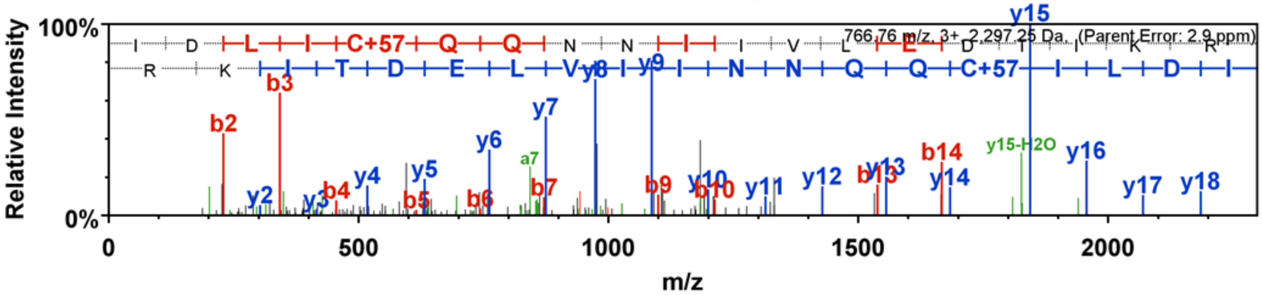


Synthetic:
02383a_BC4-TUM_unmod_proteogenom_4_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 1090.0514@hcd28.00
[100.00−2000.00] 43644 165.12

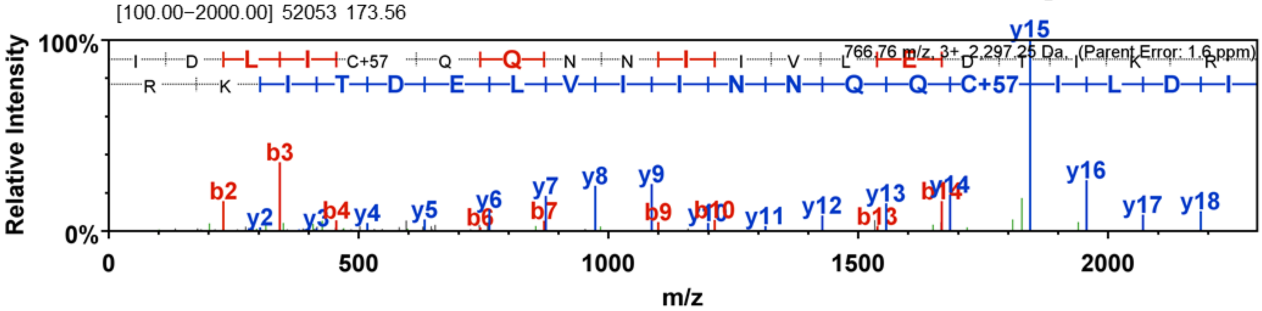


IDLICQQNNIIVLEDTIKR

Sample:
01320_B02_P013559_S00_N10_R1 FTMS + c NSI d Full ms2 767.09@hcd25.00 [157.67−2365.00] 42464 104.48

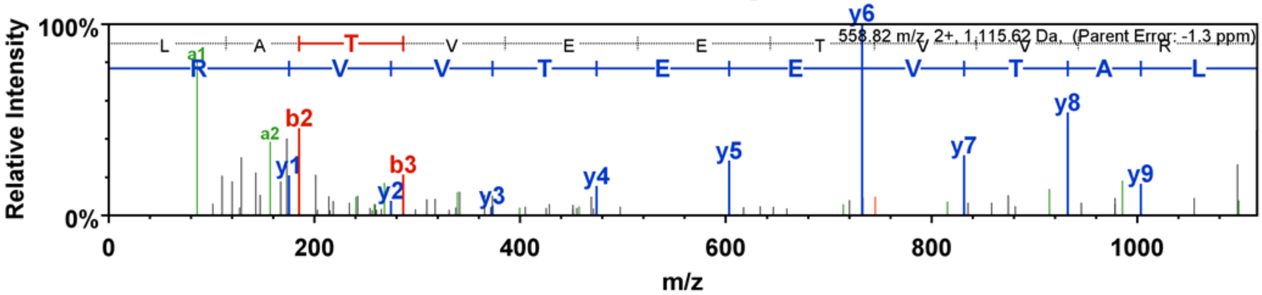


Synthetic:
02383a_BC3-TUM_unmod_proteogenom_3_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 767.0897@hcd28.00
[100.00−2000.00] 52053 173.56

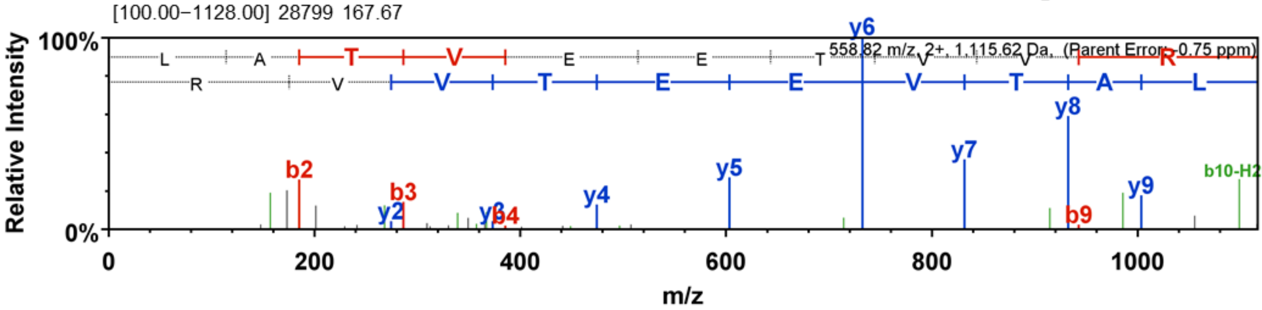


LATVEETVVR

Sample:
01087_C02_P010739_S00_N11_R1 FTMS + c NSI d Full ms2 558.82@hcd25.00 [77.33−1160.00] 16292 42.57

