Pre-Processing of LC-MS data in Proteogenomics

Sebastian K. Kristensen (S3020029)

Supervisor: Prof. Dr. Peter Horvatovich

Examinor: Prof. Dr. Jan Kok

4th of August 2019

Table of content

Abstract

The omics field, in particular genomics, transcriptomics and proteomics, is a fundamental source of knowledge with countless applications used in life sciences. These high throughput strategies have converged synergistically into a new field, proteogenomics. The domain of proteogenomics is still in a relatively early phase compared to other omics technologies, and new innovative approaches, incorporating proteogenomics strategies, are continuously being developed. By generating customized protein databases instead of using canonical public databases, researchers are successfully identifying novel proteins and protein-coding loci, thereby refining gene models and exposing potential drug targets. However, the effectiveness, in terms of data generation and analysis, is yet to be optimized before proteogenomics can lead to standardization of personalized medicine. In this literature study current approaches for overcoming challenges associated with proteogenomics have been reviewed. Key topics addressed in this study include (I) Customized protein database generation, from nucleotide sequencing through the analysis pipeline eventually giving rise to predicted proteins. (II) Database search and false discovery rate (FDR). FDR values are correlated with the peptide score-cut off threshold, specified by the researcher, and can be estimated using the target/decoy strategy. (III) Approaches for improving proteome coverage. When analyzing highly complex protein samples it is a challenge to reach full proteome coverage, since a fraction of proteins may be masked. Several strategies can be applied in order to increase the number of positive protein identifications. Using multidimensional chromatography, the complexity of the sample can be reduced by extensive peptide separation prior to MS. Multiple peptide digestion methods may be applied in order to produce complementary data containing peptides that may not be seen when relying on a single digestion method, such as tryptic digestion. Consistent research in the proteogenomics field will extend the usage of these tools, and as technology advances the potentials are continuously reaching new heights.

Introduction

The consistent progress in the development of high-throughput techniques for macromolecular data analysis is continually changing, and improving, the ways in which cellular content and processes can be studied. Clinical trials, that aim to alleviate and eventually cure disease, highly depend on technologies that can unravel the genetic and molecular heterogeneity of disease. Advances in genomics, transcriptomics and proteomics play a major role in revolutionizing high-throughput analysis of complex biological data. Moreover, the combined synergy of these fields has led to a relatively new discipline, proteogenomics, in which multiple variants of a protein are coupled to a single gene, and new genes are annotated based on identified proteins. Proteogenomics allows powerful and thorough annotation of expression data, exposing novel drug targets, and thereby leads the way to an era of personalized medicine.

Genomics

All the information that dictates a cells dynamic regulatory behavior is embedded in the genome. Mapping of genome sequences has enabled the prediction of proteins, and has provided groundbreaking information in the fields of cellular- and evolutionary biology. The automated Sanger sequencing method is considered as a "first generation" technology. The Sanger method dominated the industry in the early days of sequencing technology, where it led to immense achievements, including the completion of the first human genome sequence[1,2]. The limitations of the Sanger method, in terms of cost efficiency and labor intensity, showed a need for improved sequencing technologies. Next Generation Sequencing (NGS) constitute various methods that depend on a combination of template preparation and amplification, sequencing technique, nucleotide detection, genome alignment and sequence assembly. For DNA sequencing, the two main options are Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES). WGS gives the complete genome sequence, while WES covers the coding part of the genome[1]. Hence, WES is often the most efficient choice. The numbers of unique reads that include a given nucleotide, the sequencing depth, determine the quality of data obtained after a sequencing experiment. A typical protocol, using the Illumina/Solexa sequencer, is shown on figure 1. The sequential events of sample preparation and sequencing are as follows[1]:

1. Extraction and isolation of the desired nucleic acids.

2. Fragmentation (enzymes or sonication) to obtain shorter pieces of nucleic acid for efficient sequencing. Fragments of desired length are selected, usually by size exclusion. Adapters (including primer binding sequence) are ligated to each fragment, which allows them to be immobilized on the surface, where the sequence amplification and sequencing takes place.

3. Amplification by bridge PCR creates clusters of the same fragment, each in a limited area.

4. Sequencing, using modified nucleotides called "labeled reversible terminators". Each base (Adenine, Cytosine, Thymine and Guanine) is labeled with a different fluorophore, and will temporarily terminate the polymerase chain reaction. Detection of the emitted wavelength tells which fluorophore, and thereby which base, was added. The dye and the terminator group are cleaved from the incorporated nucleotide, and the PCR will continue. The cycle of incorporation, termination and cleavage is repeated for a fixed number of times until the whole nucleic acid fragment sequence is read. The sequencer transforms the raw fluorescence readings into base pairs and saves it in FastaQ format.

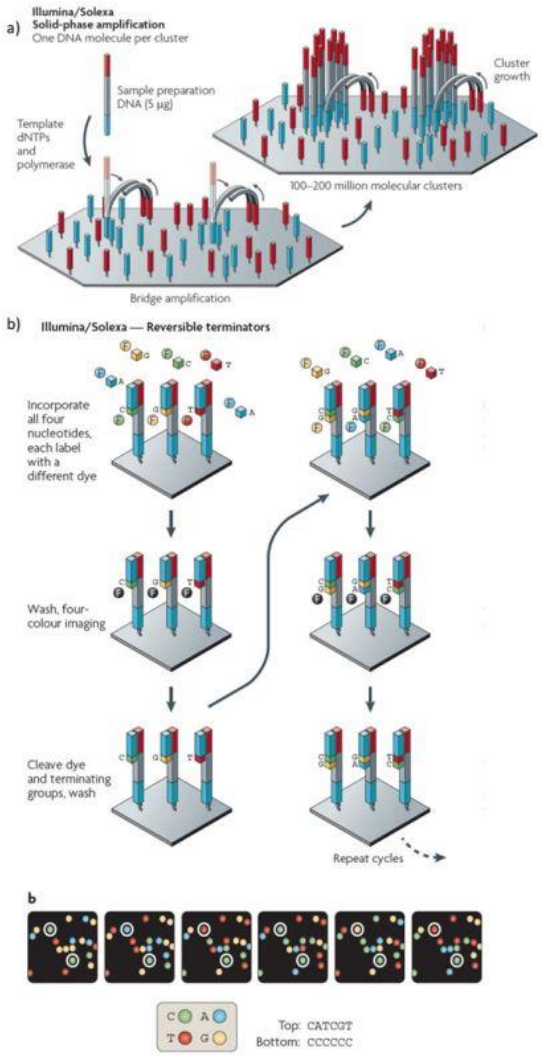

Figure 1. Sequencing of nucleotide sequences, using the Illumina/Solexa sequencer. Details of the individual steps are described in the main text. Figure adapted from [1].

