Computational Analysis of Phosphoproteomics Data in Multi-Omics Cancer Studies

Giulia Mantini, Thang V. Pham, Sander R. Piersma, and Connie R. Jimenez*

Multiple types of molecular data for the same set of clinical samples are increasingly available and may be analyzed jointly in an integrative analysis to maximize comprehensive biological insight. This analysis is important as separate analyses of individual omics data types usually do not fully explain disease phenotypes. An increasing number of studies have now been focusing on multi-omics data integration, yet not many studies have included phosphoproteomics data, an important layer for understanding signaling pathways. Multi-omics integration methods with phosphoproteomics data are reviewed in the context of cancer research as well as multi-omics methods papers that would be promising to apply to phosphoproteomics data. Analysis of individual data types is still the major approach even in large cohort proteogenomics studies. Hence, a section is dedicated on possible integrative methods for multi-omics and phosphoproteomics data. In summary, this review provides the readers with both currently used integrative methods previously applied to phosphoproteomics and multi-omics data integration and other algorithms for multi-omics data integration promising for future application to phosphoproteomics data.

1. Introduction

In the era of comprehensive, unbiased molecular profiling, global analysis of post-translational modifications is of high interest for obtaining functional insights. One of the most common post-translational modifications that is involved in cell regulation and intracellular signal transduction is reversible protein phosphorylation catalyzed by protein kinases. Through site-specific phosphorylation of protein substrates, kinases may modulate protein conformation, subcellular localization, protein–protein interaction, and signaling important for development, survival, and growth of all living beings. Importantly, signaling via tyrosine kinases is often deregulated in cancer as these enzymes mediate most growth and survival signaling in multicellular organisms. This knowledge has led to the development of selective kinase inhibitors and important clinical achievements in cancer therapy. Examples of clinically successful kinase inhibitors include imatinib to block BCR-ABL fusion kinase in chronic myeloid leukemia, trastuzumab for ERBB2 blockade in HER2-positive breast cancer, crizotinib for ALK inhibition in anaplastic lymphoma, and gefitinib to inhibit EGFR activity in lung cancer.

In recent years, focus in precision medicine has shifted from identification of single genomic events to a more comprehensive assessment of tumor biology through integration of genomic, transcriptomic, and proteomic data. Advances in the past decade in biochemical procedures for phosphopeptide enrichment, mass spectrometry technology, and computational approaches for phosphosite identification and quantification have enabled the application of phosphoproteomics for the large-scale analysis of clinical materials. Currently, phosphoproteome data analysis and its integration into a multi-omics framework are challenging. Many tools simply collapse quantitative information from multiple phosphosites into a single protein phosphorylation value. This approach does not do justice to the site-specific phosphorylation behavior involved in signaling networks. Different sites in a protein have different functional consequences (e.g., activating vs inhibiting phosphosites) that need to be preserved in multi-omics data integration.

In this review, we describe current methods and tools for phosphoproteomics analysis as well as promising new methods for the integration of phosphoproteomics with other –omics data, taking into account the enormous complexity of the data.

2. Complexity of Phosphoproteome Data

Almost 500 000 phosphorylation sites are included in databases such as PhosphoSitePlus (PSP) and approximately 500 kinases and 200 phosphatases are encoded in the human genome that enable rapid regulation. Protein...
phosphorylation leads to a reversible cascade of signaling events important for cell maintenance, proliferation, and survival. One kinase or phosphatase may have multiple substrates. These substrates can be phosphorylated at serine, threonine, or tyrosine residues. Each site may be targeted by several kinases, depending on sequence context, resulting in an intricate network of kinases, target proteins, and phosphatases. Bioinformatics tools such as PhosphoPredict, NetworKIN, and Phospho.ELM are able to predict kinase-specific substrates based on motif analyses, yet functional annotation of the majority of phosphosites is incomplete. Thus, most research to date has been restricted to clustering and enrichment methods. These two methods detect sites according to their quantitative behavior (clustering) and relate the mapped proteins to cellular processes like canonical pathways, transcriptional networks, or disease mechanisms via statistical testing against databases (enrichment). Another limitation is that modeling of phosphoproteomics networks is usually limited to few components where phosphosite information is collapsed into peptides or phosphoproteins prior to obtaining functional annotation. Thus, the phosphosite information is usually lost.

