Methods, Tools and Current Perspectives in Proteogenomics*

Kelly V. Ruggles, Karsten Krug, Xiaojing Wang, Karl R. Clauser, Jing Wang, Samuel H. Payne, David Fenyo, Bing Zhang, and D. R. Mani

With combined technological advancements in high-throughput next-generation sequencing and deep mass spectrometry-based proteomics, proteogenomics, i.e. the integrative analysis of proteomic and genomic data, has emerged as a new research field. Early efforts in the field were focused on using protein identification using sample-specific genomic and transcriptomic sequencing data. More recently, integrative analysis of quantitative measurements from genomic and proteomic studies have identified novel insights into gene expression regulation, cell signaling, and disease. Many methods and tools have been developed or adapted to enable an array of integrative proteogenomic approaches and in this article, we systematically classify published methods and tools into four major categories, (1) Sequence-centric proteogenomics; (2) Analysis of proteogenomic relationships; (3) Integrative modeling of proteogenomic data; and (4) Data sharing and visualization. We provide a comprehensive review of methods and available tools in each category and highlight their typical applications. Molecular & Cellular Proteomics 16: 10.1074/mcp.MR117.000024, 959–981, 2017.

The last decade has witnessed the rapid emergence of proteogenomics, a new research field at the interface of genomics and proteomics. The term proteogenomics came into use following a publication by George Church’s group in 2004 describing a proteogenomic mapping technique which harnessed proteomics data to improve genome annotation of Mycoplasma pneumonia (1). The reach of proteogenomics has since expanded with technological advancements enabling rapid and economical high-throughput DNA and RNA sequencing and deep mass spectrometry (MS)-based proteomics. These advancements have proved particularly useful for integrating nucleotide sequencing and MS data from the same sample, where genomic sequencing data can be used to improve protein identification through comprehensive protein sequence database construction. Proteomic data can then be used to demonstrate the validity and functional relevance of novel findings based on large scale RNA and DNA sequencing projects, including coding sequence variants and novel coding transcripts. In addition to sequence-centric proteogenomic data integration, combined quantitative analysis from genomic and proteomic studies have also been used to provide novel insights into multilevel gene expression regulation (2–13), signaling networks (14–17), disease subtypes (10, 12, 13), and clinical prediction (18–20). In this review, we subscribe to an expansive view of proteogenomics, encompassing all areas of proteomic and genomic integrative data analysis and cover the range of tools developed to tackle the associated challenges.

To complement already published review papers that focus on specific sub-domains of the broad proteogenomics research area (21–24), we systematically classified existing methods and tools for various types of integrative proteogenomic studies into four major sections. Sequence-centric Proteogenomics describes aspects of sequence-centric proteogenomics and the combined use of genomic and proteomic data to augment gene or protein annotation (Fig. 1). “Analysis of Proteogenomic Relationships” explores relationships between genomic and proteomic data using correlation, with application to deciphering the effect of mutations on signaling (Fig. 2). “Integrative Modeling of Proteogenomic Data” summarizes integrative modeling and analysis of proteogenomic data using statistical and machine learning approaches (Fig. 3). “Data Sharing and Visualization” discusses genome (Fig. 4) and network visualization (Fig. 5), along with challenges in data sharing. All four sections of the review assume tandem MS (MS/MS) as the core proteomics technology for generating peptide sequence data.
In this section we review several areas of sequence-centric proteogenomics. This includes the integrative analysis of genomic and proteomic data for exome annotation in the form of gene discovery and gene model refinement (Proteomics Aiding Genome Annotation); protein level detection of single...
amino acid variants (SAAVs), insertions, deletions, alternative splice junctions and novel gene fusions in relation to a reference genome sequence (Personalized Protein Sequence Databases); the application of proteomic sequencing to characterize antibodies (Sequencing of Antibodies); studying the effects of viral infections and transposons on gene expression in eukaryotic organisms (Viral Infections and Activations of Transposable Elements); and applications of proteogenomics to metaproteomic investigation (Metaproteogenomics).

A concept integral to all five topics in this section is the importance of an inclusive and high-quality protein sequence database for peptide identification. In a typical proteomic experiment, peptide MS/MS spectra are interpreted using a database search algorithm that matches and scores the similarity of each experimental spectrum against model spectra constructed from peptide sequences contained in a user supplied protein sequence database (25). This strategy is used, in part, because the fragmentation efficiency of current MS/MS instrumentation is unable to consistently yield spectra from which complete, unambiguous sequences can be interpreted de novo (the current state of automated de novo peptide interpretation have been reviewed elsewhere (26, 27)). To address this, researchers use a protein sequence database, ideally containing all protein sequences one expects to be present in the sample, with minimal irrelevant sequences to reduce false spectral matches and search time. Limiting the number of candidate sequences in the form of a sequence database enables the sequence ambiguity present in a spectrum to be overcome, resulting in high confidence peptide spectrum matches (PSMs) (28).

**Proteomics Aiding Genome Annotation**—The use of MS-based proteome characterization to aid genome annotation

---

1 The abbreviations used are: SAAV, single amino acid variant; DNA, deoxyribonucleic acid; RNA, ribonucleic acid; MS/MS, tandem mass spectrometry; EST, expressed sequence tag; ORF, open reading frame; TCGA, The Cancer Genome Atlas; CPTAC, Clinical Proteomic Tumor Analysis Consortium; FDR, false discovery rate; SNP, single nucleotide polymorphism; PTM, post-translational modification; PSM, peptide spectrum match; RPPA, reverse-phase protein array; GSEA, gene set enrichment analysis; ssGSEA, single-sample gene set enrichment analysis; SNV, single nucleotide variant; nsSNV, nonsynonymous single nucleotide variant; CNA, copy number aberration.

---

**Fig. 2.** Proteogenomic relationships. A, Correlation analysis of mRNA and protein pairs across samples enables the assessment of global correlation structure which typically centers between correlation coefficients of 0.3 and 0.5. B, Regulatory effects on RNA and protein expression levels caused by copy number aberrations (CNA), genetic variants (eQTL) and microRNAs (miRNAs) can be studied by different correlation-based approaches. CNA cis and trans effects on RNA, protein and PTM expression can be determined by correlating each gene copy number at a given locus to all quantified features in RNA, protein or PTM space across all samples. Expression quantitative trait loci (eQTL) analysis can be used to identify DNA sequence variants affecting RNA/protein expression levels in the sample population being studied. Global miRNA analysis accompanied with mRNA or protein profiling enables the assessment of miRNA mediated regulation of mRNA and protein expression. C, Integrative analysis of genetic variants and PTM sites like phosphorylation can identify functional consequences of genetic variants at the molecular level. Mutations that directly affect serine, threonine and tyrosine residues can result in destruction or genesis of phosphosites (I); mutations adjacent to phosphosites can result in removal or addition of phosphosites (II) or change the kinase that recognizes the phosphorylation site (III).
has been widely exploited in various organisms and has been previously reviewed by several groups (22, 29, 30). Here, we highlight some of the pioneering studies using integrative analysis of genomics and proteomics data for genome (re)annotation (See related review (31)).

Early studies integrating proteomic and genomic data date back to before the genome sequencing revolution of the early 21st century. Motivated by the lack of comprehensive and complete protein sequence databases and the emerging availability of nucleotide sequence data in the form of expressed sequence tags (ESTs), researchers interrogated peptide mass spectra using databases obtained by in-silico translation of unassembled ESTs. In 1995 Yates et al. (32) demonstrated the use of nucleotide sequences translated into all six reading frames of amino acid sequence (six-frame translation) to identify mass spectra of unmodified and phosphorylated peptides from human, bovine, E. coli, and S. cerevisiae proteins. The intrinsic ability of searching peptide mass spectra against genomic sequence databases to identify novel, unannotated genes, was employed shortly after by several groups (33–35). Choudhary et al. (36) used information from the recently released human genome draft in 2001 to query all 23 human chromosomes with tandem mass spectra and compare the results to EST database searches. From this, they concluded that MS/MS searching of genomic DNA databases were of limited utility, as the presence of introns in the database prohibited matching exon-spanning peptides and prevented identification in roughly one quarter of the

![Integrative Modeling Diagram]

Fig. 3. Integrative modeling. Overview of sub-topics in integrative modeling of proteogenomic data. A, Clustering techniques illustrating a schematic of multi-omic hierarchical clustering analysis resulting in the identification of two subtypes, B, Predictive modeling for disease diagnosis, prognosis, drug response and drug toxicity using multiple data modalities and, C, proteogenomic pathway and network modeling, including informing network composition and pathway and GO term enrichment.
spectra. Additionally, the consensus sequence chosen for the reference genome also prevented the identification of individual SAAVs with EST evidence. These limitations have since been addressed through the incorporation of sample-specific or species-specific alternative splicing and SNPs into the protein sequence database (see Personalized Protein Sequence Databases).