Transcriptomics

Profiling of gene transcripts provides unique information, which cannot be deduced from the genomic sequence. In the dynamic intracellular environment, the transcriptome represents the vast variety of genes being expressed at a specific time. Great knowledge regarding cell development, differentiation, gene function and regulatory pathways can be obtained, by mapping of mRNA transcripts to genomic loci under different conditions. Next generation sequencing has made it possible to collect genome-wide analyses of gene expression and transcript profiles, known as RNA-Seq[3]. The expression level of a gene can be quantified by the number of mRNA sequence reads that map to the gene. Moreover, RNA-Seq are used to investigate intron/exon junctions, gene fusions and alternative splice variants[3,4]. Hence, using RNA-Seq, the heterogeneity of protein expression can be predicted more accurately, compared to approaches relying exclusively on the genomic sequence. Prior to sequencing, mRNA transcript fragments are reverse-transcribed into cDNA. The fragments are subsequently sequenced as shown in figure 1.

Proteomics

Proteomics makes it possible to directly assess protein expression profiles, without relying on gene expression as a proxy. Mass spectrometry (MS) is used for high-throughput protein identification and quantification. Data are generated by analysis of peptide m/z values and ion intensities. A widely used approach is the bottom-up method, also called the shotgun method, which consists of liquid chromatography (LC) separation of a digested protein mixture followed by tandem mass spectrometry (MS/MS). A typical MS protocol is shown on figure 2. The workflow steps are as follows[1,5].

1. Extraction of proteins from cell lysate.
2. Enzymatic digestion of proteins into peptides, usually by trypsin.
3. LC separation of peptides.
4. Peptide ionization by ESI or MALDI.
5. m/z separation of peptide ions.
6. Ion fragmentation by collision induced dissociation (CID), where peptide ions are accelerated in vacuum and collided with neutral gas, or by electron transfer dissociation, which uses a negatively charged poly-aromatic compound to transfer energy to the peptide ion, leading to fragmentation of the peptide back-bone. Three different bonds can lead to peptide fragmentation, resulting in six types of fragment ions: a,b,c ions containing the N-terminal and x,y,z ions containing the C-terminal.
7. MS/MS, based on either data-dependent acquisition (DDA) or data-independent acquisition (DIA). In DDA, peptide ions that rise above a noise threshold are subjected to tandem MS and is therefore biased towards high abundant peptides. In DIA, all peptides within a specified m/z window are subjected to tandem MS. The analysis is repeated until the full m/z range of peptides is covered[1,6.]
8. Detection

Following a shotgun MS experiment, there are several approaches that can be used for protein identification (figure 3). A widely used method is comparison of acquired MS spectra to a reference database of proteins expected to be present in the original sample, digested *in silico*. Potential fragments are generated and the theoretical spectra are compared to the experimental MS/MS spectra. A commonly used protein reference database is UniProtKB/SwissProt, which is quality-controlled manually, and contains one predicted protein entry per gene. When comparing experimental data to a database, representing the predicted average proteome, identification of low abundant protein variants can be a challenge. Another protein identification approach is spectral library search, where acquired MS data is compared to previously identified spectra. Sequence tag-assisted database search can be used for protein identification, if a partial peptide sequence is available from the experimental data. *De novo* sequencing can also be used for protein identification, in case the fragment ions, resulting from tandem MS, are sufficient for a full peptide sequence[1,5].

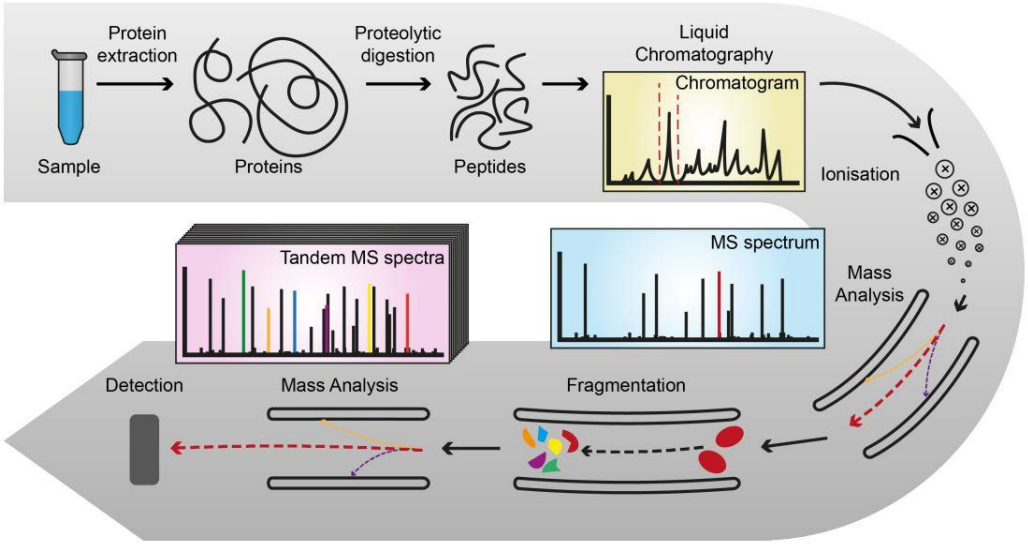

Figure 2. Typical workflow of a tandem MS experiment. Details of the individual steps are described in the main text. Figure adapted from [5].

