HHS Public Access

Author manuscript

Annu Rev Anal Chem (Palo Alto Calif). Author manuscript; available in PMC 2016 August 19.

Published in final edited form as:

Annu Rev Anal Chem (Palo Alto Calif). 2016 June 12; 9(1): 521–545. doi:10.1146/annurev-anchem-071015-041722.

Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation

Gloria M. Sheynkman[1,2,3], **Michael R. Shortreed**[3], **Anthony J. Cesnik**[3], and **Lloyd M. Smith**[3,4]

Gloria M. Sheynkman: gloriam_sheynkman@dfci.harvard.edu; Michael R. Shortreed: mshort@chem.wisc.edu; Anthony J. Cesnik: cesnik@wisc.edu; Lloyd M. Smith: smith@chem.wisc.edu

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215

[2]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115

[3]Department of Chemistry, University of Wisconsin, Madison, Wisconsin 53706

[4]Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin 53706

Abstract

Mass spectrometry–based proteomics has emerged as the leading method for detection, quantification, and characterization of proteins. Nearly all proteomic workflows rely on proteomic databases to identify peptides and proteins, but these databases typically contain a generic set of proteins that lack variations unique to a given sample, precluding their detection. Fortunately, proteogenomics enables the detection of such proteomic variations and can be defined, broadly, as the use of nucleotide sequences to generate candidate protein sequences for mass spectrometry database searching. Proteogenomics is experiencing heightened significance due to two developments: (*a*) advances in DNA sequencing technologies that have made complete sequencing of human genomes and transcriptomes routine, and (*b*) the unveiling of the tremendous complexity of the human proteome as expressed at the levels of genes, cells, tissues, individuals, and populations. We review here the field of human proteogenomics, with an emphasis on its history, current implementations, the types of proteomic variations it reveals, and several important applications.

Keywords

customized protein databases; genetic variation; isoforms; polymorphism; proteomics; sample-specific databases; proteoform; single amino acid variant; novel splice junction; alternative splicing

Disclosure Statement: The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

1. Introduction

Mass spectrometry (MS)-based proteomics has become the leading method for comprehensive detection and characterization of proteins and is ubiquitous throughout biology and medicine (1). To identify peptides or proteins, most proteomics workflows rely on database searching in which experimental peptide mass spectra are scored against theoretical mass spectra derived from a generic protein database. Thus, proteins whose exact sequences are absent from the generic databases to which they are matched remain undetected. Frequently, these missing sequences are of fundamental biological significance —novel or unannotated proteins, variations specific to individuals, mutations underlying disease—and their characterization is critical to an understanding of human biology.

Proteogenomics provides a solution to this detection problem. It may be defined as the use of genomic or transcriptomic nucleotide sequencing data to create customized or augmented proteomic databases for MS-based proteomics database searching, and as the employment of these databases to enable or improve the detection of protein variations unique to a sample. In the past decade, proteogenomics has experienced heightened significance due to two cumulative developments: (*a*) advances in DNA sequencing technologies, making complete sequencing of human genomes and transcriptomes routine, and (*b*) the realization of the tremendous complexity of the human proteome. Indeed, it is the unprecedented evolution of DNA sequencing technologies, which has been reviewed in detail elsewhere (2), that has been largely responsible for revealing the tremendous complexity encoded in the human proteome. Current proteogenomic strategies are now harnessing sequencing technologies, in the form of widely and easily accessible nucleotide sequence data, to maximize the potential of MS-based proteomics to characterize these variants at the protein level.

The benefits of proteogenomics are manifold. Proteogenomics can provide empirical evidence for the existence of proteins and protein variations, which can help delineate the set of protein-coding genes in the human genome. Though this has been a key goal of genome annotation, made evident by intensive efforts involving computational gene predictors (3), collection of troves of transcript data, and manual annotation, there is still no definitive set of protein-coding genes (4– 6). Proteogenomics can be used to study the effect of genetic variations on the ultimate proteins they encode, providing a richer understanding of genotype-phenotype relationships as proteins are more direct determinants of function. Proteogenomics can help us understand the underlying mechanisms of disease, discover therapeutic targets, or generate biomarkers for diagnosis or tailored therapies. And finally, proteogenomics can improve the efficiency and accuracy of both nucleotide and proteomics analyses. For example, an optimal protein database could in theory be adaptively constructed for each sample type for improved peptide identification accuracy.

Most proteogenomic workflows involve several main steps. First, nucleotide data relevant to the sample of interest are obtained, such as sequences corresponding to the human genome or sequences from a set of transcripts assembled from raw RNA-sequencing (RNA-Seq) reads. Second, the nucleotide sequence is translated into amino acid sequence. Translation is typically done in one (if the frame is known), three (if the strand is known), or six frames in order to create a customized protein database. Third, fragmentation mass spectra are

searched against the protein database. Finally, the peptide identifications are statistically validated and evaluated to annotate novel genes, detect disease variants, or systematically analyze protein variations.

In this review, we begin by providing a historical perspective on the long-standing relationship between nucleotide sequencing data and MS-based proteomics. Next, we outline sources of nucleotide data amenable to proteogenomic database construction, describe methods for construction of protein databases, and discuss issues relevant to proteogenomics. We enumerate human proteomic variation types thus far detected by proteogenomics, highlight emerging applications, and conclude with future directions of the field.

2. Proteomics and DNA Sequencing: A Historical Perspective

The use of nucleotide sequence data for proteomics database search has a rich history that can be traced back to the development of the first computer MS search algorithms in the early 1990s. Prior to that time, the main method used for protein analysis was Edman degradation, which sequences the N terminus of a polypeptide but requires large amounts of purified protein (7). MS was being employed in the analysis of peptides; however, the ionization modes used to introduce peptides into the mass analyzer were harsh and limited analysis to short or chemically derivatized peptides (8). Both methods were low-throughput and labor-intensive. Later development of soft ionization techniques such as plasma desorption enabled analysis of larger peptides and foreshadowed the potential of MS-based methods. This potential was realized when two soft ionization techniques, electrospray ionization (ESI) (9) and matrix-assisted laser desorption ionization (MALDI) (10), were invented, enabling the facile analysis of intact peptides and proteins at high-throughput. Shortly after, the elegant integration of liquid chromatography (LC), ESI, and mass spectrometry (LC-ESI/MS) enabled the collection of hundreds of peptide fragmentation mass spectra, a volume of data beyond what could be manually analyzed, thus precipitating a need for methods to rapidly and automatically identify peptides. In select cases, peptide sequences could be deduced de novo by extracting a peptide ladder, a series of peptide fragment peaks with spacing corresponding to the mass of amino acids in the original peptide (11). However, most mass spectra were noisy and lacked a definitive peptide ladder. Therefore, researchers quickly turned to a new method of peptide identification: protein database searching.

In the protein database search approach, peptides are identified by matching experimental fragmentation spectra to theoretical mass spectra derived from a protein database. In 1994, this method was pioneered by Eng and coworkers (12) using a limited set of available human protein sequences. In the meantime, a large effort was under way to sequence human expressed sequence tags (ESTs)—Sanger-sequenced 5′ and 3′ ends of complementary DNAs created from reverse transcription of mRNAs—that represented a systematic survey of protein-coding transcripts (13). Taking advantage of this development, Yates and coworkers (14) performed a six-frame translation of available human ESTs (~60,000) and used the candidate protein sequences for database searching against peptide fragmentation spectra. An analogous approach was developed for peptide mass fingerprinting data (15, 16).

These studies of protein database searching against EST-translated protein databases were the start of a new interdependent relationship between nucleotide sequencing and MS-based proteomics technologies that still prevails today. Initially, nucleotide and amino acid sequences could be directly determined using analytical methods: nucleotides could be “read out” by Sanger chain termination (17); amino acids could be read out by Edman degradation (7). Both methods had similar throughput. However, in 1988, nucleotide sequencing technologies leapt forward upon the invention of automated Sanger sequencing by Smith and coworkers (18), enabling the interrogation of DNA and RNA sequences at a massive scale. Meanwhile, Edman degradation remained labor-intensive and was gradually replaced by MS-based proteomics (19). But MS operated in a different mode. It did not directly sequence proteins; rather, it measured the intact and fragment mass-to-charge ratios of polypeptides, where it was not straightforward to deduce the sequence solely from the data. However, mass spectra could be analyzed with the aid of a protein database derived from translation of widely available nucleotide-based sequences. It is this interdependency between nucleotide sequencing and MS-based proteomics technologies that set the foundation of proteogenomics.

During the 1990s and 2000s, shotgun proteomics became widespread and database searching against a protein database became standard, with the exception of a few proteogenomic-like studies that directly utilized EST or genomic sequence database searching. In 1998, Neubauer and coworkers (20) constructed protein databases from public ESTs to identify components of the human spliceosome complex. In 2001, Choudhary and coworkers (21) searched human MS data against translations of the draft human genome, showing that unbiased proteomic data sets could be used to discover new protein-coding regions in the human genome. Though the proteogenomics concept was in practice before, the term proteogenomics was officially coined in 2004 when Jaffe and coworkers searched a shotgun proteomics data set against a six-frame translation of the *Mycoplasma* genome. This genome-based proteogenomic search strategy was subsequently applied to increasingly complex organisms: *Drosophila melanogaster* (22), *Arabidopsis thaliana* (23), and *Caenorhabditis elegans* (24). Collectively, these studies showed that although these species had deep-coverage EST databases and were subject to intense gene annotation efforts, there were still many novel protein-coding genes and errors in the protein annotations that could be uncovered by genome-based proteogenomic strategies. Thus, in following the evolving definition of proteogenomics, here it meant that MS could provide valuable experimental evidence confirming the existence of the protein sequences that are expressed in an organism.