In 2004, Jaffe et al. introduced the concept of a proteogenomic map as a complementary method for genome annotation, which used evidence of protein expression to predict ORFs in *Mycoplasma pneumoniae* (1). Since then, a proteomics-based approach to gene annotation model refinement has been successfully applied in both model and nonmodel organisms (37–43). Proteogenomic gene annotation is most

---

**Fig. 4. Genome-based visualization, using proBAM as an example.** proBAM is a data format to integrate mass spectrometry data with the genome. In this example, we show the visualization of 10 colorectal cancer cell lines in proBAM format. The PSMs result from a search against a customized database built from matched RNA-Seq data, which are also incorporated into the visualization. A, Integrative Genomics Viewer (IGV) snapshot visualizes peptides and RNA-Seq reads mapped to KRAS in one window. The upper panel shows proteomic data from 10 colon cancer cell lines indicated by different colors. The bottom three panels illustrate RNA-Seq data from cell lines HCT15, Caco-2 and SW480, respectively. B, Zoomed-in view of an exon region in KRAS. Similar to RNA-Seq reads (three bottom panels), peptides mapped to the genome can be classified into within exon peptides and junction peptides in the proBAM file (upper panel). C, The upper panel shows a zoomed-in view of mutations confirmed by both RNA-Seq and proteomic data in KRAS. A G13D mutation in HCT15 and a G12V mutation in SW480 are observed in both transcriptomic (second and fourth panel) and proteomics (first panel) data, whereas wild type peptide is observed in Caco-2 (third panel).
often carried out by searching peptide mass spectra against a six-frame translation of an associated reference genome sequence database. Peptides identified by this search are then mapped to the existing gene annotation model (Fig. 1). The detection and identification of these peptides provides direct and valuable evidence of protein translation, and have

Fig. 5. NetGestalt-based analysis of colorectal cancer proteomics data. A, NetGestalt created a one-dimensional (1-D), linear order of all 12,112 genes in a protein-protein interaction network based on the hierarchical modular organization of the network. The ruler indicates coordinates of the genes in the resulting linear order. B, Each bar represents a module identified from the network and alternating bar colors (green and orange) are used to distinguish neighboring modules. C, Colorectal cancer proteomics data visualized as a heat map with each row representing a sample. Red and blue colors in the heat map represent relative over- and under-expression, respectively. All 12,112 genes in the network are visualized in the heat map, and missing values in the proteomic data are indicated in gray color. The samples (rows) of the data are ordered based on the five subtypes visualized beside the heat map. D, Signed minus log10 transformed p values of the difference between subtype 3 and other subtypes visualized by the bar plot. Missing values are indicated by yellow bars in the bar plot. E, Significantly over-expressed genes in subtype 3 (FDR < 0.05). F–H, Zoomed-in view of a region corresponding to one of the over-represented modules. I, Genes in (H) visualized in a node-link diagram. The edges in the diagram represent protein-protein interactions.
been used to train algorithms for gene model prediction (43).

A crucial prerequisite for genome refinement using MS-based proteomics is sufficient proteome coverage, which became feasible following major improvements in MS instrumentation (ion traps) as well as sample preparation protocols (multidimensional fractionation at the peptide level). In fact, Liquid Chromatography (LC)-MS/MS based proteomics utilizing sensitive and fast scanning ion-trap mass analyzers dominated the field of proteogenomics for several years despite the low resolution and mass accuracy of the acquired spectra (37, 38, 40, 44–47). However, the low mass accuracy data acquired by these instruments required large mass tolerances when searching six-frame translation databases, resulting in prohibitively large search spaces, long search times and high proportions of false positive peptide identifications using conventional target-decoy FDR thresholding (48).

The invention of a new generation of mass spectrometers revolutionized the field of proteomics, providing high-resolution and high-accuracy MS data at an expanded dynamic range (49, 50). High mass accuracy data has the intrinsic capability to reduce the database search space by allowing for small mass tolerances and an associated decrease in plausible candidate sequences, something of importance when searching large genomic six-frame translation databases (31). Further improvements in MS technology enabled the acquisition of high-resolution data at both the MS and MS/MS level without compromising sequencing speed (51–53).

Despite these technological improvements, proteomics still suffers from low sequence coverage even in simple prokaryotic genomes (42). A typical LC-MS/MS proteomics strategy employs the enzyme trypsin to digest proteins by specific cleavage after lysine (K) and arginine (R) amino acids. Any sequence overlap in the resulting pool of peptides will be an artifact of cleavage sites missed by trypsin. Regions of a protein with spacing of K, R residues that are <6 AA’s or >30 AA’s will tend not to be observed by the mass spectrometer, and peptides with extremes of hydrophilicity or hydrophobicity will not be readily bound or eluted from the LC column, thereby limiting sequence coverage. Recent CPTAC studies report median protein sequence coverage of about 25% across >12,000 proteins in a human cancer cohorts (12) using extensive sample fractionation. Sequence coverage can be increased by use of multiple proteases generating different peptide species at the expense of additional experimental costs. Nucleotide sequencing-based methods to measure gene expression, such as RNA-Seq and Ribo-Seq, typically achieve higher genome coverage and are routinely used to complement the annotation of newly sequenced genomes (54–57).

Personalized Protein Sequence Databases—Reference protein sequence databases, such as those from Ensembl or RefSeq, are typically used to identify mass spectra through peptide spectrum matching. Because these databases lack sample specific sequence variation, including single amino acid variants (SAAVs), insertions, deletions, alternative splice junctions and novel gene fusions, studies using this approach are unable to identify the corresponding variant peptides present in the MS/MS data. This is a particularly important limitation to consider in cancer studies, where patients acquire tumor specific somatic variation. Analyzing nonsynonymous somatic mutations at the proteome level has the potential to yield novel insights into tumor biology (58). To do this, genome and RNA sequencing have been used to generate personalized protein sequence databases by incorporating nonsynonymous variants into reference protein sequences. Several informatics pipelines have emerged in recent years for generating these databases (59–66). Fig. 1 illustrates the core processes these pipelines perform, and Table I provides a list of typical, currently available software.

Because the likelihood of experimentally observing a peptide decreases as one progresses from a reference database to each variant database (SAAV; novel splice junctions; and novel coding loci in putative intergenic regions), MS/MS data sets are sensibly searched in an iterative fashion through individual databases with separate FDR estimations for PSMs (12, 13, 62, 67). As the total size of the protein database increases, identifying high confidence PSMs requires increased spectral quality with increasingly complete peptide fragmentation. Detailed statistical considerations for iterative search strategy design and FDR estimation in a proteogenomic paradigm have recently been thoroughly described (22).

