Clinica Chimica Acta 498 (2019) 38–46

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/cca

Review

Proteogenomics: From next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine

Mia Yang Ang[a,1,2], Teck Yew Low[a,*,2], Pey Yee Lee[a], Wan Fahmi Wan Mohamad Nazarie[a], Victor Guryev[b], Rahman Jamal[a]

[a] UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, 56000 Kuala Lumpur, Malaysia
[b] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen 9713AD, The Netherlands

ARTICLE INFO

Keywords:
Proteogenomics
Genomics
Proteomics
Next-generation sequencing (NGS)
Mass spectrometry (MS)
Genomic variant

ABSTRACT

One of the best-established area within multi-omics is proteogenomics, whereby the underpinning technologies are next-generation sequencing (NGS) and mass spectrometry (MS). Proteogenomics has contributed significantly to genome (re)-annotation, whereby novel coding sequences (CDS) are identified and confirmed. By incorporating *in-silico* translated genome variants in protein database, single amino acid variants (SAAV) and splice proteoforms can be identified and quantified at peptide level. The application of proteogenomics in cancer research potentially enables the identification of patient-specific proteoforms, as well as the association of the efficacy or resistance of cancer therapy to different mutations. Here, we discuss how NGS/TGS data are analyzed and incorporated into the proteogenomic framework. These sequence data mainly originate from whole genome sequencing (WGS), whole exome sequencing (WES) and RNA-Seq. We explain two major strategies for sequence analysis *i.e.*, *de novo* assembly and reads mapping, followed by construction of customized protein databases using such data. Besides, we also elaborate on the procedures of spectrum to peptide sequence matching in proteogenomics, and the relationship between database size on the false discovery rate (FDR). Finally, we discuss the latest development in proteogenomics-assisted precision oncology and also challenges and opportunities in proteogenomics research.

1. Introduction

Although mass spectrometry (MS)-based proteomics is closely linked to genomics, these two disciplines have remained in relative isolation. It is not until recently that the breaking down of the silos took place, and the fusion of proteomics with genomics contributed to an emerging field called “proteogenomics” [1,2]. With proteogenomics, biologists have produced insightful research that can neither be achieved by genomics or proteomics alone. Among these, the best-documented contribution of proteogenomics lies in the annotation of newly sequenced non-model organisms [3]. Besides, proteogenomics is applied in genome re-annotation whereby it has helped to correct mis-annotated genes [4], and to detect novel coding sequences (CDS) previously thought to be non-coding; such as pseudogenes, short open reading frames (sORFs) and other non-coding RNA genes [5,6]. Apart from that, it has assisted in unraveling the undiscovered protein counterparts of annotated genes, the so-named “missing proteins” [7]. Importantly, by incorporating *in-silico* translated coding variants in protein FASTAs, single amino acid variants (SAAV), post-transcriptional modifications and splice proteoforms [8–10]; in addition to products of alternative frames of translation initiation and termination can be identified and quantified at peptide level [11,12].

Precision medicine is frequently associated with genomic approaches. Notably, in precision oncology, clinical sequencing has led to cancer genomics whereby genome sequence data are not only used for diagnosis, prognosis, management and stratification of patients based on cancer sub-types; but also for guiding the selection of appropriate therapeutic regimes, as well as the development of new combinatorial and molecular targeted therapies [13]. However, it was documented that not all cancer patients responded to targeted therapies that are

[*] Corresponding author at: UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia (UKM), UKM Medical Centre, Jalan Yaacob Latiff, Bandar Tun Razak, 56000 Kuala Lumpur, Malaysia.

E-mail address: lowteckyew@ppukm.ukm.edu.my (T.Y. Low).

[1] Current addresses: a) Department of Gene Diagnostics and Therapeutics, National Center for Global Health and Medicine, Tokyo 162-8655, Japan b) Department of Clinical Genome Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan

[2] These authors contributed equally.

https://doi.org/10.1016/j.cca.2019.08.010
Received 22 May 2019; Received in revised form 13 August 2019; Accepted 13 August 2019
Available online 14 August 2019

tailored based on individual tumor genome profile [14]. Furthermore, genomics alone does not fully connect genotypes to disease phenotypes, as it cannot relate to protein modification, signaling pathways and the microbiome [15]. Thus mutational profiles of tumors cannnot fully explain or predict patient outcomes. It is precisely at this juncture that proteogenomics can play a role in precision medicine and clinical diagnostics. By combining proteomics with genomics, the following issues can potentially be addressed: (i) the validation of whether novel, cancer-specific mutations are translated into proteins, (ii) the poor correlation of mRNA abundance to protein abundance, and (iii) the identification and quantification of post-translational modifications that are known to involve in cancer development.

2. Next-generation sequencing (NGS)

Lying at the upstream of proteogenomics are next-generation sequencing (NGS) technologies. NGS is performed to decipher the genomes and the transcriptomes, two major data modalities in proteogenomics; but has since been expanded to the epigenomes, transcription factor-bound or ribosome-bound transcripts and 3D chromosomal landscape. To decipher the genomes, whole genome sequencing (WGS) is used to sequence the total DNA content; so that the context and complexity of genomic alterations including point mutations, insertion–deletion mutations (INDELs) and copy number variations (CNVs) can be identified [16]. Whereas, in whole exome sequencing (WES), only protein-coding exonic sequences are sequenced, as only the exonic fragments are enriched before sequencing [17]. As for the transcriptomes, RNA-Seq sequences allows the detection of co/post-transcriptional modifications, such as alternative spliced variants (ASVs), RNA editing and fusion transcripts; as well as the quantification of gene expression levels [18].

3. Analysis of NGS data

Post sequencing and data quality control, two major strategies are applied for analyzing genome data: (i) *de novo* assembly or (ii) alignment (reads mapping). In strategy (i), NGS reads are assembled into longer sequences without the aid of a pre-existing reference [19]. Once fully assembled genomes become available, a reference genome can be built from one or multiple of these genomes, taking into consideration the complexity and heterozygosity of multiple genomes such as the presence of SNPs, INDELs, and copy number polymorphisms (CNPs). Once a reference genome is available, it can serve as a template for the second strategy, against which sequencing reads are mapped.

3.1. *De novo assembly of genome or transcriptome sequences*

De novo assembly of genome data is performed using "*de novo* sequence assemblers". At present, most *de novo* assemblers, such as SPAdes [20] and Velvet [21] perform well on bacteria and smaller eukaryotic genomes but are challenged when assembling larger ones. These challenges arise mainly from repeats, gene duplications and sequencing errors that cause high error rates and imprecise assembly. Recently, third-generation sequencing (TGS) has increasingly been applied to generate reference genomes for small-to moderate-sized genome (< 1 Gbp). Nevertheless, generating sufficient coverage for larger genomes is still expensive, and therefore a hybrid approach that combine NGS and TGS is used. There are two main types of hybrid assemblies: (i) the assembly of long reads corrected with the support of short reads [22,23], and (ii) the contig assembly by short reads and the scaffolding by long reads [24–27]. Fig. 1 provides a schematic summary of the NGS data analysis workflow.

RNA-Seq data can also be *de novo* assembled into a transcriptome. However, sequence assemblers that are used for WGS data are not applicable in transcriptome assembly [28]. This is due to the varying sequencing depth of transcripts depending on the level of expression, and the fact that highly abundant transcripts harbor many discrepant bases, resulting in problems when ddetermining sequencing errors. Additionally, RNAseq can be strand-specific, besides sharing of exons from the same gene by spliced variants can be difficult to resolve. To counter challenges in *de novo* transcriptome assembly, a number of *de novo* transcriptome assemblers have been made available, such as Velvet [21], Mira [29] Trans-AbySS [30], Trinity [31], Oases [32] and SOAPdenovo-Trans [33]. Generally, these "assemblers" are based on either the "overlap graph" algorithm or the "De-Bruijin graph" algorithm [34]. Reference transcriptomes that are constructed can then be used translated *in silico* into protein FASTA database for querying MS-based proteomics data. This strategy is referred to as "proteomics informed by transcriptomics" (PIT) [35].

3.2. *Gene prediction*

For a newly-sequenced and assembled genome, gene prediction (gene finding) is performed to identify regions that encode protein-coding genes and other functional elements within an assembled genome. Identification of coding sequences is an essential step for the construction of a protein FASTA for proteogenomics study. Methodology-wise, algorithms for gene predictions perform gene discovery based mainly on (i) sequence similarity search and (ii) *ab initio* gene prediction.

In the first method, similarity in gene sequences is discovered by comparing a query sequence against expressed sequence tags (ESTs), gene sequences, proteins or other genomes. Sequence similarities between certain regions are used to interpret gene structure or the functions of the region. The software used in this approach are the BLAST family of programs including DIAMOND [36] and BLAT [37]. While gene prediction for prokaryotes is less complicated due to the higher gene similarity and the absence of introns, gene discovery in eukaryotes is complicated by introns that interrupt open reading frames (ORFs).

On the other hand, *ab initio* methods rely on signal and content sensors [38]. Signal sensors refer to short sequence motifs, such as splice sites, branch points, poly-pyrimidine tracts, start and stop codons. Exon detection must rely on the content sensors, which refer to the patterns of codon usage that are unique to a species and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms. Notable programs in this category includes Hidden Markov Model (HMM) [39].