Proteogenomics

Proteogenomics is a field that interconnects proteomics and genomics in a synergetic fashion. In this approach, customized protein sequence databases are generated using sample-specific genomic and transcriptomic data. These databases are used to identify novel proteins, which are not present in reference protein sequence databases. In turn, the proteomic insight can provide protein-level evidence needed for gene model refinement (figure 3). In recent years, Improvements in the depth and throughput of RNA-seq and mass spectrometry, has led to drastically accelerated proteogenomics research[7].

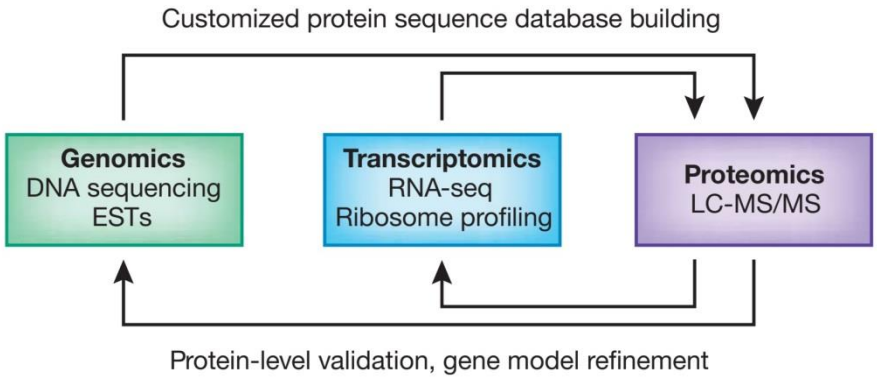

Figure 3. Concept of proteogenomics. Genomic and transcriptomic data are used to generate a customized protein database. Proteomics data are then used to identify novel proteins and protein coding loci, thereby contributing to gene model refinement. Figure adapted from [7].

Generating customized protein databases

Sample specific protein databases are essential in order to reveal the vast variation of protein products coupled to each gene, and to identify novel genomic loci. Experimental acquired MS spectra are compared to the predicted customized database, and protein isoforms, such as splice variants, single nucleotide variants (SNVs) and copy number variants (CNVs) are identified. Customized protein databases can be built using genomic data or RNA-seq data but challenges arise in the transformation of raw transcripts into reliable data used for protein prediction.

Quality assessment[8]

Although commercial sequencing platforms provide a quality control (QC) pipeline, several sequence artifacts, such as read errors and primer/adaptor contamination, often remain in the output data. Therefore, it is good practice to perform QC and filtering at the end-user level. The quality of sequence data is of high importance for subsequent analyses, such as sequence assembly and variant calling. There are several software tools, for example NGS QC Toolkit[8] and FastQC[9], which can be used for QC analysis. Parameters can be specified by the user, and at the end of analysis, High Quality (HQ) filtered data is exported alongside a detailed QC report, including the statistics for each step of processing. Figure 4 shows the workflow of the QC tools for Illumina data.

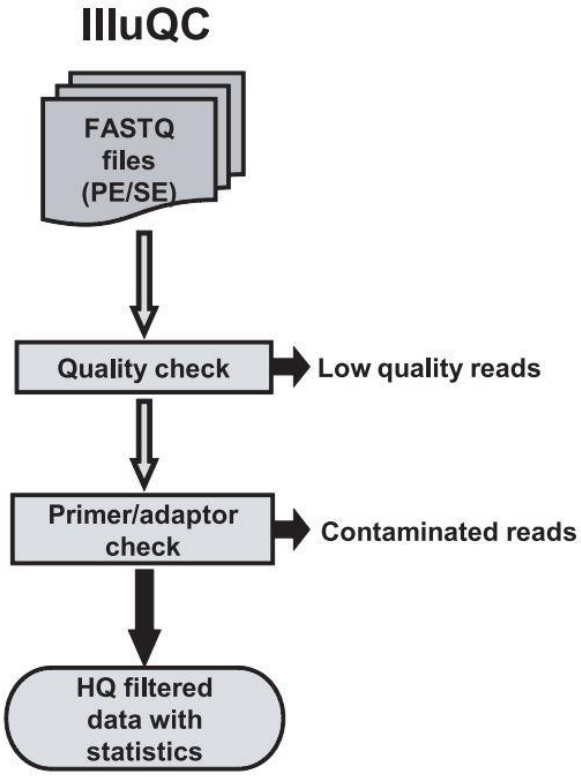

Figure 4. Quality assessment workflow, using raw Illumina data. Checking for low quality and contaminated reads results in high quality filtered data. Figure adapted from [8].

When processing huge amount of data, such as NGS data, computational power is often a bottleneck. Depending on the resources assigned to the analysis, it is advisable to consider the use of parallelization in the initial phase of a QC workflow. One type of parallelization is multiprocessing, where multiple files are processed simultaneously, with one CPU allocated to each file. Another type is multithreading, where a file is split into parts, and each part is processed by a separate CPU. If the input data is paired-end, it is important to maintain the pairing information while running the QC analysis, in order to achieve the most accurate and reliable output. Primer/adaptor contamination should be removed, which can be accomplished by software toolkits. Storage and handling of NGS data through an analysis pipeline can be an obstacle, which can be alleviated by running the QC analysis with a toolkit that supports NGS data in its compressed form (gzip), instead of decompressing the data for analysis.

Alignment and assembly

Reconstruction of full-length transcripts from short reads poses considerable challenges. (**1**) Differences in transcript expression levels, (**2**) Uneven coverage across the transcripts, due to sequencing biases, (**3**) reads with sequencing errors may be derived from highly expressed transcripts, while correct reads may be derived from poorly expressed transcripts, (**4**) Transcripts encoded by adjacent loci can overlap and fuse to form a chimeric transcript, (**5**) Data structures need to handle multiple transcript assignments per locus, due to alternative splicing. Several computational toolkits can be used in an assembly workflow to solve these challenges[10].

There exist two main computational strategies for transcriptome reconstruction. “Mapping first”, first align all the reads to a reference genome and then combine sequences with overlapping alignment. However, reference genomes are often incomplete or lacking, and the use of a reference genome also limits the possibility to discover novel transcripts[1]. “Assembly first”, assemble the reads into full-length transcripts directly, and subsequently align it to a genome if desired [10].