Sequencing of Antibodies—In addition to its role in novel peptide identification and gene annotation, sequence-centric proteogenomics has played a significant role in antibody sequencing (68–71). In vertebrates, antibodies provide the ability for an organism to differentiate between itself and the environment, and a mechanism to fight a diverse range of infections. In order to meet these needs, antibodies possess a tremendous level of sequence diversity, which is achieved through a combination of antibody sequence options and the introduction of mutations. Three types of antibody genes are encoded in the genome: Variable (V), Diversity (D), and Joining (J), and these are combined through a process called V(D)J recombination (72) to produce a diverse library of heavy and light chains that are combined pairwise, and together provide the antigen binding specificity. The affinity maturation process (73) further optimizes antibodies that recognize foreign objects by allowing for a high rate of point mutation introductions, followed by a selection for the strongest binders. At the same time, antibodies recognizing self are eliminated—a process that is defective in autoimmune diseases. Although antibody sequences are encoded in the genome, because of the process by which they are combined and matured, it is not possible to predict the final repertoire of antibody sequences for an individual from the genome alone. However, RNA-Seq...
| Name                  | URL                                                                 | Reference | Remarks                                                                                                                                 |
|-----------------------|----------------------------------------------------------------------|-----------|-----------------------------------------------------------------------------------------------------------------------------------------|
| customProDB           | https://www.bioconductor.org/packages/release/bioc/html/customProDB.html | (63)      | R-package to generate customized protein database from NGS data, including SNVs, INDELs and novel splice junctions                        |
| Galaxy Integrated Omics | https://bessantlab.org/software/gio/                                         | (60)      | Curated collection of Galaxy-based tools to generate sample-specific databases from RNA-seq data                                        |
| Galaxy-P              | https://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_mseq_sap_db_workflow/3a11830963e3 | (201)     | Collection of Galaxy-P workflows for customized database construction from RNA-seq data                                               |
| QUILTS                | http://quilts.fenyolab.org                                           | (62)      | Web-based tool to generate sample-specific databases with special focus on novel splice isoforms                                         |
| MutationDB            | http://proteomics.ucsd.edu/software-tools/cancer-proteogenomics-4/          | (65)      | Proteogenomic database construction from all types of mutational variants driven from large-scale RNA-seq data                          |
| SpliceDB              | http://proteomics.ucsd.edu/software-tools/cancer-proteogenomics-4/          | (65)      | Proteogenomic database construction with special focus on novel splice junction identifications driven from large-scale RNA-seq data |
| PGA                   | http://bioconductor.org/packages/PGA/                                 | (64)      | R-package to build customized protein databases based on RNA-Seq data with or without a reference genome guide                           |
| JUMPg                 | https://github.com/gatechatl/JUMPg                                   | (61)      | Perl-based pipeline to build search customized databases and to perform subsequent filtering and visualization of search results       |
| MSProGene             | http://sourceforge.net/projects/msprogene/                            | (66)      | Java program that constructs customized protein databases from RNA-seq data                                                            |
| PPLine                | https://sourceforge.net/projects/ppline/                              | (202)     | Python-based suite to process raw RNA-seq or Exome-seq data for customized database construction                                       |
can be applied to sequencing the variable region of the light and heavy chains to obtain a sampling of the antibody diversity of an individual.

MS has also been applied to sequencing both recombinant and circulating antibodies. For recombinant antibodies where the sequence is known, MS can be used to confirm the sequence and check for purity. For studying circulating antibodies, the most widely used approach is to use the antigen as bait to enrich for the circulating high affinity antibodies and analyze them with MS/MS. Multiple aliquots are digested with proteases of different specificity to generate comprehensive coverage with overlapping peptides; the spectra can then be interpreted using de novo sequencing approaches (74–77). This requires high quality data, and limits the number of antibodies that can be sequenced in a mixture. An alternative is to use a proteogenomics approach: performing targeted RNA-Seq of the variable region of the light and heavy chains, assembling the reads and translating the assemblies to create a protein sequence database for searching. This approach has been employed to identify high affinity circulating antibodies in infected individuals against HIV (68) and malaria (70) surface proteins that are potentially broadly neutralizing. The approach has also been applied to produce single chain llama antibodies to be used as reagents (71). Llamas and camels produce single chain antibodies in addition to paired heavy-light chain antibodies. The advantage of the single chain antibodies as reagents is that they are small (~15kDa), robust, and once sequenced, they can easily be expressed in E. coli to provide a reproducible resource. These single chain antibodies can be humanized (71) and conjugated with drugs (79); they are, therefore, highly promising candidates for developing therapeutics.

**Viral Infections and Activation of Transposable Elements**—Beyond expression of host genes, viral infections and transposons contribute potential protein-level expression in eukaryotic organisms that can be studied using proteogenomic techniques. Viral infections alter gene expression and protein production as the virus highjacks cellular processes that allow for self-replication (80–83), and can lead to cell death or cancerous growth of the host cells (84). Eukaryotic genomes also contain mobile elements or transposons that are remnants of ancient viral infections that were incorporated in the host germline genome and then spread throughout the genome through gene copying. It is estimated that about half of the human genome is transposon sequence albeit most are no longer active (85). There is, however, a small subset of the LINE-1 (Long Interspersed Elements) retrotransposons that are capable of autonomous retrotransposition through a copy and paste mechanism using an RNA intermediary, but they are most commonly inactive in somatic cells. However, increased retrotransposition activity has been observed in some disease states including many cancer types (86). The role of retrotransposition in cancer biology is currently unclear and it is not known if it promotes or suppresses tumors, or if it is just an effect of genome instability (87). Also, the mechanism for suppression and activation of retrotransposition is unknown (88) but could provide important insights into cancer biology. Human LINE-1 has two open reading frames: ORF1 and ORF2. ORF1 is an RNA binding protein and ORF2 is an endonuclease and a reverse transcriptase. ORF1 can be quantified using MS-based proteomics, and has been observed by deep MS-based proteomics in many tumors including breast, prostate and ovarian tumors (89). Interesting proteogenomics questions for future research include: Do higher levels of LINE-1 transcripts and ORF1 protein concentrations correlate with tumor progression? Which human transcripts and protein levels correlate with ORF1 protein concentration? Does ORF1 protein concentration correlate with a higher number of somatic LINE-1 insertions?

**Metaproteogenomics**—Metaproteomics (90, 91), or proteomic investigation of multiorganism communities, represents a unique application for proteomic and genomic data integration. Metagenomic studies typically collect genome data only, representing the functional potential of organisms present in a community, whereas metaproteogenomics adds the additional layer of proteomics data to elucidate what cellular functions are being expressed and utilized by community members. Unfortunately, reference genome sequence databases lack many protein sequences in natural consortia samples, as many organisms within the biological sample may not have a sequenced genome. Despite the meteoric rise in the number of sequenced genomes, there remain large swaths of bacterial diversity that are unrepresented in databases like GenBank and UniProt. A recent paper by the Banfield group (92) highlights numerous phyla which were previously unknown (not just unsequenced). Indeed, they observe a new phyla radiation with dozens of bacterial phyla entirely separate from known taxonomy. Recent predictions for bacterial diversity on the planet approach 1 trillion distinct species (93). Thus, we expect that reference genome sequence databases will be incomplete at the species level for the foreseeable future.

Therefore, a driving need for utilizing genomic data in metaproteomics experiments is the inaccuracy of annotating genomes for novel species. Proteomics has frequently been used not only to improve the set of known proteins in a single organism, but as training and/or testing data for improving gene calling algorithms (94). Moreover, protein annotation is less accurate for genes which have not been previously characterized, a common observation for samples in natural (not laboratory) conditions.

Absent algorithmic advances which allow spectra to be identified without an exact sequence match to a database (95–97), the path forward for metaproteomics involves obtaining sequencing data for the biological sample at hand. This can be either metagenomic or metatranscriptomic data. Each provides crucial information about the potential protein sequences that should be considered when searching tandem
mass spectra. Although this introduces additional costs to the project, it has so far been the best way to improve the number of identified peptides in a metaproteomic analysis (98).

ANALYSIS OF PROTEOGENOMIC RELATIONSHIPS

In this section we cover the integration of profiling data from different omics platforms to help elucidate information flow from DNA to RNA to proteins and, most importantly, to phenotype.

This includes studies aimed at understanding whether protein abundance can be reliably predicted from mRNA measurements (mRNA-Protein Correlation); assessing the genetic control of mRNA on protein abundance (Genetic Control of mRNA and Protein Abundance); and the impact of genetic aberrations on post-translational protein modification (PTM) and signaling (Relating Mutations to PTM and Signaling).

**mRNA-Protein Correlation**—Correlation between mRNA and protein profiling data has been a topic of considerable research during the past decade, and an excellent review on this has recently been published (2). Early studies focused on the correlation between steady state mRNA and protein abundance for all genes in a single sample, and it was noted in various organisms that relative abundance of proteins in a sample cannot be adequately explained by the corresponding mRNA abundance (3). This can be explained, at least in part, by our understanding that protein abundance is determined by a combination of mRNA abundance, translational regulation, and protein degradation (3). With the availability of paired mRNA and protein data for large sample cohorts, studies on gene-wise correlations between mRNA and protein abundance across many samples also reported modest correlations (5, 8, 99). More recently, the CPTAC consortium has explored mRNA-protein correlations in breast, colorectal, and ovarian cancer samples (10, 12, 13) finding predominantly positive mRNA-protein correlations for all genes with moderate (0.3–0.45) median correlations (Fig. 2A).