3. Current State-of-the-Art in Phosphoproteomics Data Analysis

3.1. Quantitative Phosphoproteomics

To enable optimal detection in large-scale mass spectrometry-based phosphoproteomics experiments, phosphopeptides need to be enriched from the cell lysate after tryptic digestion. Global phosphopeptide enrichment (phosphoserine, phosphothreonine, and phosphotyrosine) is achieved using affinity purification materials, including titanium dioxide beads (TiO₂) or immobilized metal affinity chromatography (IMAC) resins like Ti-IMAC and Fe(III)-IMAC. Phosphotyrosine-containing peptides can be enriched specifically using selectiv sequence-context independent antibodies including 4G10, pY20, and pTyr-1000. After enrichment, phosphopeptides are submitted to C18 nano-liquid chromatography coupled on-line to high-resolution/high mass accuracy tandem mass spectrometry for phosphopeptide identification and phosphosite localization. For phosphoproteomics, several fragmentation methods have been reported, including collision-induced (CID) fragmentation and higher-energy dissociation (HCD) that both suffer from neutral loss of threonine and serine phosphopeptides and electron-transfer dissociation (ETD) and combinations like ETD-HCD that result in more complete phosphopeptide sequence coverage. Following LC-MS/MS data acquisition, peptides and phosphopeptides are identified by a search engine like MaxQuant or Mascot. Correct phosphosite localization, assigning a phosphorylation probability to each potential phosphosite (serine, threonine, or tyrosine) in a given phosphopeptide, requires optimized algorithms like A-score, MaxQuant PTM-score, PhosphoRS, or Mascot delta score. Upon phosphopeptide identification and phosphosite localization, differential analysis or pathway mapping requires phosphopeptide or site quantification. Several proteomics strategies are available for accurate and precise phosphorylation quantification, including label-free quantitation, stable-isotope in cell culture (SILAC), and isotopic labeling like iTRAQ and TMT, and their relative performance has been assessed showing label-free quantitation and SILAC to be most accurate in a mixed-species analysis while TMT MS2 resulted in the deepest data at the expense of accuracy due to ratio-compression. Label-free quantitation is advantageous for large-scale clinical proteomics of heterogeneous samples and allows for single sample kinase activity analysis methods like integrative inferred kinase activity (INKA) discussed in Section 3.2.2 of this review. Recently, label-free phosphosite quantitation has been developed for data-independent analysis (DIA) and now includes a dedicated site localization algorithm that shows superior performance in phosphosite depth and quantitative accuracy over data-dependent analysis (DDA).

3.2. Strategies for Kinase Activity Analysis

To determine kinase activities and nominate drug targets, several kinases ranking approaches have been previously described and can be characterized as group-based or single sample approach. For both approaches, dedicated data visualization is key.

3.2.1. Group-Based Analysis Tools

A substrate-centered approach to infer upstream kinase activity named kinase-enrichment analysis (KEA) was first reported by Lachmann et al. in 2007 and further improved into kinase-substrate enrichment analysis (KSEA) to systematically infer the activation of given kinase pathways by Casado et al. in 2013. This computational approach was then converted into a web-based tool by Wiredja et al. in 2017. To identify the substrates for each kinase, the KSEA App sources K-S annotations from PSP by default, but users have the option to include NetworKIN annotations adjusting the inclusion stringency. In 2014, another tool named PSEA (phosphorylation set enrichment analysis) was released from Suo et al. where kinases characteristic of all collected disease-related phosphorylation substrates were analyzed using a kinase-specific prediction method. In the same year, Weidner et al. developed PHOXTRACK (phosphosite-X-Tracing analysis of casual kinases). In contrast to the other tools, this user-friendly software uses the full set of quantitative proteomics data and applies non-parametric statistics to calculate differences in phosphosites based on different biological conditions. Recently, Mischkiewicz and colleagues introduced a machine-learning approach named inference of kinase activities from phosphoproteomics (IKAP) to estimate the activities of all kinases that are known to phosphorylate at least one phosphosite in a phosphoproteomics dataset. In contrast to KSEA, the output of IKAP is not only a score for the activity of a kinase, but also a value representing the strength of a kinase–substrate interaction in the investigated cell type.