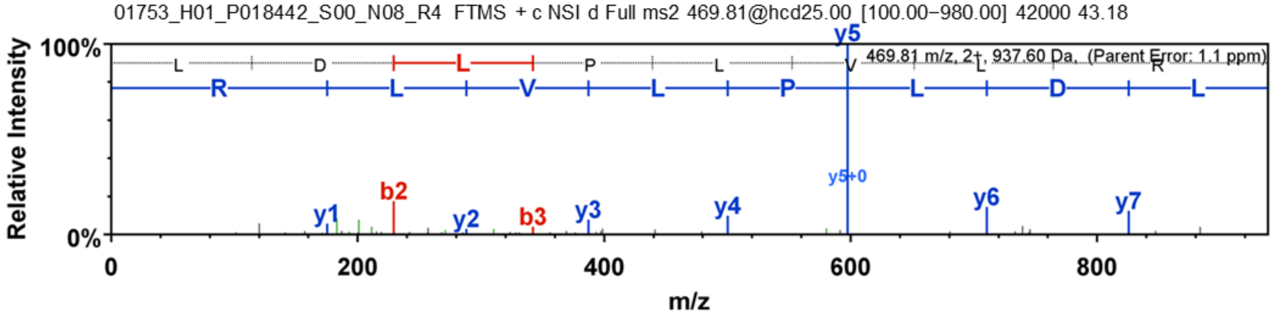


Synthetic:
02383a_BC1-TUM_unmod_proteogenom_1_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 558.8163@hcd28.00
[100.00−1128.00] 28799 167.67

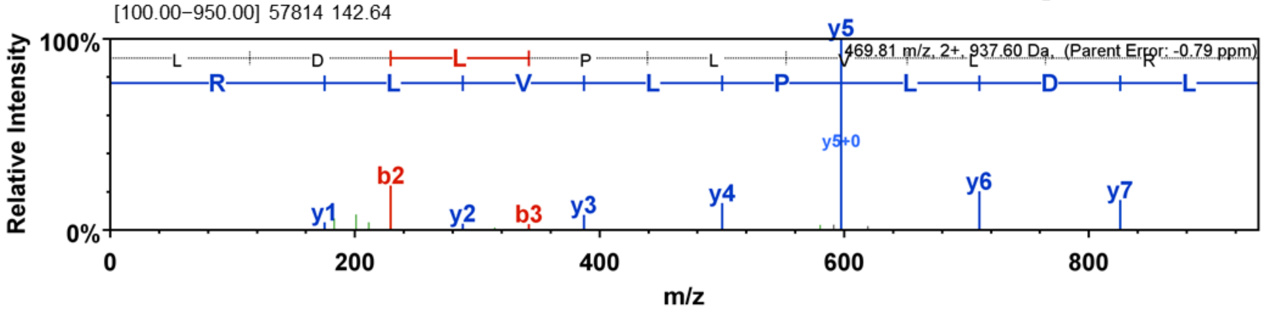


LDLPLVLR

Sample:
01753_H01_P018442_S00_N08_R4 FTMS + c NSI d Full ms2 469.81@hcd25.00 [100.00−980.00] 42000 43.18

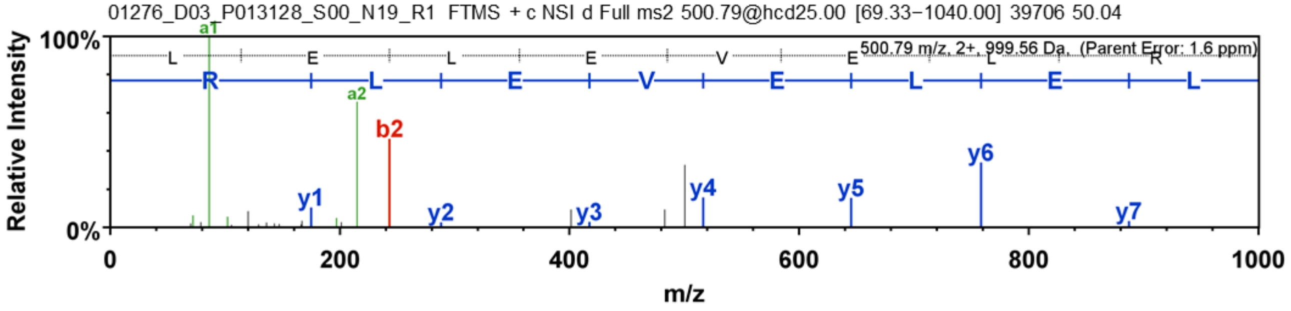


Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 469.8051@hcd28.00
[100.00−950.00] 57814 142.64