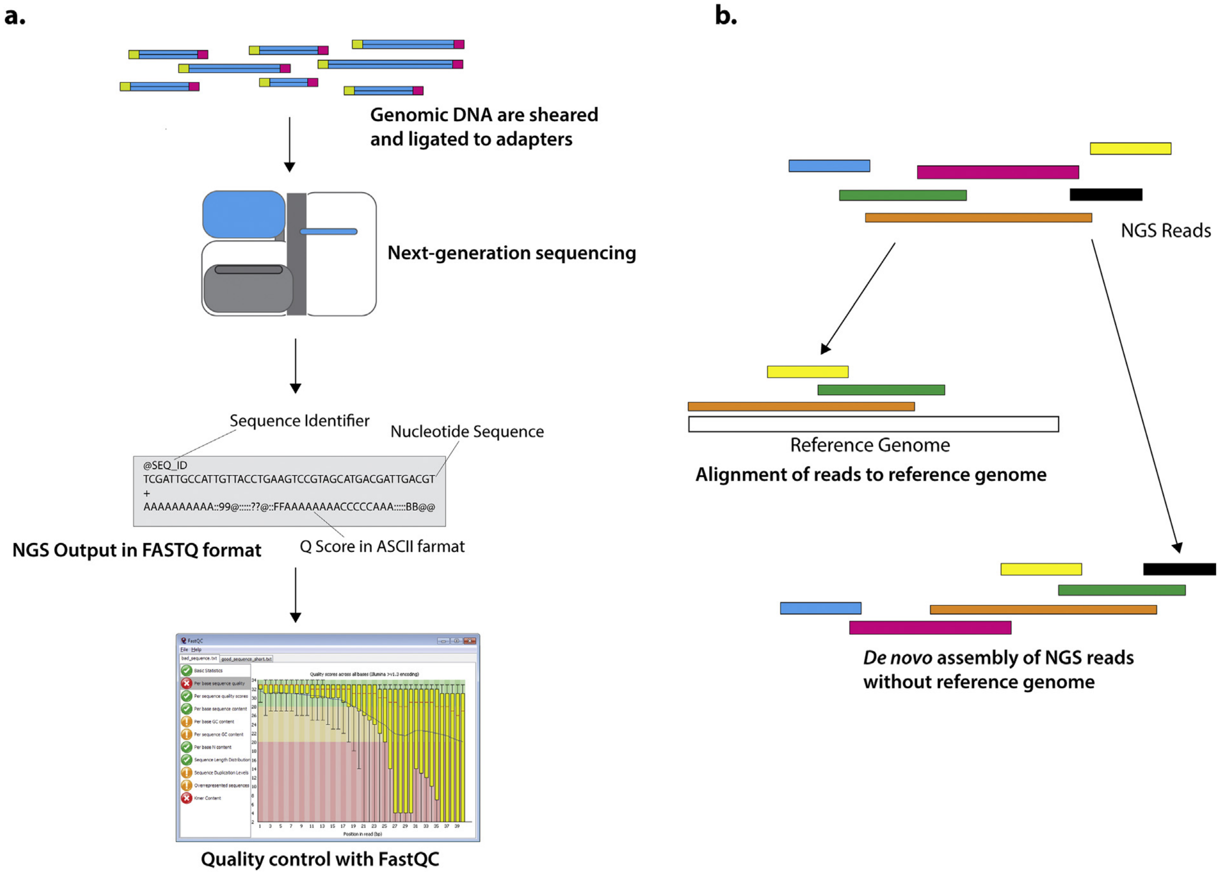
3.3. *Alignment (reads mapping) of genome and transcriptome sequences*

WGS and WES are also used to re-sequence organisms with pre-existing reference genomes. These efforts serve to identify and distinguish a variety of genomic variants of an organism against the reference. The process of mapping NGS reads against a reference genome is called alignment. The alignment process starts with the selection of an appropriate reference genome or a pre-existing assembly. The selected reference genome provides well-annotated gene models to guide the alignment process. Widely-used software tools for alignment are SOAP2 [40], BWA [41] and Bowtie2 [42] that generate alignment results in the SAM or BAM formats [43], which require some degree of pre-processing before variant calling.

3.4. *Variant calling and annotation*

Variant calling aims to identify genomic variants from sequence data that is compared against the reference genomes. The majority of genomic variants are SNVs and small INDELs, that underlie most genetic diseases and the results are summarized in the Variant Call Format (VCF) file [44]. The variant quality scores for SNPs and INDELs are then recalibrated separately with their recalibrated quality scores (VQSLOD) based on known/true variants that are available in the public domains.

Subsequently, variant annotation is performed to predict the effects


a.
Genomic DNA are sheared and ligated to adapters
Next-generation sequencing
Sequence Identifier
Nucleotide Sequence
NGS Output in FASTQ format
Q Score in ASCII farmat
Quality control with FastQC
b.
NGS Reads
Reference Genome
Alignment of reads to reference genome
De novo assembly of NGS reads without reference genome

```
@SEQ_ID
TCGATTGCCATTGTTACCTGAAGTCCGTAGCATGACGATTGACGT
+
AAAAAAAAAA::99@:::::??@::FFAAAAAAAAACCCCCAAA:::::BB@@
```

Fig. 1. Workflow in next-generation sequencing (NGS). (a.) First, genomics DNA and RNA are extracted from the biological samples of interest. They are then randomly sheared to ~100–500 nucleotides length. RNA samples, meanwhile, are converted to complementary DNA (cDNA) with reverse transcriptase. These short DNA fragments are then ligated to sequencing adaptors, before being sequenced with NGS or TGS platforms. Post-sequencing, proprietary raw data files are converted to the FASTQ files the *de facto* standard for NGS output. Within the FASTQ, there are three pieces of important information, *i.e.* (i) the identity if the sequences, (ii) the actual sequence of the stretch of nucleotides and (iii) the Q score which represents base calling error probability. Following that, a 3-step QC procedure is performed based on the Q scores to ensure that the FASTQ files have sufficiently high quality. In this figure, diagnostic graphs from FastQC are shown. (b.) Finally NGS/TRS reads (DNA fragments) are put together with two methods. If an existing reference genome is unavailable, these fragments of DNA are aligned to one another to produce a longer sequence. This strategy is called *de novo* assembly. Whereas, if a reference genome is available, NGS reads can simply be aligned to this reference.

of DNA variants based on existing information. Variants can be annotated using ANNOVAR [45] based on 3 categories, *i.e.* (a) gene-based annotation, (b) region-based annotation and (c) filter-based annotation. In gene-based annotation, variants that affect known genes are identified by inferring the following information: (i) whether the variant is exonic, intronic, affects splicing, located at 3′-untranslated region (UTR) or 5′-UTR, or intergenic; (ii) what functional roles the variant has on protein coding capacity *i.e.* synonymous, non-synonymous, frameshift or insertion/deletion; and finally (iii) which transcripts are affected and what changes occur in the corresponding amino acids residues. Meanwhile, region-based annotation is useful for identifying whether the variants overlap with a certain regions of interest, such as conserved genomic elements [46], cytogenetic bands, miRNA target sites [47] and Encyclopedia of DNA Elements (ENCODE)-annotated regions [48]. Finally, filter-based annotation helps to annotate and filter variants based on certain criteria such as (i) the alternative allele frequency for variants in the 1000 Genomes Project [49], (ii) the US National Institutes of Health–National Heart, Lung, and Blood Institute (NIH-NHLBI) ESP6500 exome-sequencing project [50], (iii) phylogenetic conservation or effect on protein 3D structure using SIFT [51] and PolyPhen [52] tools that score non-synonymous variants; and (iv) identifying the presence or absence of a variant in the dbSNP database.

RNA-Seq data can also be mapped onto reference genomes. However, since mature transcripts consist of exonic sequences that are spliced together, the transcript structures are therefore non-contiguous. Consequently, mapping them to the reference genomes requires specialized software that are capable of identifying exon-exon spliced junctions, such as STAR [53] or HISAT [54] that consider transcripts that define intron/exon boundaries. Alignment of RNA-Seq data potentially reveals co/post-transcriptionally modified transcripts such as spliced junctions, RNA edits, as well as alternative events of transcription initiation and poly-adenylation.

4. Mass spectrometry-based proteomics

4.1. Sample preparation

The second component for proteogenomics is bottom-up proteomics. Although a number of MS data acquisition strategies exist in bottom-up proteomics, all these strategies invariably involve similar sample preparation procedures encompassing (i) sample lysis and protein extraction, (ii) proteolytic digestion and optionally (iii) peptide fractionation. To extract proteins from cells or tissues, the integrity of biological samples is first disrupted with sonication or homogenization, followed by the solubilization of proteins with a mixture containing detergents (such as SDS, NP40 or Triton-X 100), chaotropes (urea or thiourea) and reducing agents (β -mercaptoethanol or DTT). Whereas, protease inhibitors and phosphatase inhibitors are included to avoid proteolysis and de-phosphorylation.