Next to scoring kinase activity, visualizing kinases in a signaling pathway is also of great relevance. Pathway analysis of post-translational modifications (PTM) data is typically performed at a gene-centric level because of the lack of appropriately curated PTM signature databases. In 2019, Krug and colleagues presented the first version of PTM Signature Database (PTMSigDB). The database contains a rich data source consisting of site-specific signatures of perturbations, kinase activities,
and signaling pathways curated from more than 2500 publications. The authors adapted the single sample Gene Set Enrichment Analysis approach to utilize PTMSigDB for a PTM Signature Enrichment Analysis (PTM-SEA) of quantitative MS data. PTMSigDB is the only available tool able to run a pathway enrichment analysis related to phosphorylated sites. The result is similar to a regular gene-set enrichment analysis where enriched pathways are ranked based on false discovery rate (FDR) score. The tool can only be applied for a group-comparison experiment.

3.2.2. Single-Sample Analysis Tools

To pinpoint hyperphosphorylated kinases as drug targets, Rikova et al.[21] performed large-scale phosphoproteomics of ∼200 lung cancer cell lines and clinical samples and sorted tyrosine kinases in single cancer samples on the basis of the sum of the spectral counts for all phosphopeptides attributed to a given kinase. The phosphokinases were then ranked by dividing the sum of phosphopeptide spectral counts by the number of samples with “non-zero” counts. Using this simple but effective approach, this pioneering 2007 study identified known and novel oncogenic kinases in lung cancer. This type of analysis can be performed in individual samples, but is limited by a focus on phosphorylation of the kinase itself, and does not include its substrates.

Recently, two tools have been reported to infer kinases activities in a single-sample manner: kinase activity ranking using phosphoproteomics data (KARP) (Wilkes et al., 2017) and INKA (Beekhof et al., 2019). KARP measures the net kinase activity as the sum of the intensities of its known substrates relative to the sum of the intensities of all phosphorylation sites present in the dataset. The result of such analysis is called K-score. INKA combines label-free kinase-centric (phosphokinase and activation loop data) and substrate-centric information from NetworKIN and PhosphoSitePlus to rank kinase activity (INKA score) in a single sample manner. INKA has been extensively benchmarked and applied, demonstrating that INKA scoring can i) identify oncogenes in cancer cell lines with known genomic aberrations, ii) identify differential kinase activity in models (wild-type vs mutant, +/- drug) and clinical samples (pre- and on-treatment tumor needle biopsies), iii) be correlated to drug sensitivity data, and iv) enable drug selection. The latter was shown for patient-derived tumor xenografts with functional testing in organoids by Beekhof et al. 2019[31] and for AML cell lines by Van Alphen et al. 2020.[52] The algorithm, easily accessible by inkascore.org website or in R language, requires MaxQuant output tables and enables a per-sample visualization of ranked kinase activities as bar graph and kinase–substrate network. Moreover, benchmarking INKA and KARP on a set of cancer cell lines with drug sensitivity data, INKA outperformed KARP in assigning a highest score to the known drivers.

Besides the extensive attention to kinase–substrate enrichment and kinase activity ranking, the functional role of many phosphosites remains to be elucidated. Recently, Ochoa et al.[53] applied a machine learning approach to public datasets and constructed a curated functional landscape of > 100 K phosphosites. Moreover, Ochoa condensed 59 features indicative of proteomic, structural, regulatory, or evolutionary relevance into a single functional score that can be used to identify regulatory phosphosites across different molecular mechanisms, processes, and diseases. This first study can be the basis of future functional studies in the phosphoproteomics field.

3.2.3. Tools for Phospho-Data Visualization

Cytoscape[54] is the most widely used software for network analysis and visualization with a well maintained dedicated app store[55] with more than 500 applications, among which the new published Cytoscape app “Omnis Visualizer”[56] is particularly useful for visualizing phosphoproteomics. Researchers and developers around the world have contributed new functionalities and tailored solutions for a broad range of network-related projects.

Raaijmakers et al. developed a Cytoscape plugin named “PhospoPath” in 2015.[57] The app was designed to visualize and analyze quantitative proteome and phosphoproteome datasets by importing publicly available data and allowing user to visualize specific phosphosites. However, this app can only be used for group-comparison experiments and data input needs to be ready upfront following the rigorous app-guidelines. In the following years, Narushima et al. developed another Cytoscape app “PTMapper” to integrate PPI data with publicly available kinase–substrate relations at the resolution of phosphorylated amino acid residues.[58] One year later, “PTMOracle” was released from Tay et al.[59] allowing the user to visualize and analyze their own PTM data.

Besides Cytoscape, aforementioned tools such as INKA and PTM-SEA provide both phospho-data analysis and visualization (see Section 3.2.2 for details). Additional tools for phospho-data visualization and databases sources are listed in Table 1.