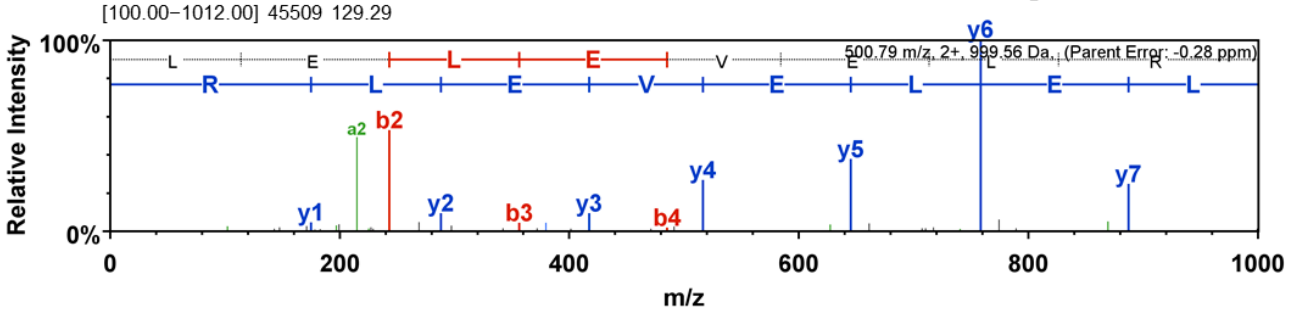


LELEVELR

Sample:
01276_D03_P013128_S00_N19_R1 FTMS + c NSI d Full ms2 500.79@hcd25.00 [69.33−1040.00] 39706 50.04

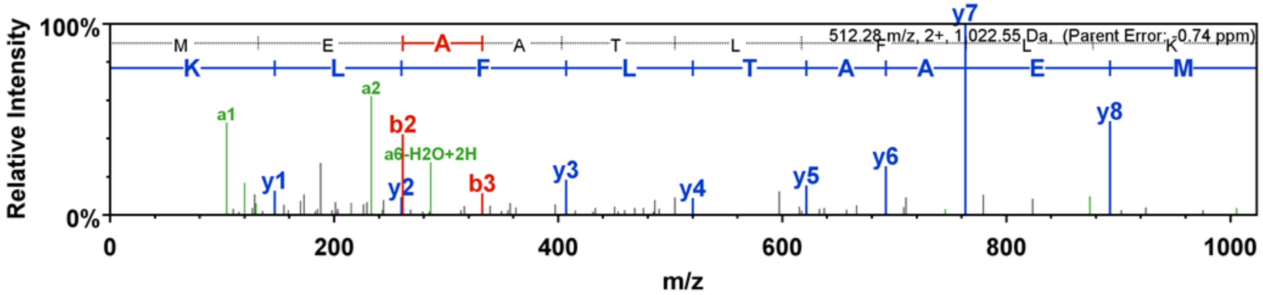


Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 500.7873@hcd28.00
[100.00−1012.00] 45509 129.29

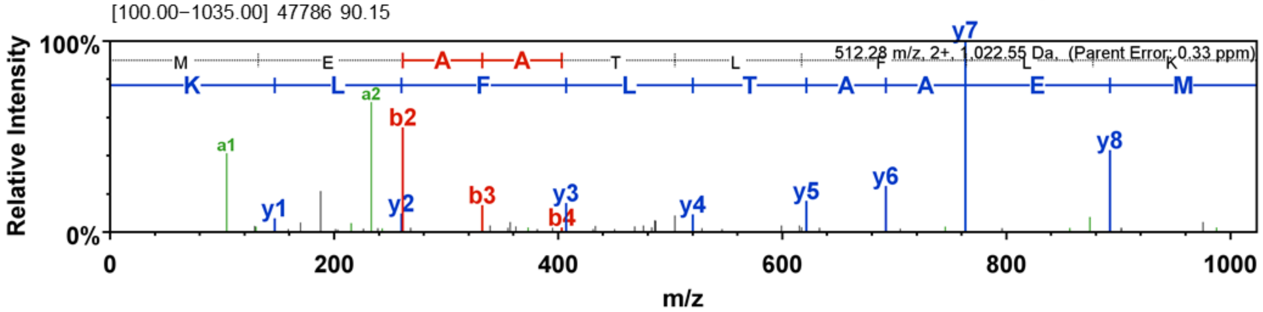


MEAATLFLK

Sample:
01752_A02_P018440_S00_N09_R1 FTMS + c NSI d Full ms2 512.28@hcd25.00 [100.00−1065.00] 33275 43.32

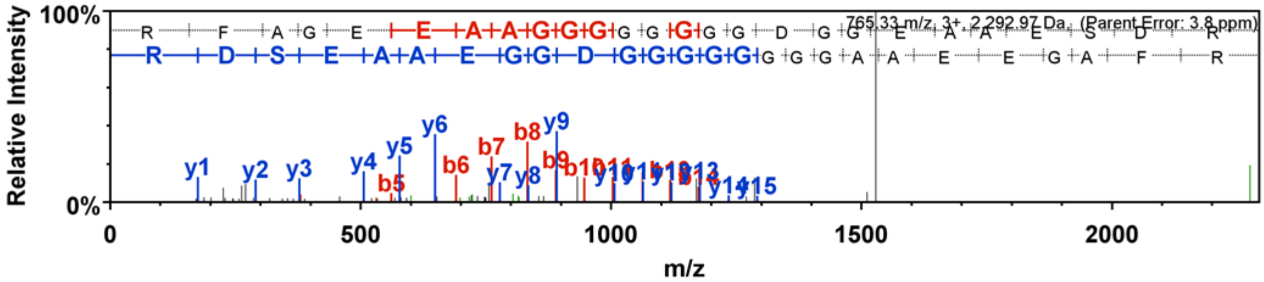


Synthetic:
02383a_BC1−TUM_unmod_proteogenom_1_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 512.2813@hcd28.00
[100.00−1035.00] 47786 90.15

