OXFORD

Briefings in Bioinformatics, 20(3), 2019, 767–777

doi: 10.1093/bib/bbx141
Advance Access Publication Date: 25 October 2017
Paper

Proteomics and phosphoproteomics in precision medicine: applications and challenges

Girolamo Giudice and Evangelia Petsalaki

Corresponding author: Evangelia Petsalaki, European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom. Tel.: +44 (0)1223 492665; E-mail: petsalaki@ebi.ac.uk

Abstract

Recent advances in proteomics allow the accurate measurement of abundances for thousands of proteins and phosphoproteins from multiple samples in parallel. Therefore, for the first time, we have the opportunity to measure the proteomic profiles of thousands of patient samples or disease model cell lines in a systematic way, to identify the precise underlying molecular mechanism and discover personalized biomarkers, networks and treatments. Here, we review examples of successful use of proteomics and phosphoproteomics data sets in as well as their integration other omics data sets with the aim of precision medicine. We will discuss the bioinformatics challenges posed by the generation, analysis and integration of such large data sets and present potential reasons why proteomics profiling and biomarkers are not currently widely used in the clinical setting. We will finally discuss ways to contribute to the better use of proteomics data in precision medicine and the clinical setting.

Key words: proteomics; phosphoproteomics; data integration; precision medicine

Introduction

Precision medicine refers to the use of diagnostic, therapeutic and monitoring strategies for individual patients based on their molecular profiles [1]. While there has been one promising example of monitoring molecular data from a single individual for a long term to assess their health and disease status [2], in practice, the focus of the community lies mainly in the stratification of diseases into subtypes, based on molecular biomarkers or signatures, i.e. in the molecular taxonomy of disease [3]. The aim is to use these signatures to assign patients to specific disease subgroups and administer the most effective therapy for them. For example, patients with certain variants of TPMT, a thiopurine methyltransferase, are known to exhibit severe toxicity to the most common leukemia chemotherapy drug, thiopurine [4]. The dosage of the drug for their treatment is thus currently adjusted, based on TPMT variant screening, to avoid the toxicity and treat leukemia effectively [5]. Extensive molecular characterization of gene expression signatures in breast cancer [6–8] has allowed the development of multigenes assays that are currently undergoing clinical trials for routine use in the clinic to guide patient treatment and monitoring [9].

Most efforts to molecularly characterize diseases use genomic-based methodologies to identify genetic variants, including copy number variations [10] and differential gene expression [6] associated with specific disease subtypes [11] (Figure 1). While significant progress has been made in stratifying patients and diseases, there has been limited success in using this information in the clinic. In a recent meta-analysis study of a Phase 1 trial for treating refractory malignant neoplasms, they found that, while the response rate using the 'precision' biomarker was significantly higher than in its absence, the median response rate was still only ∼30% [12]. Systems biology [13] has shown that focusing only on the genomic and transcriptomic layers of cell function regulation leaves us blind to other important regulators of cell phenotypes and outcomes. For example, metabolomics data provide information regarding the metabolism and energy balance regulation of the cell, and epigenomics can reflect the regulation of the gene expression and the

Girolamo Giudice is a postdoc in the Petsalaki group at the EMBL-EBI since April 2017 working on the development of methods for the characterization of context-specific signaling networks. He acquired his PhD in Biomedicine from the University Autónoma de Madrid on March 2017.
Evangelia Petsalaki is a Group Leader at the EMBL-EBI since February 2017 studying cell signaling using whole cell models. She did her post doc (2010–16) at the LTRI in Toronto with Tony Pawson and Fritz Roth and her PhD with Rob Russell at the EMBL in Heidelberg (2009).
Submitted: 1 July 2017; **Received (in revised form):** 21 September 2017

effect of environmental factors on the cell. The use of these data sets in precision medicine has been reviewed elsewhere [14, 15].

It is well known that changes in gene expression do not always reflect changes in protein abundance [16–18]. Proteins are the major effectors of cell functions through changes in their posttranslational modifications (PTMs) and abundance, reflected also on changes in their interactome with effects on cell phenotypes. It is therefore critical to also consider proteomics, phosphoproteomics and other PTM-'omics' data sets in our studies to understand disease development and subtypes, as they can better capture the functional state and dynamic properties of a cell. However, these data sets have not been extensively used in the precision medicine field because of the time required to run samples, complexity and dynamic range of proteomics samples, lack of reproducibility among laboratories, differences between quantification methods and other confounding factors [19, 20].

Recently, technological developments in instrumentation, sample preparation and data analysis [20–23] and initiatives to develop standards for the generation and evaluation of these data [24–30] have resulted in the availability of high-quality, reproducible and comprehensive proteomics and phosphoproteomics data sets and protocols to generate such data. For example, Sharma and colleagues [31] were able to detect 50 000 phosphopeptides in a single human cancer cell line, and scientists can routinely and accurately measure thousands of peptides within short time frames: Hebert *et al.* [32] were able to measure the entire yeast proteome comprising peptides from ~3980 proteins in just over 1 h. Hundreds of targeted and global proteomics data sets are also collected by the CPTAC (Clinical Proteomic Tumor Analysis Consortium) to contribute to the study of cancer [33]. Therefore, the bioinformatics community must currently address the challenge of taking advantage of this new layer of information and integrating it with other valuable omics layers to study the mechanism of human disease and translate it into actionable insight in the clinic. Targeted proteomics methods such as SRM/MRM (Selected/Multiple Reaction Monitoring; [34]) and data-independent acquisition methods such as SWATH-MS (Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra) also allow significant reduction in variability during data acquisition and improved data set quality [35]. For details on the technological advances that have allowed this revolution in proteomics and PTM-omics data acquisition, we redirect the reader to numerous existing publications [21, 22, 36–38]. Recent reviews have discussed proteomics and phosphoproteomics in the context of precision medicine [39, 40]. In this review, we will present an overview of bioinformatics approaches used to analyze these data individually as well as integrated with other omics data sets and will discuss challenges that should be tackled to gain insight into disease mechanisms and advance the field of precision medicine.

Proteomics-derived precision biomarkers and signatures

A major application of proteomics is for the identification of biomarkers for disease. Biomarkers can be divided in (i) diagnostic to identify a given type of disease (ii) prognostic to measure the disease status and (iii) predictive to measure a response to a treatment [41]. Ideally, a biomarker should distinguish the disease unambiguously and should be detected in an accessible body fluid such as plasma, blood, serum urine, saliva or cerebrospinal fluids [42]. For example, the prostate-specific antigen (PSA) is one of the most famous noninvasive screening biomarkers and is used to detect prostate cancer [43]. However, a high concentration of PSA in the blood is also associated with benign prostatic hyperplasia and prostatitis [44–46]. Thus, even though PSA provides sufficient sensitivity, it fails in the discrimination between prostate cancer and other prostate pathologies because of its poor specificity [47]. In recent years, to improve biomarker sensitivity and specificity, researchers have turned to a combination of biomarkers, i.e. a disease signature, instead of pursuing an ideal biomarker [48].