Another turning point in the evolution of proteogenomics coincided with the development of next-generation sequencing (NGS) methods. NGS platforms harnessed massively parallel sequencing to allow for the shotgun sequencing of millions of short fragments en masse. In 2009, RNA-Seq, in which fragments from a eukaryotic transcriptome are sequenced to great depth, was invented (25). NGS data illuminated a newfound vastness of human proteomic variation encoded in the genome, such as variations arising from nucleotide polymorphisms (26) and alternative splicing (27, 28). It became clear that there were more proteomic variations than were cataloged in standard protein databases. Catalyzed by NGS, a new type of proteogenomics emerged, in which sample-specific nucleotide and proteomic data were

collected from the same sample to create customized protein databases for detection of novel variations (29). Today, this NGS-driven proteogenomic strategy is being increasingly applied to detect and study human protein variations in basic and disease biology.

Proteogenomics operates at the interface of genomics and proteomics and has evolved in the past two decades. From the first EST-derived database to genome-based searching to the latest NGS-based methods, proteogenomics will undoubtedly play a key role in the integration of genomic, transcriptomic, and proteomic data for the improved understanding of cellular biology.

3. Proteogenomic Database Construction

3.1. Standard Human Proteomic Databases

The main protein databases used in MS-based proteomics searching include UniProt, RefSeq, and Gencode. UniProt has become one of the leading proteomic databases because it provides manual human protein annotations supplemented with known functional information (30). RefSeq is a cDNA-centric database that aims to provide a conservative, manually annotated set of proteins (31). Gencode is another database and contains both manual annotation (Havana group) and all automatic annotations predicted by Ensembl (4). Gencode is a genome-centric database; all transcript and protein sequences can be directly mapped to the reference genome and there is perfect DNA-RNA-protein concordance.

Common to most protein databases is the idea of nonredundancy. In the early days of protein annotation, the high number of overlapping or similar sequences was a known problem, leading to efforts to remove redundant sequences. Though this solved the problem of redundancy, it also resulted in the loss of true biological variations. Whereas the concept of nonredundancy has been slowly reversing and databases such as UniProt and Gencode now strive to include known variations, such as isoforms or single-nucleotide polymorphisms (SNPs), the protein databases simply do not include all measured and yet-to-be measured protein variations extant in the human population.

3.2. DNA Sequencing Platforms and Sources of Nucleotide Sequence Data

Capillary-based Sanger sequencing was the main method for the initial sequencing of the human genome and transcriptome. With the development of NGS methods, many (millions to billions) short reads could be obtained at great depth (2). Although the exact mechanisms for sequencing differ between the platforms, what they have in common is the ability to produce millions to billions of short DNA reads, providing ample data from which to build proteomic databases.

The type of data relevant to proteogenomics can be defined as any nucleotide sequence that has the potential to encode a protein expressed in a sample, which includes sequences from the genome, exome, transcriptome, and translatome (Figure 1). Genome sequence contains predominantly noncoding regions but is comprehensive in that it contains the original backbone of all protein sequences. Exome sequence comprises the 1% of the genome that codes for protein. These sequences are obtained through exome sequencing where the exons of a genome are enriched through hybridization capture and sequenced (32). Transcriptome

sequence represents the cumulative output of gene transcription and can either be noncoding or coding. Most RNA-Seq data are derived from the 1–3% of protein-coding mRNAs remaining after removal of ribosomal RNA (25). Translatome sequence represents the portions of the transcriptome that are bound by ribosomes and thus have a high likelihood of coding for protein. These data sets are generated through ribosomal sequencing (Ribo-Seq), where the portions of the mRNAs that are bound by ribosomes are captured and sequenced to provide a global snapshot of transcripts actively being translated into protein (33).

3.3. Deriving Candidate Protein Sequences for Database Searches

As described above, there are different types of nucleotide data that can be used for proteogenomic database construction (Figure 1). Ultimately, the effectiveness of a protein database derived from nucleotide data depends on how closely the predicted protein sequences match the real protein sequences expressed in the sample under analysis. In the following sections, we discuss the different mechanisms through which protein databases can be generated and the pros and cons of each method.

3.3.1. Three- or six-frame expressed sequence tag translation—A majority of the predicted protein sequences that populate current protein databases have been derived from ESTs, full or partial sequences of mRNAs. ESTs can be translated in three—if the original 5′-3′ orientation is known—or six frames. There are currently 9 million human ESTs present in GenBank ranging in size from 100–2,000 bases (13). The first direct use of nucleotide sequence data to identify proteins from MS data utilized publicly available ESTs. This includes the pioneering studies by Yates and coworkers (14), who searched six-frame translated data against a proteomic data set, and Neubauer and coworkers (20), who did similar searches but using a tag-based peptide search method. These studies were motivated by the incompleteness of the protein databases at the time. Later, after protein databases became more complete, ESTs were primarily used to detect variations [e.g., single amino acid variants (SAVs), splice variants] that may not have been included in the generic protein databases. Two groups accomplished this by collapsing ESTs into a compact splice graph structure, where exons are represented as nodes and the junctions are represented as edges, and translating unique nucleotide stretches within the splice graph (34, 35).

3.3.2. Six-frame genome translation—The genome of an organism can be translated in six frames to create an all-inclusive set of possible protein-coding sequences. The six-frame genome proteogenomic method was first comprehensively demonstrated for *Mycoplasma*, where shotgun proteomics data were searched against translated sequences and identified peptides were mapped back to the genome to delineate protein-coding genes (36, 37). Though demonstrably successful, *Mycoplasma* is a prokaryote with a small genome and uninterrupted ORFs. The six-frame search is more challenging for the human genome, which has a higher proportion of noncoding sequence and genes with complex exon-intron structures. Nevertheless, the method has been applied to humans in several studies. The initial proof of concept was demonstrated in 2001 in two studies that identified and mapped peptides to the draft human genome (38, 39). In 2006, Fermin and coworkers (39) used the growing plasma proteome data sets provided by the Human Plasma Project to discover novel plasma-specific genes. More recent studies have combatted the issue of high false positives

in six-frame genome searches by first searching mass spectra against a protein and translated transcriptome database and searching the remaining unidentified spectra against the genome. Using this multitiered search on an ENCODE cell line proteomics dataset, Khatun and coworkers (40) detected peptides supporting the existence of novel translational start sites and even the translation of sequences annotated as untranslated regions (UTRs). Similarly, Kim and coworkers (5) detected unannotated protein sequences from data representing a draft map of the human proteome. Overall, despite the large size and higher false positive rate associated with six-frame genome searches, this strategy enables identification of novel peptides that may be missing from protein or transcriptome-based databases.

3.3.3. Protein databases from RNA sequencing—Similarly to ESTs, RNA-Seq reads represent fragments of transcript sequences, but at much greater depth due to the unprecedented throughput of NGS. As RNA-Seq read lengths are typically much shorter (50–250 bp) than the lengths of ESTs (500–800 bp), reads must be assembled into the longer transcripts from which they originated. Assembly is accomplished in two main ways, either through alignment-based methods or de novo assembly. In alignment-based methods, the RNA-Seq reads are aligned to the human reference genome, which acts as a scaffold so that reads derived from the same transcript are essentially pieced back together. In de novo assembly, the RNA-Seq reads themselves are used to assemble a full-length transcript, relying on many partially overlapping reads to build a contig. As the human reference genome is of high quality, virtually all RNA-Seq methods utilize alignment-based transcript reconstruction.

RNA-Seq provides information on the abundance of and nucleotide variations encoded within transcripts. Of significance for the implementation of proteogenomic studies, this information may be classified into three areas: transcript abundance, nucleotide-level variations (SNPs, small indels, etc.), and large structural variations (alternative splicing, large insertions, etc.). How this information is utilized in protein database construction and in augmentation differs; therefore, we consider them separately. In addition, we discuss the databases created from de novo assembled RNA sequences.

3.3.3.1. Transcriptional expression: RNA-Seq can provide an estimate of transcript abundance as the number of reads sequenced for each transcript is proportional to that transcript's concentration. Assuming that transcript expression is a prerequisite for protein expression, some proteogenomic studies have created reduced protein databases comprising only proteins with transcriptional evidence. Theoretically, this has the beneficial effect of removing noise, that is, protein sequences that are in the database but not present in the sample. Indeed, reduced databases have been shown to increase the peptide identification rate by 5% for moderate-coverage proteomic data sets (29). However, this improvement vanishes for deeper-coverage proteomic data (41). In fact, there may be danger of removing proteins with low RNA-protein abundance correlations or with transcripts that are undersampled in RNA-Seq, such as mRNAs without polyA tails (42).

A promising use of transcriptional abundance data in proteomics is the incorporation of isoform expression to improve protein inference. In protein inference, peptides identified from a shotgun proteomics experiment are mapped to all annotated protein isoforms for a

gene, and the identities of the expressed proteins are inferred through such mappings (43). Frequently, peptides map to more than one protein isoform because isoforms of the same gene share many exons. Knowledge of which isoforms are expressed at the transcript level can help eliminate unlikely protein isoforms during protein inference, leading to a less ambiguous or more accurate set of protein isoform identifications.

3.3.3.2. Small nucleotide variations extracted from the RNA-Seq reads: RNA-Seq data represent global sampling of the sequences encoded in a particular transcriptome. These sequences can be aligned to the reference genome, and the RNA sequences can be compared to the reference genome sequences to identify small sequence differences, specifically SNPs and short indels. These variations can be directly translated into protein and appended to a protein database, or the information about the variations can be used to guide amendment or alternation of the standard protein database. For example, nonsynonymous SNPs identified from sample-matched RNA-Seq data were translated into polypeptide sequences containing the amino acid polymorphisms and appended to a protein database for MS database searching to identify variant-containing peptides (29, 44).