However, none of the above-mentioned tools reflect sufficiently the complexity of the phosphoproteomics layer when analyzing multi-omics data. In many –omics integration studies, results from phosphoproteomics data are usually shown alone to avoid complexity in the figure. After accurately describing the data integration method, a small and centered signaling pathway is usually shown, highlighting the expression of phosphoproteins but more often losing the phosphosite information. A rather complete description and visualization of multi-omics data centered to the phosphoproteome level and more in detail to the phosphosite information is shown in Figure 1. The figure represents a hypothetical interaction of kinase, phosphatase, and substrates with multi-omics information (protein and gene abundances, phosphoproteomics signaling, mutations, and methylations) making use of kinase activity annotation from already existing tools such as INKA.

4. Phosphoproteomics Analysis in a Multi-Omics Context

Recently, multiple studies showed that proteogenomic integration not only prioritized known genomics targets but also yielded to novel findings for therapeutic hypothesis generation.[11–15] It stands to reason that the integration of phosphoproteomics, with global proteomics and genomics, may provide greater clues to these signaling events than single omics.
Table 1. Repositories and visualization tools for phosphoproteomics data.

| Name            | Computational platform | Description                                                                 | Availability                      | Reference |
|-----------------|------------------------|----------------------------------------------------------------------------|------------------------------------|-----------|
| PRIDE           | Web                    | Proteomics                                                                 | https://www.ebi.ac.uk/pride/archive/ | EMBL-EBI & Elixir |
| LinkedOmics     | Web                    | TCGA and CPTAC web portal                                                | http://www.linkedomics.org/login.php | [97]      |
| OmicsDI         | Web                    | Contains data sets from 11 repositories in a common data structure        | https://www.omicsdi.org/           |           |
| cBioPortal      | Web                    | Data repository, analysis and visualization                               | https://www.cbiointerface.org      | [96]      |
| TCPA            | Web                    | Data repository, analysis and visualization                               | https://tcpaportal.org             |           |
| PeptideAtlas repository | Web               | Proteomics                                                                 | http://www.peptideatlas.org/repository/ | Institute for System Biology |
| PathVisio       | Local                  | Integrative visualization of pathways                                     | https://github.com/PathVisio/pathvisio/releases | Maastricht University |
| G6G             | Web                    | Collection of commercially available and free tools for omics analysis    | https://www.g6g-softwaredirectory.com | Al Shapiro G6G Tech |
| qPhos           | Web                    | Curated quantitative phosphoproteome datasets in human tissues and cell lines from published literatures | http://qphos.cancerbio.info/index.php | [112]      |
| PhosphoSitePlus | Web                    | Multiple type of phosphosite information                                 | https://www.phosphosite.org/homeAction | Cell Signalling Technology |
| dBPAF           | Web                    | Integrative database of protein phosphorylation in animals and fungi      | http://dbpaf.biocuckoo.org/index.php | [113]      |
| PTMapper        | Cytoscape              | Integrate PPI data with publicly available kinase–substrate relations at the resolution of phosphorylated amino acid residues | https://github.com/y-narushima/PTMapper | [18]       |
| iPTMnet         | Web                    | Integrative bioinformatics approach combining text mining, data mining, and ontological representation to capture rich PTM information, including PTM enzyme–substrate–site relationships, PTM-specific protein–protein interactions (PPIs), and PTM conservation across species | https://research.bioinformatics.udel.edu/iptmnet/ | [102]      |
| x2kweb          | Web                    | X2K Web infers upstream regulatory networks from signatures of differentially expressed genes. X2K Web produces inferred networks of transcription factors, proteins, and kinases predicted to regulate the expression of the inputted gene list. | https://amp.pharm.mssm.edu/X2K/ | [114]      |
| phosphoPath     | Cytoscape              | Visualization and analysis of quantitative proteome and phosphoproteome datasets. PhosphoPath imports publically available data from BioGrid, PhosphositePlus, and Wikipathways. PhosphoPath further enables PTM site-specific visualization of information from quantitative time series. | http://apps.cytoscape.org/apps/phosphopath | [57]       |
| PoGo            | Local                  | PoGo allows the visualization of protein sequences and the corresponding gene with mutation, CNV, PTM, and isoforms in one single environment (IGV) | http://www.sanger.ac.uk/science/tools/pogo | [115]      |
Landmark cancer molecular profiling programs such as genomics (TCGA)\(^6\) and proteomics (CPTAC,\(^61\) TCPAv3.0\(^62\)) provide multi-omics characterization of individual patients. Combining data from proteomics, phosphoproteomics, and metabolomics with genomics and transcriptomics helps to prioritize disease drivers and channel the attention to specific signaling that can explain the investigated phenotype with the common goal to improve diagnosis, treatment, and cancer prevention. Application studies to the analysis of cancer that include phosphoproteomics are summarized in Table 2.