Using proteomics characterization of samples from different stages of luminal-type breast cancer progression, Pozniak *et al.* [49] identified differences in components of protein homeostasis and metabolic regulation that can differentiate healthy, from primary or lymph node-metastasized tumor tissues, and lymph node-positive and negative breast cancers. Proteomics-based subtyping of colon and rectal cancer patients by the CPTAC was also more fine-grained than that based on transcriptomics data leading to better prediction of patient prognosis [50]. Combining protein with phosphoprotein abundance measurements using reverse phase protein arrays has also been used, e.g. for the prediction of ovarian cancer recurrence [51]. Numerous studies have showcased the value of phosphoproteomics data in providing mechanistic information underlying the disease mechanism [52–54]. For example, phosphoproteomics data have been used to discover the mechanism of resistance of melanoma cells to *BRAF* inhibitors [52] and of glioblastoma to *mTOR* (mechanistic target of rapamycin) inhibitors, leading to the discovery of a novel combination therapy for the latter [53]. Casado and collegues [55] used phosphoproteomics data on hematological cancer cell lines to assign them to specific tumor types and potential treatments. They also studied acute myeloid leukemia primary cells to identify the differential activation of kinases in cells that presented different drug resistance profiles [56]. Excitingly, cell-specific

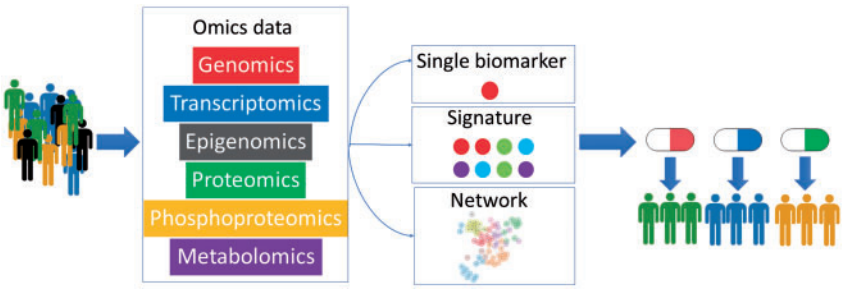


Figure 1. Example workflow for precision medicine. Multi-omics data are initially collected from patients and integrated to create their individual molecular profiles. These profiles are then matched to previously defined disease profiles that can guide the selection of treatment. This is achieved either through a match to known biomarkers, omics signatures or network/pathway signatures. The appropriate drug is then selected based on this match, to improve the chance of successful treatment and reduce the probability of side effects.

phosphoproteomics has also been used to study bidirectional signaling between endothelial cells and tumor cells to understand metastatic mechanisms of tumor cells [54]. Recently, phosphoproteomics data were used to create mechanistic models of colorectal cancer cell line-specific drug resistance, suggesting that this could be a viable option also for patients [57]. It is therefore clear that the proteomics and phosphoproteomics layer of omics information can provide valuable insight in our quest towards precision medicine.

Extracting relevant and reliable features (proteins) from high-throughput proteomics data is the main challenge for the biomarkers identification process. One approach is to use those proteins that are differentially expressed between normal and disease state [58–62]. More sophisticated methods such as machine learning and network-based approaches are also used. Machine learning methods such as support vector machine [63, 64] (SVM), neural networks [65–69], decision tree [67], random forest [70, 71] and genetic algorithms [72] have been successfully applied to proteomics data to identify biomarkers for several cancer types, heart failure and other conditions. Ahn *et al.* [73] constructed a 29-plex array platform comprising 29 potential biomarkers associated with gastric adenocarcinoma. A total of 13 candidate biomarkers were selected by random forest feature selection algorithm. Random forest and SVM were used to classify individuals as patients with gastric adenocarcinoma or controls. The algorithms tested on an independent blinded set of 95 gastric adenocarcinoma sera and 51 controls reached a mean accuracy of 89.2 and 85.6%, respectively. Random forest generally outperformed SVM, regardless of stage or tumor size; however, the SVM algorithm performed well for diagnosing small tumors. Rogers *et al.* [66] trained a neural network on either presence/absence of peaks or peak intensity values in a cohort of patients affected by renal cell carcinoma. Their model reaches sensitivity and specificity values of 98.3–100%. However, in an independent validation cohort of 80 cases, the performances were significantly weaker (sensitivities and specificities ranged from 41.0 to 76.6%). This highlights the frequent tendency of machine learning approaches to overfit their functions to noise inherent to the data set rather than the signal. Appropriate consideration regarding the complexity of the model and control data sets should thus always be used to avoid this issue when using such approaches.

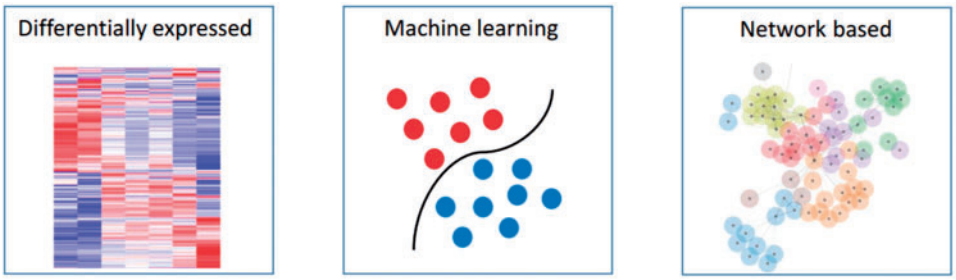
High-throughput proteomics data sets are characterized by a high number of variables/features compared with the total number of samples available. Hence, the input space includes many irrelevant or noisy features, which, coupled with the wide heterogeneity commonly found in biological samples, make it difficult to identify the truly important biomarkers. To tackle this problem, dimensionality reduction methods [74], such as PAM (Prediction Analysis for Microarrays) [75], SVM-RFE (Support Vector Machine-Recursive Feature Elimination) [48], SAM (Significance Analysis of Microarrays) [76], are used, in combination with machine learning methods, to reduce the noise in the data sets. This is achieved by discarding irrelevant features and enhances the generalization and the prediction performance. For reviews of feature selection algorithms, we redirect the reader elsewhere [77–79].

The lack of reproducibility across different data sets, technical issues such as the overfitting problem in machine learning approaches and the intrinsic complexity of human diseases often prevent promising biomarkers from reaching clinical application [80]. A promising idea to improve the reproducibility and the interpretation of the results is to incorporate prior biological knowledge and different high-throughput data sets to facilitate our understanding of biological processes at a mechanistic level.

From lists to integrated networks

Uncovering the individual mechanisms of disease development and progression in different patients will be key to designing accurate precision therapy strategies. As a first step in that direction, omics data analysis approaches typically attempt to identify affected biological processes and functions [49] by using Gene Ontology [81] or pathway (or other features) enrichment analyses [82] on the differentially regulated entities of each data set (e.g. genes, proteins or phosphopeptides). These differentially regulated entities can also be mapped onto existing interaction networks or pathway maps to provide a better picture of the cell processes affected in a specific sample. For example in the tumor endothelial bidirectional signaling study mentioned above [54], the authors mapped the affected phosphopeptides onto KEGG pathway maps [83], to understand the pathways involved in the transendothelial metastasis of tumors. More recently, a collection of methods, mostly developed for and applied to genomics and transcriptomics data sets, has been developed that take into consideration also the protein interaction network and pathway structure to identify patient-specific disease-perturbed pathways [84]. The SPIA algorithm (Signalling Pathway Impact Analysis) combines information on the differential expression of genes with their influence in a pathway based on their placement in a pathway topology [85]. HotNet2 [86] and Tied Diffusion Through Interacting Events (TieDIE; [87]) use slightly varied diffusion-based approaches that include a form of random walk and weighting according to the connection strength and network topology to propagate the effect of the perturbation in a given network [88]. There are many other methods available (the most widely used are reviewed here [84]) using, for example, network propagation [89] and clustering [90], current flow through the network [91], random walk [92, 93], pathway models [57] or other approaches for identifying perturbed functional modules or pathways in a network and using these as signatures to stratify patients or differentiate cancer model cell lines (Figure 2).