Because both mRNA and protein profiling data are noisy (100, 101), it is unclear how much of the reported low correlation between mRNA and protein expression is because of technological issues versus underlying biology. Statistical methods that attempt to model stochastic and systematic errors in mRNA and protein profiling data have produced higher mRNA-protein correlations (7, 102), and thus it has been suggested that transcription predominantly determines protein abundance (103). Recent studies by the CPTAC consortium have reported nonrandom associations between the level of mRNA-protein correlation and biological functions of the genes (10, 12, 13). For example, metabolic functions such as amino acid, fatty acid and nucleotide metabolism are enriched for genes with high mRNA-protein correlations, whereas ribosomal and mRNA splicing functions are enriched for genes with low or negative mRNA-protein correlations. A more systematic study using mRNA and protein profiling data from the three CPTAC cancer types showed that proteomic data strengthened the link between gene expression and function for at least 75% of Gene Ontology (GO) biological processes and 90% of KEGG pathways (104). Thus, mRNA-protein discrepancy cannot be simply explained by experimental errors, and biological functions arise from both mRNA- and protein-level regulations.

**Genetic Control of mRNA and Protein Abundance**—Genetic variation plays an important role in determining mRNA and protein abundance. mRNA and protein expression data from a cohort of samples can be integrated with DNA variation information to study the underlying genetic determinants of gene expression variation. This type of analysis is an extension of the traditional quantitative trait locus (QTL) mapping, in which a section of DNA (the locus) is correlated with variation in a phenotype (i.e., quantitative trait). When expression levels of mRNAs are treated as quantitative traits, the QTL analysis is named eQTL analysis, a method that has become well-established in the field of genetics (105) (Fig. 2B). eQTLs may be cis- or trans- acting, determined by their physical distance from the gene they regulate. Specifically, cis-eQTLs affect gene expression at the same locus of the genotype, whereas trans-eQTLs affect gene expression at a different locus. Although many cis-eQTLs have been reported, mapping trans-eQTLs has been less successful (106). It remains unclear whether the difficulty in mapping trans-eQTLs reflects true biology (i.e., eQTLs primarily act in cis) or computational and statistical challenges. More recently, ribosome occupancy and protein abundance have been used as quantitative traits to identify ribosome occupancy QTLs (rQTLs) and protein abundance QTLs (pQTLs), respectively (8).

An integrative multi-omics study on a set of HapMap Yoruba lymphoblastoid cell lines found that most QTLs were associated with mRNA expression levels, but their impact on protein expression levels were significantly reduced (4). This buffering of protein levels may allow cells to cope with noisy genetic variations and attenuate their impact on downstream phenotypes. Interestingly, a set of cis QTLs that affect protein abundance showed little or no effect on messenger RNA or ribosome levels, suggesting their potential roles in post-translational regulation. Both the buffering effect and protein abundance specific QTLs have been reported in earlier studies in yeast (5, 6), Arabidopsis (7), mouse (8), and human (9). These studies all suggest that integrating high-throughput proteomic data into QTL analysis could provide new insights into gene expression regulation.

Similarly, analysis of the correlation between copy number alteration (CNA) and mRNA or protein abundance has been used to infer the impact of CNAs on mRNA and protein abundance, including both cis-effects on the abundance of genes in the same loci and trans-effects on the abundance of genes at other loci in the genome. Visualization of the resulting correlation matrix in a heatmap can help highlight statistically significant cis- and trans- correlations. Furthermore, visually and statistically comparing the correlation heatmaps
for mRNA and protein can reveal relationships between these profiles: cis- and trans-effects in protein (and also phosphoprotein) are generally subsets of mRNA cis- and trans-effects respectively, with more directionally uniform effects at the protein level (10, 12, 13) (Fig. 2B). These correlation matrices can also be used to identify candidate driver genes whose copy number alterations directly drive significant trans-effects by comparing with functional knockdown data in large public databases like LINC5 (Library of Integrated Network-based Cellular Signatures) (12).

Efforts have also been made to study the roles of miRNAs in gene expression regulation. miRNAs are small noncoding RNAs that pair to the messenger RNAs (mRNAs) of protein-coding genes to suppress their expression (107). Several studies have shown that in addition to downregulating mRNA levels, miRNAs also directly repress translation of hundreds of genes (108–110). To investigate all miRNAs simultaneously in their endogenous context, Liu et al. performed an integrative analysis of global miRNA, mRNA, and protein profiles in nine colorectal cancer cell lines using a correlation-based method (11) (Fig. 2B). This study showed that translational repression was involved in more than half, and played a major role in a third of all predicted miRNA-target interactions. These predicted miRNA-target interactions can be further confirmed by more focused miRNA perturbation studies. Interestingly, sequence features known to drive site efficacy in mRNA decay, such as 8mer seed site, site positioning within 3’ UTR, local AU-rich context, and additional 3’ pairing, are generally not applicable to translational repression (11). A key unanswered question is what sequence features determine selectivity for miRNA-mediated translational repression.

Relating Mutations to Post-translational Modifications and Signaling—Millions of nonsynonymous single nucleotide polymorphisms (nsSNPs) identified by next-generation sequencing (NGS) and genome-wide association (GWAS) studies have been correlated with certain phenotypes and diseases (111, 112). However, the functional mechanisms of these associations are often barely understood or completely unknown. One likely explanation is that a subset of these SNPs result in amino acid changes in PTM targets, including targets of phosphorylation (specific to serines, threonines, and tyrosines), or acetylation and ubiquitination of lysines, directly perturbing cell signaling networks (14, 113–116). Because these four amino acids account for 22.2% of all amino acids in the human proteome (117), they are expected to be disproportionately affected by missense mutations. Substitutions of amino acids that are targets of PTMs can result in destruction, genesis, or constitutive activation of PTM sites (114). Moreover, mutations affecting proximal flanking positions of PTM sites might alter the recognition motif for corresponding transferases, e.g., protein kinases recognize, besides other factors, specific motifs on its substrate protein (14, 15).

To address this, several studies have assessed the effect of SNP-induced changes to PTM sites, predominantly serine, threonine and tyrosine phosphorylation. In 2008 Ryu et al. (14) used data from Swiss-Prot and Swiss-variant (117) databases to develop software predicting phosphorylation sites accompanied by a database for human phosphovariants, which the authors defined as genetic variations that change phosphorylation sites or their interacting kinases. In this study variants were classified into three groups depending on whether the variant directly affects a phosphorylation site, the flanking region or the kinase itself (Fig. 2C). Two years later, Yang et al. (116) used phosphosites annotated in Phosho.ELM (118), the Human Protein Reference Database (HPRD) (119) and Swiss-Prot (117) together with SNPs annotated in the NCBI dbSNP database (120), to identify 64 phosphorylation sites that potentially result in a disease phenotype, including schizophrenia and hypertension, when substituted by a nonphosphorylatable amino acid. In total 1451 nsSNPs which were present in dbSNP (downloaded May 2007) occurred in a ± 7 amino acid flanking region of a phosphosite, thereby potentially influencing the recognition of a kinase toward its preferred substrates. In a related study, Ren et al. (15) carried out a genome-wide analysis of SNPs that potentially influenced protein phosphorylation status. The authors used a combination of dbSNP predicted kinase-specific phosphosites and experimentally detected phosphosites to identify and classify SNPs affecting phosphosignaling. Based on the predicted phosphosites the authors estimated that ~70% of nsSNPs have the potential to affect phosphosignaling, suggesting that a large portion of nsSNPs play an important role in rewiring biological pathways. Creixell et al. (16) described a similar computational approach (ReKINect) to systematically classify and interpret such network-attacking mutations (NAMs) specifically in phosphosignaling. The authors used exome sequencing, bioinformatics and phosphoproteomics to demonstrate as a proof-of-principle the existence of six types of NAMs in human cancer cell lines.

Additionally, Reimand et al. (121) developed a computational method (ActiveDriver) based on a gene-centric generalized linear regression model to detect aberrant mutation rates proximal to phosphorylation sites. This method was used to analyze 800 genomes spanning eight cancer types to detect mutations specifically targeting the phosphorylation machinery, identifying 44 genes with significantly higher mutation rates in regions with detected phosphosites compared with the gene sequence given its structured and disordered regions. Mutations identified were comprised of both known driver mutations and novel candidates. The authors then extended their approach to the TCGA pan cancer data set, containing more than 3000 genomes from 12 cancer types and found mutations affecting phosphosignaling in about 90% of all tumors (17).