Next, solublized proteins are digested with trypsin to produce peptides that are amenable to MS. Nonetheless, trypsin comes with a number of pitfalls that prevent the full sequence coverage of proteins. Consequently, alternative proteases such as LysN, GluC, AspN, or chymotrypsin are used and it has been demonstrated to provide a good complement for proteomics analysis [55,56]. Another challenge in bottom-up proteomics arises from the complexity and dynamic range of peptide mixtures. To be able to detect peptides within the limit of detection and loading capacity of mass analyzers, peptide mixtures can be fractionated with chromatographic techniques such as ion-exchange (IEX) chromatography, reverse phase chromatography, hydrophilic interaction liquid chromatography (HILIC), or in-solution isoelectric focusing [57].

4.2. Mass spectrometry analysis

Among the MS data acquisition strategies used for bottom-up proteomics, the most popular is shotgun proteomics. Shotgun proteomics comprises experiments in which both the identities of target proteins and their amino acid sequences are not known beforehand. This approach is often associated with data dependent acquisition (DDA) scheme [58]. This MS acquisition mode first performs a survey scan of precursor ions eluting from the LC, generating a list of *m/z* ratios and intensities for detected parent (peptide) ions. Subsequently, the instrument selects a parent ion from the list, isolates and fragmentats it; and the resulting daughter ions are measured at the MS2 scan. Typically, fixed number of MS2 scans (for example 10) are performed following one survery scan; and the selection of parent ions is based on pre-determined rules (for example in the order of decreasing signal intensity), often with dynamic exclusion. As such, shotgun experiments are capable of sequencing of large number of peptides, albeit at the expense of run-to-run reproducibility of peptide identification and precision of quantification. Hence, shotgun experiments are ideal for screening and exploratory purpose.

In contrast to the discovery nature of shotgun proteomics, targeted proteomics is applied primarily for testing hypotheses for a subset of *a priori* target proteins [59,60]. It aims to quantify a selected set of proteotypic peptides, which uniquely represent target proteins or protein isoforms of interest, and this set of target proteins can be derived either from genomics, transcriptomics or shotgun proteomics studies. Targeted proteomics mainly adopts the selected reaction monitoring (SRM) method and is typically performed with triple quadrupole MS instruments [61]. An SRM assay monitors a targeted peptide not only by its chromatopgraphic retention time, but more specifically the transitions, *i.e.* the selected precursor/fragment ion pairs [62], rendering the assay highly specific. In the SRM mode, the Q1 quadrupole mass analyzer of the MS first isolates a specific peptide precursor ion with a narrow isolation window. The isolated peptide precursor is subsequently fragmented in the collision cell (Q2); followed by the monitoring of one of the resulting daughter ions by the third quadrupole (Q3) which is set to filter a certain *m/z* ratio. Effectively, an SRM assay monitors three to five fragment ions per peptide to establish the identity and quantity of the peptide of interest. Due to the double filtering at Q1 and Q3, SRM has improved sensitivity and signal-to-noise ratio. Besides, since it repeats the sampling cycle on a selected peptide and monitors the same transitions, this results in better reproducibility of peptide identification and more precise quantification.

Meanwhile, data-independent acquisition (DIA) refers to several recently developed MS acquisition methods in which MS2 scans are acquired in a continuous and unbiased manner for all precursor ions falling within a specified mass range [63–65]. Thus, the MS2 scans in DIA are not dependent on: (a) peptide precursor ions with well-defined *m/z* ratios from the MS1 survey scan (DDA) or (b) *a priori* knowledge about peptide precursor m/z values (SRM). In one of the more popular DIA schemes, *i.e.* the SWATH-MS, single precursor ion (MS1) spectrum is first recorded for currently eluting peptides, followed by a series of MS2 scans [64]. However, these MS2 scans are not triggered by a mass list following certain pre-determined rules, as in DDA. Instead, it is triggered for a specified mass range (for instance 25 *m/z*), and this step is repeated for a number of consecutive, slightly overlapping precursor isolation windows (with 25 m/z width). Doing this over a defined mass range allows the accumulation of a continuous data set on all detectable fragment and precursor ions. Due to co-fragmentation of co-eluting peptides within the precursor isolation window, the DIA method produces highly multiplexed and complex fragment ion spectra, and several data analysis strategies have been proposed to resolve this complexity [66]. As DIA is able to quantify huge number of proteins with high reproducibility and accuracy, it is best-suited for analyzing huge sample size [67].

4.3. Peptide and protein identification from MS data

Due to space constraint, we only discuss peptide and protein identification for shotgun proteomics data here, which normally comprises two steps. In the first, a MS/MS spectrum is matched to a peptide sequence, known as a peptide spectrum match (PSM). Subsequently, peptide sequences obtained from PSMs are used to infer proteins.

4.3.1. Peptide to spectrum matching

Computational algorithms that are currently used to assign fragment ion spectra to peptide sequences can be classified into three categories [68]. The first category depends on “database searching”, whereby peptide sequences are usually identified by correlating an MS2 spectrum with a theoretical spectrum predicted from peptides obtained from *in-silico* digestion of a reference protein FASTA database that is *in silico* translated from the reference genome and is often downloadable from the public domain. In this method, such correlation is usually performed with a variety of computatonal tools that are known collectively as “database search engines” [69]. Alternatively, the same MS2 spectrum can be correlated with a library of experimental MS2 spectra that had been identified in previous experiments, the so called “spectral-library search” approach [70]. The second category is called “*de novo* sequencing”. In this method, peptide sequences are directly inferred from fragment ion spectra by calculating mass differences among fragment ions, without any reference to any database [71]. Finally, the last category features hybrid approaches, which are based on the extraction of short “sequence tags” of 3–5 residues in length, followed by “error-tolerant” database searching [72]. Currently, database searching strategies are most widely-used for high-throughput peptide and protein identification.

4.3.2. Protein inference from peptides

Having assigned peptide sequences, the next step is to map and assemble these peptide sequences into proteins. This process is not straightforward because peptide sequences are often shared among more than one protein due to sequence homology or alternative splicing. A common solution is to simply report the parsimonious set of proteins to explain the observed peptides. The drawback for this solution is that, even small variations among the peptides may generate a different protein group, resulting in discrepancies. One solution is by increasing protein sequence coverage with multiple enzymatic digestion. By identifying more unique (non-shared) peptides for a given protein, it becomes easier to resolve protein isoforms, contributing to higher confidence.

4.3.3. Statistical analysis of peptide assignments and protein inference

The shotgun approach generates massive amounts of data with certain degree of noise that affects peptide and protein identification. Therefore, it is essential to estimate confidence and error rates for these large-scale experiments. For estimating the false discovery rate in peptide to spectrum matching, a target-decoy analysis is most widely-used [73]. On the other hand, PeptideProphet is a machine learning method that can separate correct PSMs from incorrect ones by combining multiple search score-related parameters into a single discriminant score, followed by estimation of posterior probabilities of correct identifications as inferential indicators [74]. Meanwhile, Percolator uses a semi-supervised support vector machine (SVM) to discriminate between correct and decoy spectrum identifications [75].

As for protein inference, ProteinProphet computes a probability that a protein is present in the sample by combining the probabilities of peptides assigned to this protein [76]; while MAYU extends the target-decoy strategy to the level of protein identification [77]. The Fido algorithm uses a Bayesian method to infer the protein posterior error probabilities when given a set of PSMs [78].

5. Connecting genome data to proteome data

5.1. Deriving customized FASTA fom genome sequences

Similar to proteomics, identification of proteins in proteogenomics is inferred from a set of peptide-to-spectra matches (PSM) by assigning fragment ion mass spectra to peptide sequences, but this time database search engines are used to query a customized protein FASTA instead, that contain a variety of genome variants, mutations, co- or post-transcriptionally modified sequences, as well as novel protein coding sequences. These customized databases can be contructed from from six-frame translation of genome sequences, predicted gene models or specialized databases containing annotated RNA sequences or genome variants (Fig. 2). Recently, a number of bioinformatic tools have been made available for proteogenomics. For instance, customProBD [79], a Galaxy-P based workflows [80,81] and PGA [82] have been developed to generate protein FASTA from RNA-seq data, but only PGA can do so without a reference genome guide. On the other hand, sapFinder automatically incorporate variants obtained from public SNV repositories or sample-specific next-generation sequencing (NGS) data, and allow the identification of single amino acid polymorphisms (SAPs) through database searching, post-processing and generation of HTML-based report with visualized interface [83].

5.2. Database search engines and data validation

Naturally, the main objective for proteogenomics is to maximize the identification of novel peptides, albeit within an well-controlled false discovery rates (FDR) at both the peptide and protein levels. As such, optimum parameters for database search in proteogenomics is always a compromise between the database size and error rates [84]. Recently, an integrated proteomic pipeline (IPP) was recently introduced by consolidating different search engines (SEQUEST, MASCOT and MS-GF +) and this combination managed to improve the sensitivity of novel peptide identifications, including alternative splicing variants and missing proteins within a 1% protein FDR [85]. Meanwhile, a “cascade search” approach was developed by Japtap et al. [86]. In this method, MS2 spectra are first searched against a large custom database; followed by the creation of a subset database from the matches of the search. Subsequently, decoy proteins were added to this subset database, which was then merged with a host database for the second search. This two-step method successfully doubled the high confidence hits in comparison to the conventional method. Besides, by building real and simulated proteogenomic databases followed by testing with two database search engines, as well as two validation approaches (target-decoy search strategy and a mixture model-based method), Li et al. discovered that separate filtering of known and novel peptides increases the sensitivity and reliability in proteogenomic search [87].