The concepts and methodologies can also be applicable to proteomics data sets; however, there are some issues that should be considered both when using these methods for transcriptomics/genomics data and when attempting to apply them to proteomics and phosphoproteomics data sets. Specifically, most of them tend to use existing interactome data and annotated pathway data, which are currently incomplete and biased toward highly expressed proteins [94–96]. This issue is further exacerbated, when trying to apply them to proteomics data sets, by the fact that these also inherently contain this bias. Moreover, our knowledge of tissue-specific interactions and their rewiring in different cellular states or conditions is currently limited [97, 98]. Such rewiring also occurs in disease and may vary across patients, and therefore, the use of generic networks and pathways for precision medicine applications may not be ideal. Finally, another issue to consider when applying such methods to proteomics and phosphoproteomics data sets is that they tend to have a much smaller coverage of the entire proteome than other respective omics data sets, depending on the instrument or technology used and the dynamic range of the abundances in the sample [99, 100]. It would therefore be useful to develop computational approaches that are tailored specifically to proteomics and phosphoproteomics data sets to account for these associated data characteristics.



Method	Study	Sample	Ref
DE (a)	Various stages of pancreatic ductal adenocarcinoma	Mouse pancreas	58
DE (a)	Pancreatic ductal adenocarcinoma	Pancreas	59
DE (a)	Gastric cancer	Saliva	60
DE (a)	Breast Cancer progression	Urine	61
DE (a)	Kawasaki disease	Serum	62
SVM (b)	Heart failure	Blood	63
SVM (b)	Different forms of arthritis	Urine	64
SVM and RF (b)	Gastric adenocarcinoma	Serum	73
NN (b)	Breast Cancer	Plasma	65
NN (b)	Renal Cancer	Urine	66
NN (b)	Colorectal Cancer	Serum	67
NN (b)	Liver cancer	Liver	68
NN (b)	Colon cancer	Serum	69
DT (b)	Liver Cancer	Liver	68
RF(b)	Fasting 36h vs 12h	Biofluids, blood cells	70
RF (b)	Mesothelioma	Blood	71
GA (b)	Ovarian Cancer	Serum	72
SPIA (c)	colorectal cancer, spontaneous labor vs without labor, Vessels	colonic mucosa, cervix, umbilical, myometrium	85
HotNet2 (c)	16 significantly mutated subnetworks	TCGA (12 cancer types)	86
TieDie (c)	Breast Cancer	TCGA	87
PRINCE (c)	Prostate cancer, Alzheimer's disease and type 2 diabetes	OMIM	89
NBS (c)	Ovarian cancer, uterine cancer and lung cancer	TCGA	90
Current flow (c)	Glioblastoma multiforme	Glioma Molecular Diagnostic Initiative	91
Rw (c)	110 disease-gene families	OMIM	92
Rw (c)	Disease families	OMIM	93
Pathway models (c)	14 phosphoproteins under 43 different perturbed conditions in 14 colorectal cancer cell lines	Colorectal Cancer	57

Figure 2. Different methods used in biomarker discovery. (**A**) Differentially expressed method, (**B**) machine learning method, (**C**) network-based method DE, differentially expressed; NN, neural network; RF, random forest; DT, decision tree; GA, genetic algorithm; NBS, network-based stratification; RW, random walk.

In the past few years, there have been a number of such methods developed. They mainly focus on accurately estimating the activity of diverse kinases in the systems under study to highlight the context-specific signaling networks that are active in each context. The most widely used method is the kinase–substrate enrichment analysis method [56], which calculates the kinases activity based on the differential abundance of their known substrates. Other methods include IKAP (inference of kinase activities from phosphoproteomics; [101]), which uses a machine learning approach, KARP (kinase activity ranking using phoshphoproteomics data; [102]), which calculates the relative phosphorylation of a kinase's substrates versus the total phosphorylation in the data and KinasePA (Perturbation analysis; [103]) and CLUE (CLUster Evaluation; [104]), which require perturbation or time series data. A few different approaches have been extensively benchmarked by Hernandez-Armenta and colleagues [105].

As proteomics and phosphoprteomics data sets provide a direct picture of the cell's functional state, inclusion of prior knowledge in these methods, such as motifs or interaction interfaces, known enzyme–substrate relationships and effects of mutations on protein structure and function, can also help better understand the effect of perturbations on the functional network.

Data integration approaches

Despite the wealth of information that proteomics and phosphoproteomics data can provide, it still represents only one layer of cell function and regulation. Thus, to truly understand

cell function in-depth, it is critical to consider as many as possible layers of cell function regulation [13]. This is especially true in the context of precision medicine where different layers of cell regulation may be important for each patient, and additional clinical information must also be included in the analysis. Therefore, one major challenge that our community is currently trying to solve is that of effective data integration of the less mature proteomics and phosphoproteomics layers of information with other omics data sets that have been more extensively studied and integrated in recent years.

There is currently no standard or optimal approach to data integration, and several methods have been developed (for reviews, see [106–108]). Here, we will focus on the main approaches used thus far to integrate proteomics data sets with other omics or clinical data. Depending on the data sets they integrate, methods can be divided into homogenous, where the data sets contain the same type of data but from different sources, and heterogeneous, where multiple data sets with different data types are integrated. These methods can either integrate the layers of information in a step-wise fashion or in a single step to generate an integrated model of the system under study.

For example Drake *et al.* [109], used a step-wise approach to integrate genomic, transcriptomic and phosphoproteomics data to identify patient-specific networks that are affected in prostate cancer and suggest potential precision treatments for these patients. Specifically they first used the data sets to broadly identify the pathways, transcription factors and kinases that are likely active in their samples and then applied their diffusion-based algorithm, TieDie [87], to pinpoint the different functional modules and pathways that are affected in the different patients. In this study, they also showed that the integration of phosphoproteomics was able to uncover pathways that would have otherwise been missed underlining the importance of including this level of information in precision medicine approaches. By applying this pipeline on three different prostate cell lines as validation, they were able to support their results either through evaluation of their predicted drug sensitivity or through gene essentiality studies.

Rudolph and colleagues [110] integrate protein interaction networks with phosphoproteomics data and evolutionary conservation to define signaling functionalities for proteins in a data set and delineate the active signaling pathways in a given phosphoproteomics data set. A recent systematic search for algorithms to reconstruct signaling pathways from phosphoproteomics [111] has shown that integration with prior knowledge yields the best results.

The most promising methods that integrate data sets in a single step include principle component analysis (PCA) [112] (or factor analysis)-based and nonnegative matrix factorization (NMF)-based [113, 114] approaches, as they are able to integrate diverse and large data sets and perform effective dimensionality reduction to allow easy downstream machine learning [115] or network-based [113] analyses and creation of models that represent the system under study.