Two databases that are valuable for PTM proteogenomic analysis that have not yet been described are PhosphoSite-
Plus (PSP) (122) and g2pDB (115). PSP is a manually curated database of mammalian PTM sites containing over 330,000 nonredundant PTMs (122). In its latest release (2014), PSP introduced the “PTMVar” data set, which intersects missense mutations and PTMs detailing 25,000 PTMs impacted by known variants, about 75% of which relate to phosphorylation. The remaining PTM sites comprise ubiquitylation, acetylation, mono-methylation, and succinylation sites. These additional modifications, despite their low coverage, enable researchers to interrogate genomic mutations with PTMs beyond phosphorylation. g2pDB is a database mapping protein PTMs to genomic coordinates for all phosphorylations, acetylations and ubiquitylations that are available in the Global Proteome Machine Database (GPMDB) (123, 124). Overlaying the genome-mapped PTM sites with genome coordinates of known, disease-associated SNPs might reveal a role of these PTM sites in the respective disease. A list of all relevant tools and databases can be found in Table II.

All the aforementioned studies focused on the classification of mutations as either directly or indirectly (i.e. in a proximal flanking region) affecting the PTM site. However, there is no consensus on the length of flanking regions, number of classes, and their nomenclature, making it difficult to directly compare the findings of different studies. The integrative analysis of genomic mutations and PTM-mediated signaling shows great promise in providing insights into the mode of action of disease-associated mutations. This type of analysis can aid in discriminating tumor driver mutations from functionally neutral passenger mutations, and ultimately lead to novel personalized treatments. More importantly, the analysis of PTMs is not accessible to genomics sequencing technologies and the ever-increasing collection of published, global PTM-omes at single amino acid resolution demonstrates the indispensable value of state-of-the art MS-based proteomics in the era of precision medicine.

### INTEGRATIVE MODELING OF PROTEOGENOMIC DATA

Integrative modeling involves the application of statistical, machine learning and network-modeling tools to data obtained from one or more omics platforms. In this section, we focus on the application of integrative modeling to proteogenomic analyses. Models can be developed on combined omics data sets (e.g. genomics and proteomics), or applied to each omics data type separately, and the results comparatively analyzed. We review clustering (Unsupervised Clustering) and predictive modeling (Predictive Modeling)—usually termed class discovery and class prediction, respectively—which are both orthogonal approaches to gaining insight from biological data; and network modeling (Pathway and Network Modeling), which interprets data in the context of prior biological knowledge and promotes understanding at the level of pathways and cellular mechanisms.

**Unsupervised Clustering**—Clustering is a method of grouping similar entities—e.g. samples, genes, proteins, etc.—together based on a similarity metric. Because meta-data about the entities—like phenotypes, mutations, disease type, etc.—are not used in the clustering process, the algorithms are termed “unsupervised,” and are primarily used to discover new groups or classes, in addition to computationally validating known biology. Most proteogenomic analysis includes unsupervised clustering of proteome and/or phosphoproteome data, followed by comparison of the resulting clusters to known subgroups, cluster labels derived from genomic data, or other mutation, survival, or clinical data.

Clustering of proteome data is performed using a variety of algorithms including hierarchical (10), k-means (12) and model-based clustering (13). Although the clustering algorithms can vary, consensus clustering (125) is a common approach used to assess cluster stability and define the natural number of clusters in the data. Visualization of the consensus matrix, along with the delta-area plot and silhou-

### TABLE II  
**Computational frameworks and resources for intersecting PTMs and mutations**

| Name          | URL                                      | Reference | Remarks                                                                 |
|---------------|------------------------------------------|-----------|-------------------------------------------------------------------------|
| PhosphoPOINT  | http://kinase.bioinformatics.tw*         | (116)     | Human kinase interactome and phospho-protein database                    |
| PhosSNP       | http://phosnp.biocuckoo.org/             | (15)      | Database of mutations predicted to impact phosphorylation status of proteins |
| PhosphoVariant| NA                                       | (14)      | Database for definite and possible variants changing phosphosites        |
| PTMvar        | phosphosite.org                          | (122)     | Database intersecting non-synonymous SNPs and PTM sites                  |
| ActiveDriver  | http://individual.utoronto.ca/reimand/ActiveDriver/ | (121)     | Prediction of ‘active’ phosphosites in proteins that are specifically and significantly mutated in cancer genomes |
| ReKiNect      | http://rekinect.science/home             | (16)      | Prediction of network attacking mutations (NAMs) from NGS data           |
| MIMP          | http://mimp.baderlab.org/                | (203)     | Characterization of genetic variants that specifically alter kinase-binding sites in proteins |
| g2pDB         | www.g2pdb.org                            | (115)     | Database of auto-curated PTM sites mapped to their genomic locations     |

* link not working at the time of writing.
ette plots (126) are an effective way to determine the number of clusters in the data.

Once the proteome or phosphoproteome clusters are identified, the samples constituting these clusters can be characterized by enrichment tests for known subgroups (e.g., PAM-50 classification or RPPA groups in breast cancer, methylation subtype in colon cancer, mutation status for relevant genes, or other clinical or survival data). In addition, supervised marker selection methods combined with pathway enrichment analysis (e.g., SAM (127) marker selection followed by Gene Set Enrichment Analysis (GSEA) (128) for pathway enrichment) can also be used to characterize proteome or phosphoproteome clusters by identifying pathways or gene sets that are selectively up or down regulated in each cluster.

An alternative approach to clustering proteome or phosphoproteome data is to project the original data to pathway space and then cluster the projected data. This approach is used by Mertins et al. (12) to cluster phosphoproteome pathways, resulting in a unique cluster not directly observed either in the proteome or phosphoproteome data. The projection to pathway space is performed using single-sample gene set enrichment analysis (ssGSEA) (129), where the enrichment of curated pathways (MSigDB C2 gene sets, http://software.broadinstitute.org/gsea/msigdb) in each sample is evaluated. The enrichment scores are then subject to unsupervised clustering, followed by characterization of the derived clusters using the pathways constituting the data set.

Coclustering—In coclustering, data from multiple modalities (e.g., mRNA and proteome) are treated as independent “samples,” and clustering is performed over the collection of disparate omics profiles. The key here is to either transform the data so that different modalities are comparable (e.g. z-scores), or to use a similarity metric that is agnostic to the scale of values in the data (e.g. Spearman correlation (130)). Coclustering mRNA and proteome data (using hierarchical clustering) after filtering to retain genes or proteins with moderate to high correlation is used to show that mRNA profiles of samples are closest to their corresponding proteome profiles, thereby validating sample quality and mitigating concerns regarding tumor heterogeneity in (12) (Fig. 3A).

Multi-omic Clustering—Clustering over sample profiles obtained from two or more omic platforms is referred to as multi-omic clustering. Unlike coclustering, where the multi-omic data from each sample provides independent items that are clustered simultaneously, multi-omic clustering attempts to derive an integrative clustering to assign each sample to a single cluster based on combined evidence from multi-omic data. An overall review of multi-omic clustering methods is presented in (131) where algorithms are grouped by strategy:

- Direct integrative clustering methods use a combined multi-omics data set as input to the clustering analysis. Examples in this category include iCluster+ (132), LRACluster (133) and moCluster (134).
- Clustering of clusters is an approach where clustering is initially performed on each omics data set and the results integrated into final cluster assignments. Examples include COCA (135) and SNF (136).
- Regulatory integrative clustering harnesses molecular regulatory structures and/or networks to integrate different omics data sets in a robust manner. Examples in this group include PARADIGM (137) and iRafNet (138).

Many of the clustering algorithms use de novo or regulatory network graphs to model interactions in each omics domain, and to drive integration across different omics data sets. A review of clustering methods from this orthogonal perspective is covered in (139).

Predictive Modeling—Predictive modeling is a statistical approach in which models are built to predict a future outcome based on data attributes. Machine learning, pattern recognition and predictive analytics all lie within the umbrella of predictive modeling and this method of analysis has been rapidly gaining traction across most scientific disciplines. Predictive modeling and machine learning techniques applied to proteogenomics can greatly improve our ability to accurately diagnose, guide prognosis, and treat disease. For example, global molecular profiling of tissues and tumors enables a shift from nonspecific treatment strategies toward a more targeted, personalized approach based on the presence or absence of predictive genetic and/or protein signatures. Typical supervised classification methods used for predictive modeling from omics data include Support Vector Machines (SVMs) (140), Bayesian logistic regression (141), and random forests (142).