Using the “mapping first” approach, short transcript reads are first aligned to a reference genome, using software, such as STAR[11] or Hisat[12], and the alignment output is stored in a Binary Alignment/Map (BAM) file. The aligned transcripts are subsequently assembled via an assembly toolkit, for example Cufflinks[13]. Moreover, the BAM file can be used as input for a genomic viewer tool, such as IGV (integrative Genome Viewer). If the aim of the study is to reveal novel transcript variants, *de novo* assembly, or “assembly first”, should be applied, in order to overcome the consensus entries in reference genomes. Trinity[10] and Abyss[14] are widely used to assemble transcriptomes and to reveal the variance in isoforms[1]. The assembling algorithm builds on a “De Brujin Graph”, which divides all the sequence reads into directories of *k* length. The algorithm then searches for alignments with *k*-1 overlap. Resulting contigs are again checked for *k-1* overlaps. Eventually, full-length transcripts are compared and analyzed in order to identify isoforms.

Protein prediction

The final procedure in generating a customized protein database is to obtain the sequence of proteins that are most likely to be present in the sample. Protein sequences are obtained from prediction of transcripts, using tools such as TransDecoder[15]. TransDecoder accepts files in General Transfer Format (GTF), which is used to describe genomic entities, and the output file is a Fasta protein sequence consisting of an identifier line followed by the amino acid sequence. Prediction of proteins is based on identifying protein coding regions within the transcripts by deduction of exon structures and translation initiation and termination sites. TransDecoder is able to predict protein sequence from *de novo* assembled RNA transcripts as well as transcripts assembled via genomic alignment.

An Alternative tool for prediction of protein coding regions is the relatively recent developed GeneMarkS-T[16]. GeneMarkS-T is an extension of GeneMarkS, but whereas GeneMarkS predicts only prokaryotic proteins, the predictor in GeneMarkS-T has been modified to translate eukaryotic transcipts. The algorithm parameters used in GeneMarkS-T are estimated by model training, which produces robust and accurate results[16]. A study published in 2014 compared the accuracy of TransDecoder, GeneMarkS-T and a third predicting tool named Prodigal[16] on five sets of *D. melanogaster* transcripts derived from different sources (Figure 5). In all cases GeneMarkS-T Shows a considerable lower fraction of False positives (FP) compared to Transdecoder and Prodigal.

Transcripts built by	No. of transcripts	GeneMarkS-T		Prodigal		TransDecoder	
		TP	FP	TP	FP	TP	FP
Cufflinks	7222	**7162**	**60**	7098	232	7046	432
Augustus	9444	**9423**	**21**	9383	246	9332	480
Exonerate	6971	**6953**	**18**	6940	190	6915	454
Velvet	7344	**7146**	**198**	7096	312	7030	429
Oases	13 869	**13 769**	**100**	13 659	347	13 598	582

Figure 5. Comparative protein prediction using GeneMarkS-T, Prodigal and Transdecoder. Each protein prediction tool is tested with five different nucleotide sources. GeneMarkS-T shows the highest rate of true positive (TP) predictions, independently of the nucleotide source. Figure adapted from [16].

Following the prediction of all amino acid sequences in a specific sample, a customized database can be generated and used in database search (DBS) to identify peptides and proteins in raw shotgun proteomics derived data.

Database search, scoring and false discovery rate

Database search and peptide identification

Following acquisition of experimental MS data, computational approaches are needed for analysis. The main goal is to identify the peptides that produced the measured MS/MS spectra. A common procedure for peptide identification is to correlate experimental peptide MS/MS spectra with theoretical spectra predicted from a protein sequence database. In the field of proteogenomics a sample specific protein database, as previously described, is used. Alternatively, peptide sequences can be derived directly from the spectra using *de novo* sequencing. The advantage of *de novo* sequencing is the possibility to identify peptides with a sequence not present in the database. However, for large scale data analysis this approach requires intensive computational power and high quality MS/MS spectra. Consequently, DBS is often sufficient for researchers to identify peptides and only in some cases, if desired; *de novo* sequencing can be applied to retrieve peptide sequences belong to unassigned MS/MS spectra.[17]

Various computational tools are available for matching peptide sequences to MS/MS spectra by DBS. The program takes MS/MS spectra as input and compares it against theoretical spectra generated by *in silico* fragmentation of peptides from the protein under investigation. The candidate peptide list is generated by the program by applying criteria such as parent ion mass, post translational- or chemical modifications and enzyme digestion constraints (for example tryptic cleavage). Constraining the peptide list has the advantage of reducing the number of spectra comparisons that has to be made, thus increasing analysis efficiency.

The generated output is a list of peptides, each assigned to a MS/MS spectrum. The peptide list is sorted according to search score and the highest scoring peptide/spectrum match (PSM) for each MS/MS spectrum is selected as a potential peptide identification match. Search scores given by the search program is calculated based on the similarity between experimental MS/MS spectra and theoretical database spectra. A variety of scoring algorithms have been implemented in the different database search tools. These include spectral correlation functions in programs such as SEQUEST and MASCOT. The score can either be on an arbitrary scale or be represented as a statistical measure, such as *p*-value.[17]

Score confidence and false discovery rate (FDR) for peptide/spectrum matches

Only a fraction of the peptide/spectrum matches (PSMs) that are listed after MS/MS database search are true matches. There will always be some degree of error and it is essential to estimate the error rate that lies in the data. Error assessment can be addressed on the single PSM level as well as globally in larger datasets. The score that is assigned to each PSM after a database search can be on an arbitrary scale but the search program is able to conveniently convert the score to *p*- or *E*-values, which provides statistical confidence information. These measures, unlike arbitrary scores, do not depend on experimental conditions nor search algorithms, making them comparable between different MS/MS analyses. Nonetheless, when multiple MS/MS spectra are being processed, confidence values representing single spectra are not sufficient. Therefore, large PSM datasets are subjected to additional error estimation.[17]

.