The major issue with PCA-based approaches is the difficulty in interpretation of the biological mechanism underlying the different factor associations. Therefore, different supervised [116] or unsupervised [117] approaches can be used to choose the appropriate factors and help the results interpretation. These can include implementation of linear discriminant analysis [116], Bayesian classifiers [118], SVMs [119] and *K*-nearest neighbor [120] approaches after the PCA analysis. Liu *et al.* [118] integrated microRNA, mRNA and proteomics data into a joint matrix. They then used factor analysis and linear discriminant analysis to extract the molecular mechanism of cancer in different cell lines. The integrated approach identified clinically relevant markers and outperformed the analyses performed on the separate data sets.

While matrix factorization methods such as NMF and variations have been routinely applied to genomics and transcriptomics data [113, 121, 122], they have only recently been applied to proteomics data sets. For example, Yuan *et al.* [123] used pairwise NMF between omics data sets and clinical data to study the utility of using these omics data integration approaches in the clinic. In the subgroups, which they identified by combining proteomics and clinical data, they were able to identify—among other biomarkers and activated pathways—an additional patient subgroup that might also benefit from *MEK* (Mitogen-activated protein kinase kinase) targeting therapies.

A great advantage of matrix factorization approaches for proteomics and phosphoproteomics data sets is that they can also be used to impute missing data points [124]. This can be valuable for these data sets, as they inherently do not provide comprehensive measurements of all the components that might be present in other omics data types such as transcriptomic or genomic data sets. Other approaches for data imputation that can be applied in proteomics and phosphoproteomics data use nonlinear optimization approaches [125, 126].

Another integration approach that has been applied to the proteomics data is based on a multiple extension co-inertia analysis to identify the relationships among different omics data sets. Meng *et al.* [127], for example, integrated the transcriptome and proteome profiles of cells in the NCI-60 cancer cells. Using the integrated model, they found that the extravasation signaling pathway plays a fundamental role in leukemia; the same pathway was not identified in the single data set analyses.

Other than the missing data points that were discussed above, one of the major challenges for integrating proteomics and phosphoproteomics with other omics data sets is the inconsistent annotation and reporting of such data sets and analysis pipelines. This, in combination with the dynamic nature of the proteome and phosphoproteome, can result in the introduction of noise to the integrative models used to study a disease or a patient. As unified data collection and standardization processes are being developed for use of these data in the clinic, consistent methods to record the associated meta-data for this information that can be used in conjunction with existing methods for genomic and other omic data sets need to also be developed.

In recent years, there have been bioinformatics platforms and methods developed to reduce the variability from the data acquisition and analysis processes. Examples for this are the ProHits [128] and OpenMS [129]. ProHits is a software platform that is used mainly for interaction proteomics and provides a variety of options for data management and analysis that are systematically tracked to ensure the downstream reproducibility of the analysis pipelines. OpenMS is an open-source suite of analysis software for mass spectrometry data allowing the implementation of different pipelines and analyses procedures in a transparent and scalable way. These kinds of platforms ensure the reproducibility of the analyses pipelines. Methods for ensuring reproducibility during data acquisition are also important. For example, the TRIC (Transfer of Identification Confidence; [130]) algorithm, developed for SWATH-MS-targeted proteomics, uses a clever alignment approach to reduce the variability in peak picking and quantification across mass

spectrometry runs. Other similar software has been previously compared by Navarro *et al.* [131].

The inherent variability of proteomics and phosphoproteomics data sets can also be a confounding factor in data integration efforts. It has been shown in single-cell studies that the noise and sample variability significantly decrease when a specific cell response is activated compared with a static state, because of regulatory coordination [132]. Therefore, acquiring nonstatic data points, where possible, will reduce data variability and increase the signal to noise ratio. Additionally, single-cell technologies, providing single-cell measurements of protein or phosphoprotein abundance, have the potential to mitigate the data variability issue and improve the use of these data for understanding disease development.

From networks to mechanistic models

For use of proteomics and phosphoproteomics in the clinic, it is important to provide mechanistic information for a disease beyond the pathways and functional modules that have been affected. Halasz *et al.* [133] used phosphoproteomics data sets and a probabilistic framework to create a mechanistic and executable model of the rewiring that occurs in signal transduction pathways in cancer cells. They were able to identify a cell line-specific feedback loop for inhibition of IRS1 by p70S6K in colorectal cell lines and to perform stimulations to identify ways to increase their sensitivity to TRIC (*TCP-1* ring complex) inhibitors. Eduati *et al.* [57] used dynamic logic models and phosphoproteomics data to study the colorectal cell line-specific mechanism of drug resistance and a identified novel drug combination that can be used to overcome it.

Such models can be invaluable in the clinic to not only understand the mechanism of disease but also to simulate and predict the outcome of a treatment on specific patient groups or even individuals, depending on the available models.

Challenges for clinical application

While the proteomics and phosphoproteomics layers of functional regulation provide valuable insight into disease development and mechanism, there are still some challenges that need to be tackled before they can be readily applied for stratification of patients, even if data quality and bioinformatics challenges discussed above are tackled.

One of the major challenges is that most current 'omics' data analyses provide results that are not readily interpretable or actionable. For example, while identifying that a handful of pathways are affected in a specific patient subgroup may suggest the administration of specific kinase inhibitors as therapy, it does not necessarily uncover the full mechanism of a disease. There have been successful examples, such as the work of Zeevi and colleagues [134] that used omics data, clinical data and machine learning to devise an actionable change in personalized nutrition to regulate post-meal glucose levels, without an in-depth understanding of the mechanism at play. However, in most situations, lack of mechanistic information regarding a disease's development, makes it difficult to identify the causal targets for therapy at a reliability level that is appropriate for precision therapies in the clinic. As new methods for proteomics data analysis develop, our community needs to take this into consideration: rather than providing 'big picture' representations of affected cell processes in a disease, there is a need for producing reliable ranked targets or biomarkers by probability of being effective [135–137] or ranked testable hypotheses to help decide on one, alongside an easy-to-interpret explanation for their selection. This requires an in-depth understanding of cell processes and their interactions.

Recent years have seen the collaborations between computational biologists and clinicians or basic-science biologists dramatically increase, because of the advent of large-scale data sets and systems biology. The importance, however, of understanding basic biological processes in-depth to be able to understand disease mechanisms underlines the need for increased collaboration also between clinicians and basic research scientists. Interdisciplinary collaborations, including clinical data to take snapshots of the disease 'omics' profile, and iterations of computational analysis and basic biology for in-depth mechanistic studies of relevant cell processes, can lead to a detailed understanding and models of disease development, thus helping better stratify patients according to their disease subtype mechanism and design more knowledge-based treatments. Proteomics and phosphoproteomics data sets, as described above, can provide mechanistic insight into cell processes and are therefore ideal for inclusion in such studies to provide testable mechanistic hypotheses. Of course, the major disadvantage of such three-level approaches is that it takes time to perform in-depth studies of cell processes; however, as our knowledgebase of cell processes, their cross-talk and their role in different diseases increases, this will prove to be a worthwhile investment in the long run and might be the only way to truly achieve the goal of precision medicine across multiple diseases.