Machine Learning—To date, machine learning and statistical modeling techniques applied to genomics and transcriptomics data have identified genetic profiles predictive of disease diagnoses (143) and drug response (18, 144–146). One would expect the predictive analysis of proteome and phosphoproteome data to be more informative regarding clinical outcomes compared with NGS data, as these data modalities are more proximal to the disease. These techniques have been applied to proteomics data to classify clinically relevant disease subtypes in cancer (147–149), to define prognosis (150), and to identify biomarkers predicting drug sensitivity (150–152).

Despite the use of predictive modeling in genomics and proteomics independently, studies integrating proteomics and genomics are less common. Several studies using “multimodal” integration of data types including RNA-Seq, exon expression, and Reverse Phase Protein Array (RPPA) data to predict clinical phenotypes and drug response found no advantage to combining data modalities compared with individual platform analysis and showed gene expression data to be consistently more predictive than RPPA-based proteomics (19, 144). Similarly, Ma et al. (20) found that in machine learning models predicting ten-year survival from 77 breast tumors (12), fusion of four data types (genome, transcriptome, MS/
MS-based proteome and phosphoproteome did not improve the predictive performance of the model. However, they did find proteomics to outperform models based on genomics and transcriptomics data in survival prediction (20). As this is still fairly uncharted territory in proteogenomics, we anticipate to see a wealth of studies focused on assessing the predictive power of proteomics, and phosphoproteomics in disease prognosis, diagnosis and drug response in the future (Fig. 3B).

**Supervised Analysis for Marker Selection**—Aside from machine learning, supervised analysis has been used to derive markers for a variety of distinctions including intrinsic disease subtypes (e.g. PAM-50 subtype in breast cancer or HRD status in ovarian cancer), subtypes identified by clustering, samples with and without mutations in genes of interest (e.g. PIK3CA or TP53 mutations) and survival analysis. For examples, see (12) and (13).

Common marker selection methods used include the t test, ANOVA (F-test), moderated tests (153) and SAM (127), in addition to nonparametric tests like the Mann-Whitney test and the Kruskal-Wallis test. Although these tests are in most cases applied to specific types of omics data, marker ranking from these tests can be combined across multiple omics data sets to derive a global overall rank using rank aggregation algorithms (154, 155).

**Pathway and Network Modeling**—Historically, the field of biomedicine has operated under the “molecular biology paradigm,” in which it is assumed that biological function can be explained through the comprehensive knowledge of genes and their associated proteins, and that these proteins operate in linear pathways (156). Despite large-scale efforts to link genotype and phenotype under this paradigm, the relationships between the two are still wholly unresolved and surprisingly complex. Instead, the systems or network biology approach attempts to consider these complex relationships to better understand this genotype-phenotype connection (157–159). Studies in proteogenomics can build upon current models of network biology, contributing to both network annotation and using established pathway and gene ontology tools in gene-protein enrichment analyses.

**Network Annotation**—In network biology, nodes represent the molecules of interest (gene, protein, metabolite) and edges represent a function, physical or enzymatic relationship. Genetic and physical interaction networks are commonly used models for studying complex systems and disease. These networks can reflect a static system, built from information in a single condition or a differential system, highlighting changes in network connections in two distinct states, revealing state-specific and disease-specific interactions (see reviews (160, 161)).

Biological networks are typically built in three ways: (1) curation of available physical or biochemical interaction data, (2) computational predictions based on sequence similarity, gene cooccurrence, or gene coexpression; and (3) comprehensive assessment of whole genomes or proteomes (158). MS-based proteomics and PTM-omics can be layered atop these scaffold networks to both fine tune the biological network representation and identify network rewiring in a disease state (162) (Fig. 3C). For example, Zhang et al. identified protein-protein interaction network modules that were enriched in down-regulated proteins in a poor-prognosis colorectal cancer subtype (10). Similarly, analysis of gene-protein coexpression found differential interaction patterns in a subset of network modules in basal-enriched and luminal-enriched breast cancer subgroups (12).

**Pathway and Gene Ontology (GO) Enrichment**—Several approaches for pathway and GO enrichment analysis have been developed, including over-representation and Gene Set Enrichment Analysis (GSEA). Over-representation analysis uses the Fisher’s exact test to identify pathways and GO terms with significant over-representation in a gene or protein list of interest, which should be predefined based on differential expression, clustering, or other upstream analyses. Representative tools in this category include DAVID (163) and WebGestalt (164) (Table III). GSEA (128) ranks genes or proteins in the entire data set based on differential expression or association to a continuous phenotype, and then uses a modified version of the Kolmogorov-Smirnov test (165) to identify pathways, signatures and GO terms in which the gene members are enriched at the top or bottom of the ranked list (Fig. 3C). As both the over-representation method and the GSEA approach ignore pathway topology when performing enrichment analysis, an additional tool, SPIA (166), was established to address this limitation. Further, because these methods all perform enrichment analysis at the gene level, they do not allow for phosphosite-level enrichment analysis, which is critical for understanding kinase-phosphate signal transduction in phosphoproteome profiling studies. PHOXTRACK (167) was developed for this purpose, and modifies the GSEA approach to search for an enrichment of known kinase targets in an uploaded phosphoproteomics profile data set (Table III).

**DATA SHARING AND VISUALIZATION**

In this section we describe useful frameworks for organizing, sharing, and visualizing multi-omics data in the context of the genome (Genome-based Data Sharing and Visualization) or protein interactome (Network-based Data Sharing and Visualization). Efficient data sharing and visualization plays a central role in making complicated proteogenomic data directly available and useful to the broad biological community. One successful example is the cBioPortal (168), which allows easy retrieval and visualization of multi-omics data from many studies for user-selected genes or pathways. The simple user query interface of cBioPortal hides data complexity from the users, and is self-explanatory and well suited for answering focused questions. Herein we will describe additional tools relating to data sharing and visualization allowing
for both exploratory analyses and focused queries of proteogenomics data.

**Genome-based Data Sharing and Visualization**—The genomic sequence provides a natural platform for the dissemination and visualization of genome-anchored information and as such, genome browsers, such as the UCSC genome browser and the Ensembl Genome Browser (169), have long been used to present sequence alignment and gene annotation information. Building upon these, tools such as the Integrative Genomics Viewer (IGV) (170) have been developed to extend the capacity of genome browsers to allow integrative visualization of genomic and transcriptomic sequencing data in the context of the genomic sequence. The addition of proteomic data into genome browsers enables covisualization with gene annotation information as well as genomic and transcriptomic abundance and coverage information, and thus facilitates proteogenomic data integration. A major obstacle lies in mapping peptides to their genomic locations, and many tools have been developed for both genome-centric mapping and visualization. The mapping process is either based on genome sequence alone or the composite of the genome and existing annotation. An early study by Kalume et al. used the TBLASTN algorithm to determine the corresponding regions in the genome for each peptide (171) which could then be visualized in the Ensembl Genome Browser. Annotation arising from this analysis was shared with the community using The Distributed Annotation System provided by Ensembl (171). Similar tools including Pepline (172) and The Proteomic Genomic Mapping Pipeline (173) use string searching algorithms to map peptides against a given genome translated in all six reading frames. However, these tools were not designed for visualization purposes, creating output files incompatible with genome browser visualization tools.

Other tools take advantage of extra gene annotation information to speed up this mapping process and facilitate visualization of peptides in genome browsers (UCSC, Ensembl, Integrative Genomics Viewer (IGV)), including PeptideAtlas (174), iPIG (175), PG Nexus (176), proBAMsuite (177) and PGx (178), among others (179–181). PG-Nexus (176) adopts a simplified SAM file so that proteomic data can be visualized in alignment with genome and transcriptome. The recently introduced proBAM format stores PSMs within the context of the genome. The proteomic identifications from three published proteomics data sets, including colorectal cancer samples from CPTAC (13) and cancer cell line samples from two different sources (182, 183), have been converted to proBAM format and made available in a JBrowse-based genome browser (http://proteogenomics.zhang-lab.org/), allowing for dissemination of proteomics data to a general audience beyond the proteomics community (177). Fig. 4 shows an example of proBAM file visualization using data from colorectal cancer cell lines (182). PGx (178) maps identified peptides onto their putative genomic coordinates based on a gene model and outputs a Browser Extensible Data (BED) and BEDgraph file for both peptide location and relative quantitation visualization, and has been applied to mapping and visualization of CPTAC data in the UCSC Genome Browser (http://fenyolab.org/ucsc_cptac_v1). Beyond the tools mentioned above, additional tools like VESPA (184) provide a stand-alone visualization solution. Tools connecting the peptide identifications to their underlying evidence, and providing easily accessible browser-based interactive visualization of large sets of MS data, are also being developed (185). With the increasing availability of genomic and transcriptomic data from high-throughput sequencing technologies, putative gene structure could be inferred for un-annotated genomes. These approaches are expected to promote the development of additional mapping applications.