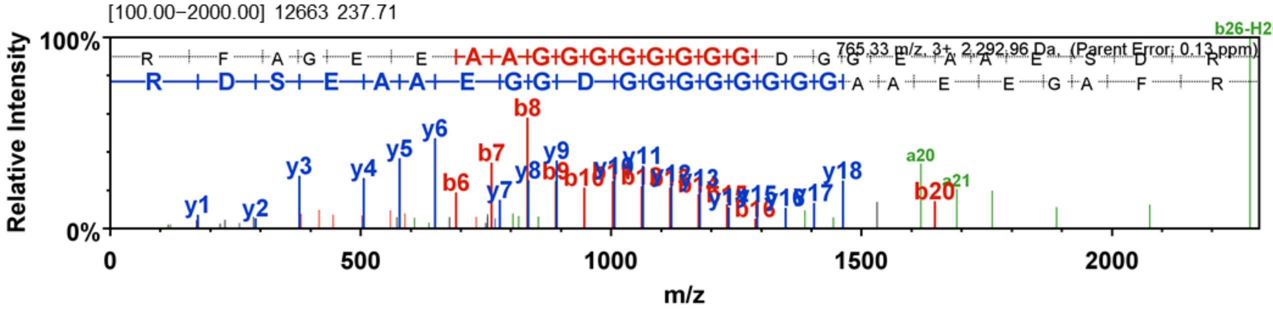


RFAGEEAAGGGGGGGGDGGEAAESDR

Sample:
01320_E03_P013559_S00_N21_R1 FTMS + c NSI d Full ms2 765.33@hcd25.00 [157.33−2360.00] 7562 108.91

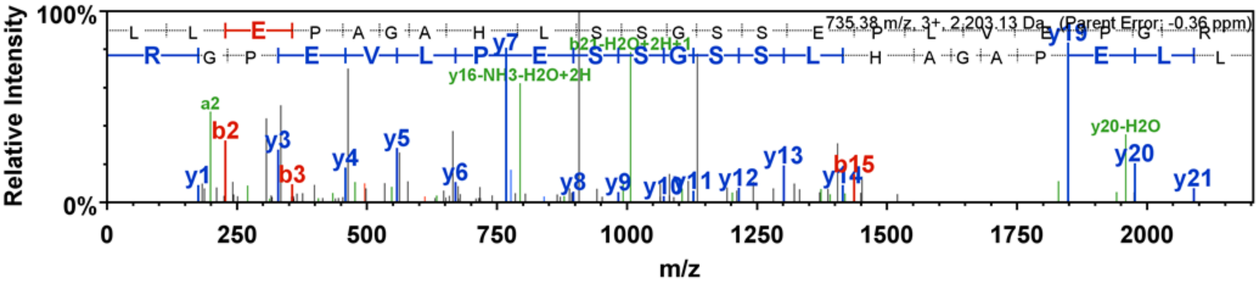


Synthetic:
02383a_BC4−TUM_unmod_proteogenom_4_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 765.6615@hcd28.00
[100.00−2000.00] 12663 237.71

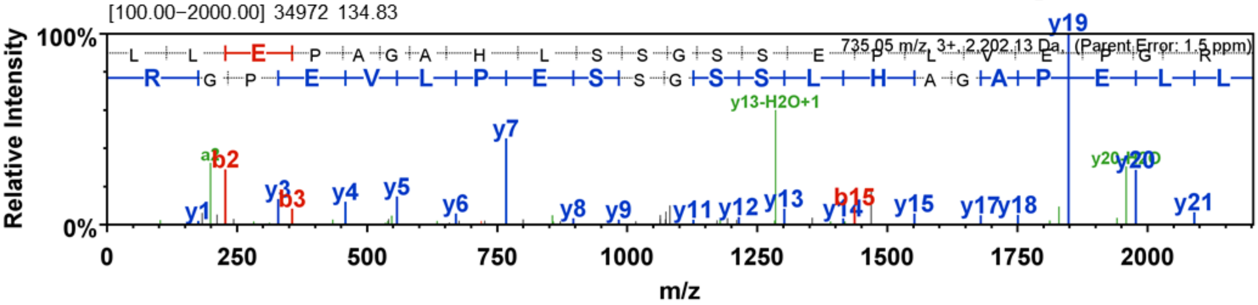


LLEPAGAHLSSGSSEPLVEPGR

Sample:
01280_E02_P013164_S00_N13_R1 FTMS + c NSI d Full ms2 735.38@hcd25.00 [151.33−2270.00] 18355 68.73

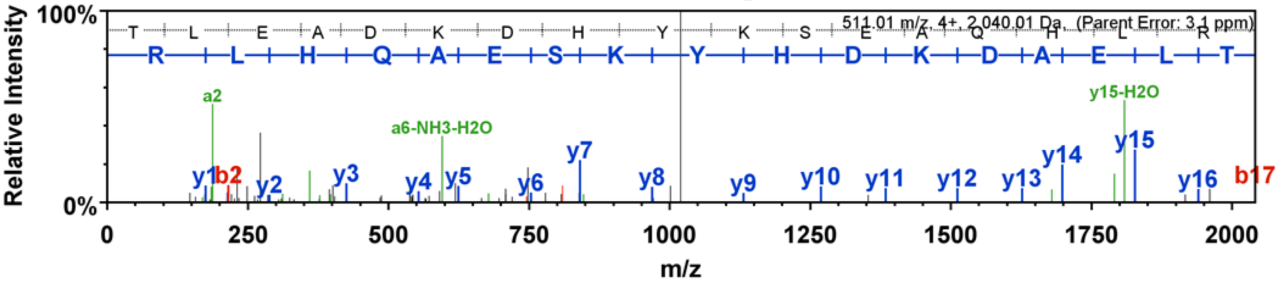


Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 735.3846@hcd28.00 [100.00−2000.00] 34972 134.83