An additional challenge is presented when associating identified affected cell processes with specific disease phenotypes or clinical data. Currently, most studies use patient survival data as the patient phenotype and associate omics signatures with remission or survival rates [138]. More detailed and standardized phenotyping of patients can provide a better understanding of the causal cell processes of a disease and can improve diagnosis and tracking both of its progression and the effects of treatment and other issues that might affect a patient's quality of life [138]. As more omics data from patients are being generated, standardized protocols for systematically recording the phenotype of the relevant cells—if possible—and wider availability of in-depth patient clinical characteristics to data scientists beyond survival rate will also provide a significant contribution toward our community's goal of precision medicine. Ethical considerations to ensure patient anonymity and privacy need to also be taken into account in the development of these protocols as well as in the process of data sharing [140, 141].

The standardization of analysis pipelines and representation of results also present an issue for the routine application of proteomics protocols in the clinic. Whether the outcome of patient data analysis is the identification of a biomarker or a disease signature, robust quality control and analysis tools needs to be readily available to clinicians as well as accurate protocols for sample acquisition and results interpretation. This is critical to provide reproducible, high-quality precision care for patients across different hospitals and treatment centers. Proteomics and phosphoproteomics-specific data analysis pipelines have only recently started to be systematically developed and included in precision medicine studies [109]. Therefore, as the field matures, we expect to see significant progress in their standardized use across laboratories, institutes and eventually in the clinics.

Future directions

Proteomics and phosphoproteomics have recently emerged as a new layer of patient omics information in the field of precision medicine. Technological advancements and community efforts to standardize protocols and achieve robust and reproducible results [24–30] have contributed greatly to the utility of this data type in large-scale studies of disease and patient stratification. Their major strength lies in the fact that they give a picture of the actual workforce of the cell and are thus highly suited for studying the mechanism of disease development and progression. Other than the data reproducibility issue that the community is now efficiently tackling, one of the main challenges from a bioinformatics perspective that still prevents the wide-spread use of proteomics and phosphoproteomics data is the need for effective, data type-specific methods to extract the valuable knowledge it encodes and to integrate it efficiently with other large-scale data sets and prior knowledge. There are significant efforts made in this direction, and as the field matures, and more PTMs are also included, we expect it to provide great insight into the development of disease and help improve stratification of patients and design of precision approaches to their treatment and monitoring. Additionally, proteomics and phosphoproteomics data, like transcriptomics, encode highly dynamic information. Therefore, to accurately highlight differences in disease mechanisms and functional networks, and to reduce data variation, it is optimal to collect data sets on stimulation or perturbation rather than in a static state. This is currently impractical in a clinical setting, where we rely on a single sample from a usually untreated patient, but it could prove useful when performing, for example, window of opportunity trials where novel drugs are tested on patients before the standard treatment to evaluate their effect on untreated individuals [142, 143].

Currently, the bulk of population-level omics data is being collected to study cancer for precision oncology applications. Clearly, for precision medicine to become widely applicable more focus should be placed on characterizing also other diseases and their subtypes. These cancer studies, nevertheless, provide a unique learning opportunity for our community: we can use this rich data set to define what is the best way to maximize the orthogonal information we acquire from all these different omics layers, to estimate how many data sets are sufficient for characterizing a disease and potentially to identify the minimal components that one needs to measure in a cell to get the global signaling, gene regulation and metabolic status from a sample. From a proteomics perspective, such information can dramatically reduce the cost and variability of a study, making it even more applicable for clinical applications, for example through an educated design of targeted proteomics or phosphoproteomics approaches.

Of the drugs that are tested in clinical trials only 1 in 10 successfully go to the market [144]. This presents a huge financial burden for the pharmaceutical companies and the public. Bioinformatics approaches that effectively integrate omics data with in-depth clinical data can help guide many aspects of clinical trials to improve the chances of their success (for a recent review, see [145]): analysis of patients' omics data can help to guide the selection of targets and associated drugs and the appropriate group to which a drug can be administered with improved chance of success. Bioinformatics data storage and automated analysis pipelines can also make this knowledge available to future studies. At later stages, side effects or outcomes of the trial can be associated with specific molecular signatures in the patients to understand their mechanisms and design approaches to circumvent them. Indeed, these methods are already in use, and there are already guidelines in place to guide the design of clinical trials using omics data sets [146]. Thus, as an increased amount of clinical records and associated omics data sets become available to scientists, bioinformatics approaches will play an important role in guiding clinical trials with an increased success rate.

In an ideal precision medicine scenario, we would be able to create a widely used and robust clinical tool that can guide doctors with respect to the data required from a patient to provide his subdisease mechanism and guide the choice of therapy and monitoring. While we are several decades away from such a tool, and indeed from widespread use of any precision medicine approaches at all, it is nevertheless becoming increasingly clear that understanding at the molecular level and creating dynamic mechanistic models of cell functions during disease development and progression are critical for the success of precision medicine.

Precision medicine for all is still a long-term goal for our community. However, the field is rapidly progressing, and it certainly does not seem as far-fetched as it did 10 years ago. Even not taking into consideration the improvement in global quality of life, studies have demonstrated the cost–benefit of applying such approaches in the clinic [147].

Programs such as the St Jude Children's Research Hospital Pharmacogenomics of Anticancer Agents Research 4Kids (PG4Kids) program [148] and the Icahn School of Medicine at Mount Sinai Clinical Implementation of Personalized Medicine through Electronic Health Records and Genomics-Pharmacogenomics (CLIPMERGE PGx) program [149] can provide valuable knowledge regarding the practical prerequisites for real life precision medicine implementation. Additionally, exciting developments in preclinical studies include the use of patient-derived xenograph mouse models of disease (e.g. at the Jackson Laboratory), for testing precision therapies. We expect current and future advances in proteomics and phosphoproteomics data collection and analysis to greatly improve our understanding of disease development and progression also contributing to improved implementation of precision medicine in real world applications.

Key Points

- Precision medicine aims to tailor diagnostic, therapeutic and monitoring approaches to specific patient subgroups.
- Proteomics and phosphoproteomics data sets can provide mechanistic insight into disease development and are thus valuable for precision medicine approaches.
- Major challenges presented by these data include the lack of data robustness and standardization as well as the limited proteome and phosphoproteome coverage.
- Methods that are developed specifically for these data types as well as their effective integration with other data sets can mitigate the issues.

Funding

Funding for open access publication fees was provided by the European Molecular Biology Laboratory.

References

1. Huang BE, Mulyasasmita W, Rajagopal G. The path from big data to precision medicine. *Expert Rev Precis Med Drug Dev* 2016;1(2):129–43.