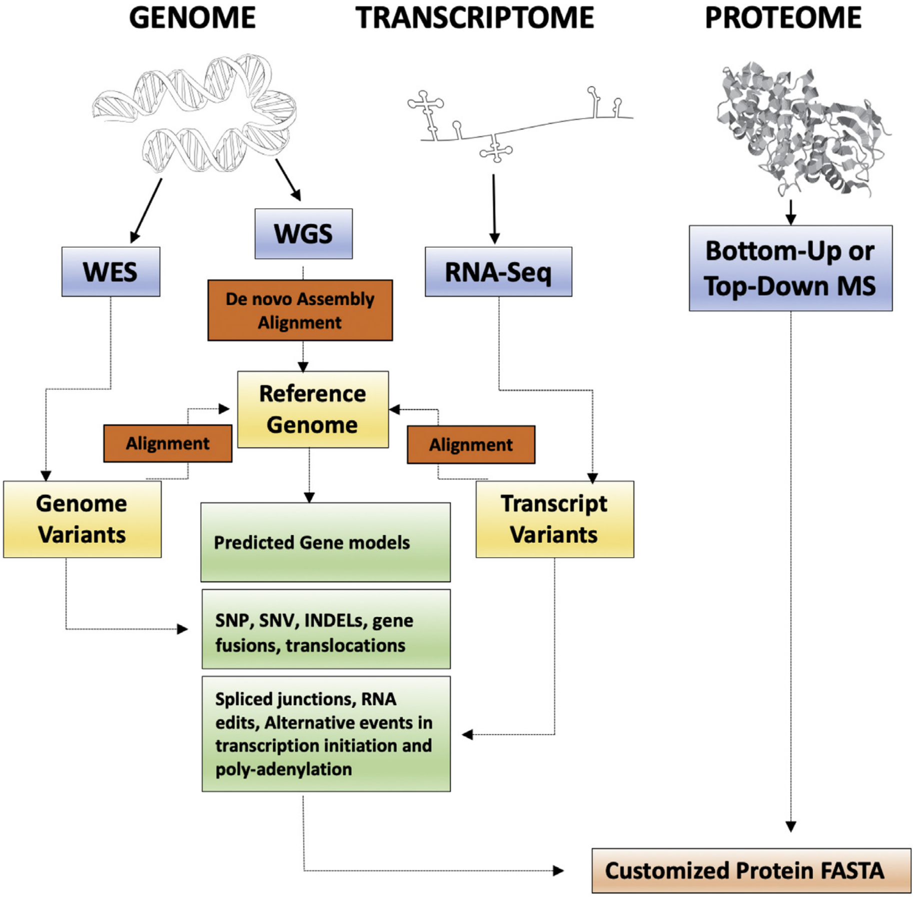


Fig. 2. The integrated view of a proteogenomic workflow. The total DNA content of a genome can be sequenced by using whole genome sequencing (WGS). WGS data can be *de novo* assembled to produce a draft genome. One or several fully assembled genomes can be built into a reference genome. Once a reference genome ia available, it can be used as a template onto which NGS reads obtained from different modes of sequencing can be aligned. These include whole exome sequencing (WES), a targeted sequencing approach that only sequences the exonic regions and is useful for revealing genome variants such as SNV, SNP, INDELs and other structural variants. Meanwhile, in transcriptomics, RNA-Seq is applied for sequencing mRNAs. It is also useful for unraveling co-post transcriptional modifications such as gene splicing and RNA edits. Finally, these NGS data can be *in silico* translated to customized protein FASTAs for database search using data protein sequence data obtained from bottom-up or top-down proteomics.

6. The application of proteogenomics in clinical research – precision oncology

A widely-applied area for proteogenomics is precision oncology; giving rise to onco-proteogenomics [88,89]. One ambitious onco-proteogenomic effort is The Clinical Proteomic Tumor Analysis Consortium (CPTAC) that was launched almost simultaneously with the Cancer Genome Atlas (TCGA). Among the objectives of CPTAC are to unravel the different proteogenomic sub-types of tumors; to correlate copy number alterations and protein abundance so as to prioritize driver mutations; and to study changes in post-translational modifications (PTMs) and associate them with cancer-relevant pathways [90]. So far, cancer types that have been analyzed by CPTAC are colorectal, breast and ovarian tumor samples that had been analyzed previously by TCGA [91–93]. In general, the Spearman correlation between mRNAs and proteins in these three studies was found to be < 0.5. In the CRC study, only two out of five proteomic sub-types overlapped with genomic subtypes reported by TCGA; whereas copy number alterations (CNV) showed strong cis- and trans-effects on mRNA, but not protein abundance [91]. As for the breast cancer study, Mertins et al., performed quantitative mass-spectrometry-based proteomic and phosphoproteomics experiments on 105 genomically annotated samples [92]. They interrogated the trans-effects of chromosomal 5q deletion and managed to connect the loss of CETN3 and SKP1 to elevated expression of EGFR; and SKP1 loss to increased SRC tyrosine kinase. In addition, phosphoproteome analysis revealed a G-protein-coupled receptor cluster that was not readily identified at the mRNA level. Finally, the ovarian cancer study identified specific protein acetylations that are associated with deficiency in homologous recombination, suggesting a potential means for stratifying patients for therapy [93].

Public repositories are increasingly leveraged to confirm the expression of novel, cancer-specific peptides. For instance, a unified genomic variant protein database had been created from TCGA that contained novel splice junctions, deletions, insertions and substitutions [94]. This allowed Woo et al. to identify 524 novel peptides from a single ovarian carcinoma sample. By compiling variants from dbSNP and UniProt, somatic variations in cancer (COSMIC) along with sample-specific genomic and transcriptomic data, 7.3 million novel variant peptides and 4771 mutations were identified from in 59 NCI60 cancer cell-lines [95]. Likewise, a fusion peptide database was created by including reported and predicted gene fusion pairs using CanProFu, allowing 11 unique fusion peptides to be identified in non-small cell lung cancers (NSCLC) [96]. On the other hand, the alterations in PI3K/AKT/mTOR pathway were investigated by re-analyzing TCGA genome and proteome data obtained from 11,219 human cancers (32 major types) and it was discovered that a number of oncogenic pathways such as the PI3K/AKT/mTOR pathway converged on similar sets of downstream transcriptional targets [97].

The primary aim of precision oncology is to aid the selection of the best course of treatment based on patient's tumor biology. However, it is also important to study drug toxicity and the acquisition of resistance to treatments. Therefore, proteogenomics is also used to associate the efficacy or resistance of cancer therapy with different mutations. For example, proteogenomic methods were applied to identify and quantify gatekeeper mutations so as to investigate the varied efficacy of tyrosine kinase inhibitors on different racial populations in lung cancer [98]. Proteogenomics was also performed to examine in gastrointestinal stromal tumors (GISTs), the acquisition of resistance against tyrosine kinase inhibitors and to identify driver mechanisms of secondary resistance [99]. Similarly, proteogenomics has provided new insights into the most aggressive forms of medulloblastoma. By analyzing CNV data, clinical studies and mass spectrometry-proteomics, c-MYC amplification was identified as an important risk factor in this study [100]. Immunotherapy is a new form of cancer therapy that is very promising. By examining cells derived from 13 volunteers with proteogenomic techniques, it was discovered that out of > 6000 minor histocompatibility antigens (MiHAs), a set of 39 'optimal MiHAs' shared optimal features for immunotherapy of HCs [101].

7. Challenges and opportunities

The bottleneck for proteogenomics mainly stems from proteomics. Although state-of-the-art mass spectrometers have far improved sensitivity, mass accuracy, resolution and speed, there is so far no amplification techniques that is equivalent to PCR that has allowed sequencing to be performed with a single cell. As a result, this affects the minimum sample size and throughput in MS-based proteomics [102]. This is in addition to the lack of reproducibility in the DDA scheme that are adopted by most proteomic researchers. Another challenge is the dynamic range of samples especially that of the body fluids, that are often used for biomarker discovery and have a dynamic range of 10 orders of magnitude [103]. However, this has presented some new opportunities, and an increasing number of researchers have begun to adopt new methods such as SRM and DIA so as to optimize the reproducibility, sample throughput and limit of detection in proteogenomic experiments. Besides, identification of spurrious peptides is especially a problem for proteogenomics not only due to the inflated database size, but also because assigning sequences for novel peptides that contain all sort of alterations require much more stringent filtering and validation [84]. Apart from that, there are currently few standardized data analysis workflow, file formats and bioinformatics tools that allow proteogenomics to be commonly adopted by biologists [104]. This is particularly crucial for massive projects that involve international collaboration. Furtuitiously, there are already efforts within the communities to develop proteogenomics pipelines to address these challenges [1,105]. Finally, there are also concerns relating to data storage, data safety and data transfer, besides ethical and privacy concerns with respect to patients' information and disease-related proteogenomics variants [106].