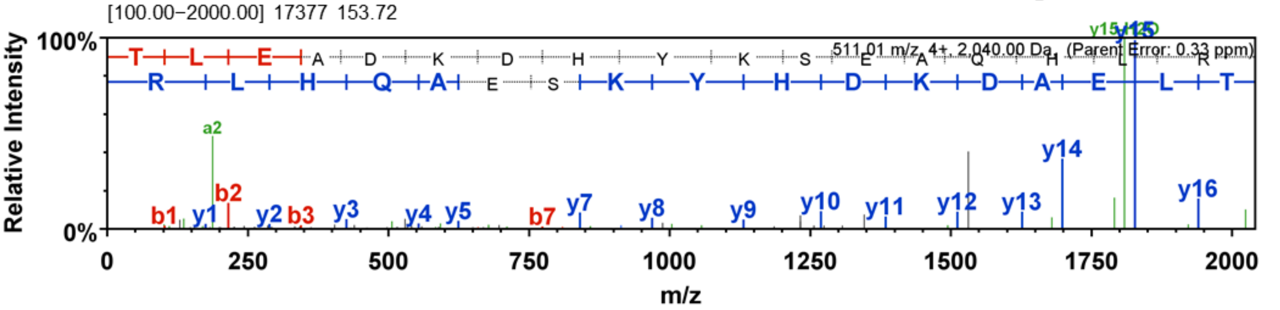


TLEADKDHYKSEAQHLR

Sample:
01320_E02_P013559_S00_N13_R1 FTMS + c NSI d Full ms2 511.01@hcd25.00 [140.33−2105.00] 8332 85.25

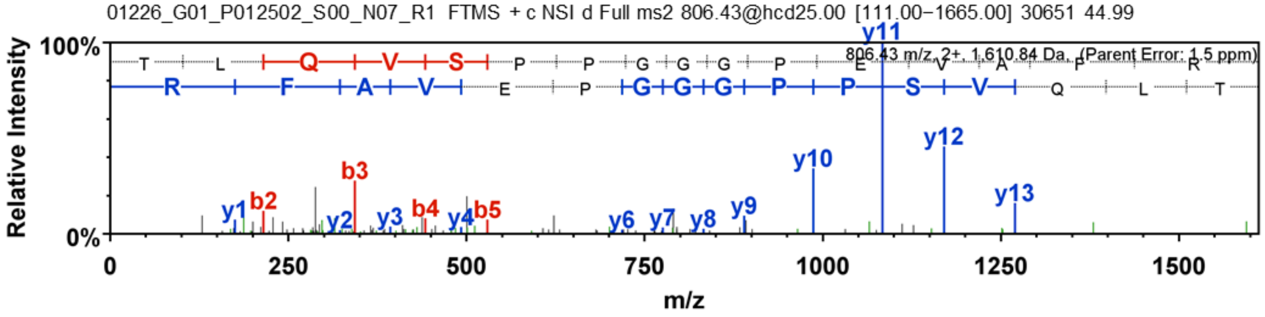


Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 511.2588@hcd28.00 [100.00−2000.00] 17377 153.72

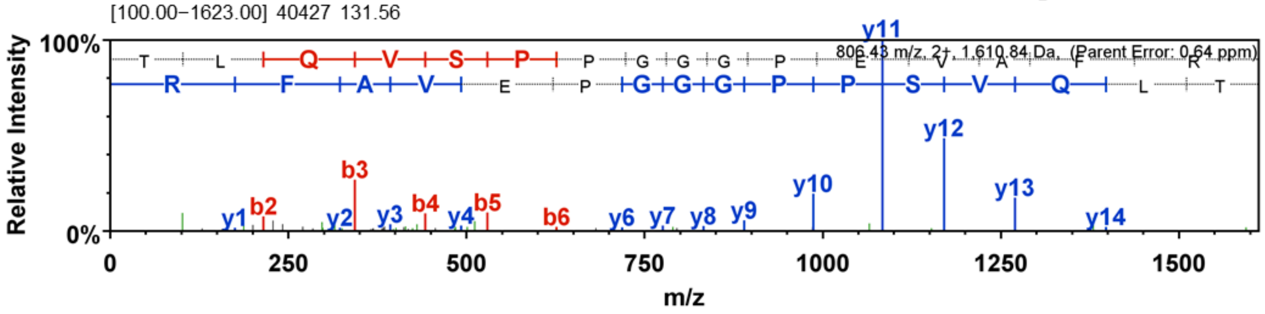


TLQVSPPGGGPEVAFR

Sample:
01226_G01_P012502_S00_N07_R1 FTMS + c NSI d Full ms2 806.43@hcd25.00 [111.00−1665.00] 30651 44.99

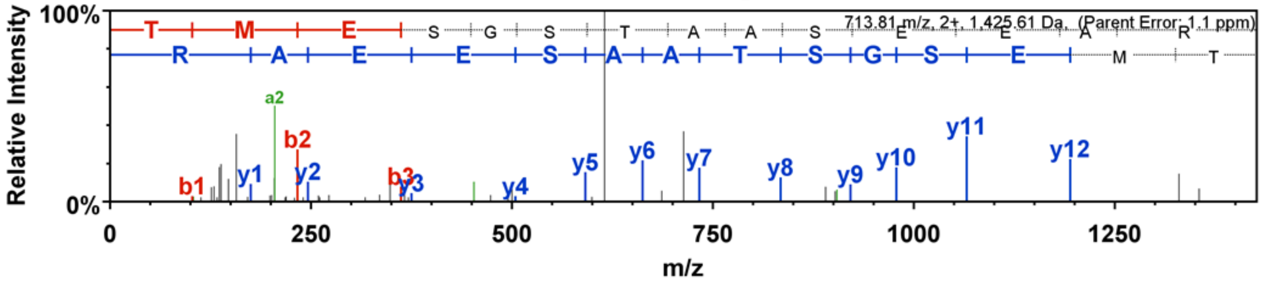


Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 806.4286@hcd28.00 [100.00−1623.00] 40427 131.56