3.3.3.3. Large structural variations inferred from RNA-Seq data: RNA-Seq data contain information about large structural variations such as gene fusions and alternative splice junctions; however, to get this information, the read alignment and interpretation is different from the extraction of smaller nucleotide variations. When RNA-Seq reads are aligned to the genome, reads that span fusions or splice junctions must be split across the breakpoints and thus be aligned using a splice-aware aligner. Once aligned, the locations of the junctions are then compared to the locations of known junctions or fusions and novel sites are retained for database construction. Regions of the aligned RNA-Seq reads that contain the nucleotide breakpoint are then translated into protein sequences, using one, three, or six frames depending on prior knowledge. Using this strategy, the identification of peptides spanning novel splice junctions (45, 46) and chimeric transcripts (47) has been demonstrated. In some cases, splice junctions detected from RNA-Seq data can be converted into a splice graph representation to create a compact splice database for MS searching (48, 49).

3.3.3.4. Three-frame translation of assembled RNA-Seq reads: Though RNA-Seq reads are short, they can still be aligned to the genome or assembled de novo to produce contigs, inferred sequences of partial or full-length transcripts. These sequences can then be translated in three frames to produce candidate proteins that may harbor many variation types, from SAVs to splice junctions. Such protein sequences are directly utilized in MS searching. Examples of this strategy include those employing RNA-Seq data that underwent transcript reconstruction (50) or de novo assembly (51).

3.3.4. Ribosomal sequencing-guided database construction—Ribo-Seq is a method for sequencing portions of mRNA molecules bound by ribosomes, providing a global picture of the actively translated components of the transcriptome. Ribo-Seq provides information about the mRNA locations that are subject to translation, including novel translational start sites and coding regions, so that the underlying RNA sequence for these regions may be translated to create a highly specific proteogenomic database. Such a

strategy has been demonstrated to enable discovery of novel translation initiation sites (TISs) using noncanonical (CTG instead of ATG) initiation codons, novel short peptides residing within 5′ UTRs, and unannotated proteins (52, 53).

3.3.5. Protein databases from exhaustive combinations of all possible variations—Another way to create databases that encode putative variations is to include all theoretical combinations of the variants, such as all possible SAVs or splice junctions. As all these methods suffer from a large search space that includes inordinately high numbers of specious sequences unlikely to be expressed in the sample, they have been replaced by newer proteogenomic methods. Early MS search algorithms, such as X!Tandem and Mascot, allowed for the searching of spectra against all possible SNP combinations (54, 55). Early studies in splice detection created “exhaustive” exon-exon and exon-intron databases that created polypeptide sequences spanning all possible junctions from known or predicted exons (56–60).

3.3.6. Specialized databases focused on variation types—Beyond standard proteomic databases, several specialized databases have been created that focus on a particular variation type, such as SNPs (61–67), splice variants (68–70), or chimeric transcripts (47, 71). In many cases, these databases were designed so that the relevant amino acid sequences corresponding to the variation could be downloaded in a format amenable to MS database searching. Nucleotide databases of note for characterization of human variations include dbSNP, a catalog of all detected human SNPs (72); dbVar, a catalog of human structural variations, such as large insertions and deletions (73); COSMIC (Catalogue of Somatic Mutations in Cancer), a database of cancer-specific variants (67); and HGMD (Human Gene Mutation Database), a database of human disease variants.

3.3.7. Cases where protein variations can be directly extracted from the MS data—In some cases, protein variations can be detected directly from features in the mass spectra, given that the spectra are of high enough quality. For example, error-tolerant search methods detect variations directly from the MS data. Peptide sequence tag–based database search algorithms make allowances for mass shifts corresponding to amino acid polymorphisms, and thus provide a partial de novo approach for finding amino acid variations without prior knowledge of their identities (74, 75). A similar concept is the use of template proteogenomics (76), where spectra that represent partial matches to the protein database are subject to de novo sequencing techniques to “fill in” the rest of the sequence that does not match the database, thus enabling the identification of amino acid or splice variants that may not be directly encoded by the database. In these methods, multiple enzymatic digests can be performed to supply multiple spectra overlapping the same variation-containing region. This method was recently used to fully regenerate an antibody sequence de novo without knowledge of the DNA sequence (77). We note that although there are cases of de novo and partial de novo sequencing methods that can utilize MS data for the detection of variations, they are limited to specialized cases, require high-quality spectra, and are in efficient compared to database searching.

4. Key Proteogenomics Issues Relevant to Human Proteomics

4.1. Controlling for False Positive Identifications

Most proteogenomic databases contain many more protein candidate sequences than traditional protein databases; the computational aspects of this are reviewed in detail in Reference 78. Given such large protein search spaces with many more protein sequences in the database but not necessarily in the sample, there is a higher chance of false positive identifications (79). However, because proteogenomics enables the detection of unannotated proteins and those unique to an individual or disease, there is undeniable value to the strategy and several approaches have been described to improve the quality of proteogenomic search results.

4.2. Reducing Nucleotide/Peptide Candidates

How can we improve the specificity and accuracy of proteogenomic identifications? One clear solution is database reduction: the use of outside knowledge beyond the raw nucleotide sequences to pinpoint the sequences most likely translated and thus remove random sequence unlikely to code for protein (80). This can be accomplished on either the nucleotide side or the MS-based proteomics side. On the nucleotide side, extrinsic and intrinsic information can be used to reduce the set of amino acid candidates to those most likely to code for protein. Similarly, on the MS side, extrinsic and intrinsic information can be used to either reduce the protein candidate space or provide increased confidence for certain peptide identifications.

The use of transcriptome sequence dramatically reduces the proteogenomic search space; instead of translating a 3 Gb genome, one can translate the processed transcriptome, which is less than 1% of the size of the human genome. Further, though the transcriptome is a much smaller portion of the genome, it still contains both noncoding RNAs and coding RNAs, which have UTRs. Thus, an even further reduction in search space can be attained if there is knowledge of the actively translated portions of the transcriptome. Ribo-Seq data, which show which transcripts are actively translated, can be employed to this end. And lastly, if there is knowledge of the abundances of the transcripts, this information can be used to increase confidence in sequences with higher RNA-Seq read coverage.

Strategies to reduce the search space in proteogenomic databases are not limited to the nucleotide side. Adjustments in the way MS database searching is performed can help reduce the number of false positives. For instance, instead of searching mass spectra against the entire proteogenomic database, a series of first-pass searches can be performed to reduce the initially large list of candidate proteins. This multitiered search strategy was applied to a six-frame human genome translation (81). Methods have been published that describe how to accomplish this in a target-decoy search framework (81, 82). An even more common strategy is to first search higher-confidence protein databases and subsequently search the remaining unmatched spectra against proteogenomic databases (83, 84). For example, ordered searches against a generic protein database, three-frame translation of a transcriptome, and lastly, six-frame translation of the genome, where the unmatched spectra

are carried through to the next level of search, has been implemented in the most recent human proteogenomic studies (5, 40).

Another strategy for the increase in peptide identification efficiency is to reduce the candidate peptide search space. This can be accomplished using extrinsic or intrinsic MS information. An example of extrinsic information is the use of experimental data regarding the physicochemical properties of the peptides, such as isoelectric point (pI) and protein molecular weight. For example, MS workflows that employed an initial fraction of the peptide digest separated by pI used this information about the peptide to reduce the candidate peptide search space during database search (85, 86). In these studies, the pI information of each peptide was used to eliminate peptide search candidates not falling within the expected pI range, which dramatically reduced the peptide search space. Methods of using intrinsic MS features for peptide candidate reduction center around extracting information from the mass spectra themselves. For example, peptide tags can be extracted from the mass spectra to filter out peptide candidates lacking the tag. This strategy was used in early six-frame genome searches to reduce candidate search spaces by orders of magnitude (87–89). Another example is to incorporate the intrinsic mass spectral qualities in the proteogenomic search as some spectra are of much higher quality and their peptide identifications could be more reliable (90). And lastly, a strategy to increase the overall quality (i.e., signal-to-noise ratios of fragment peaks) of mass spectra is to use clustering of related spectra either within an experiment or across spectral library databases (91).

5. Other Proteogenomic Issues

5.1. False Positives

A current practical strategy for handling proteogenomic search results is to require different score thresholds for different categories of peptides. In general, peptides matching to the generic protein database have, on average, higher scores than novel peptides. This means that the group of novel peptides contains a higher rate of false positives and should require a higher score cutoff. An understanding of the uniquely different positive and negative score distributions for different classes of peptides could help in building a framework for statistical validation of proteogenomic results.

Artifacts are another source of false positives, such as when a chemically modified peptide is misidentified as a biological variation. Sources of artifacts include abundant peptides that contain a chemical modification. A recent study found that a subset of the variant peptides identified in a cancer proteogenomic study was not a biological variation but an artifact resulting from conversion of methionine to isothreonine during iodoacetamide treatment of the peptides, a common processing step in most shotgun proteomics experiments (92). Other sources of artifacts may include common chemical modifications such as oxidation or deamidation of peptides. Therefore, an increasingly important requirement should be the validation of novel peptides using synthetic peptide standards or, at minimum, the careful consideration of known and possible sources of chemical artifacts.

5.2. False Negatives

False negatives can also be a problem in proteogenomics, notably when the nucleotide data do not adequately reflect the protein content in the sample. This can occur for RNA-Seq data sets, which, though comprehensive, sometimes do not capture all protein sequences in a sample. For example, most RNA-Seq data sets that employ polyA enrichments before sequencing miss proteins derived from mRNAs lacking a polyA tail, such as histones and some zinc-finger proteins (42). False negatives can also result from transcripts that are out of phase with their protein counterparts in time or space. In time, RNA may be rapidly degraded or the protein may be unusually stable so that the transcript sequence is undetected while the protein is present. In space, secreted or extracellular proteins may be far from the original cell containing the genome or transcript sequence encoding the protein. For example, a recent proteogenomic study on *Xenopus* eggs showed that hundreds of proteins were transported into the egg during maturation, and these proteins were not detected in the RNA-Seq data (93). Similar scenarios occur for human samples, such as excreted proteins in serum or antibodies secreted by B cells. A solution to these issues is to identify sources of transcript-protein discrepancies and for cases in which the nucleotide data cannot capture all protein sequences, to supplement the database with these special proteins.