2. Chen R, Mias G, Li-Pook-Than IJ, *et al*. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;**148**(6):1293–307.
3. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: National Academies Press, 2011.
4. Rudin S, Marable M, Huang RS. The promise of pharmacogenomics in reducing toxicity during acute lymphoblastic leukemia maintenance treatment. *Genomics Proteomics Bioinformatics* 2017;**15**(2):82–93.
5. Drew L. Pharmacogenetics: the right drug for you. *Nature* 2016;**537**(7619):S60–2.
6. Perou C, Sørlie MT, Eisen MB, *et al*. Molecular portraits of human breast tumours. *Nature* 2000;**406**(6797):747–52.
7. Curtis C, Shah SP, Chin SF, *et al*. The genomic and transcriptomic architecture of 2, 000 breast tumours reveals novel subgroups. *Nature* 2012;**486**(7403):346–52.
8. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**(7418):61–70
9. KwaMakris MA, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. *Nat Rev Clin Oncol* 2017;**14**:595–610.
10. Nik-Zainal S, Davies H, Staaf J, *et al*. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;**534**(7605):47–54.
11. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, *et al*. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* 2016;**17**(5):257–71.
12. Schwaederle M, Zhao M, Lee JJ, *et al*. Association of biomarker-based treatment strategies with response rates and progression-free survival in refractory malignant neoplasms: a meta-analysis. *JAMA Oncol* 2016;**2**(11):1452–9.
13. Kitano H. Systems biology: a brief overview. *Science* 2002;**295**(5560):1662–4.
14. Kronfol MM, Dozmorov MG, Huang R, *et al*. The role of epigenomics in personalized medicine. *Expert Rev Precis Med Drug Dev* 2017;**2**(1):33–45.
15. Clish CB. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb Mol Case Stud* 2015;**1**(1):a000588.
16. Tchourine K, Poultney CS, Wang L, *et al*. One third of dynamic protein expression profiles can be predicted by simple rate equations. *Mol Biosyst* 2014;**10**(11):2850–62.
17. Vogel C, Abreu RS, Ko D, *et al*. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* 2010;**6**(1):400.
18. Gholami AM, Moghaddas A, Hahne H, *et al*. Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* 2013;**4**(3):609–20.
19. Bell AW, Deutsch EW, Au CE, *et al*. A HUPO test sample study reveals common problems in mass spectrometry–based proteomics. *Nat Methods* 2009;**6**(6):423–30.
20. Nilsson T, Mann M, Aebersold R, *et al*. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods* 2010;**7**(9):681–5.
21. Zhou L, Wang K, Li Q, *et al*. Clinical proteomics-driven precision medicine for targeted cancer therapy: current overview and future perspectives. *Expert Rev Proteomics* 2016;**13**(4):367–81.
22. Guerin M, Gonçalves A, Toiron Y, *et al*. How may targeted proteomics complement genomic data in breast cancer? *Expert Rev Proteomics* 2017;**14**(1):43–54.
23. Mitchell P. Proteomics retrenches. *Nat Biotechnol* 2010;**28**(7):665–70.
24. Searle BC. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* 2010;**10**(6):1265–9.
25. Varjosalo M, Sacco R, Stukalov A, *et al*. Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat Methods* 2013;**10**(4):307–14.
26. Mann M. Comparative analysis to guide quality improvements in proteomics. *Nat Methods* 2009;**6**(10):717–9.
27. Stead DA, Paton NW, Missier P, *et al*. Information quality in proteomics. *Brief Bioinform* 2008;**9**(2):174–88.
28. Tabb DL. Quality assessment for clinical proteomics. *Clin Biochem* 2013;**46**(6):411–20.
29. Wang X. Statistical assessment of QC metrics on raw LC-MS/MS data. In: L Comai, JE Katz, and P Mallick (eds), *Proteomics*. Springer New York, 2017, pp. 325–37.
30. Whiteaker JR, Halusa GN, Hoofnagle AN, *et al*. Using the CPTAC Assay Portal to identify and implement highly characterized targeted proteomics assays. *Methods MolBiol* 2016;**1410**:223–36.
31. Sharma K, D'Souza RCJ, Tyanova S, *et al*. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep* 2014;**8**(5):1583–94.
32. Hebert AS, Richards AL, Bailey DJ, *et al*. The one hour yeast proteome. *Mol Cell Proteomics* 2014;**13**(1):339–47.
33. Edwards NJ, Oberti M, Thangudu RR, *et al*. The CPTAC data portal: a resource for cancer proteomics research. *J Proteome Res* 2015;**14**(6):2707–13.
34. Elschenbroich S, Kislinger T. Targeted proteomics by selected reaction monitoring mass spectrometry: applications to systems biology and biomarker discovery. *Mol Biosyst* 2011;**7**(2):292–303.
35. Collins BC, Hunter CL, Liu Y, *et al*. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat Commun* 2017;**8**:291.
36. Jain K. Role of pharmacoproteomics in the development of personalized medicine. *Pharmacogenomics* 2004;**5**(3):331–6.
37. Duarte TT, Spencer CT. Personalized proteomics: the future of precision medicine. *Proteomes* 2016;**4**(4):29.
38. Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol* 2010;**11**(6):427–39.
39. Casado P, *et al*. Impact of phosphoproteomics in the translation of kinase-targeted therapies. *Proteomics* 2017;**17**(6):1600235.
40. Cutillas PR. Role of phosphoproteomics in the development of personalized cancer therapies. *Proteomics Clin Appl* 2015;**9**(3–4):383–95.
41. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat Rev Clin Oncol* 2008;**5**(10):588–99.
42. Kienzl-Wagner K, Pratschke J, Brandacher G. Proteomics—a blessing or a curse? Application of proteomics technology to transplant medicine. *Transplantation* 2011;**92**(5):499–509.
43. Papsidero LD, Wang MC, Valenzuela LA, *et al*. A prostate antigen in sera of prostatic cancer patients | cancer research. *Cancer Res* 1980;**40**(7):2428–32.
44. Ilyin SE, Belkowski SM, Plata-Salamán CR. Biomarker discovery and validation: technologies and integrative approaches. *Trends Biotechnol* 2004;**22**(8):411–6.