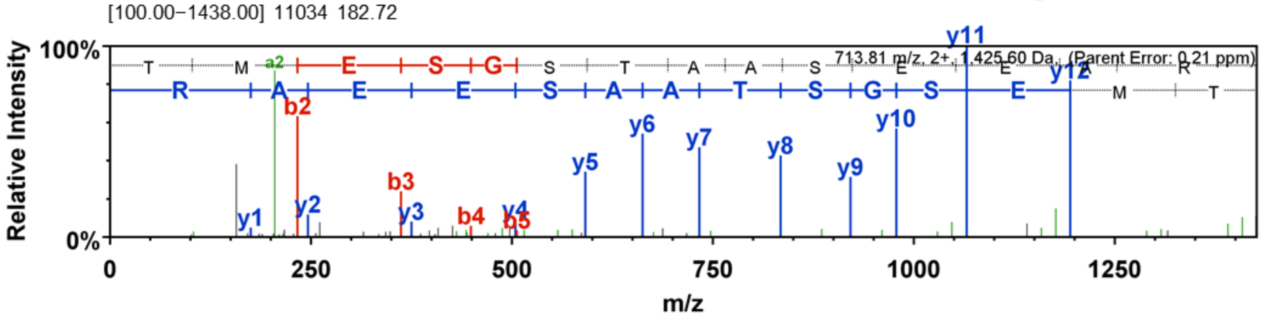


TMESGSTAASEEAR

Sample:
01697_F02_P018020_S00_N14_R2 FTMS + c NSI d Full ms2 713.81@hcd25.00 [100.00−1475.00] 11218 56.22

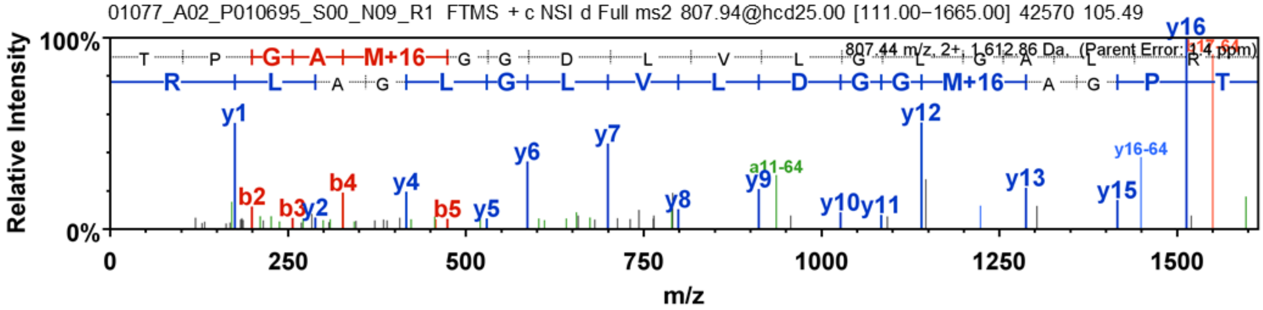


Synthetic:
02383a_BC2−TUM_unmod_proteogenom_2_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 713.8093@hcd28.00 [100.00−1438.00] 11034 182.72

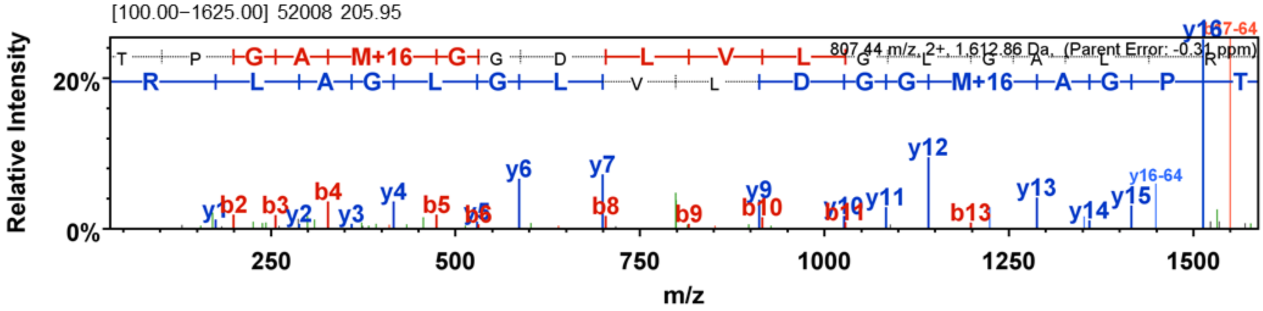


TPGAM(ox)GGDLVLGLGALR

Sample:
01077_A02_P010695_S00_N09_R1 FTMS + c NSI d Full ms2 807.94@hcd25.00 [111.00−1665.00] 42570 105.49

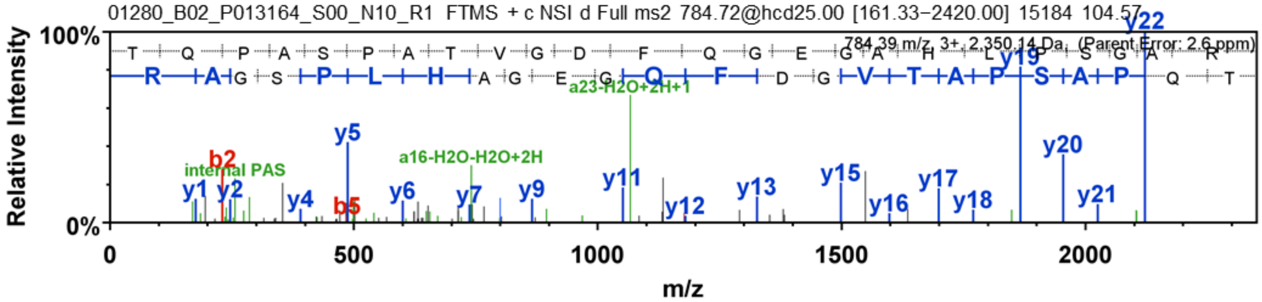


Synthetic:
02383a_BC3−TUM_unmod_proteogenom_3_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 807.4388@hcd28.00
[100.00−1625.00] 52008 205.95