5.3. Challenges in Detecting Low-Abundance Novel Peptides

Another challenge in proteogenomics is that many novel protein variations may be of lower endogenous abundance or only expressed in certain tissues. This issue is exemplified by the difficulty in detecting alternatively spliced protein isoforms (94). The reference isoform may be many orders of magnitude higher in abundance than the alternative isoform and the high sequence overlap of the two isoforms makes unambiguous identification of the alternative isoform difficult. Some solutions to this problem include the use of multiple enzymatic digestions to produce a complementary set of peptides, as was recently demonstrated for HeLa cells (95); to rely on massive spectral clustering databases, which could group as-of-yet unidentified peptides from unusual samples (96); or to employ targeted proteomics methods such as selected reaction monitoring (SRM) to detect possible peptide variants at higher sensitivity.

5.4. Need for Bioinformatic Tools in Proteogenomics

As the size and complexity of both NGS and MS-based proteomics data sets increase, there is a pressing need for efficient and easy-to-use tools for their bioinformatic analysis. In a proteogenomic workflow, the raw nucleotide data must be aligned or assembled, compared to existing gene annotations, and translated into polypeptide sequence. The proteomics data sets also require complex workflows, including database searching and proper interpretation of the statistical results. Several groups have developed tools to aid in proteogenomic database construction (41, 97–102) and visualization of peptides on the genome or in comparison to gene models (103–107). Both NGS and proteomic technologies and tools are rapidly evolving, requiring constant parallel efforts to develop adaptable bioinformatic tools for their analysis.

6. Complexity of the Human Proteome

6.1. Human Protein Variations Encoded in the Primary Sequences of Proteins

The advent of NGS technologies has unlocked a trove of data leading to the realization of the tremendous complexity of the human transcriptome, and by proxy, the potential human proteome. Genetic variations that affect protein sequences can be encoded at the level of the genome, transcriptome, or translatome (ribosomal-bound portion of the RNA) (Figure 1). In addition, the annotation of the human proteome is incomplete, so there may still be proteins to discover. Here, we enumerate possible sources of human variation that can affect the primary sequence of proteins and describe studies that utilized proteogenomic strategies for their detection. Though posttranslational phenomena such as proteolytic cleavage and many posttranslational modifications (e.g., phosphorylation, glycosylation) are an important source of proteomic variations, we do not cover these variations in this review.

6.2. DNA-Level Human Variations

The human genome is the ultimate information carrier as it contains the raw sequence from which, after transcription and translation, proteins are derived. Thus, genomic nucleotide variations that change the sequence of the ultimate protein encoded have a direct effect on the composition of the human proteome. An understanding of how genome-encoded variants ultimately affect the proteome is crucial to understand human genetic variation and disease. Genomic variations range from small point mutations (e.g., SNPs) to large structural changes (e.g., gene fusions), and proteogenomic studies have shown the ability to detect the corresponding protein variations.

6.2.1. Single-nucleotide polymorphisms—The importance of point mutations that change the amino acid in the corresponding protein (single amino acid variant, or SAV) has been known ever since the discovery that sickle-cell anemia is caused by a glutamic acid to valine mutation in the beta subunit of hemoglobin (108). SNPs are found in the human genome at an average frequency of 1 every approximately 800 bp (26). SNPs residing in protein-coding regions of a gene can be nonsynonymous (changing the encoded amino acid) or nonsense (producing an early stop codon). Because SNPs are a major source of variation in human disease biology, there has been intense research in this area, and computational and experimental tools to predict the functional effects of SAVs (109, 110) have been developed. Proteogenomics enables the direct detection of proteins containing amino acid variations, which is crucial to the study of the functional effects of variants.

6.2.2. Nonsynonymous single-nucleotide polymorphisms—Information about nonsynonymous SNPs and their corresponding SAVs can be used to extend or amend existing protein databases, which can subsequently be used to identify peptides containing these amino acid variations. Though rarely used today, one of the earliest methods for SNP detection via MS-based proteomics was performed by exhaustively generating all possible DNA mutations and searching MS data against all possible peptide variants (111). Another approach commonly used is to convert nucleotide variant information cataloged in human SNP databases, such as the National Center for Biotechnology Information's dbSNP or COSMIC, into a SAV-containing database for MS search (61, 62, 112–114). In the most

recent SAV detection pipelines, variants that are specific to a cell line or cancer type are extracted from exome or RNA-Seq data to create a more specific, compact database of candidate SAVs (29, 44, 115, 116). These studies show the increased quality of variant peptides identified from the more compact, sample-specific SAV databases.

6.2.3. Allele-specific expression variations—As the human genome is diploid, heterozygous SNPs can result in the expression of two different protein variants from the same gene (117). Furthermore, the expression of the two variant proteins may differ due to *cis*-regulatory effects operating at the level of the genome or transcriptome. These differences in variant protein abundances are called allele-specific expression differences and have been measured via MS (44).

6.2.4. Nonsense single-nucleotide polymorphisms—Nonsense SNPs introduce an early stop codon in the middle of the protein, resulting in a so-called loss-of-function mutation, which is predicted to disrupt the function of the protein as it causes a truncated protein product that is unlikely to fold properly. Through genome-wide surveys, it was shown that each individual carries approximately 100 protein-disabling loss-of-function mutations (118). The predominant notion is that all nonsense SNPs are loss-of-function mutations, but a recent investigation into the mechanisms by which SNPs can disrupt transcript structure suggests a more complex picture (119). Thus, proteogenomic studies that attempt to measure the proteins corresponding to these mutations, including their quantitative levels, may provide a better understanding of their effects.

6.2.5. Large structural chromosomal variations—In contrast to point nucleotide differences, structural chromosomal variations affect large portions of the genome up to hundreds of megabases in length. Large regions of the genome can invert, translocate to another region of the chromosome or another chromosome, be amplified to produce multiple copies, or be deleted. These variations affect protein-coding genes by increasing or decreasing the copy numbers of the genes in integer increments [i.e., copy number variations (CNVs)] or causing one portion of a gene to be fused to another gene (i.e., gene fusion). Proteogenomic studies can measure the proteins corresponding to these variations. For example, in a quantitative proteomics study employing SILAC (stable isotope labeling by amino acids in cell culture)-labeled human cells, the effect of CNVs on the levels of protein abundances was measured, showing that although the number of CNVs roughly correlated with protein levels, this was not the case for proteins that were part of protein complexes that require strict stoichiometry of subunits (120). Gene fusions, especially those that occur during cancer progression (121), have been measured at the protein level through use of a database of gene fusions (122, 123). Proteins coded from the fusion of genes are referred to as either fusion or chimeric proteins. Customized fusion databases directly derived from NGS-derived genomic or transcriptomic sequence to detect gene fusions from RNA-Seq data have also been developed (124).

6.3. RNA-Level Human Variations

There are many variations encoded in the genome, but RNAs directly transcribed from the genome are variably processed, contributing to an even higher complexity of the

transcriptome. Alternative promoters could cause the N terminal ends of the proteins to differ; processing by the spliceosomal machinery produces multiple alternatively spliced mature RNAs; and in some unusual cases, there can even be *trans*-splicing to produce a chimeric transcript (125). And lastly, after the processed mRNA is formed, there is also the possibility of RNA editing events (126). All these sources of RNA variation that lie within the protein-coding regions will affect the protein sequence and are thus relevant to proteogenomic strategies.

6.3.1. Alternative splicing—Alternative splicing is a pervasive mechanism that creates multiple distinct mRNA molecules from a single genetic locus (27, 28). To identify protein isoforms, the generic protein database must include the relevant isoform sequences. In the case where an isoform is not in the database, one can use a comprehensive splice database such as ECGene or SpliceProt (68, 70) to identify peptides corresponding to novel isoforms specific to a sample or relevant to a disease (68–70, 105, 127). However, though the number and size of splice-specific databases is high, RNA-Seq data collected on hundreds of sample types have shown that the catalog of human isoforms is still incomplete (128). Therefore, proteogenomic strategies that use sample-specific or sample-related RNA-Seq data can be used to create both a compact and near complete—in terms of the sample of analysis—database of candidate splice junction sequences for database searching and identification of splice junction peptides (46, 129). Another strength of proteomics data in analyzing spliced proteins is the ability to confirm the expression of aberrantly or out-of-frame spliced transcripts, which are typically thought to undergo nonsense-mediated decay.

6.3.2. RNA editing—The central dogma theorem assumes that the RNA sequence perfectly matches the genome sequence from which it came; however, the discovery of widespread RNA editing events reveals an exception to this rule. The transcripts can be subject to editing events that are catalyzed by enzymes that recode the underlying protein sequence. A recent survey of DNA-RNA discordances in humans aimed to provide estimates of RNA editing events, and peptides were identified that corresponded to both the unedited and edited versions of the transcripts (126). A proteogenomic study in two rat strains also detected variant peptides from RNA editing events, which were distinguished from SNPs by the joint analysis of genomic and transcriptomic sequences (130). As it is largely unknown what the functional effects of RNA editing are, proteogenomics can help by providing detection of the affected proteins.

6.4. Translation-Level Human Variations

Knowledge of the transcript sequences in a given sample is only the first step toward knowledge of the proteome. Although a simplistic model of translation assumes that the ribosome creates a polypeptide from the first occurring ATG to the first stop codon, there are many deviations from this model (131). For instance, the ribosome may exhibit leaky scanning and initiate translation from the second ATG, creating a new N terminus. Because ribosomes' initiation sites cannot be predicted reliably from transcript sequence, proteogenomics is invaluable in delineating the translated regions. Indeed, several recent proteogenomic studies, which we highlight below, have demonstrated unbiased detection of such translation-based variations.