Below we discuss large scale proteogenomics studies focused on multi-omics data integration with phosphoproteomics expanding on specific methods based on network modules and correlation analysis. Moreover, we report on other methods for multi-omics data integration. Those methods are currently not applied to phosphoproteomics and multi-omics data and provide novel opportunities to integrate this data type. Lastly, we mention a list of public multi-omics databases to analyze genomics and proteomics data without any computational knowledge, thus, easily accessible to biologists.

### 4.1. Proteogenomics

In a typical proteomics study, mass spectrometry data are matched against peptide sequences in a reference proteome database. However, mutated peptides cannot be identified when using a reference proteomics database. This problem can be avoided using an updated database that includes human protein cancer variants\(^61\) or through the integration of multiple data types, collectively called “proteogenomics”. Proteogenomics integrates genome, transcriptome, and proteome datasets allowing for correlations of mRNA and protein pairs across samples. Having a general overview of upstream data (genomics and/or...
Table 2. Application studies with phosphoproteomics data integration.

| Data type                          | Title                                                                 | Algorithm       | Type of cancer | Comments                                                                                                                                                                                                 | Phosho-data level | Type of integration | Reference |
|-----------------------------------|-----------------------------------------------------------------------|-----------------|----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|--------------------|-----------|
| CNV, mutation, mRNA, phosphoproteins | Phosphoproteome Integration reveals Patient-Specific Networks in Prostate Cancer | TieDIE          | Prostate cancer | The TieDIE algorithm computes a subnetwork of gene and protein interactions. The pathways derived in this way predict interlinking genes that may correspond to essential components of cancer signaling                                                                 | p-Phospho level   | Network based      | [11]      |
| SCNA, mutation, mRNA, proteins, phosphoproteins | Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities | /               | Colon cancer   | Proteogenomic integration not only prioritized genomically inferred targets, such as copy-number drivers and mutation-derived neoantigens, but also yielded novel findings. This study demonstrated the utility of proteogenomics in therapeutic hypothesis generation                                                                 | p-Site level and association with mutational status | /                  | [12]      |
| Mutation, CNA, mRNA, proteins, phosphoproteins | Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma | /               | HBV-related hepatocellular carcinoma | Joint random forest to join two co-expression networks                                                                                                                                            | p-Protein level   | /                  | [13]      |
| mRNA, epigenomics, proteins, phosphoproteins | Identification of Therapeutic Targets in Rhabdomyosarcoma through Integrated Genomic, Epigenomic, and Proteomic Analyses | WGCNA           | Rhabdomyosarcoma | WGCNA with DE protein and DE phosphosite                                                                                                                                                    | p-Site level      | ad hoc gene ranking | [14]      |
| DNA, mRNA, proteins, phosphoproteins | Proteogenomic Characterization of Human Early-Onset Gastric Cancer | iClusterPlus, orthogonal non-negative matrix factorization (ONMF), PLS-DA | Gastric cancer | iClusterPlus on DE mRNA, phosphoproteins, and PLS-DA to identify association of phosphopeptide mutations                                                                                          | MAD score on p-peptide level | matrix-factorization/cluster modeling | [15]      |
| mRNA, phosphoproteins | Integrated Phosphoproteome and Transcriptome Analysis Reveals Chlamydia-Induced Epithelial-to-Mesenchymal Transition in Host Cells | /               | Chlamydia<sup>a</sup> | p-Peptides in MotifX for kinase upstream regulators. Kinase-substrate prediction for up- and down-regulated phosphosite                                                                                   | p-Peptide and p-site | /                  | [99]      |

(Continued)
Table 2. Continued.