Despite the potential of genome browser-based visualization, these tools typically require specific input formats such as Sequence Alignment/Map (SAM) and BED, most of which were designed for genomic but not proteomic data. A standard file format including both proteomics-specific information and genome mapping coordinates would be useful for future applications.

**TABLE III**

| Name       | URL                                      | Reference | Remarks                                           |
|------------|------------------------------------------|-----------|---------------------------------------------------|
| DAVID      | http://david.abcc.ncifcrf.gov/           | (163)     | GO/Pathway annotation and enrichment               |
| GoMiner    | http://discover.nci.nih.gov/gominer/     | (204)     | GO analysis                                       |
| GSEA       | http://software.broadinstitute.org/gsea/ | (123)     | Identifies pathways/GO terms with gene enrichment based on gene/protein ranking |
| InnateDB   | http://www.innatedb.com/                | (205)     | GO/Pathway annotation and enrichment, visualization |
| KEGG Atlas | http://www.kegg.jp/kegg/atlas/          | (206)     | Pathway visualization, enrichment                  |
| LINCS      | http://www.lincsproject.org/             | (207)     | Identifies common perturbation networks based on drug treated human cell lines |
| PHOXTRACK  | http://phoxtrack.molgen.mpg.de/          | (167)     | Modified GSEA approach focused on phosphosite-level profiling |
| SPIA       | http://vortex.cs.wayne.edu/ontoexpress/  | (166)     | GO/pathway enrichment modified to take into account pathway topology |
| WebGestalt | http://www.webgestalt.org/               | (164)     | GO/Pathway annotation and enrichment               |
proteogenomics studies. To facilitate the visualization and sharing of the rapidly growing body of proteogenomics data, the Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) is working on two new proteogenomics data formats, proBAM (http://www.psidev.info/proBAM) and proBED (http://www.psidev.info/probed). The basic concepts of the proBAM format are: (1) to hold more proteomics information; (2) provide functionality highly analogous to BAM, which connects proteomics data to a large number of bioinformatics tools that have been developed for BAM files; (3) to serve as a well-defined interface between PSM identification and downstream analyses. proBAM takes advantage of existing data formats and imports additional information from the peptide identification process to the genome browser visualization. As a highly compact file format, proBAM also facilitates public sharing and re-use of MS-based proteomics data. Similarly, proBED adopts the BED format and can be loaded directly to the UCSC genome browser as an annotation track. The proBAM format contains more detail and structure than proBED and thus it is expected that a proBAM to proBED conversion should be possible under most circumstances.

Network-based Data Sharing and Visualization—Network-based data visualization plays a central role in biological knowledge discovery and a wide range of tools have been developed to visualize and analyze omics data (Table IV) in the context of biological networks (186). Most tools visualize a network as a node-link diagram in which nodes represent molecular components and edges represent relationships between the components. Typically, four types of data can be visualized in a node-link diagram based on various approaches that are available in different tools.

(I) Single binary data, such as functional annotation data for genes or significant calls from statistical analysis. A single set of binary data can be visualized in a network by changing one of the node attributes, including node color, node border color, node label color, node size, node label size and node shape. Tools such as Cytoscape (187) and VisANT (188) support this function.

(II) Single continuous data, such as fold changes or p values for all genes in an omics study. For this type of data, a color gradient or size gradient can be used to color or size the nodes or node labels in the network according to the scores. This function is also available in Cytoscape and VisANT.

(III) Composite binary data, such as mutation status for genes in multiple samples. The approach described in type I can map the measurements from each sample to the network but only one sample at a time (185, 188). To visualize the measurements from more than one sample, Cerebral (190) developed the “small multiples” method that can view the data in parallel by creating multiple versions of the same network in a grid, and each version represents a different sample. Some other tools, e.g. VisANT (188) and VistaClara (191), use animation-like visualization to update node attributes sequentially in the network.

(IV) Composite continuous data, such as gene expression data from multiple conditions. Beside animation and small multiples, some other tools have been developed to visualize the complete data set simultaneously. For example, GENeVis (192), VisANT and VANTED (193) can replace each node in the network by a bar plot in which each bar represents a measurement from one condition, whereas VistaClara adds the bar plot below the node to visualize the data.

The above tools are very efficient for small networks. However, when the number of nodes goes beyond a few hundred, it is difficult to show the details of the embedded visualizations. Furthermore, simultaneous visualization of multi-omics data from multiple experimental conditions is almost impossible. The challenge cannot be addressed by developing graph layout algorithms (194) or extending the two-dimensional representation of the network into three-dimension (195).

To address this challenge, Shi et al. developed NetGestalt (196), a web-based tool that exploits the inherent hierarchical modular architecture of a biological network to create a one-dimensional layout of the network nodes (e.g. genes or proteins), with network-module information displayed below the linearly ordered nodes. Notably, this transformation makes it possible to visualize the above four types of data below the linearly ordered nodes by barcode plots, bar plots and heat maps. Thus, diverse experimental data such as DNA methylation, gene mutation, copy number variation, gene expression, and protein expression and modification, can be simul-

| Name          | URL                                      | Reference | Pros                                | Cons                                |
|---------------|------------------------------------------|-----------|-------------------------------------|-------------------------------------|
| Cytoscape     | http://www.cytoscape.org                | (187)     | Visualize single binary or continuous data in node-link diagrams | Difficult to integrate multi-omics data in large networks |
| VisANT        | http://visant.bu.edu/                   | (188)     | Visualize four types of data in node-link diagrams |                                     |
| Cerebral      | http://www.pathogenomics.ca/cerebral    | (190)     |                                     |                                     |
| GENeVis       | http://www.win.tue.nl/~rmwestenb/genevis/ | (192)     |                                     |                                     |
| VANTED        | http://www.wanted.org                   | (193)     |                                     |                                     |
| VistaClara    | http://chiiant.ucsd.edu/cyto_web/plugins/index.php | (191)     |                                     |                                     |
| NetGestalt    | http://www.netgestalt.org/              | (196)     | Visualize multi-omics data in large networks based on the one-dimensional layout | Do not support node-link diagram visualization for large networks |
aneously visualized in the context of the network. In addition, annotations from Gene Ontology, pathway databases, and other resources can also be visualized concurrently to facilitate the interpretation of experimental data. Fig. 5 shows a case study using NetGestalt to identify subtype specific modules from the human protein-protein interaction network (197) based on the CPTAC colorectal proteomics data (10).

Although network-based data visualization has advanced rapidly in the last decade, communicating network-based findings between scientists remains a big challenge. Web-based visualization tools such as CellMaps (198) provide a possible platform for sharing visualization results online for remote access and editing. However, sharing is not currently available in web applications allowing network-based data visualization. NDEx (the network data exchange) (199) allows scientists to create communities and to share networks but does not support data visualization. A network-based data visualization tool with the capability to share results would be particularly useful in helping scientists identify, record and communicate their findings.

CONCLUSION

Rapid technological developments occurring over the last decade have made it possible to generate large proteomic data sets, thereby driving the development of new methods for proteogenomic data analysis. It is now quite common to see genomics data informing proteomics data analysis or vice versa. Further, significant progress has been made in our understanding of the complex relationships between DNA, RNA and protein, and these developments are being used to both understand basic biological processes, and improve molecular markers and drug target discovery. There are, however, still many challenges in proteogenomics including the limited coverage and dynamic range of MS-based proteomics (200), and the difficulty of generating proteogenomic data sets with large numbers of samples because of cost and sample accessibility. This mismatched combination of limited sample numbers and measurements with large variability remains challenging for most analysis methods and can lead to ungeneralizable findings. The large number of quantitative measurements across multiple data modalities also poses a serious challenge for visualizing the data in a meaningful way. For example, visualizing large-scale gene expression analysis alongside proteomics and phosphoproteomics data requires that we present gene-level quantitation while also maintaining information on protein isoforms and expression differences across multiple phosphosites. Moreover, our knowledge of basic biological processes and pathways is still limited, clouding our ability to explain many of the proteogenomic relationships we identify. Despite these remaining challenges, we foresee continued and rapid development in this field both from a data generation perspective—with deeper, faster and cheaper genomic and proteomic profiling technologies—and a data analysis perspective, with the continued development of more sophisticated computational tools for data integration, modeling and visualization.