6.4.1. Alternative translation initiation sites—Variations in the N terminal ends of proteins can arise from the ribosomes' differential usage of TISs. These can include initiation sites either upstream or downstream of the annotated start ATG or even use of a noncanonical initiation site such as CTG. Proteogenomic studies that characterize TISs typically utilize MS data sets that specifically enrich for N terminal peptides [e.g., combined fractional diagonal chromatography (COFRADIC)] (132). A recent study using MS data from various human cell lines combined COFRADIC data with predicted TISs from a publically available Ribo-Seq data set and showed that 20% of all N terminal identifications were from alternative TISs (53). This result strongly suggests that the understanding of annotated protein start sites is largely incomplete and more experimental data is needed to explore the full N terminal landscape.

6.4.2. Short open reading frames—Short open reading frames (sORFs) can be defined as those protein-coding sequences that are shorter than 100 nucleotides or approximately 33 amino acids (133). These sORFs are present at much higher frequencies than longer ORFs and are thus much harder to predict based on nucleotide sequences alone. Therefore, it came as a surprise when proteogenomic studies utilizing peptidomics data (i.e., the direct MS analysis of endogenous peptides) showed the widespread presence of translated sORFs (134–137). So far, peptides have been detected for (*a*) sORFs that lie upstream of known protein-coding regions, where they are thought to regulate the expression of the downstream protein; (*b*) sORFs that lie within RNAs annotated as noncoding (i.e., long noncoding RNAs); and (*c*) sORFs that lie in protein-coding transcripts but use an alternative frame of translation. The discovery of this new class of peptides has created a new line of investigation to elucidate their function.

6.4.3. Alternative open reading frames—It has been known for some time that bacterial genomes harbor overlapping ORFs, thus producing dually encoded proteins that are vastly different from each other in sequence. Only very recently has this phenomenon been shown to also exist in humans. A recent proteogenomic study focusing on the detection of peptides corresponding to alternative ORFs (altORFs) showed that there may be up to several hundred altORFs in different human cell types, especially for cells of the immune system (138, 139). As with the discovery of translated sORFs the functions of altORF products remain unknown. However, additional proteogenomic workflows that are designed to maximize the detection of peptides corresponding to altORFs can at least provide a global catalog of the tissue- and cell-specific products of these unique protein variations.

6.4.4. Ribosomal recoding events—During active translation of the mRNA message, the ribosome can deviate from the rules of the genetic code. Several dozen human proteins contain a stop codon to selenocysteine translational recoding event, where guided by special sequence signals in the 3′ UTR, the ribosome incorporates a selenocysteine at a UGA stop codon instead of terminating translation (4). Selenocysteine-containing peptides expressed in human cell lines have been detected by MS (140). Other translational recoding events include ribosomal frame shifting, which has been shown to occur for at least one human protein (141), and ribosomal read-through, where the ribosome continues translation after

the stop codon. These types of recoding events give rise to proteins that differ in sequence and that can be detected in a shotgun proteomics experiment.

6.5. Detecting Proteins that Contradict Annotations

One of the strengths of proteogenomic methods is that they produce an unbiased snapshot of expressed proteins, empirical evidence independent of existing protein annotations. In fact, previous studies using proteogenomic strategies have provided experimental support for the translation of unannotated genes, pseudogenes, and annotated noncoding RNAs. The translation of novel genes has been confirmed in large-scale proteogenomic studies (5, 86), a surprise given the assumed completeness of human genome annotations. Another study showed that certain annotated pseudogenes that are supposed to be dormant or nonfunctional actually produce protein product, contributing to the idea that what may have once been a pseudogene in the evolutionary past could be refunctionalized at a later time (142). And lastly, peptides have been detected that correspond to sORFs encoded in lncRNAs, creating new questions about their putative function (143–145). Common to all these studies are the use of protein databases derived from nucleotide sequence without discriminating between annotated and unannotated protein-coding regions. The only way to discover proteins without ascertainment bias or limitation is through such a proteogenomic technique that searches mass spectra against the entire genome or transcriptome.

7. Proteogenomic Applications

7.1. Genomic Effects on Protein Expression: Protein Quantitative Trait Loci, Allele-Specific Expression

Quantitative trait loci (QTL) mapping has been used to find associations between genetic variations and molecules upon which they have a regulatory effect. For example, in a *cis*-QTL, the presence of a certain SNP within a promoter of a gene that disrupts a transcription factor binding site can be found to downregulate expression of the linked transcript. However, in many cases, the protein is the ultimate driver of phenotype. In some cases, the RNA is uncorrelated to protein levels. Other times, the protein levels are subject to a buffering effect not experienced by the mRNAs (146). Driven by advances in quantitative MS–based proteomics, the concept of QTL mapping has recently been extended to protein (147). The idea is to test for associations of genomic variations lying near genes, within 5′/3′ UTRs, within introns, or within coding regions. Two large-scale protein quantitative trait loci (pQTL) studies looked for associations between genomic variants and the levels of protein expression in lymphoblast cell lines derived from a multiracial cohort. The pQTLs found tended to modulate splicing or lie in the UTRs, suggesting that genetic effects that exclusively influenced protein expression, but not transcript expression, may specifically modulate translation or the stability of proteins (148, 149).

In the special case where a gene is heterozygous for a nonsynonymous SNP, the gene produces two different proteins that differ by an amino acid. Proteogenomic databases that include such SAVs can enable identification of the variant peptides. Quantitative proteomics methods can be developed to measure the levels of the variant pair, thus estimating the effect of allele-specific variations on protein abundance. For example, SRM methods were

developed for pairs of SAV peptides corresponding to heterozygous alleles and used to quantify the relative allelic abundances of these genes across an Asian population (150). The feasibility of quantifying allelic expression was also demonstrated in a study that detected such variants from an RNA-Seq–derived SAV database (44).

7.2. Cancer Proteogenomics

Cancer is a disease characterized by a progression of somatic genomic mutations. Generally speaking, it is the gradual or punctuated accumulation of mutations such as SNPs, insertions, deletions, and myriad gross chromosomal rearrangements or CNVs that rewire normal cellular networks to be tumorigenic. The subfield of onco-proteogenomics is the application of proteogenomics toward characterizing cancer proteomes and has been reviewed recently (151). The abundance of genomic and transcriptomic sequencing data and specialized cancer databases (152, 153) that have created catalogs of cancer-specific mutations has provided a wealth of information about the possible protein mutations expressed in cancer samples. For example, The Cancer Genome Atlas consortium is an ongoing effort to sequence the exomes and transcriptomes of hundreds of ovarian and breast cancer tumors; the Clinical Proteomic Tumor Analysis Consortium is another effort to collect proteomics data from the same samples, which provides information about how the genomic alterations affect the proteome (154). By using proteogenomics to create sample-specific or cancer-specific databases for MS search, the corresponding variations can be detected at the protein level. Proteomic detection of cancer-associated somatic mutations can help define the mutations expressed as protein, provide estimates of the stability of the mutant proteins, and facilitate identification of clinical biomarkers or actionable drug targets. Recent onco-proteogenomics studies have used custom databases to detect proteins arising from SNPs (116, 155–157), aberrant splicing (158, 159), and gene fusions (122, 123).

7.3. Biomarkers

Biomarkers are molecules measured, typically, in easily available patient samples, such as blood, plasma, saliva, and urine, and that give an indication of a patient's health or disease state or reflect the course of therapy response. Many diseases such as cancer are characterized by highly individualized mutations; though similar proteins or pathways may be affected, the precise set of mutations that affect an individual may vary. Thus, proteogenomics has played an increasingly significant role in the specification of biomarkers, potentially increasing the sensitivity and accuracy with which one can diagnose a patient.

In recent years, the use of patient-specific protein databases derived from nucleotide sequencing data to enable highly-targeted monitoring of mutant peptides has shown promise in the field of personalized medicine. Serving as one of the earliest proofs of concept, Wang and coworkers (160) demonstrated the quantification of immuno-enriched *KRAS* (Kirsten rat sarcoma) mutants from colorectal and pancreatic tumor samples using an SRM method that monitored the corresponding mutant peptides. Other studies that measure cancer-specific mutant peptides have also been published (156, 161). Mathivanan and coworkers (162) later demonstrated the detection of mutant peptides excreted from colon cells in culture. Proteogenomics has shown the most potential for the detection of peptides highly

specific to patients. For example, Barnidge and coworkers (163) generated patient-specific protein databases to detect and monitor monoclonal immunoglobulins (i.e., M-proteins) secreted by plasma cells in multiple myeloma, and as M-proteins differ from patient to patient, the proteogenomic approach enabled the specific monitoring of patient response to treatment. Using a similar concept, Dasari and coworkers (164) monitored clonotypic light chain peptide sequences in patients with amyloidosis, and as these sequences are distinct for each patient, there were no other methods besides proteogenomics that allow for such detection. In this new era of personalized medicine, we predict that proteogenomic strategies will experience new prominence in translational and drug research.

7.4. Antibody Characterization

The use of proteogenomics for improved characterization of antibodies has been extensively reviewed elsewhere (165); however, we briefly mention that, driven by the pharmaceutical industry and developments in NGS, B cell receptor sequencing (BCR-Seq) has been used increasingly to create antibody-specific or template databases for detection of immunoglobulin peptides. Some of the earliest proteogenomic methods for sequencing antibodies have used a so-called template-based approach, where the antibody protein database employed for MS searching is not a perfect match with the sample of analysis, but through the use of partial spectral matches and extracted peptide sequence tags to fill in the missing sequence, the antibody protein sequence can be inferred. Now, with the development of BCR-Seq, several studies (described above in Section 7.3) have employed highly customized antibody databases containing all candidate antibody sequences.