8. Conclusion

In summary, proteogenomics is an emerging dicispline that is increasingly being adopted by the clinical science community. Although at the moment, international collaboration and research funding for protegenomics mainly focus on precision oncology, it should ideally be extended to other prevalent non-communicable diseases such as cardiovascular diseases; or communicable diseases especially those related to the development of antobiotic resistance in superbugs. With the forthcoming 4th Industrial Revolution (4IR), artificial intelligence (AI) and machine learning, combined with robotics are especially useful for implemeting and interpreting large scale OMIC experiments [107]. Besides clinical data, wearable technology allows continuous monitoring and collection of biophysical data from individuals [108]. With the rapidly reducing cost for clinical sequencing and and clinical proteomics, we envisage that proteogenomics may be implemented as a routine clinical lab test in the future.

Fundings

This work is supported by the Fundamental Research Grant Scheme (FRGS/1/2018/STG04/UKM/02/7) awarded to T.Y. Low by the Ministry of Education Malaysia.

Declaration of Competing Interest

The authors have declared no conflict of interest.

Acknowledgement

The authors would like to thank Miss Chui Li Leaw (for fruitful discussion and proofreading of the manuscript) and Dr. Mao Peng (for fruitful discussion and contribution of ideas). They would also like to thank SCIEX Malaysia for techincal and research support.

References

[1] T.Y. Low, M.A. Mohtar, M.Y. Ang, R. Jamal, Connecting proteomics to next-generation sequencing: proteogenomics and its current applications in biology, Proteomics 1800235 (2018), https://doi.org/10.1002/pmic.201800235.

[2] K.V. Ruggles, K. Krug, X. Wang, K.R. Clauser, J. Wang, S.H. Payne, D. Fenyö, B. Zhang, D.R. Mani, Methods, tools and current perspectives in proteogenomics, Mol. Cell. Proteomics 16 (2017) 959–981, https://doi.org/10.1074/mcp.MR117.000024.

[3] H. Marx, H. Hahne, S.E. Ulbrich, A. Schnieke, O. Rottmann, D. Frishman, B. Kuster, Annotation of the domestic pig genome by quantitative proteogenomics, J. Proteome Res. 16 (2017) 2887–2898, https://doi.org/10.1021/acs.jproteome.7b00184.

[4] Y. Mao, X. Yang, Y. Liu, Y. Yan, Z. Du, Y. Han, Y. Song, L. Zhou, Y. Cui, R. Yang, Reannotation of Yersinia pestis strain 91001 based on omics data, Am. J. Trop. Med. Hyg. 95 (2016) 562–570, https://doi.org/10.4269/ajtmh.16-0215.

[5] N. Ucciferri, S. Rocchiccioli, Proteomics techniques for the detection of translated pseudogenes, Methods Mol. Biol. 1167 (2014) 187–195, https://doi.org/10.1007/978-1-4939-0835-6_12.

[6] V. Olexiouk, G. Menschaert, Identification of small novel coding sequences, a proteogenomics endeavor, Adv. Exp. Med. Biol. 926 (2016) 49–64, https://doi.org/10.1007/978-3-319-42316-6_4.

[7] S.S. Manda, R.S. Nirujogi, S.M. Pinto, M.-S. Kim, K.K. Datta, R. Sirdeshmukh, T.S.K. Prasad, V. Thongboonkerd, A. Pandey, H. Gowda, Identification and characterization of proteins encoded by chromosome 12 as part of chromosome-centric human proteome project, J. Proteome Res. 13 (2014) 3166–3177, https://doi.org/10.1021/pr401123v.

[8] Y.-I. Kim, J. Lee, Y.-J. Choi, J. Seo, J. Park, S.-Y. Lee, J.-Y. Cho, Proteogenomic study beyond chromosome 9: new insight into expressed variant proteome and transcriptome in human lung adenocarcinoma tissues, J. Proteome Res. 14 (2015) 5007–5016, https://doi.org/10.1021/acs.jproteome.5b00544.

[9] K.G. Kuznetsova, I.Y. Ilina, A.L. Chernobrovkin, S.E. Novikova, Proteogenomics of adenosine-to-inosine RNA editing in fruit fly, BioRxiv, 2017.

[10] R. Tavares, G. Wajnberg, N.M. de Scherer, B.A. Pauletti, J.S. Cassoli, C.G. Ferreira, A.F. Paes Leme, P.S. de Araujo-Souza, D. Martins-de-Souza, F. Passetti, Unveiling alterative splice diversity from human oligodendrocyte proteome data, J. Proteome 151 (2017) 293–301, https://doi.org/10.1016/j.jprot.2016.05.023.

[11] A. Koch, D. Gawron, S. Steyaert, E. Ndah, J. Crappé, S. De Keulenaer, E. De Meester, M. Ma, B. Shen, K. Gevaert, W. Van Criekinge, P. Van Damme, G. Menschaert, 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 (2014) 2688–2698, https://doi.org/10.1002/pmic.201400180.

[12] J. Armengaud, O. Pible, J.-C. Gaillard, A. Cian, N. Gantois, K.S.W. Tan, M. Chabe, E. Viscogliosi, Proteogenomic insights into the intestinal parasite *Blastocystis* sp. subtype 4 isolate WR1, Proteomics 17 (2017), https://doi.org/10.1002/pmic.201700211.

[13] G.M. Blumenthal, E. Mansfield, R. Pazdur, Next-generation sequencing in oncology in the era of precision medicine, JAMA Oncol. 2 (2016) 13, https://doi.org/10.1001/jamaoncol.2015.4503.

[14] V. Subbiah, R. Kurzrock, Debunking the delusion that precision oncology is an illusion, Oncologist 22 (2017) 881–882, https://doi.org/10.1634/theoncologist.2017-0040.

[15] H. Rodriguez, S.R. Pennington, Revolutionizing precision oncology through collaborative proteogenomics and data sharing, Cell. 173 (2018) 535–539, https://doi.org/10.1016/j.cell.2018.04.008.

[16] B.-S. Petersen, B. Fredrich, M.P. Hoeppner, D. Ellinghaus, A. Franke, Opportunities and challenges of whole-genome and -exome sequencing, BMC Genet. 18 (2017) 14, https://doi.org/10.1186/s12863-017-0479-5.

[17] A. Warr, C. Robert, D. Hume, A. Archibald, N. Deeb, M. Watson, Exome sequencing: current and future perspectives, G3 (Bethesda) 5 (2015) 1543–1550, https://doi.org/10.1534/g3.115.018564.

[18] R. Hrdlickova, M. Toloue, B. Tian, RNA-Seq methods for transcriptome analysis, Wiley Interdiscip. Rev. RNA 8 (2017) e1364, , https://doi.org/10.1002/wrna.1364.

[19] H. Du, Y. Yu, Y. Ma, Q. Gao, Y. Cao, Z. Chen, B. Ma, M. Qi, Y. Li, X. Zhao, J. Wang, K. Liu, P. Qin, X. Yang, L. Zhu, S. Li, C. Liang, Sequencing and de novo assembly of a near complete indica rice genome, Nat. Commun. 8 (2017) 15324, https://doi.org/10.1038/ncomms15324.

[20] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477, https://doi.org/10.1089/cmb.2012.0021.

[21] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829, https://doi.org/10.1101/gr.074492.107.

[22] S. Koren, M.C. Schatz, B.P. Walenz, J. Martin, J.T. Howard, G. Ganapathy, Z. Wang, D.A. Rasko, W.R. McCombie, E.D. Jarvis, A.M. Phillippy, Hybrid error correction and de novo assembly of single-molecule sequencing reads, Nat. Biotechnol. 30 (2012) 693–700, https://doi.org/10.1038/nbt.2280.

[23] S. Goodwin, J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M.C. Schatz, W.R. McCombie, Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome, Genome Res. 25 (2015) 1750–1756, https://doi.org/10.1101/gr.191395.115.

[24] V. Deshpande, E.D. Fung, S. Pham, V. Bafna, Cerulean: A Hybrid Assembly Using High Throughput Short and Long Reads, http://arxiv.org/abs/1307.7933, (2013) , Accessed date: 20 March 2019.

[25] D. Antipov, A. Korobeynikov, J.S. McLean, P.A. Pevzner, hybridSPAdes: an algorithm for hybrid assembly of short and long reads, Bioinformatics 32 (2016) 1009–1015, https://doi.org/10.1093/bioinformatics/btv688.

[26] A. Bashir, A. Klammer, W.P. Robins, C.-S. Chin, D. Webster, E. Paxinos, D. Hsu, M. Ashby, S. Wang, P. Peluso, R. Sebra, J. Sorenson, J. Bullard, J. Yen, M. Valdovino, E. Mollova, K. Luong, S. Lin, B. LaMay, A. Joshi, L. Rowe, M. Frace, C.L. Tarr, M. Turnsek, B.M. Davis, A. Kasarskis, J.J. Mekalanos, M.K. Waldor, E.E. Schadt, A hybrid approach for the automated finishing of bacterial genomes, Nat. Biotechnol. 30 (2012) 701–707, https://doi.org/10.1038/nbt.2288.

[27] C. Ye, C.M. Hill, S. Wu, J. Ruan, Z.S. Ma, DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies, Sci. Rep. 6 (2016) 31900, https://doi.org/10.1038/srep31900.