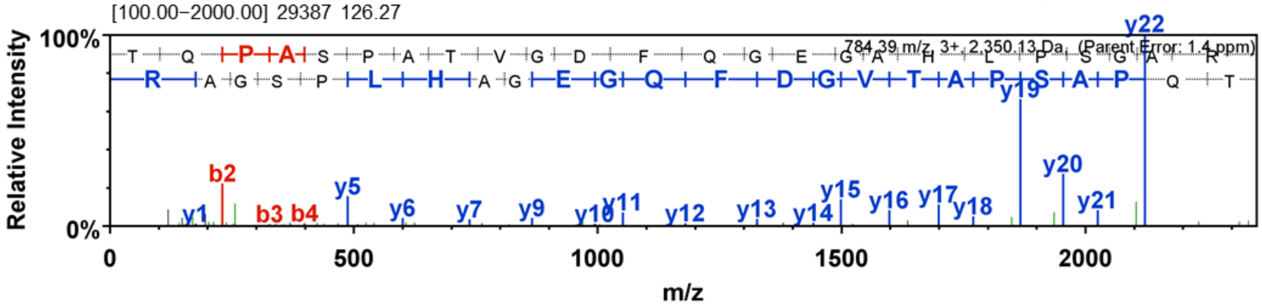


TQPASPATVGDFQGEGAHLPSGAR

Sample:
01280_B02_P013164_S00_N10_R1 FTMS + c NSI d Full ms2 784.72@hcd25.00 [161.33−2420.00] 15184 104.57

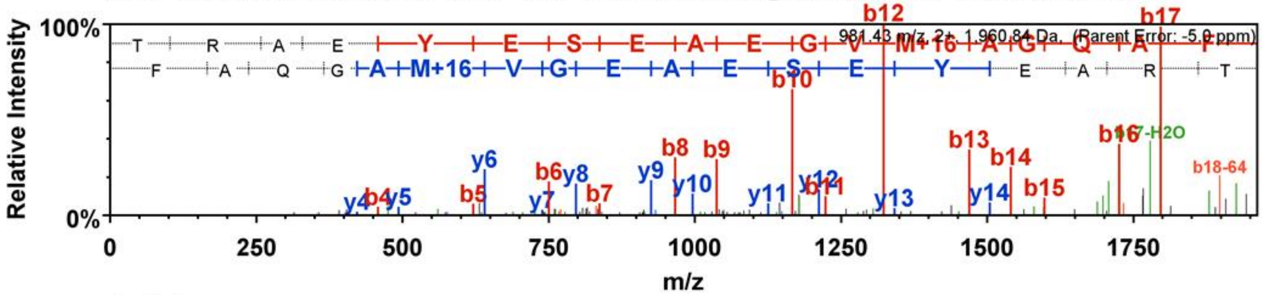


Synthetic:
02383a_BC4−TUM_unmod_proteogenom_4_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 784.7189@hcd28.00
[100.00−2000.00] 29387 126.27

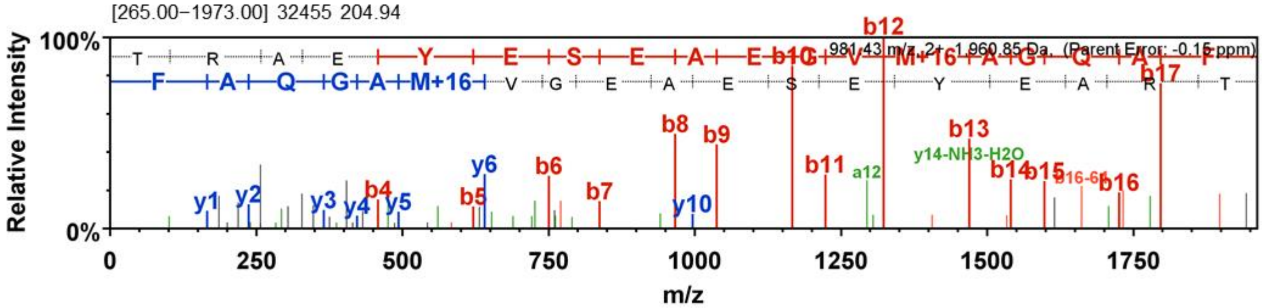


TRAEYESEAEGVM(ox)AGQAF

Sample:
01781_A01_P018699_S00_N23_R3 FTMS + c NSI d Full ms2 981.4308@cid35.00 [265.00−1973.00] 21762 109.14

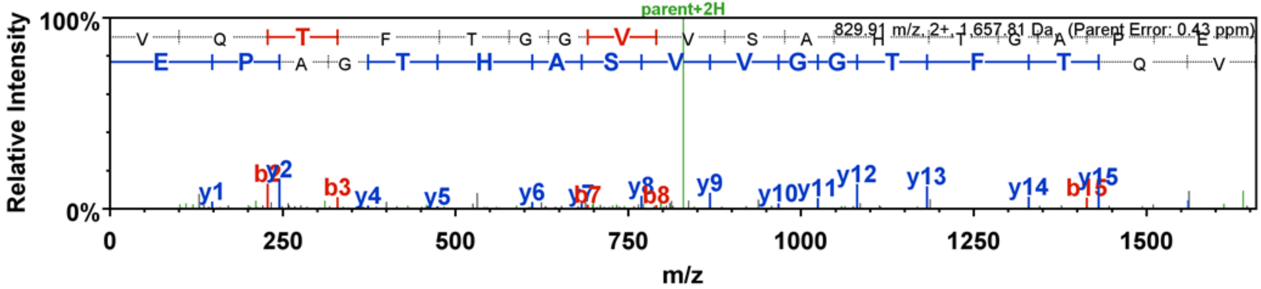


Synthetic:
02383a_BC5−TUM_unmod_proteogenom_5_01_01−DDA−1h−R1 ITMS + c NSI r d Full ms2 981.4319@cid35.00 [265.00−1973.00] 32455 204.94