* KVR, SHP, and DF were funded by National Cancer Institute (NCI) CPTAC award U24 CA210972. KVR and DF were funded by contract 13XS068 from Leidos Biomedical Research, Inc. SHP was funded by an Early Career Award from the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. JW, JW, and BZ were funded by National Cancer Institute (NCI) CPTAC awards U24CA159988 and U24CA210954 and by contract 13XS029 from Leidos Biomedical Research, Inc. BZ is Cancer Prevention & Research Institutes of Texas (CPRIT RR160027) Scholar and McNair Medical Institute Scholar. KK, KRC and DRM were funded by National Cancer Institute (NCI) CPTAC awards U24CA160034, U24CA210986 and U24CA210979. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

To whom correspondence may be addressed: The Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, Tel.: +1-617-714-7483, E-mail: manidr@broadinstitute.org; Tel.: +1-713-798-1443, E-mail: bing.zhang@bcm.edu; or Tel.: +1-212-263-2216, E-mail: David.Fenyo@nyumc.org.

** Cofirst authors.
82. Jean Beltran, P. M., Mathias, R. A., and Cristea, I. M. (2016) A portrait of the human organelle proteome in space and time during cytomegalovirus infection. Cell Syst. 3, 361–373.e6

83. Luo, Y., Jacobs, E. Y., Greco, T. M., Mohammed, K. D., Tong, T., Keegan, S., Binley, J. M., Cristea, I. M., Fenyo, D., Rout, M. P., Chait, B. T., and Muesing, M. A. (2016) HIV-host interactome revealed directly from infected cells. Nat. Microbiol. 1, 16088

84. Crawford, D. H., Johannessen, I., and Rickinson, A. B. (2014) Cancer Virus: the discovery of the Epstein-Barr Virus, 1 edition, Oxford University Press, Oxford, United Kingdom

85. Huang, C. R. L., Burns, K. H., and Boeke, J. D. (2012) Active transposition in genomes. Annu. Rev. Genet. 46, 651–675

86. Rodic, N., Sharma, R., Sharma, R., Zampella, J., Bai, L., Taylor, M. S., Hruban, R. H., Iacobuzio-Donahue, C. A., Maia, A., Torbenson, M. S., Glimm, E., and Bond, P. L. (2010) Open MS/MS spectral library search to identify unanticipated metabolic pathways. J. Proteome Res. 9, 301–333

87. Ardeljan, D., Taylor, M. S., Burns, K. H., Boeke, J. D., Espey, M. G., Orr, B. A., Gallia, G. L., Eberhart, C. G., Boeke, J. D., Harris, C. R., and Thomas, B. C. (2016) Toward a standard in structural genome annotation for prokaryotes. Stand. Genomic Sci. 10, 45

88. Burns, K. H. (2014) Long interspersed element-1 protein expression is a hallmark of many human cancers. Am. J. Pathol. 184, 1280–1286

89. Ardeljan, D., Taylor, M. S., Burns, K. H., Boeke, J. D., Espey, M. G., Woodhouse, E. C., and Howcroft, T. K. (2016) Meeting report: the role of the mobiome in cancer. Cancer Res. 76, 4316–4319

90. Na, S., Payne, S. H., and Bandeira, N. (2016) Multi-species identification of polymorphic peptide variants via propagation in spectral networks. Mol. Cell. Proteomics 15, 3501–3512

91. Wilmes, P., and Bond, P. L. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ. Microbiol. 6, 911–920

92. Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Thomas, B. C., and Bond, P. L. (2016) A new view of the tree of life. Proc. Natl. Acad. Sci. U.S.A. 113, 5970–5975

93. Tripp, H. J., Sutton, G., White, O., Wortman, J., Pati, A., Mikhailova, N., Ovchinnikova, G., Payne, S. H., Kyrpides, N. C., and Ivanova, N. (2015) Toward a standard in structural genome annotation for prokaryotes. Stand. Genomic Sci. 10, 45

94. Horlacher, O., Lisacek, F., and Müller, M. (2016) Mining large scale tandem mass spectrometry data for protein modifications using spectral libraries. J. Proteome Res. 15, 721–731

95. Na, S., Payne, S. H., and Bandeira, N. (2016) Multi-species identification of polymorphic peptide variants via propagation in spectral networks. Mol. Cell. Proteomics 15, 3501–3512

96. Ye, D., Fu, Y., Sun, R.-X., Wang, H.-P., Yuan, Z.-F., Chi, H., and He, S.-M. (2010) Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. Bioinformatics 26, i399–i406

97. Wilmes, P., Heinrich-Buschart, A., and Bond, P. L. (2015) A decade of metaproteomics: where we stand and what the future holds. Proteomics 15, 3409–3417

98. Dietz, W., Kuras, K., and Sahl, H. W. (2015) Improved genomic context for genome studies. Annu. Rev. Genet. 49, 283–301

99. Keshava Prasad, T. S., Gopinath, K., Keerthi Kumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Anbaraj, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmad, M., Khashyap, M. K., Mohommad, R., Ramachandra, Y. L., Krishna, V., Rahman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Ncera, R., Cardongal, M., and Pandey, A. (2009) Human Protein Reference Database—2009 update. Nucleic Acids Res. 37, D767–D772

100. Shen, L., Zhang, B., M Raleigh, M., and Beavis, R. C. (2006) Quantitative proteomics: where we stand and what the future holds. Proteomics 6, 3038–3047

101. Debabrata, G., and Bader, G. D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol. Syst. Biol. 9, 637

102. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and reca- librations. Nucleic Acids Res. 43, D152–D250

103. Craig, R., Cortens, J. P., and Beavis, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data. J. Proteome Res. 3, 1234–1242
197. Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2011) Interaction databases on the same page. *Nat. Biotechnol.* **29**, 391–393

198. Salavert, F., García-Alonso, L., Sánchez, R., Alonso, R., Bléda, M., Medina, I., and Dopazo, J. (2016) Web-based network analysis and visualization using CellMaps. *Bioinformatics* 10.1093/bioinformatics/btw332

199. Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., Stojmirovic, A., Dobrin, R., Braxenthaler, M., Kuentzer, J., Demchak, B., and Ideker, T. (2015) NDEx, the Network Data Exchange. *Cell Syst.* **1**, 302–305

200. Eriksson, J., and Fenyo, D. (2007) Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat. Biotechnol.* **25**, 651–655

201. Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., Griffin, T. J., and Smith, L. M. (2014) Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* **15**, 703

202. Krasnov, G. S., Dmitriev, A. A., Kudryavtseva, A. V., Shargunov, A. V., Karpov, D. S., Urosthev, L. A., Melnikova, N. V., Blinov, V. M., Poverennaya, E. V., Archakov, A. I., Lielitsa, A. V., and Ponomarenko, E. A. (2015) PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics. *J. Proteome Res.* **14**, 3729–3737

203. Wagih, O., Reimand, J., and Bader, G. D. (2015) MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods* **12**, 531–533

204. Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28

205. Lynn, D. J., Winsor, G. L., Chan, C., Richard, N., Laird, M. R., Barsky, A., Gardy, J. L., Roche, F. M., Chan, T. H. W., Shah, N., Lo, R., Naseer, M., Que, J., Yau, M., Acab, M., Tulpan, D., Whiteside, M. D., Chikatamarla, A., Mah, B., Munzner, T., Hokamp, K., Hancock, R. E. W., and Brinkman, F. S. L. (2008) InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* **4**, 218

206. Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **36**, W423–426

207. Ma’ayan, A., Rouillard, A. D., Clark, N. R., Wang, Z., Duan, Q., and Kou, Y. (2014) Lean Big Data Integration in Systems Biology and Systems Pharmacology. *Trends Pharmacol. Sci.* **35**, 450–460