[28] Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T.W. Lam, Y. Li, X. Xu, G.K. Wong, J. Wang, SOAPdenovo-trans: de novo transcriptome assembly with short RNA-Seq reads, Bioinformatics. 30 (2014) 1660–1666, https://doi.org/10.1093/bioinformatics/btu077.

[29] B. Chevreux, T. Pfisterer, B. Drescher, A.J. Driesel, W.E.G. Müller, T. Wetter, S. Suhai, Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs, Genome Res. 14 (2004) 1147–1159, https://doi.org/10.1101/gr.1917404.

[30] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S.D. Jackman, K. Mungall, S. Lee, H.M. Okada, J.Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y.S. Butterfield, R. Newsome, S.K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R.A. Moore, M. Hirst, M.A. Marra, S.J.M. Jones, P.A. Hoodless, I. Birol, De novo assembly and analysis of RNA-seq data, Nat. Methods 7 (2010) 909–912, https://doi.org/10.1038/nmeth.1517.

[31] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. LeDuc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–1512, https://doi.org/10.1038/nprot.2013.084.

[32] M.H. Schulz, D.R. Zerbino, M. Vingron, E. Birney, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, Bioinformatics 28 (2012) 1086–1092, https://doi.org/10.1093/bioinformatics/bts094.

[33] Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T.-W. Lam, Y. Li, X. Xu, G.K.-S. Wong, J. Wang, SOAPdenovo-trans: de novo transcriptome assembly with short RNA-Seq reads, Bioinformatics. 30 (2014) 1660–1666, https://doi.org/10.1093/bioinformatics/btu077.

[34] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, W. Fan, Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph, Brief. Funct. Genomics 11 (2012) 25–37, https://doi.org/10.1093/bfgp/elr035.

[35] T. Luge, C. Fischer, S. Sauer, Efficient application of De Novo RNA assemblers for proteomics informed by transcriptomics, J. Proteome Res. 15 (2016) 3938–3943, https://doi.org/10.1021/acs.jproteome.6b00301.

[36] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (2015) 59–60, https://doi.org/10.1038/nmeth.3176.

[37] W.J. Kent, BLAT—the BLAST-like alignment tool, Genome Res. 12 (2002) 656, https://doi.org/10.1101/GR.229202.

[38] Y. Huang, S.-Y. Chen, F. Deng, Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction, Comput. Struct. Biotechnol. J. 14 (2016) 298–303, https://doi.org/10.1016/j.csbj.2016.07.002.

[39] R. Guigo, P. Agarwal, J.F. Abril, M. Burset, J.W. Fickett, An assessment of gene prediction accuracy in large DNA sequences, Genome Res. 10 (2000) 1631–1642 https://www.ncbi.nlm.nih.gov/pubmed/11042160.

[40] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment, Bioinformatics. 25 (2009) 1966–1967, https://doi.org/10.1093/bioinformatics/btp336.

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

[42] W.B. Langdon, Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks, BioData Min. 8 (2015) 1, https://doi.org/10.1186/s13040-014-0034-0.

[43] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, Genome project data processing, the sequence alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

[44] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, R. Durbin, G. Genomes Project Analysis, The variant call format and VCFtools, Bioinformatics 27 (2011) 2156–2158, https://doi.org/10.1093/bioinformatics/btr330.

[45] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010) e164, , https://doi.org/10.1093/nar/gkq603.

[46] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome Res. 15 (2005)

1034–1050, https://doi.org/10.1101/gr.3715005.

[47] B.P. Lewis, I.H. Shih, M.W. Jones-Rhoades, D.P. Bartel, C.B. Burge, Prediction of mammalian microRNA targets, Cell. 115 (2003) 787–798 https://www.ncbi.nlm.nih.gov/pubmed/14697198.

[48] E.P. Consortium, E. Birney, J.A. Stamatoyannopoulos, A. Dutta, R. Guigo, T.R. Gingeras, E.H. Margulies, Z. Weng, M. Snyder, E.T. Dermitzakis, R.E. Thurman, M.S. Kuehn, C.M. Taylor, S. Neph, C.M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J.A. Greenbaum, R.M. Andrews, P. Flicek, P.J. Boyle, H. Cao, N.P. Carter, G.K. Clelland, S. Davis, N. Day, P. Dhami, S.C. Dillon, M.O. Dorschner, H. Fiegler, P.G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K.D. James, B.E. Johnson, E.M. Johnson, T.T. Frum, E.R. Rosenzweig, N. Karnani, K. Lee, G.C. Lefebvre, P.A. Navas, F. Neri, S.C. Parker, P.J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F.S. Collins, J. Dekker, J.D. Lieb, T.D. Tullius, G.E. Crawford, S. Sunyaev, W.S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I.L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H.A. Hirsch, E.A. Sekinger, J. Lagarde, J.F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J.S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M.C. Dickson, D.J. Thomas, M.T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K.G. Srinivasan, W.K. Sung, H.S. Ooi, K.P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M.L. Tress, A. Valencia, S.W. Choo, C.Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T.G. Clark, J.B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C.N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J.S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R.M. Myers, J. Rogers, P.F. Stadler, T.M. Lowe, C.L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S.E. Antonarakis, Y. Fu, E.D. Green, U. Karaoz, A. Siepel, J. Taylor, L.A. Liefer, K.A. Wetterstrand, P.J. Good, E.A. Feingold, M.S. Guyer, G.M. Cooper, G. Asimenos, C.N. Dewey, M. Hou, S. Nikolaev, J.I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N.R. Zhang, I. Holmes, J.C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W.J. Kent, E.A. Stone, N.C.S. Program, C. Baylor College of Medicine Human Genome Sequencing, C. Washington University Genome Sequencing, I. Broad, I. Children's Hospital Oakland Research, S. Batzoglou, N. Goldman, R.C. Hardison, D. Haussler, W. Miller, A. Sidow, N.D. Trinklein, Z.D. Zhang, L. Barrera, R. Stuart, D.C. King, A. Ameur, S. Enroth, M.C. Bieda, J. Kim, A.A. Bhinge, N. Jiang, J. Liu, F. Yao, V.B. Vega, C.W. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M.J. Oberley, D. Inman, M.A. Singer, T.A. Richmond, K.J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J.C. Fowler, P. Couttet, A.W. Bruce, O.M. Dovey, P.D. Ellis, C.F. Langford, D.A. Nix, G. Euskirchen, S. Hartman, A.E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T.H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C.K. Glass, M.G. Rosenfeld, S.F. Aldred, S.J. Cooper, A. Halees, J.M. Lin, H.P. Shulha, X. Zhang, M. Xu, J.N. Haidar, Y. Yu, Y. Ruan, V.R. Iyer, R.D. Green, C. Wadelius, P.J. Farnham, B. Ren, R.A. Harte, A.S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A.S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R.M. Kuhn, D. Karolchik, L. Armengol, C.P. Bird, P.I. de Bakker, A.D. Kern, N. Lopez-Bigas, J.D. Martin, B.E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I.B. Hallgrimsdottir, J. Huppert, M.C. Zody, G.R. Abecasis, X. Estivill, G.G. Bouffard, X. Guan, N.F. Hansen, J.R. Idol, V.V. Maduro, B. Maskeri, J.C. McDowell, M. Park, P.J. Thomas, A.C. Young, R.W. Blakesley, D.M. Muzny, E. Sodergren, D.A. Wheeler, K.C. Worley, H. Jiang, G.M. Weinstock, R.A. Gibbs, T. Graves, R. Fulton, E.R. Mardis, R.K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D.B. Jaffe, J.L. Chang, K. Lindblad-Toh, E.S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, P.J. de Jong, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447 (2007) 799–816, https://doi.org/10.1038/nature05874.

[49] C. Genomes Project, G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth, G.A. McVean, An integrated map of genetic variation from 1,092 human genomes, Nature. 491 (2012) 56–65, https://doi.org/10.1038/nature11632.

[50] W. Fu, T.D. O'Connor, G. Jun, H.M. Kang, G. Abecasis, S.M. Leal, S. Gabriel, M.J. Rieder, D. Altshuler, J. Shendure, D.A. Nickerson, M.J. Bamshad, N.E.S. Project, J.M. Akey, Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants, Nature. 493 (2013) 216–220, https://doi.org/10.1038/nature11690.

[51] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812–3814 https://www.ncbi.nlm.nih.gov/pubmed/12824425.

[52] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249, https://doi.org/10.1038/nmeth0410-248.

[53] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics. 29 (2013) 15–21, https://doi.org/10.1093/bioinformatics/bts635.

[54] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (2015) 357–360, https://doi.org/10.1038/nmeth.3317.

[55] P. Giansanti, L. Tsiatsiani, T.Y. Low, A.J.R. Heck, Six alternative proteases for mass spectrometry–based proteomics beyond trypsin, Nat. Protoc. 11 (2016) 993–1006, https://doi.org/10.1038/nprot.2016.057.

[56] T.Y. Low, S. van Heesch, H. vanden Toorn, P. Giansanti, A. Cristobal, P. Toonen, S. Schafer, N. Hübner, B. van Breukelen, S. Mohammed, E. Cuppen, A.J.R. Heck, V. Guryev, et al., Cell Rep. 5 (2013) 1469–1478, https://doi.org/10.1016/j.celrep.2013.10.041.