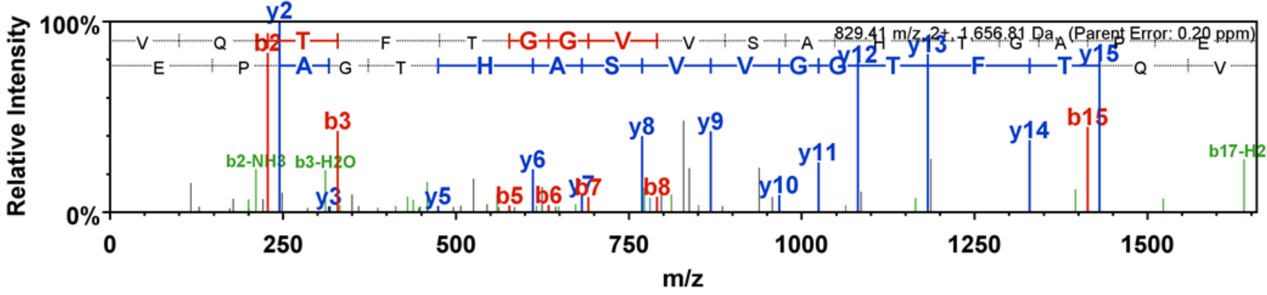


VQTFTGGVVSAHTGAPE

Sample:
01752_F11_P018441_S00_N11_R1 FTMS + c NSI d Full ms2 829.91@hcd25.00 [100.00−1710.00] 23083 74.38

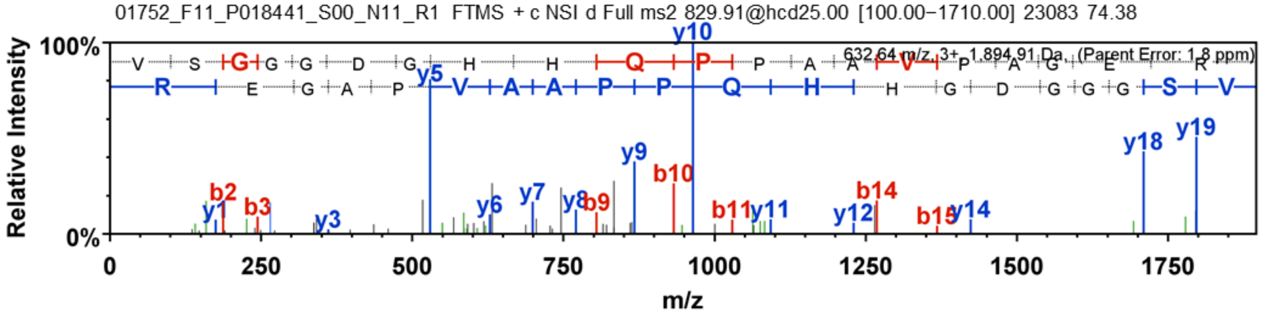


Synthetic:
02383a_BC5−TUM_unmod_proteogenom_5_01_01−DDA−1h−R1 FTMS + c NSI d Full ms2 829.4130@hcd28.00 [100.00−1669.00] 29403 223.14

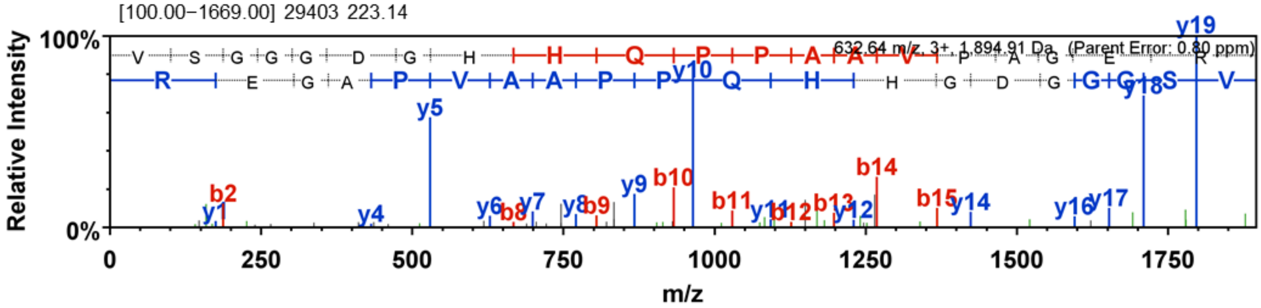


VSGGGDGHHQPPAAVPAGER

Sample:
01752_F11_P018441_S00_N11_R1 FTMS + c NSI d Full ms2 829.91@hcd25.00 [100.00−1710.00] 23083 74.38



Synthetic:
02383a_BC5-TUM_unmod_proteogenom_5_01_01-DDA-1h-R1 FTMS + c NSI d Full ms2 829.4130@hcd28.00 [100.00−1669.00] 29403 223.14


Relative Intensity
632.64 m/z, 3+, 1,894.91 Da, (Parent Error: 0.80 ppm)
V S G G G D G H H Q P P A A V P A G E R
R E G A P V A A P P Q H H G D G G G S V
y19
y10
y18
y5
b14
b10
y9
b2
y1
y4
y6
b8
y7
y8
b9
b11
y11
b12
b13
y12
b15
y14
y16
y17
m/z