8. Future of Proteogenomics

Much has been learned about human proteomic variations from the integration of NGS and MS-based proteomics, but as these complementary technologies evolve, there is even more potential for unbiased proteomic discovery. Current proteogenomic strategies have relied heavily on data sets derived from deep, shotgun-like sampling of the transcriptome and proteome. RNA-Seq is a sampling of transcript fragments, and MS-based proteomics is a sampling of enzymatically digested peptides. Though both techniques can deeply sample fragments, a major drawback is the inability to know with certainty the sequence of the intact transcript or protein from which these fragments were derived (166). However, both techniques are experiencing nascent, but tangible, improvements in the ability to sequence or detect intact transcripts and proteins, the de facto biological unit. Improvements in third-generation DNA sequencing from platforms such as PacBio have enabled the sequencing of full-length transcripts (167). Similarly, improvements in top-down proteomics, which includes advances in sample fractionation as well as MS instrumentation, have enabled more widespread detection of intact proteins (168). In fact, the ability to measure intact proteins has uncovered myriad protein forms resulting from both posttranscriptional processing (e.g., splicing) and posttranslational modifications (e.g., phosphorylations, proteolytic cleavage), leading to the coining of the word proteoform to describe each molecularly distinct protein arising from the unique combination of such variations (169). The twin improvements in third-generation sequencing and top-down proteomics hint at the future of proteogenomics, where transcriptomes and proteomes could be interrogated at both high sensitivity and

resolution. For instance, with third-generation sequencing data on a human transcriptome, an entire sample-specific, full-length protein database could be predicted, imparting greater clarity about which precise protein isoforms may be expressed in a sample. Undoubtedly, the synergistic relationship between nucleotide sequencing and proteomics will continue to evolve and will be key for the complete characterization of the human proteome in the coming decades.

Acknowledgments

This work was supported by National Institutes of Health grants 1R01GM114292 and 1R01CA193481 and National Science Foundation grant DBI-1458524. A.J.C. was supported by National Library of Medicine Training Grant 5T15LM007359, and G.M.S. was supported by National Institutes of Health Training Grant T32CA009361.

Literature Cited

1. Yates JR. The revolution and evolution of shotgun proteomics for large-scale proteome analysis. J Am Chem Soc. 2013; 135:1629–40. [PubMed: 23294060]
2. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem. 2013; 6:287–303.
3. Brent MR. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet. 2008; 9:62–73. [PubMed: 18087260]
4. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–74. [PubMed: 22955987]
5. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. A draft map of the human proteome. Nature. 2014; 509:575–81. [PubMed: 24870542]
6. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014; 509:582–87. [PubMed: 24870543]
7. Edman P. Method for determination of the amino acid sequence in peptides. Acta Chem Scand. 1950; 4:283–93.
8. Biemann K. Mass-spectrometry of peptides and proteins. Annu Rev Biochem. 1992; 61:977–1010. [PubMed: 1497328]
9. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass-spectrometry of large biomolecules. Science. 1989; 246:64–71. [PubMed: 2675315]
10. Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption ionization mass-spectrometry of biopolymers. Anal Chem. 1991; 63:A1193–202.
11. Mann M, Wilm M. Error tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem. 1994; 66:4390–99. [PubMed: 7847635]
12. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J Am Soc Mass Spectrom. 1994; 5:976–89. [PubMed: 24226387]
13. Boguski MS, Lowe TMJ, Tolstoshev CM. dbEST—database for “expressed sequence tags”. Nat Genet. 1993; 4:332–33. [PubMed: 8401577]
14. Yates JR, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem. 1995; 67:3202–10. [PubMed: 8686885]
15. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C. Identifying proteins from 2-dimensional gels by molecular mass searching of peptide-fragments in protein-sequence databases. PNAS. 1993; 90:5011–15. [PubMed: 8506346]
16. James P, Quadroni M, Carafoli E, Gonnet G. Protein identification in DNA databases by peptide mass fingerprinting. Protein Sci. 1994; 3:1347–50. [PubMed: 7987229]
17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. PNAS. 1977; 74:5463–67. [PubMed: 271968]

18. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, et al. Fluorescence detection in automated DNA-sequence analysis. Nature. 1986; 321:674–79. [PubMed: 3713851]
19. Pandey A, Mann M. Proteomics to study genes and genomes. Nature. 2000; 405:837–46. [PubMed: 10866210]
20. Neubauer G, King A, Rappsilber J, Calvio C, Watson M, et al. Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. Nat Genet. 1998; 20:46–50. [PubMed: 9731529]
21. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS. Interrogating the human genome using uninterpreted mass spectrometry data. Proteomics. 2001; 1:651–67. [PubMed: 11678035]
22. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, et al. A high-quality catalog of the *Drosophila melanogaster* proteome. Nat Biotechnol. 2007; 25:576–83. [PubMed: 17450130]
23. Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, et al. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science. 2008; 320:938–41. [PubMed: 18436743]
24. Merrihew GE, Davis C, Ewing B, Williams G, Kall L, et al. Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. Genome Res. 2008; 18:1660–69. [PubMed: 18653799]
25. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]
26. Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]
27. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–76. [PubMed: 18978772]
28. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321:956–60. [PubMed: 18599741]
29. Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res. 2012; 11:1009–17. [PubMed: 22103967]
30. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, et al. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–12. [PubMed: 25348405]
31. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014; 42:D756–63. [PubMed: 24259432]
32. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–76. [PubMed: 19684571]
33. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009; 324:218–23. [PubMed: 19213877]
34. Edwards NJ. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol. 2007; 3:102. [PubMed: 17437027]
35. Tanner S, Shen ZX, Ng J, Florea L, Guigo R, et al. Improving gene annotation using peptide mass spectrometry. Genome Res. 2007; 17:231–39. [PubMed: 17189379]
36. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics. 2004; 4:59–77. [PubMed: 14730672]
37. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, et al. The complete genome and proteome of *Mycoplasma mobile*. Genome Res. 2004; 14:1447–61. [PubMed: 15289470]
38. Kuster B, Mortensen P, Andersen JS, Mann M. Mass spectrometry allows direct identification of proteins in large genomes. Proteomics. 2001; 1:641–50. [PubMed: 11678034]
39. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, et al. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol. 2006; 7:R35. [PubMed: 16646984]

40. Khatun J, Yu YB, Wrobel JA, Risk BA, Gunawardena HP, et al. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. BMC Genom. 2013; 14:141.
41. Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genom. 2014; 15:9.
42. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. Genomewide characterization of non-polyadenylated RNAs. Genome Biol. 2011; 12:14.
43. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data—the protein inference problem. Mol Cell Proteom. 2005; 4:1419–40.
44. Sheynkman GM, Shortreed MR, Frey BL, Scalf M, Smith LM. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. J Proteome Res. 2014; 13:228–40. [PubMed: 24175627]
45. Woo S, Cha SW, Merrihew G, He Y, Castellana N, et al. Proteogenomic database construction driven from large scale RNA-Seq data. J Proteome Res. 2013; 13:21–28. [PubMed: 23802565]
46. Sheynkman GM, Shortreed MR, Frey BL, Smith LM. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Mol Cell Proteom. 2013; 12:2341–53.
47. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, et al. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. Genome Res. 2012; 22:1231–42. [PubMed: 22588898]
48. Kim H, Park H, Paek E. NextSearch: a search engine for mass spectrometry data against a compact nucleotide exon graph. J Proteome Res. 2015; 14:2784–91. [PubMed: 26004133]
49. Woo S, Cha SW, Na S, Guest C, Liu T, et al. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics. 2014; 14:2719–30. [PubMed: 25263569]
50. Zickmann F, Renard BY. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. Bioinformatics. 2015; 31:106–15.
51. Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods. 2012; 9:1207–11. [PubMed: 23142869]
52. Koch A, Gawron D, Steyaert S, Ndah E, Crappe J, et al. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. Proteomics. 2014; 14:2688–98. [PubMed: 25156699]
53. VanDamme P, Gawron D, Van Criekinge W, Menschaert G. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. Mol Cell Proteom. 2014; 13:1245–61.
54. Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics. 2002; 2:1426–34. [PubMed: 12422359]
55. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004; 20:1466–67. [PubMed: 14976030]
56. Lin BY, Mo F, Hong X, Gao F, Du L, et al. A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. BMC Bioinform. 2008; 9:537.
57. Power KA, McRedmond JP, de Stefani A, Gallagher WM, Gaora PO. High-throughput proteomics detection of novel splice isoforms in human platelets. PLOS ONE. 2009; 4:e5001. [PubMed: 19308253]
58. Zhang F, Drabier R. SASD: the Synthetic Alternative Splicing Database for identifying novel isoform from proteomics. BMC Bioinform. 2013; 14:S13.
59. Roos FF, Jacob R, Grossmann J, Fischer B, Buhmann JM, et al. PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. Bioinformatics. 2007; 23:3016–23. [PubMed: 17768164]
60. Zhou A, Zhang F, Chen JY. PEPPI: a peptidomic database of human protein isoforms for proteomics experiments. BMC Bioinform. 2010; 11(Suppl. 6):S7.