Error estimation in large collections of PSMs is usually quantified by the false discovery rate (FDR). FDR is a global measure of the expected fraction of false PSMs among all accepted PSMs in the dataset. The posterior error probability (PEP) is often referred to as local FDR and indicates the probability of a single PSM to be incorrect.[7,17,18]

An article published in *Nature*[19] assesses the target-decoy search strategy for FDR estimation. In the target-decoy strategy, experimental MS/MS spectra are searched against a database consisting of the original target database of predicted peptides along with a decoy database. The decoy database is essentially a reversed or stochastically randomized version of the target database, including the same number of entries with the same length. The fundamental assumption in this strategy is that accepted matches to the decoy database and incorrect matches to the target database follow the same distribution.

In order to trust that decoy hits are indeed incorrect matches, another assumption must be made; that the overlap between target database and decoy database is negligible. This was tested *in silico* by generation of a peptide list resulting from tryptic digestion of a target proteome, and comparing that list to a correspondent decoy database subjected to the same proteolytic digestion (figure 6). It appears from the figure that only very short peptides were found in both target and decoy database. According to the study the similarity between target and decoy database is negligible when the peptide length exceeds 8 amino acids. It is also stated that all examined databases in the study showed an average peptide length greater than nine amino acids after tryptic digestion, suggesting the assumption of no overlap to be reasonable.[19]

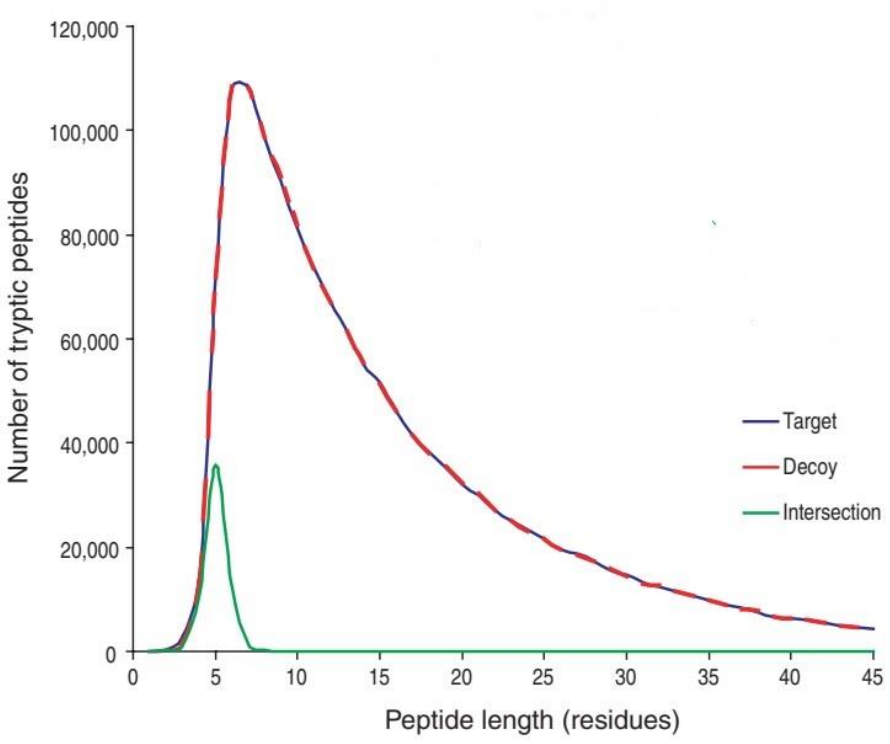

Figure 6. Tryptic digestion of a corresponding target and decoy database. Peptide overlap between the two databases is negligible. Figure adapted from [19].

PSMs are filtered using score cut offs and the FDR is calculated for the correspondent cut off value as D/T, where D is the number of decoy PSMs and T is the number of target PSMs. Referring back to the first assumption for the target-decoy strategy, the FDR can be estimated as the number of decoy hits relatively to target hits. Figure 7 illustrates the target/decoy strategy.[17,19]

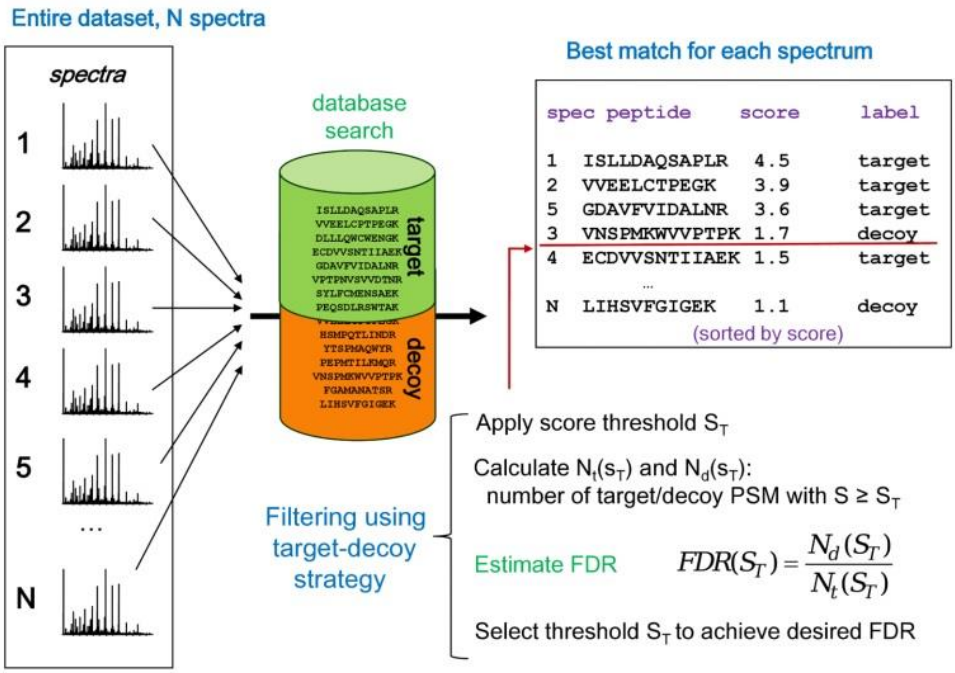

Figure 7. Target/decoy strategy. MS/MS spectra are matches to a target/decoy database. Peptide matches are sorted by score and the desired FDR value can be achieved by adjusting the peptide score cut/off, as the two values are directly correlated. Figure adapted from [17].