[57] E. Mostovenko, C. Hassan, J. Rattke, A.M. Deelder, P.A. van Veelen, M. Palmblad, Comparison of peptide and protein fractionation methods in proteomics, EuPA Open Proteomics 1 (2013) 30–37, https://doi.org/10.1016/J.EUPROT.2013.09.001.

[58] N.W. Bateman, S.P. Goulding, N.J. Shulman, A.K. Gadok, K.K. Szumlinski, M.J. MacCoss, C.C. Wu, Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA), Mol. Cell. Proteomics 13 (2014) 329–338, https://doi.org/10.1074/mcp.M112.026500.

[59] R. Aebersold, A. Bensimon, B.C. Collins, C. Ludwig, E. Sabido, Applications and developments in targeted proteomics: from SRM to DIA/SWATH, Proteomics. 16 (2016) 2065–2067, https://doi.org/10.1002/pmic.201600203.

[60] T.T. Aye, T.Y. Low, Y. Bjørlykke, H. Barsnes, A.J.R. Heck, F.S. Berven, Use of stable isotope dimethyl labeling coupled to selected reaction monitoring to enhance throughput by multiplexing relative quantitation of targeted proteins, Anal. Chem. 84 (2012) 4999–5006, https://doi.org/10.1021/ac300596r.

[61] E. Borràs, E. Sabidó, What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry, Proteomics 17 (2017) 1700180, https://doi.org/10.1002/pmic.201700180.

[62] P. Picotti, O. Rinner, R. Stallmach, F. Dautel, T. Farrah, B. Domon, H. Wenschuh, R. Aebersold, High-throughput generation of selected reaction-monitoring assays for proteins and proteomes, Nat. Methods 7 (2010) 43–46, https://doi.org/10.1038/nmeth.1408.

[63] J.D. Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin, J.R. Yates, Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra, Nat. Methods 1 (2004) 39–45, https://doi.org/10.1038/nmeth705.

[64] L.C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis, Mol. Cell. Proteomics 11 (2012), https://doi.org/10.1074/mcp.O111.016717 O111.016717.

[65] J.D. Egertson, A. Kuehn, G.E. Merrihew, N.W. Bateman, B.X. MacLean, Y.S. Ting, J.D. Canterbury, D.M. Marsh, M. Kellmann, V. Zabrouskov, C.C. Wu, M.J. MacCoss, et al., Nat. Methods 10 (2013) 744–746, https://doi.org/10.1038/nmeth.2528.

[66] A. Bilbao, E. Varesio, J. Luban, C. Strambio-De-Castillia, G. Hopfgartner, M. Müller, F. Lisacek, Processing strategies and software solutions for data-independent acquisition in mass spectrometry, Proteomics. 15 (2015) 964–980, https://doi.org/10.1002/pmic.201400323.

[67] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B.C. Collins, R. Aebersold, et al., Mol. Syst. Biol. 14 (2018) e8126, https://doi.org/10.15252/msb.20178126.

[68] A.I. Nesvizhskii, O. Vitek, R. Aebersold, Analysis and validation of proteomic data generated by tandem mass spectrometry, Nat. Methods 4 (2007) 787–797, https://doi.org/10.1038/nmeth1088.

[69] K. Verheggen, L. Martens, F.S. Berven, H. Barsnes, M. Vaudel, Database search engines: paradigms, challenges and solutions, Adv. Exp. Med. Biol. (2016) 147–156, https://doi.org/10.1007/978-3-319-41448-5_6.

[70] E.W. Deutsch, Y. Perez-Riverol, R.J. Chalkley, M. Wilhelm, S. Tate, T. Sachsenberg, M. Walzer, L. Käll, B. Delanghe, S. Böcker, E.L. Schymanski, P. Wilmes, V. Dorfer, B. Kuster, P.-J. Volders, N. Jehmlich, J.P.C. Vissers, D.W. Wolan, A.Y. Wang, L. Mendoza, J. Shofstahl, A.W. Dowsey, J. Griss, R.M. Salek, S. Neumann, P.-A. Binz, H. Lam, J.A. Vizcaíno, N. Bandeira, H. Röst, Expanding the use of spectral libraries in proteomics, J. Proteome Res. 17 (2018) 4051–4060, https://doi.org/10.1021/acs.jproteome.8b00485.

[71] T. Muth, B.Y. Renard, Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief. Bioinform. 19 (2018) 954–970, https://doi.org/10.1093/bib/bbx033.

[72] S. Kim, N. Bandeira, P.A. Pevzner, Spectral profiles, a novel representation of tandem mass spectra and their applications for *de Novo* peptide sequencing and identification, Mol. Cell. Proteomics 8 (2009) 1391–1400, https://doi.org/10.1074/mcp.M800535-MCP200.

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

[74] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. 74 (2002) 5383–5392, https://doi.org/10.1021/ac025747h.

[75] L. Käll, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets, Nat. Methods 4 (2007) 923–925, https://doi.org/10.1038/nmeth1113.

[76] A.I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, A statistical model for identifying proteins by tandem mass spectrometry, Anal. Chem. 75 (2003) 4646–4658, https://doi.org/10.1021/ac0341261.

[77] L. Reiter, M. Claassen, S.P. Schrimpf, M. Jovanovic, A. Schmidt, J.M. Buhmann, M.O. Hengartner, R. Aebersold, Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry, Mol. Cell. Proteomics 8 (2009) 2405–2417, https://doi.org/10.1074/mcp.M900317-MCP200.

[78] O. Serang, M.J. MacCoss, W.S. Noble, Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data, J. Proteome Res. 9 (2010) 5346–5357, https://doi.org/10.1021/pr100594k.

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

[80] G.M. Sheynkman, J.E. Johnson, P.D. Jagtap, M.R. Shortreed, G. Onsongo, B.L. Frey, T.J. Griffin, L.M. Smith, Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations, BMC Genomics 15 (2014) 703, https://doi.org/10.1186/1471-2164-15-703.

[81] M.C. Chambers, P.D. Jagtap, J.E. Johnson, T. McGowan, P. Kumar, G. Onsongo, C.R. Guerrero, H. Barsnes, M. Vaudel, L. Martens, B. Gruning, I.R. Cooke, M. Heydarian, K.L. Reddy, T.J. Griffin, An accessible proteogenomics informatics resource for cancer researchers, Cancer Res. 77 (2017) e43–e46, https://doi.org/10.1158/0008-5472.CAN-17-0331.

[82] B. Wen, S. Xu, R. Zhou, B. Zhang, X. Wang, X. Liu, X. Xu, S. Liu, PGA: an R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq, BMC Bioinformatics. 17 (2016) 244, https://doi.org/10.1186/s12859-016-1133-3.

[83] B. Wen, S. Xu, G.M. Sheynkman, Q. Feng, L. Lin, Q. Wang, X. Xu, J. Wang, S. Liu, sapFinder: an R/Bioconductor package for detection of variant peptides in shotgun proteomics experiments, Bioinformatics 30 (2014) 3136–3138, https://doi.org/10.1093/bioinformatics/btu397.

[84] A.I. Nesvizhskii, Proteogenomics: concepts, applications and computational strategies, Nat. Methods 11 (2014) 1114–1125, https://doi.org/10.1038/nmeth.3144.

[85] G.W. Park, H. Hwang, K.H. Kim, J.Y. Lee, H.K. Lee, J.Y. Park, E.S. Ji, S.-K.R. Park, J.R. Yates 3rd, K.-H. Kwon, Y.M. Park, H.-J. Lee, Y.-K. Paik, J.Y. Kim, J.S. Yoo, et al., J. Proteome Res. 15 (2016) 4082–4090, https://doi.org/10.1021/acs.jproteome.6b00376.

[86] P. Jagtap, J. Goslinga, J.A. Kooren, T. McGowan, M.S. Wroblewski, S.L. Seymour, T.J. Griffin, A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies, Proteomics. 13 (2013) 1352–1357, https://doi.org/10.1002/pmic.201200352.

[87] H. Li, Y.S. Joh, H. Kim, E. Paek, S.-W. Lee, K.-B. Hwang, Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification, BMC Genomics 17 (2016) 1031, https://doi.org/10.1186/s12864-016-3327-5.

[88] M. Helmy, N. Sugiyama, M. Tomita, Y. Ishihama, Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing, Genome Biol. 11 (2010) P17, https://doi.org/10.1186/gb-2010-11-s1-p17.

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

[90] R.C. Rivers, C. Kinsinger, E.S. Boja, T. Hiltke, M. Mesri, H. Rodriguez, Linking cancer genome to proteome: NCI's investment into proteogenomics, Proteomics. 14 (2014) 2633–2636, https://doi.org/10.1002/pmic.201400193.