61. Schandorff S, Olsen JV, Bunkenborg J, Blagoev B, Zhang Y, et al. A mass spectrometry-friendly database for cSNP identification. Nat Methods. 2007; 4:465–66. [PubMed: 17538625]
62. Li J, Su ZL, Ma ZQ, Slebos RJC, Halvey P, et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics. Mol Cell Proteom. 2011; 10 M110.006536.
63. Alves G, Ogurtsov AY, Yu YK. RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration. BMC Genom. 2008; 9:505.
64. Xi H, Park JS, Ding GH, Lee YH, Li YX. SysPIMP: the web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry. Nucleic Acids Res. 2009; 37:D913–20. [PubMed: 19036792]
65. Yip YL, Famiglietti M, Gos A, Duek PD, David FP, et al. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat. 2008; 29:361–66. [PubMed: 18175334]
66. Li J, Duncan DT, Zhang B. CanProVar: a human cancer proteome variation database. Hum Mutat. 2010; 31:219–28. [PubMed: 20052754]
67. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011; 39:D945–50. [PubMed: 20952405]
68. Menon, R.; Omenn, GS. Identification of alternatively spliced transcripts using a proteomic informatics approach. In: Hamacher, M.; Eisenacher, M.; Stephan, C., editors. Data Mining in Proteomics: From Standards to Applications. New York: Humana; 2011. p. 319-26.
69. Kroll JE, Galante PAF, Ohara DT, Navarro FCP, Ohno-Machado L, de Souza SJ. A new portal for the analysis of human splicing variants. RNA Biol. 2012; 9:1339–43. [PubMed: 23064119]
70. Tavares R, de Miranda Scherer N, Pauletti BA, Araujo E, Folador EL, et al. SpliceProt: a protein sequence repository of predicted human splice variants. Proteomics. 2014; 14:181–85. [PubMed: 24273012]
71. Frenkel-Morgenstern M, Gorohovski A, Vucenovic D, Maestre L, Valencia A. ChiTaRS 2.1—an improved database of the chimeric transcripts and RNA-Seq data with novel sense-antisense chimeric RNA transcripts. Nucleic Acids Res. 2015; 43:D68–75. [PubMed: 25414346]
72. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000; 28:352–55. [PubMed: 10592272]
73. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, et al. dbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013; 41:D936–41. [PubMed: 23193291]
74. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJL, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res. 2010; 9:1716–26. [PubMed: 20131910]
75. Su ZD, Sheng QH, Li QR, Chi H, Jiang X, et al. De novo identification and quantification of single amino-acid variants in human brain. J Mol Cell Biol. 2014; 6:421–33. [PubMed: 25007923]
76. Castellana NE, Pham V, Arnott D, Lill JR, Bafna V. Template proteogenomics: sequencing whole proteins using an imperfect database. Mol Cell Proteom. 2010; 9:1260–70.
77. Castellana NE, McCutcheon K, Pham VC, Harden K, Nguyen A, et al. Resurrection of a clinical antibody: template proteogenomic de novo proteomic sequencing and reverse engineering of an anti-lymphotoxin-α antibody. Proteomics. 2011; 11:395–405. [PubMed: 21268269]
78. Castellana N, Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteom. 2010; 73:2124–35.
79. Krug K, Carpy A, Behrends G, Matic K, Soares NC, Macek B. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. Mol Cell Proteom. 2013; 12:3420–30.
80. Blakeley P, Overton IM, Hubbard SJ. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. J Proteome Res. 2012; 11:5221–34. [PubMed: 23025403]
81. Bitton DA, Smith DL, Connolly Y, Scutt PJ, Miller CJ. An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. PLOS ONE. 2010; 5:e8949. [PubMed: 20126623]

82. Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics. 2013; 13:1352–57. [PubMed: 23412978]
83. Ning K, Fermin D, Nesvizhskii AI. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. Proteomics. 2010; 10:2712–18. [PubMed: 20455209]
84. Helmy M, Sugiyama N, Tomita M, Ishihama Y. Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. Genes Cells. 2012; 17:633–44. [PubMed: 22686349]
85. Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, et al. Whole genome searching with shotgun proteomic data: applications for genome annotation. J Proteome Res. 2008; 7:80–88. [PubMed: 18062665]
86. Branca RM, Orre LM, Johansson HJ, Granholm V, Huss M, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods. 2014; 11:59–62. [PubMed: 24240322]
87. Kim S, Gupta N, Bandeira N, Pevzner PA. Spectral dictionaries integrating de novo peptide sequencing with database search of tandem mass spectra. Mol Cell Proteom. 2009; 8:53–69.
88. Ferro M, Tardift M, Reguer E, Cahuzac R, Bndey C, et al. PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. J Proteome Res. 2008; 7:1873–83. [PubMed: 18348511]
89. Tanner S, Shu HJ, Frank A, Wang LC, Zandi E, et al. InsPecT: identification of posttransitionally modified peptides from tandem mass spectra. Anal Chem. 2005; 77:4626–39. [PubMed: 16013882]
90. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data—toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteom. 2006; 5:652–70.
91. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, et al. Clustering millions of tandem mass spectra. J Proteome Res. 2008; 7:113–22. [PubMed: 18067247]
92. Chernobrovkin AL, Kopylav AT, Zgoda VG, Moysa AA, Pyatnitskiy MA, et al. Methionine to isothreonine conversion as a source of false discovery identifications of genetically encoded variants in proteogenomics. J Proteom. 2015; 120:169–78.
93. Wuehr M, Freeman RM, Presler M, Horb ME, Peshkin L, et al. Deep proteomics of the *Xenopus laevis* egg using an mRNA-derived reference database. Curr Biol. 2014; 24:1467–75. [PubMed: 24954049]
94. Blakeley P, Siepen JA, Lawless C, Hubbard SJ. Investigating protein isoforms via proteomics: a feasibility study. Proteomics. 2010; 10:1127–40. [PubMed: 20077415]
95. Guo X, Trudgian DC, Lemoff A, Yadavalli S, Mirzaei H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. Mol Cell Proteom. 2014; 13:1573–84.
96. Frank AM, Monroe ME, Shah AR, Carver JJ, Bandeira N, et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. Nat Methods. 2011; 8:587–591. [PubMed: 21572408]
97. Risk BA, Spitzer WJ, Giddings MC. Peppy: proteogenomic search software. J Proteome Res. 2013; 12:3019–25. [PubMed: 23614390]
98. Wen B, Xu SH, Sheynkman GM, Feng Q, Lin L, et al. sapFinder: an R/Bioconductor package for detection of variant peptides in shotgun proteomics experiments. Bioinformatics. 2014; 30:3136–38. [PubMed: 25053745]
99. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics. 2013; 29:3235–37. [PubMed: 24058055]
100. Ghali F, Krishna R, Perkins S, Collins A, Xia D, et al. ProteoAnnotator—open source proteogenomics annotation software supporting PSI standards. Proteomics. 2014; 14:2731–41. [PubMed: 25297486]
101. Jagtap PD, Johnson JE, Onsongo G, Sadler FW, Murray K, et al. Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. J Proteome Res. 2014; 13:5898–908. [PubMed: 25301683]

102. Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, et al. Multi-omic data analysis using Galaxy. Nat Biotechnol. 2015; 33:137–39. [PubMed: 25658277]
103. Pang CNI, Tay AP, Aya C, Twine NA, Harkness L, et al. Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. J Proteome Res. 2014; 13:84–98. [PubMed: 24152167]
104. Nagaraj SH, Waddell N, Madugundu AK, Wood S, Jones A, et al. PGTools: a software suite for proteogenomic data analysis and visualization. J Proteome Res. 2015; 14:2255–66. [PubMed: 25760677]
105. Zhu YF, Hultin-Rosenberg L, Forshed J, Branca RM, Orre LM, Lehtiö J. SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. Mol Cell Proteom. 2014; 13:1552–62.
106. Sanders WS, Wang N, Bridges SM, Malone BM, Dandass YS, et al. The proteogenomic mapping tool. BMC Bioinform. 2011; 12:7.
107. Kuhring M, Renard BY. iPiG: integrating peptide spectrum matches into genome browser visualizations. PLOS ONE. 2012; 7:e50246. [PubMed: 23226516]
108. Ingram VM. Gene mutations in human haemoglobin: chemical difference between normal and sickle cell haemoglobin. Nature. 1957; 180:326–28. [PubMed: 13464827]
109. Cavallo A, Martin ACR. Mapping SNPs to protein sequence and structure data. Bioinformatics. 2005; 21:1443–50. [PubMed: 15613399]
110. Karchin R. Next generation tools for the annotation of human SNPs. Brief Bioinform. 2009; 10:35–52. [PubMed: 19181721]
111. Gatlin CL, Eng JK, Cross ST, Detter JC, Yates JR. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. Anal Chem. 2000; 72:757–63. [PubMed: 10701260]
112. Bunger MK, Cargile BJ, Sevinsky JR, Deyanova E, Yates NA, et al. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. J Proteome Res. 2007; 6:2331–40. [PubMed: 17488105]
113. Chen M, Yang B, Ying WT, He FC, Qian XH. Annotation of non-synonymous single polymorphisms in human liver proteome by mass spectrometry. Protein Pept Lett. 2010; 17:277–86. [PubMed: 19508201]
114. Song C, Wang F, Cheng K, Wei X, Bian Y, et al. Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. J Proteome Res. 2013; 13:241–48. [PubMed: 24237036]
115. Krug K, Popic S, Carpy A, Taumer C, Macek B. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. Proteomics. 2014; 14:2699–708. [PubMed: 25251379]
116. Zhang B, Wang J, Wang X, Zhu J, Liu Q, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014; 513:382–87. [PubMed: 25043054]
117. Yan H, Yuan WS, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. Science. 2002; 297:1143. [PubMed: 12183620]
118. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012; 335:823–28. [PubMed: 22344438]
119. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015; 348:666–69. [PubMed: 25954003]
120. Geiger T, Cox J, Mann M. Proteomic changes resulting from gene copy number variations in cancer cells. PLOS Genet. 2010; 6:e1001090. [PubMed: 20824076]
121. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer. 2015; 15:371–81. [PubMed: 25998716]
122. Sun H, Xing X, Li J, Zhou F, Chen Y, et al. Identification of gene fusions from human lung cancer mass spectrometry data. BMC Genom. 2013; 14:S5.
123. Conlon KP, Basrur V, Rolland D, Wolfe T, Nesvizhskii AI, et al. Fusion peptides from oncogenic chimeric proteins as putative specific biomarkers of cancer. Mol Cell Proteom. 2013; 12:2714–23.

124. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011; 12:15.
125. Casado-Vela J, Lacal JC, Elortza F. Protein chimerism: novel source of protein diversity in humans adds complexity to bottom-up proteomics. Proteomics. 2013; 13:5–11. [PubMed: 23161619]
126. Li MY, Wang IX, Li Y, Bruzel A, Richards AL, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011; 333:53–58. [PubMed: 21596952]
127. Kroll JE, de Souza SJ, de Souza GA. Identification of rare alternative splicing events in MS/MS data reveals a significant fraction of alternative translation initiation sites. PeerJ. 2014; 2:e673. [PubMed: 25405079]
128. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, et al. The human transcriptome across tissues and individuals. Science. 2015; 348:660–65. [PubMed: 25954002]
129. Ning K, Nesvizhskii AI. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinform. 2010; 11(Suppl. 11):S14.
130. Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, et al. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 2013; 5:1469–78. [PubMed: 24290761]
131. Gawron D, Gevaert K, Van Damme P. The proteome under translational control. Proteomics. 2014; 14:2647–59. [PubMed: 25263132]
132. Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, et al. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol. 2003; 21:566–69. [PubMed: 12665801]
133. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, et al. The abundance of short proteins in the mammalian proteome. PLOS Genet. 2006; 2:515–28.
134. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 2014; 15:193–204. [PubMed: 24514441]
135. Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, et al. Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res. 2004; 14:2048–52. [PubMed: 15489325]
136. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol. 2013; 9:59–64. [PubMed: 23160002]
137. Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J Proteome Res. 2014; 13:1757–65. [PubMed: 24490786]
138. Vanderperre B, Lucier JF, Bissonnette C, Motard J, Tremblay G, et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLOS ONE. 2013; 8:e70698. [PubMed: 23950983]
139. Vanderperre B, Lucier JF, Roucou X. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. Database. 2012; 2012:bas025. [PubMed: 22613085]
140. Bianga J, Touat-Hamici Z, Bierla K, Mounicou S, Szpunar J, et al. Speciation analysis for trace levels of selenoproteins in cultured human cells. J Proteom. 2014; 108:316–24.
141. Belew AT, Meskauskas A, Musalgaonkar S, Advani VM, Sulima SO, et al. Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. Nature. 2014; 512:265–69. [PubMed: 25043019]
142. Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and “resurrected” pseudogenes in the mouse genome. Genome Res. 2011; 21:756–67. [PubMed: 21460061]
143. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. eLife. 2014; 3:e03523. [PubMed: 25233276]
144. Sun H, Chen C, Shi M, Wang D, Liu M, et al. Integration of mass spectrometry and RNA-Seq data to confirm human ab initio predicted genes and lncRNAs. Proteomics. 2014; 14:2760–68. [PubMed: 25339270]

145. Prabakaran S, Hemberg M, Chauhan R, Winter D, Tweedie-Cullen RY, et al. Quantitative profiling of peptides from RNAs classified as noncoding. Nat Commun. 2014; 5:10.
146. Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science. 2013; 342:1100–4. [PubMed: 24136357]
147. Horvatovich P, Franke L, Bischoff R. Proteomic studies related to genetic determinants of variability in protein concentrations. J Proteome Res. 2014; 13:5–14. [PubMed: 24237071]
148. Wu LF, Candille SI, Choi Y, Xie D, Jiang LH, et al. Variation and genetic control of protein abundance in humans. Nature. 2013; 499:79–82. [PubMed: 23676674]
149. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, et al. Impact of regulatory variation from RNA to protein. Science. 2015; 347:664–67. [PubMed: 25657249]
150. Wu JR, Zeng R. Molecular basis for population variation: from SNPs to SAPs. FEBS Lett. 2012; 586:2841–45. [PubMed: 22828278]
151. Alfaro JA, Sinha A, Kislinger T, Boutros PC. Onco-proteogenomics: Cancer proteomics joins forces with genomics. Nat Methods. 2014; 11:1107–13. [PubMed: 25357240]
152. Yang X, Lazar IM. XMAn: a *Homo sapiens* mutated-peptide database for the MS analysis of cancerous cell states. J Proteome Res. 2014; 13:5486–95. [PubMed: 25211293]
153. Huang PJ, Lee CC, Tan BCM, Yeh YM, Chu LJ, et al. CMPD: cancer mutant proteome database. Nucleic Acids Res. 2015; 43:D849–55. [PubMed: 25398898]
154. Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, et al. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. Cancer Discov. 2013; 3:1108–12. [PubMed: 24124232]
155. Halvey PJ, Wang XJ, Wang J, Bhat AA, Dhawan P, et al. Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. Cancer Res. 2014; 74:387–97. [PubMed: 24247723]
156. Nie S, Yin H, Tan Z, Anderson MA, Ruffin MT, et al. Quantitative analysis of single amino acid variant peptides associated with pancreatic cancer in serum by an isobaric labeling quantitative method. J Proteome Res. 2014; 13:6058–66. [PubMed: 25393578]
157. Karpova MA, Karpov DS, Ivanov MV, Pyatnitskiy MA, Chernobrovkin AL, et al. Exome-driven characterization of the cancer cell lines at the proteome level: the NCI-60 case study. J Proteome Res. 2014; 13:5551–60. [PubMed: 25333775]
158. Menon R, Zhang Q, Zhang Y, Fermin D, Bardeesy N, et al. Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. Cancer Res. 2009; 69:300–9. [PubMed: 19118015]
159. Menon R, Im H, Zhang E, Wu SL, Chen R, et al. Distinct splice variants and pathway enrichment in the cell-line models of aggressive human breast cancer subtypes. J Proteome Res. 2014; 13:212–27. [PubMed: 24111759]
160. Wang Q, Chaerkady R, Wu JA, Hwang HJ, Papadopoulos N, et al. Mutant proteins as cancer-specific biomarkers. PNAS. 2011; 108:2444–49. [PubMed: 21248225]
161. Olsen L, Campos B, Winther O, Sgroi DC, Karger BL, Brusic V. Tumor antigens as proteogenomic biomarkers in invasive ductal carcinomas. BMC Med Genom. 2014; 7(Suppl. 3):S2.
162. Mathivanan S, Ji H, Tauro BJ, Chen YS, Simpson RJ. Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. J Proteom. 2012; 76:141–49.
163. Barnidge DR, Tschumper RC, Theis JD, Snyder MR, Jelinek DF, et al. Monitoring M-proteins in patients with multiple myeloma using heavy-chain variable region clonotypic peptides and LC-MS/MS. J Proteome Res. 2014; 13:1905–10. [PubMed: 24552626]
164. Dasari S, Theis JD, Vrana JA, Meureta OM, Quint PS, et al. Proteomic detection of immunoglobulin light chain variable region peptides from amyloidosis patient biopsies. J Proteome Res. 2015; 14:1957–67. [PubMed: 25734799]
165. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. Curr Opin Chem Biol. 2015; 24:112–20. [PubMed: 25461729]

166. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Hubbard TJ, et al. Assessment of transcript reconstruction methods for RNA-Seq. Nat Methods. 2013; 10:1177–84. [PubMed: 24185837]

167. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013; 31:1009–14. [PubMed: 24108091]

168. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011; 480:254–58. [PubMed: 22037311]

169. Smith LM, Kelleher NL. Consort. Top Down Proteom. Proteoform: a single term describing protein complexity. Nat Methods. 2013; 10:186–87. [PubMed: 23443629]

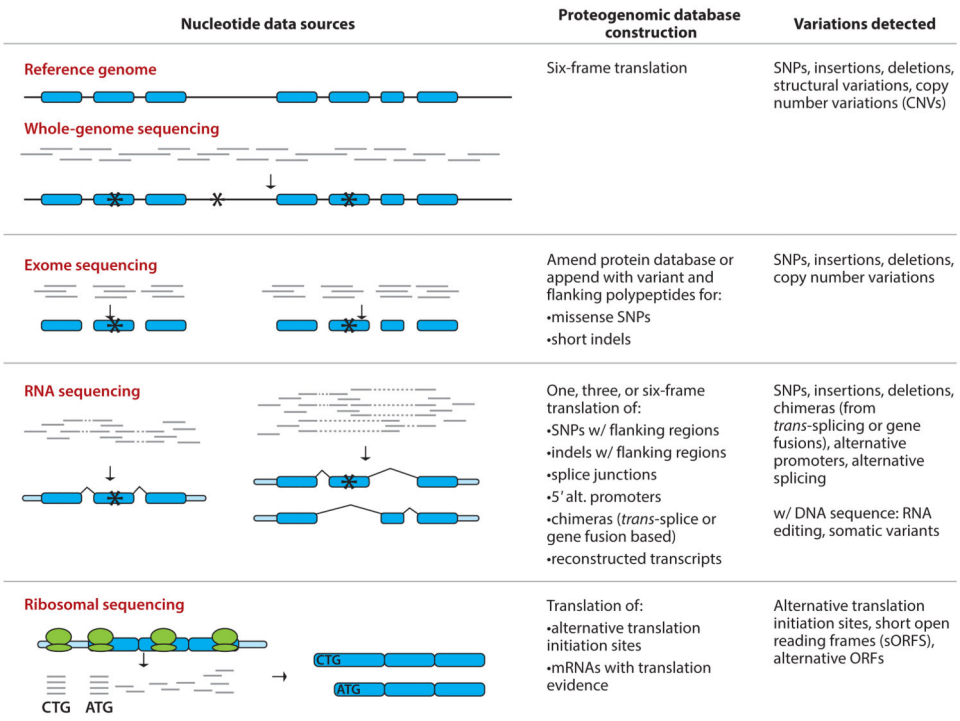

Figure 1.

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