A 2014 review article published in *Nature* emphasizes the importance of separating peptides into classes when estimating statistical confidence[7]. They argue that peptides identified using proteogenomic approaches should require a degree of evidence appropeiate to the peptide class. In this manner, peptides identifying rare events such as novel SNPs should require a more strict score-cut off value compared to peptides identifying common events. Applying a strict score cut-off for novel peptides increases the probability that the peptide will only be accepted as a PSM if it is matching correctly to the customized database, and not falsy matching to a homologous sequence. Thus, when estimating FDR, It is recommended to perform the analysis separately for each peptide class.

The paper moreover suggests that database search tools could be improved in order to explicitly assign peptides to a class. The lack of peptide class separation in several published proteogenomic studies is critically addressed in the article. It is mentioned that two recent large-scale human studies applied the same score cut-offs for all peptides. Hence, it is likely that the actual error rates among novel peptides are higher than what is reported.[7]

Improving sequence coverage

Covering the complete proteome is a huge challenge that requires improvement of conventional proteomics strategies. Rare isoforms are often in low abundance and therefore hard to detect. Especially higher eukaryotes with large genomes result in complex samples containing tens of thousands of peptides. Approaches that researchers use in order to increase separation power include multidimensional chromatography and improved protein digestion.

Multidimensional chromatography

Multidimensional chromatography combines multiple separation dimensions, thereby increasing peak capacity and dynamic range of peptides. This technology can be coupled to mass spectrometry resulting in so called LC_n-MS platforms[20]. The high separation power of these platforms increases the effectiveness of peptide identification and quantification in complex samples. A review article published in 2010 thoroughly evaluates the use and benefits of LC_n-MS platforms[20]. The article mentions two fundamental criteria that must be fulfilled for efficient multidimensional chromatography.

First of all, the separations should be chromatographically orthogonal, meaning that the mechanism in each dimension should be independent from each other. Second, separated compounds should be transferred to subsequent dimensions without interfering with the initial separation. Even though the goal is to generate as much separation efficiency as possible, maximum peak capacity is never reached. Peaks are not evenly distributed across each dimension, instead they appear in a more sporadic manner, and the separation mechanisms are never entirely orthogonal. According to the article it is common practice, when analyzing complex proteomics samples, to use low elution rate, low resolution LC columns with large internal diameter at the first dimension and use high resolution LC columns with small internal diameter for the second dimension. If the column diameter is too large in the second dimension, the sample will be diluted resulting in lower dynamic range and decreased sensitivity.

The last dimension in the LC_n-MS platform is the mass spectrometer itself, as it separates components in the eluted LC fractions based on m/z values. Multidimensional chromatography combined with mass spectrometry thereby allow for better detection of low abundant components by reducing overlapping peaks. The maximum peak capacity is directly proportional to the buffer gradient time and is inversely proportional to the peak with. Hence, it is possible to increase the separation efficiency by increasing the gradient time. Researchers usually aim to find the balance between peak capacity and time efficiency. Common techniques used in LC_n-MS platforms include reverse phase (RP), hydrophilic interaction liquid chromatography (HILC), strong cation exchange (SCX), strong anion exchange (SAX), size exclusion chromatography (SEC) and affinity chromatography (AF). Compatibility between separation dimensions must be considered when designing an experiment. An example of a common setup is strong ion exchange (SCX or SAX) in the first dimension, reverse phase in the second dimension, coupled to a mass spectrometer as the third dimension.[20]

Peptide digestion methods

Peptide digestion prior to LC separation is often targeted in order to improve sequence coverage. An article published in *Journal of Proteomics* in 2015 addresses peptide digestion methods in proteogenomics[21]. Since proteins are identified based on peptide matches, the variety of peptides subsequent to digestion is directly correlated to the degree of proteome coverage. The article presents two key effects that proteolytic specificity has on peptide fragmentation and sequence coverage.

First of all, peptide length and mass are affected by the cleavage specificity and frequency. Fragmentation efficiency depends on the peptide size, since larger peptides are more difficult to fragment, especially in collision induces dissociation (CID). Secondly, the distribution of amino acid residues defines the cleavage sites and thereby dictates the distribution of generated peptides.

Trypsin is the most widely used proteolytic enzyme for digestion in proteomics and tryptic peptides will always have arginine or lysine at the C-terminus, except for the C-terminal of the original protein. The length and composition of tryptic peptides makes them very suitable for ionization and fragmentation in mass spectrometry. The average peptide mass resulting from tryptic cleavage of the human proteome is approximately 1kDa. However, the small peptide size is a disadvantage in regards to complete protein sequence coverage.

A promising approach to obtain high sequence coverage is to combine multiple digestion methods and thereby produce complementary data. Pepsin is a nonspecific protease that produces even shorter peptides than trypsin, and is sometimes used in combination with digestion methods that produce larger peptides. The article suggests a peptide mass range between 3 and 10 kDa to be optimal for full protein coverage. This mass range can be achieved by targeting rare residues, such as methionine, tyrosine and tryptophan. Cyanogen bromide is a chemical cleavage agent that targets methionine and produces peptides in the 5 kDa range on average. An alternative and promising cleaving strategy is electrochemical oxidation of tyrosine and tryptophan, which results in peptides of approximately 2.8 kDa on average.

Data integration and applications

The potential and value of proteogenomics strategies have been proved in several studies. By building customized databases and applying methods for high separation power, proteomes have been revealed on a deep coverage level and novel peptides and protein-coding loci have been identified.