45. Thompson IM, Pauler DK, Goodman PJ, *et al*. Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per milliliter. *N Engl J Med* 2004;**350**(22):2239–46
46. Catalona WJ, Smith DS, Ratliff TL, *et al*. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med* 1991;**324**(17):1156–61.
47. Petricoin EF, Belluco C, Araujo RP, *et al*. The blood peptidome: a higher dimension of information content for cancer biomarker discovery. *Nat Rev Cancer* 2006;**6**(12):961–7.
48. Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines—Kernel Machines. *Mach Learn* 2002;**46**:389–422.
49. Pozniak Y, Balint-Lahat N, Rudolph JD, *et al*. System-wide clinical proteomics of breast cancer reveals global remodeling of tissue homeostasis. *Cell Syst* 2016;**2**(3):172–84.
50. ZhangWang B, Wang JX, *et al*. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014; **513**(7518):382–7.
51. Yang J-Y, Yoshihara K, Tanaka K, *et al*. Predicting time to ovarian carcinoma recurrence using protein markers. *J Clin Invest* 2013;**123**(9):3740–50.
52. Parker R, Vella LJ, Xavier D, *et al*. Phosphoproteomic analysis of cell-based resistance to BRAF inhibitor therapy in melanoma. *Front Oncol* 2015;**5**:95.
53. Wei W, Shin YS, Xue M, *et al*. Single-cell phosphoproteomics resolves adaptive signaling dynamics and informs targeted combination therapy in glioblastoma. *Cancer Cell* 2016;**29**(4): 563–73.
54. Locard-Paulet M, Lim L, Veluscek G, *et al*. Phosphoproteomic analysis of interacting tumor and endothelial cells identifies regulatory mechanisms of transendothelial migration. *Sci Signal* 2016;**9**(414):ra15.
55. Casado P, Alcolea MP, Iorio F, *et al*. Phosphoproteomics data classify hematological cancer cell lines according to tumor type and sensitivity to kinase inhibitors. *Genome Biol* 2013; **14**(4):R37.
56. Casado P, Rodriguez-Prados J-C, Cosulich SC, *et al*. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci Signal* 2013;**6**(268):rs6.
57. Eduati F, Doldàn-Martelli V, Klinger B, *et al*. Drug resistance mechanisms in colorectal cancer dissected with cell type–specific dynamic logic models. *Cancer Res* 2017;**77**(12): 3364–3375.
58. Kuo K-K, Kuo C-J, Chiu C-Y, *et al*. Quantitative proteomic analysis of differentially expressed protein profiles involved in pancreatic ductal adenocarcinoma. *Pancreas* 2016;**45**(1): 71–83.
59. Chung JC, Oh MJ, Choi SH, *et al*. Proteomic analysis to identify biomarker proteins in pancreatic ductal adenocarcinoma. *ANZ J Surg* 2008;**78**(4):245–251.
60. Xiao H, Zhang Y, Kim Y, *et al*. Differential proteomic analysis of human saliva using tandem mass tags quantification for gastric cancer detection. *Sci Rep* 2016;**6**(1):22165.
61. Beretov J, Wasinger VC, Millar EKA, *et al*. Proteomic analysis of urine to identify breast cancer biomarker candidates using a label-free LC-MS/MS approach. *PLoS One* 2015;**10**(11): e0141876.
62. Kimura Y, Yanagimachi M, Ino Y, *et al*. Identification of candidate diagnostic serum biomarkers for Kawasaki disease using proteomic analysis. *Sci Rep* 2017;**7**:43732.
63. Willingale R, Jones DJL, Lamb JH, *et al*. Searching for biomarkers of heart failure in the mass spectra of blood plasma. *Proteomics* 2006;**6**(22):5903–5914.
64. Siebert S, Porter D, Paterson C, *et al*. Urinary proteomics can define distinct diagnostic inflammatory arthritis subgroups. *Sci Rep* 2017;**7**:40473.
65. ZhangChen F, Wang JM, *et al*. A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer. *BMC Proc* 2013;**7**:S10.
66. Rogers M, Clarke A, Noble PJ, *et al*. Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis. *Cancer Res* 2003;**63**(20):6971–83.
67. Chen Y, Zheng S, Yu J, *et al*. Artificial neural networks analysis of surface-enhanced laser desorption/ionization mass spectra of serum protein pattern distinguishes colorectal cancer from healthy population. *Clin Cancer Res* 2004;**10**(24): 8380–85.
68. Luk JM, Lam BY, Lee NPY, *et al*. Artificial neural networks and decision tree model analysis of liver cancer proteomes. *Biochem Biophys Res Commun* 2007;**361**(1):68–73.
69. Ward DG, Suggett N, Cheng Y, *et al*. Identification of serum biomarkers for colon cancer by proteomic analysis. *Br J Cancer* 2006;**94**(12):1898–905.
70. Bouwman FG, de Roos B, Rubio-Aliaga I, *et al*. 2D-electrophoresis and multiplex immunoassay proteomic analysis of different body fluids and cellular components reveal known and novel markers for extended fasting. *BMC Med Genomics* 2011;**4**(1):24.
71. Ostroff RM, Mehan M, Stewart RA, *et al*. Early detection of malignant pleural mesothelioma in asbestos-exposed individuals with a noninvasive proteomics-based surveillance tool. *PLoS One* 2012;**7**:e46091.
72. Petricoin EF, Ardekani AM, Hitt BA, *et al*. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002; **359**(9306):572–7.
73. Ahn HS, Shin YS, Park PJ, *et al*. Serum biomarker panels for the diagnosis of gastric adenocarcinoma. *Br J Cancer* 2012; **106**(4):733–9.
74. Tan CS, Ploner A, Quandt A, *et al*. Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics* 2006;**22**(12):1515–23.
75. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 2002;**99**(10):6567–72.
76. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;**98**(9):5116–21.
77. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**(19): 2507–17.
78. Ressom HW, Varghese RS, Zhang Z, *et al*. Classification algorithms for phenotype prediction in genomics and proteomics. *Front Biosci J Virtual Libr* 2008;**13**(13):691–708.
79. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;**3**:1157–82.
80. Check E. Proteomics and cancer: running before we can walk?. *Nature* 2004;**429**(6991):496–7.
81. Ashburner M, Ball CA, Blake JA, *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**(1):25–29.
82. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**(43):15545–50.
83. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**(D1):D353–61.