[91] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M.C. Chambers, L.J. Zimmerman, K.F. Shaddox, S. Kim, S.R. Davies, S. Wang, P. Wang, C.R. Kinsinger, R.C. Rivers, H. Rodriguez, R.R. Townsend, M.J.C. Ellis, S.A. Carr, D.L. Tabb, R.J. Coffey, R.J.C. Slebos, D.C. Liebler, S.A. Carr, M.A. Gillette, K.R. Klauser, E. Kuhn, D.R. Mani, P. Mertins, K.A. Ketchum, A.G. Paulovich, J.R. Whiteaker, N.J. Edwards, P.B. McGarvey, S. Madhavan, P. Wang, D. Chan, A. Pandey, I.-M. Shih, H. Zhang, Z. Zhang, H. Zhu, G.A. Whiteley, S.J. Skates, F.M. White, D.A. Levine, E.S. Boja, C.R. Kinsinger, T. Hiltke, M. Mesri, R.C. Rivers, H. Rodriguez, K.M. Shaw, S.E. Stein, D. Fenyo, T. Liu, J.E. McDermott, S.H. Payne, K.D. Rodland, R.D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D.F. Ransohoff, A.N. Hoofnagle, D.C. Liebler, M.E. Sanders, Z. Shi, R.J.C. Slebos, D.L. Tabb, B. Zhang, L.J. Zimmerman, Y. Wang, S.R. Davies, L. Ding, M.J.C. Ellis, R. Reid Townsend, Proteogenomic characterization of human colon and rectal cancer, Nature 513 (2014) 382–387, https://doi.org/10.1038/nature13438.

[92] P. Mertins, D.R. Mani, K.V. Ruggles, M.A. Gillette, K.R. Clauser, P. Wang, X. Wang, J.W. Qiao, S. Cao, F. Petralia, E. Kawaler, F. Mundt, K. Krug, Z. Tu, J.T. Lei, M.L. Gatza, M. Wilkerson, C.M. Perou, V. Yellapantula, K. Huang, C. Lin, M.D. McLellan, P. Yan, S.R. Davies, R.R. Townsend, S.J. Skates, J. Wang, B. Zhang, C.R. Kinsinger, M. Mesri, H. Rodriguez, L. Ding, A.G. Paulovich, D. Fenyo, M.J. Ellis, S.A. Carr, Proteogenomics connects somatic mutations to signalling in breast cancer, Nature 534 (2016) 55–62, https://doi.org/10.1038/nature18003.

[93] H. Zhang, T. Liu, Z. Zhang, S.H. Payne, B. Zhang, J.E. McDermott, J.-Y. Zhou, V.A. Petyuk, L. Chen, D. Ray, S. Sun, F. Yang, L. Chen, J. Wang, P. Shah, S.W. Cha, P. Aiyetan, S. Woo, Y. Tian, M.A. Gritsenko, T.R. Clauss, C. Choi, M.E. Monroe, S. Thomas, S. Nie, C. Wu, R.J. Moore, K.-H. Yu, D.L. Tabb, D. Fenyö, V. Bafna, Y. Wang, H. Rodriguez, E.S. Boja, T. Hiltke, R.C. Rivers, L. Sokoll, H. Zhu, I.-M. Shih, L. Cope, A. Pandey, B. Zhang, M.P. Snyder, D.A. Levine, R.D. Smith, D.W. Chan, K.D. Rodland, S.A. Carr, M.A. Gillette, K.R. Klauser, E. Kuhn, D.R. Mani, P. Mertins, K.A. Ketchum, R. Thangudu, S. Cai, M. Oberti, A.G. Paulovich, J.R. Whiteaker, N.J. Edwards, P.B. McGarvey, S. Madhavan, P. Wang, D.W. Chan, A. Pandey, I.-M. Shih, H. Zhang, Z. Zhang, H. Zhu, L. Cope, G.A. Whiteley, S.J. Skates, F.M. White, D.A. Levine, E.S. Boja, C.R. Kinsinger, T. Hiltke, M. Mesri, R.C. Rivers, H. Rodriguez, K.M. Shaw, S.E. Stein, D. Fenyo, T. Liu, J.E. McDermott, S.H. Payne, K.D. Rodland, R.D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D.F. Ransohoff, A.N. Hoofnagle, D.C. Liebler, M.E. Sanders, Z. Shi, R.J.C. Slebos, D.L. Tabb, B. Zhang, L.J. Zimmerman, Y. Wang, S.R. Davies, L. Ding, M.J.C. Ellis, R.R. Townsend, Integrated proteogenomic characterization of human high-grade serous ovarian cancer, Cell 166 (2016) 755–765, https://doi.org/10.1016/j.cell.2016.05.069.

[94] S. Woo, S.W. Cha, S. Na, C. Guest, T. Liu, R.D. Smith, K.D. Rodland, S. Payne, V. Bafna, Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data, Proteomics. 14 (2014) 2719–2730, https://doi.org/10.1002/pmic.201400206.

[95] J.A. Alfaro, A. Ignatchenko, V. Ignatchenko, A. Sinha, P.C. Boutros, T. Kislinger, Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines, Genome Med. 9 (2017) 62, https://doi.org/10.1186/s13073-017-0454-9.

[96] H. Sun, X. Xing, J. Li, F. Zhou, Y. Chen, Y. He, W. Li, G. Wei, X. Chang, J. Jia, Y. Li, L. Xie, Identification of gene fusions from human lung cancer mass spectrometry data, BMC Genomics 14 (2013) S5, https://doi.org/10.1186/1471-2164-14-S8-S5.

[97] Y. Zhang, P. Kwok-Shing Ng, M. Kucherlapati, F. Chen, Y. Liu, Y.H. Tsang, G. de Velasco, K.J. Jeong, R. Akbani, A. Hadjipanayis, A. Pantazi, C.A. Bristow, E. Lee, H.S. Mahadeshwar, J. Tang, J. Zhang, L. Yang, S. Seth, S. Lee, X. Ren, X. Song, H. Sun, J. Seidman, L.J. Luquette, R. Xi, L. Chin, A. Protopopov, T.F. Westbrook, C.S. Shelley, T.K. Choueiri, M. Ittmann, C. Van Waes, J.N. Weinstein, H. Liang, E.P. Henske, A.K. Godwin, P.J. Park, R. Kucherlapati, K.L. Scott, G.B. Mills, D.J. Kwiatkowski, C.J. Creighton, A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations, Cancer Cell 31 (2017), https://doi.org/10.1016/j.ccell.2017.04.013 (820–832.e3).

[98] T. Nishimura, H. Nakamura, Developments for personalized medicine of lung cancer subtypes: mass spectrometry-based clinical proteogenomic analysis of oncogenic mutations, Adv. Exp. Med. Biol. 926 (2016) 115–137, https://doi.org/10.1007/978-3-319-42316-6_8.

[99] T. Kondo, Proteogenomics for the study of gastrointestinal stromal tumors, Adv. Exp. Med. Biol. 926 (2016) 139–151, https://doi.org/10.1007/978-3-319-42316-6_9.

[100] J.A. Staal, Y. Pei, B.R. Rood, et al., Int. J. Mol. Sci. 17 (2016), https://doi.org/10.3390/ijms17101744.

[101] D.P. Granados, A. Rodenbrock, J.-P. Laverdure, C. Cote, O. Caron-Lizotte, C. Carli, H. Pearson, V. Janelle, C. Durette, E. Bonneil, D.C. Roy, J.-S. Delisle, S. Lemieux, P. Thibault, C. Perreault, Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers, Leukemia. 30 (2016) 1344–1354, https://doi.org/10.1038/leu.2016.22.

[102] K.D. Rodland, P. Piehowski, R.D. Smith, Moonshot objectives, Cancer J. 24 (2018) 121–125, https://doi.org/10.1097/PPO.0000000000000315.

[103] P.E. Geyer, N.A. Kulak, G. Pichler, L.M. Holdt, D. Teupser, M. Mann, Plasma proteome profiling to assess human health and disease, Cell Syst. 2 (2016) 185–195, https://doi.org/10.1016/j.cels.2016.02.015.

[104] L. Dimitrakopoulos, I. Prassas, E.P. Diamandis, A. Nesvizhskii, T. Kislinger, J. Jaffe, A. Drabovich, Proteogenomics: opportunities and caveats, Clin. Chem. 62 (2016) 551–557, https://doi.org/10.1373/clinchem.2015.247858.

[105] D. Kumar, D. Dash, Proteogenomic tools and approaches to explore protein coding landscapes of eukaryotic genomes, Adv. Exp. Med. Biol. 926 (2016) 1–10, https://doi.org/10.1007/978-3-319-42316-6_1.

[106] F. Saeed, Big data proteogenomics and high performance computing: challenges and opportunities, 2015 IEEE Glob. Conf. Signal Inf. Process., IEEE, 2015, pp. 141–145, , https://doi.org/10.1109/GlobalSIP.2015.7418173.

[107] J. Powles, H. Hodson, Google DeepMind and healthcare in an age of algorithms, Health Technol. (Berl.) 7 (2017) 351–367, https://doi.org/10.1007/s12553-017-0179-1.

[108] G. Gresham, J. Schrack, L.M. Gresham, A.M. Shinde, A.E. Hendifar, R. Tuli, B.J. Rimel, R. Figlin, C.L. Meinert, S. Piantadosi, Wearable activity monitors in oncology trials: current use of an emerging technology, Contemp. Clin. Trials. 64 (2018) 13–21, https://doi.org/10.1016/j.cct.2017.11.002.