One example is a Swedish research group who published an article in *Nature* in 2013, presenting their strategy and findings[22]. Using high-resolution isoelectric focusing (HiRIEF) coupled to LC-MS, they identified 98 and 52 previously unknown human and mouse protein-coding loci respectively. Moreover, the high separation efficiency and resolution of this strategy enabled the identification of more than 13000 human proteins and more than 10000 mouse proteins. Lysate prepared from human A431 and mouse N2A cells were digested with trypsin and prefractionated based on peptide PI. pH range 3.7 – 5.0 was targeted, using the following pH interval HiRIEF gel strips; 3.70–4.05, 4.00–4.25, 4.20–4.45 and 4.39–4.99. Each interval was divided into 72 fractions and analyzed by LC-MS. The MS/MS spectra were compared to a sample specific databased, constructed from genomic sequence data and FDR rates were calculated using the previously describes target/decoy strategy. The experimental workflow is illustrated on figure 8.

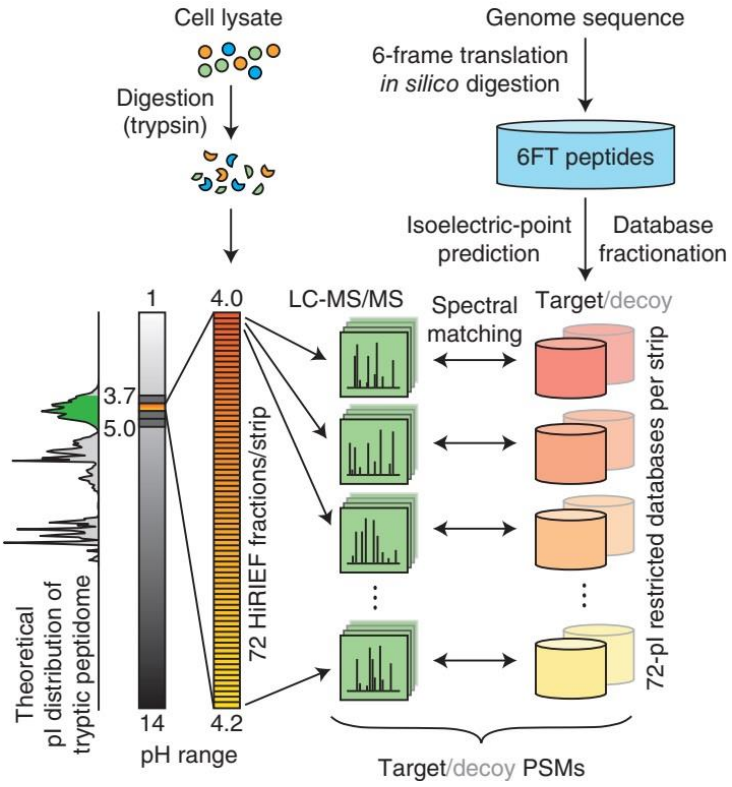

Figure 8. Workflow of a proteogenomics experiment using multidimensional peptide separation. Peptides are separated according to pI and matched to a customized target/decoy database. Figure adapted from [22].

The findings from this study did not only result in refined gene models but also showed evidence of pseudogene expression at the protein level. The protein product of pseudogene MYH16 was validated by LC-MS. MYH16 was previously thought to have lost its protein-coding potential from a premature stop-codon, but the data from this study show that, in the A431 cell line, MYH16 is actively encoding a shorter protein isoform. The A31 cell line is often used in cancer associated studies, and by comparing cell-linage expression profiles with reference databases, light can be shed on cancer-specific proteins. This study shows how innovative strategies in proteogenomics can provide new insight on cell specific protein expression. However, as it is mentioned in the article, this approach could be further improved by expanding the pH range of the HiRIEF separation.

In another study published in 2013 liver tissue from two different rat strains was subjected to proteogenomics analysis[23]. The BN-*Lx* strain is derived from the Brown Norway (BN) strain, which is commonly used as the proteome reference in proteomics studies. The spontaneously hypertensive (SHR) strain is a model often used in hypertension studies. A customized protein database was generated by extending existing rat protein databases with predictions based on genomic and transcriptomic data derived from the two strains. The database construction approach can be seen on figure 9.

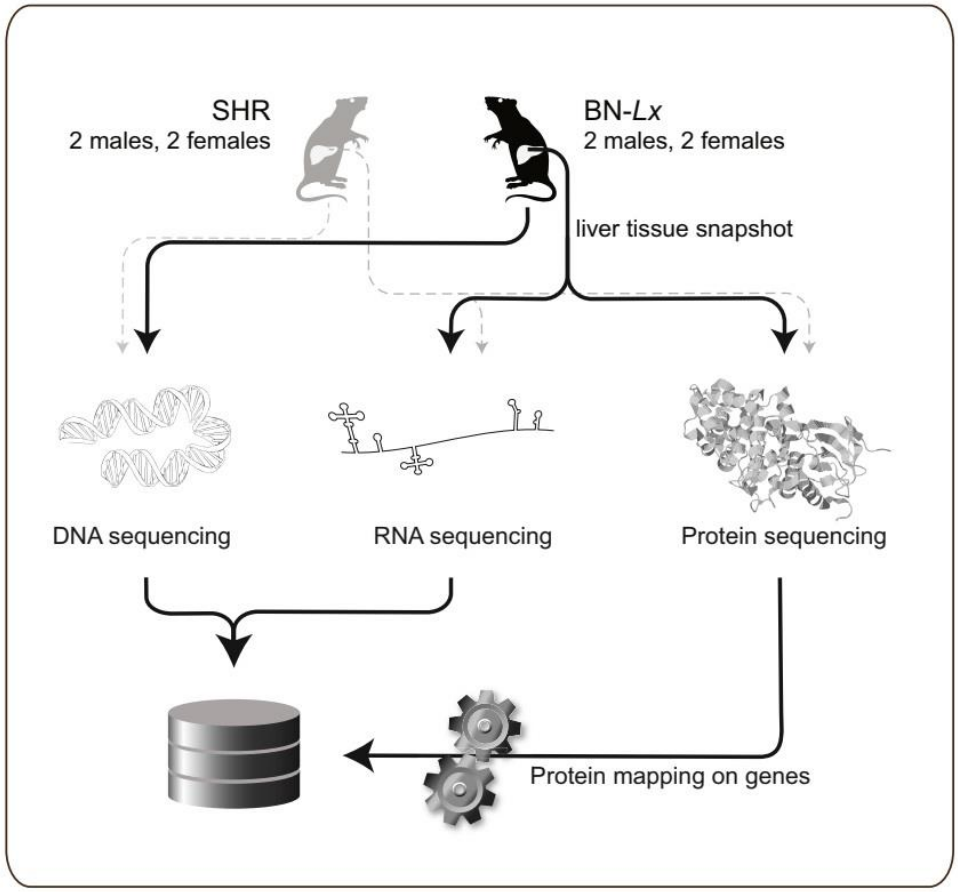

Figure 9. Generation of a customized protein database using genomic and transcriptomic data from two rat strains. Figure adapted from [23].