84. Creixell P, Reimand J, Haider S, *et al*. Pathway and network analysis of cancer genomes. *Nat Methods* 2015;**12**(7):615–21.
85. Tarca AL, Laurentiu A, Draghici S, *et al*. A novel signaling pathway impact analysis. *Bioinformatics* 2009;**25**(1):75–82.
86. Leiserson MD, Vandin MF, Wu H-T, *et al*. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015;**47**(2): 106–14.
87. Paull EO, Carlin DE, Niepel M, *et al*. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 2013;**29**(21):2757–64.
88. CowenIdeker LT, Raphael BJ, *et al*. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* 2017;**18**:551–62.
89. Vanunu O, Magger O, Ruppin E, *et al*. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010;**6**(1):e1000641.
90. Hofree M, Shen JP, Carter H, *et al*. Network-based stratification of tumor mutations. *Nat Methods* 2013;**10**(11):1108–15.
91. Kim Y-A, Wuchty S, Przytycka TM, Covert MW. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol* 2011;**7**(3):e1001095.
92. Köhler S, Bauer S, Horn D, Robinson PN. Walking the Interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**(4):949–58.
93. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 2010;**26**(8):1057–63.
94. Hakes L, Pinney JW, Robertson DL, *et al*. Protein-protein interaction networks and biology—what's the connection? *Nat Biotechnol* 2008;**26**(1):69–72.
95. Müller T, Schrötter A, Loosse C, *et al*. Sense and nonsense of pathway analysis software in proteomics. *J Proteome Res* 2011;**10**(12):5398–408.
96. Soh D, Dong D, Guo Y, *et al*. Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics* 2010;**11**:449.
97. Yeger-Lotem E, Sharan R. Human protein interaction networks across tissues and diseases. *Front. Genet* 2015;**6**:257.
98. Bossi A, Lehner B. Tissue specificity and the human protein interaction network. *Mol Syst Biol* 2009;**5**:260.
99. Meyer B, Papasotiriou DG, Karas M. 100% protein sequence coverage: a modern form of surrealism in proteomics. *Amino Acids* 2011;**41**(2):291–310.
100. Reinders J, Lewandrowski U, Moebius J, *et al*. Challenges in mass spectrometry-based proteomics. *Proteomics* 2004;**4**(12): 3686–703.
101. Mischnik M, Sacco F, Cox J, *et al*. IKAP: a heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics* 2016;**32**(3):424–31.
102. Wilkes EH, Casado P, Rajeeve V, *et al*. Kinase activity ranking using phosphoproteomics data (KARP) quantifies the contribution of protein kinases to the regulation of cell viability. *Mol Cell Proteom* 2017;**16**(9):1694–704.
103. Yang P, Patrick E, Humphrey SJ, *et al*. KinasePA: Phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis. *Proteomics* 2016;**16**(13):1868–71.
104. Yang P, Zheng X, Jayaswal V, *et al*. Knowledge-based analysis for detecting key signaling events from time-series phosphoproteomics data. *PLoS Comput Biol* 2015;**11**(8):e1004403.
105. Hernandez-Armenta C, Ochoa D, Gonçalves E, *et al*. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* 2017;**33**:1845–51.
106. Ritchie MD, Holzinger ER, Li R, *et al*. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 2015;**16**(2):85–97.
107. Bersanelli M, Mosca E, Remondini D, *et al*. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;**17**(Suppl 2):15.
108. Ruggles KV, Krug K, Wang X, *et al*. Methods, tools and current perspectives in proteogenomics. *Mol Cell Proteomics* 2017;**16**(6):959–81.
109. Drake JM, Paull EO, Graham NA, *et al*. Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell* 2016;**166**(4):1041–54.
110. Rudolph JD, de Graauw M, van de Water B, *et al*. Elucidation of signaling pathways from large-scale phosphoproteomic data using protein interaction networks. *Cell Syst* 2016;**3**(6): 585–593.e3.
111. Hill SM, *et al*. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Methods* 2016;**13**(4):310–18.
112. Jolliffe IT. *Principal Component Analysis*, 2nd edn. New York: Springer-Verlag New York, Inc., 2002.
113. Vaske CJ, Benz SC, Sanborn Z, *et al*. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 2010;**26**(12): i237–45.
114. Žitnik M, Zupan B. Data fusion by matrix factorization. *IEEE Trans Pattern Anal Mach Intell* 2015;**37**(1):41–53.
115. Fusi N, Elibol HM, Probabilistic Matrix Factorization for Automated Machine Learning. arXiv preprint arXiv:1705.05355 2017.
116. Liu Y, Devescovi V, Chen S, *et al*. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Syst Biol* 2013;**7**:14.
117. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lunch cancer subtype analysis. *Bioinformatics* 2009;**25**(22):2906–12.
118. Persson O, Krogh M, Saal LH, *et al*. Microarray analysis of gliomas reveals chromosomal position-associated gene expression patterns and identifies potential immunotherapy targets. *J Neurooncol* 2007;**85**(1):11–24.
119. Furey TS, Cristianini N, Duffy N, D, *et al*. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000; **16**(10):906–14.
120. Theilhaber J, Connolly T, Roman-Roman S, *et al*. Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome Res* 2002;**12**(1): 165–76.
121. Alexandrov LB, Nik-Zainal S, Wedge DC, *et al*. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 2013;**3**(1):246–59.
122. Zhang S, Liu C-C, Li W, *et al*. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res* 2012;**40**(19):9379–91.
123. Yuan Y, Van Allen EM, Omberg L, *et al*. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* 2014;**32**(7):644–52.
124. Li Y, Ngom A. The non-negative matrix factorization toolbox for biological data mining. *Source Code Biol Med* 2013;**8**(1):10.
125. Torres-García W, Zhang W, Runger GC, *et al*. Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics* 2009;**25**(15):1905–14.

126. Li F, Nie L, Wu G, *et al*. Prediction and characterization of missing proteomic data in *Desulfovibrio vulgaris*. *Comp Funct Genomics* 2011;**2011**:78073.
127. Meng C, Kuster B, Culhane AC, *et al*. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 2014;**15**:162,
128. Liu G, Zhang J, Larsen B, *et al*. ProHits: an integrated software platform for mass spectrometry-based interaction proteomics. *Nat Biotechnol* 2010;**28**(10):1015–17.
129. Pfeuffer J, *et al*. OpenMS – A platform for reproducible analysis of mass spectrometry data. *J Biotechnol* 2017;**261**:142–8.
130. Röst HL, Liu Y, D'Agostino G, *et al*. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat Methods* 2016;**13**(9):777–83.
131. Navarro P, Kuharev J, Gillet LC, *et al*. A multi-center study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* 2016;**34**(11):1130.
132. Martinez-Jimenez C, Pilar P, Eling CN, *et al*. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* 2017;**355**(6332):1433–6.
133. Halasz M, Kholodenko BN, Kolch W, *et al*. Integrating network reconstruction with mechanistic modeling to predict cancer therapies. *Sci Signal* 2016;**9**(455):ra114.
134. Zeevi D, Korem T, Zmora N, *et al*. Personalized nutrition by prediction of glycemic responses. *Cell* 2015;**163**(5):1079–94.
135. Katsila T, Spyroulias GA, Patrinos GP, Matsoukas M-T. Computational approaches in target identification and drug discovery. *Comput Struct Biotechnol J* 2016;**14**:177–84. May
136. Kramer R, Cohen D. Functional genomics to new drug targets. *Nat Rev Drug Discov* 2004;**3**(11):965–72.
137. Terstappen GC, Schlüpen C, Raggiaschi R, Gaviraghi G. Target deconvolution strategies in drug discovery. *Nat Rev Drug Discov* 2007;**6**(11):891–903.
138. Gerstung M, Papaemmanuil E, Martincorena I, *et al*. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat Genet* 2017;**49**(3):332–40.
139. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat* 2012;**33**(5):777–80.
140. Juengst E, McGowan ML, Fishman JR, *et al*. From 'personalized' to 'precision' medicine: the ethical and social implications of rhetorical reform in genomic medicine,. *Hastings Cent Rep* 2016;**46**(5):21–33.
141. Dzau VJ, Ginsburg GS. Realizing the full potential of precision medicine in health and health care. *JAMA* 2016;**316**(16):1659–60.
142. Glimelius B, Lahn M. Window-of-opportunity trials to evaluate clinical activity of new molecular entities in oncology. *Ann Oncol* 2011;**22**(8):1717–25.
143. Schmitz S, Duhoux F, Machiels J-P. Window of opportunity studies: do they fulfil our expectations? *Cancer Treat Rev* 2016;**43**:50–57.
144. Hay M, Thomas DW, Craighead JL, *et al*. Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014;**32**(1):40–51.
145. Gill SK, Christopher AF, Gupta V, Bansal P. Emerging role of bioinformatics tools and software in evolution of clinical research. *Perspect Clin Res* 2016;**7**(3):115–22.
146. McShane LM, Cavenagh MM, Lively TG, *et al*. Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration. *BMC Med* 2013;**11**(1):220.
147. Stenehjem DD, Bellows BK, Yager KM, *et al*.. Cost-utility of a prognostic test guiding adjuvant chemotherapy decisions in early-stage non-small cell lung cancer. *Oncologist* 2016;**21**:196–204.
148. St Jude Children's Research Hospital. St Jude's Family Advisory Council: PGEN4Kids Study Information, 2012. https://s.stjude.org/multimedia/PG4KDS/PGEN4Kid.html
149. Gottesman O, Scott SA, Ellis SB, *et al*. The CLIPMERGE PGx program: clinical implementation of personalized medicine through electronic health records and genomics - pharmacogenomics. *Clin Pharmacol Ther* 2013;**94**(2):214–17.