Liver proteins were digested with five different proteases in order to produce complementary data, and digested peptides were separated using strong cation exchange (SCX) prior to LC-MS/MS. The data yielded 12 million tandem spectra in total and provided evidence for thousands of rat liver proteins. By comparing non-synonymous protein variants from BN-*Lx* and SHR, a list of potential hypertension-associated genes could be generated. This study is an example, showing how strain-specific proteogenomics analysis can be used as a tool to expose protein expression differences between a disease-model and a control group. These advantages surpass the capacity of conventional proteomics approaches. The integration of sample-specific nucleotide sequence data, on genomic and transcriptomic level, with in-depth proteome coverage, enabled peptide/spectrum matches that otherwise could have been missed without notification.

Future aspects

Proteogenomics is still a relatively new phenomenon, yet it has shown promise in leading the path for high-throughput technologies into a synergetic new era. However, reaching full potential depends on the improvement of several technologies. Transitioning from 2nd to 3rd generation nucleotide sequencing will greatly reduce sequencing time and will also produce longer reads, thereby alleviating computational challenges associated with sequence assembly. Recent advances in machine learning and AI will secure continuous and consistent improvements in algorithms used in analysis pipelines, particularly sequence assembly, isoform identification and protein prediction. Increasing peak capacity of peptide separation platforms along with increasing raw computational power needed for MS spectra analysis, will make peptide identification significantly faster. At some point proteogenomics strategies will be efficient enough to identify huge numbers of novel drug targets in a high throughput manner. However, in order to standardize personalized treatment of disease, two challenges must be addressed and overcome. First, mass drug screening will be an essential need and must be optimized. High throughput assays should be developed in combination with chemical libraries. Second, the need for laboratory animal models will increase exponentially. Ethical committees will be an obstacle that may be overcome by futuristic advances in induced pluripotent stem cell (IPSC) research, which could alleviate the requirements for animal models. The field of proteogenomics has launched and only the future will show how far it can reach.

Conclusions

Proteogenomics strategies have enabled researchers to identify novel proteins and protein coding loci. The methods and approaches are not yet applied to its fullest capacity, since massive data generation and analysis require technological progression and optimization. Proteome sequence coverage can be improved in several ways, such as multidimensional separation and complementary peptide digestion, thereby allowing for identification of low abundant and/or poorly ionized peptides. Proteogenomics has proved itself as a valuable and very promising approach, which may change the pharmaceutical industry as we know it and leave mankind in a state of bliss.

Reference list

1. Medicine, E. (n.d.). *Proteogenomics*.

2. Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

3. Bunnik, E. M., & Le Roch, K. G. (2013). An Introduction to Functional Genomics and Systems Biology. *Advances in Wound Care*, *2*(9), 490–498. https://doi.org/10.1089/wound.2012.0379

4. Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., … Mani, D. R. (2017). Methods, Tools and Current Perspectives in Proteogenomics. *Molecular & Cellular Proteomics*, *16*(6), 959–981. https://doi.org/10.1074/mcp.mr117.000024

5. Schlaffner, C. N. (2017). *Proteogenomics for Personalised Molecular Profiling*. (June).

6. Doerr, A. (2014). DIA mass spectrometry. *Nature Methods*, *12*(1), 35. https://doi.org/10.1038/nmeth.3234

7. Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods*, *11*(11), 1114–1125. https://doi.org/10.1038/NMETH.3144

8. Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0030619

9. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

10. Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N., Friedman, & Regev, A. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

11. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

12. Kim D., Landmead B., & Salzberg SL. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat Methods, 12*(4): 357–360. doi:10.1038/nmeth.3317

13. Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2011). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

14. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

15. https://github.com/TransDecoder/TransDecoder/wiki

16. Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research*, *43*(12), 1–10. https://doi.org/10.1093/nar/gkv227

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

18. Aggarwal S & Yadav AK. (2016). False Discovery Rate Estimation in Proteomics. Statistical Analysis in Proteomics, *Methods in Molecular Biology,* vol. 1362. DOI 10.1007/978-1-4939-3106-4_7

19. Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, *4*(3), 207–214. https://doi.org/10.1038/nmeth1019

20. Horvatovich, P., Hoekman, B., Govorukhina, N., & Bischoff, R. (2010). Multidimensional chromatography coupled to mass spectrometry in analysing complex proteomics samples. *Journal of Separation Science*, *33*(10), 1421–1437. https://doi.org/10.1002/jssc.201000050

21. Bischoff, R., Permentier, H., Guryev, V., & Horvatovich, P. (2016). Genomic variability and protein species - Improving sequence coverage for proteogenomics. *Journal of Proteomics*, *134*, 25–36. https://doi.org/10.1016/j.jprot.2015.09.021

22. Branca, R. M. M., Orre, L. M., Johansson, H. J., Granholm, V., Huss, M., Pérez-Bercoff, A., … Lehtiö, J. (2014). HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature Methods*, *11*(1), 59–62. https://doi.org/10.1038/nmeth.2732

23. Low, T. Y., vanHeesch, S., vandenToorn, H., Giansanti, P., Cristobal, A., Toonen, P., … Guryev, V. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Reports*, *5*(5), 1469–1478. https://doi.org/10.1016/j.celrep.2013.10.041

24. Delorme, R., Ey, E., Toro, R., Leboyer, M., Gillberg, C., & Bourgeron, T. (2013). Progress toward treatments for synaptic defects in autism. *Nature Medicine*, *19*(6), 685–694. https://doi.org/10.1038/nm.3193