| Data type                          | Title                                                                 | Algorithm              | Type of cancer       | Comments                                                                 | Phospho-data level | Type of integration | Reference |
|-----------------------------------|-----------------------------------------------------------------------|------------------------|----------------------|--------------------------------------------------------------------------|--------------------|---------------------|-----------|
| WXS and phosphoproteins           | Proteogenomic systems analysis identifies targeted therapy resistance mechanisms in EGFR-mutated lung cancer | Dcglob                 | EGFR-mutated lung cancer | A modified version of the previously published algorithm DCglob          | Differential covariances of p-sites | /                      | [100]     |
| DNA-seq, mRNA, phosphoproteins    | Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling | Prize-collecting Steiner tree (PCST) | GBM                  | An important aspect of the PCST algorithm is that it is able to naturally account for missing data and false positives. As a result, the final networks are compact, enriched for functionally relevant proteins and the most reliable interactions that include these proteins, and can be used to guide subsequent experiments. | p-Peptide level   | Network based       | [76]      |
| mRNA, proteins, phosphoproteins   | Integrated systems biology analysis of KSHV latent infection reveals viral induction and reliance on peroxisome mediated lipid metabolism | Prize-collecting Steiner tree (PCST) | Kaposi’s Sarcoma associated Herpesvirus | This algorithm identifies Steiner nodes, which are proteins that were not implicated in the proteomic or TF analyses but form crucial connections between other important proteins. From OmicsIntegrator R package | p-Peptide level   | Network based       | [72]      |
| mutation, CNA, mRNA, proteins, phosphoproteins | Proteogenomics connects somatic mutations to signaling in breast cancer | Consensus clustering | Breast cancer | The breadth and depth of proteomic and phosphoproteomic analyses displayed in this study demonstrates the strength of mass-spectrometry-based proteomics | p-Site level       | Correlations        | [65]      |
| mRNA, proteins, phosphoproteins   | Reconstructing targetable pathways in lung cancer by integrating diverse omics data | Abundance score and Prize-Collecting Steiner Tree (PCST) | Non-small cell lung cancer | Using this approach on 13 KRAS-mutant NSCLC cell lines known to be KRAS-D or KRAS-Ind, our integrative analysis nominated 115 proteins that were differentially abundant between these two groups (Hochberg-adjusted p-value < 0.05) | p-Protein level    | Network based       | [80]      |
| Mutation, CNV, mRNA, proteins, phosphoproteins | Proteogenomic integration reveals therapeutic targets in breast cancer xenografts | /                      | Breast cancer       | Collapsed the phosphosites to gene-level phosphorylation values by averaging the phosphosite expressions observed for each gene and converted to z-score | p-Site to gene-level | /                    | [101]     |
### Table 2. Continued.

| Data type | Title | Algorithm | Type of cancer | Comments | Phospho-data level | Type of integration | Reference |
|-----------|-------|-----------|----------------|----------|--------------------|---------------------|-----------|
| mRNA, proteins, phosphoproteins | Proteomic and phosphoproteomic comparison of human ES and iPS cells | / | Pluripotent stem cells<sup>1</sup> | Grouped peptides containing multiple sites with other peptides containing the exact same combination of sites. Therefore, we presented a list of phosphorylation isoforms rather than a list of phosphorylated sites. | From p-site to p-isoforms | / | [102] |
| WES, mRNA, circRNA, miRNA, proteins, phosphoproteins | Proteogenomic Characterization of Endometrial Carcinoma | / | Endometrial carcinoma | Kinase activity measure and p-peptide collapsed to gene-level for PTM-outlier | p-Peptide collapsed to gene-level for PTM-outlier | / | [103] |
| Methylation, mRNA, proteins, phosphoproteins | Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling | Similarity network fusion (SNF) | Medulloblastoma | This method, described in (Wang et al., 2014), consists of first constructing patient-to-patient similarity networks for each of the available data types and, in a second step, efficiently integrating and fusing them into a final consensus network. | p-Peptide level | Network based | [104] |
| Methylation, mRNA, proteins, phosphoproteins | Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups | Prize-collecting Steiner tree (PCST) | Medulloblastoma | / | p-Peptide level | Network based | [16] |
| mRNA, proteins, phosphoproteins | Insights into Impact of DNA Copy Number Alteration and Methylation on the Proteogenomic Landscape of Human Ovarian Cancer via a Multi-omics Integrative Analysis | iProFun | Ovarian cancer | iProFun takes as input the association summary statistics from associating CNAs and methylations of genes to each type of cis-molecular trait, aiming to detect the joint associations of DNA variations and molecular traits in various association patterns | p-Protein level | Multiple linear regression | [17] |
| Genome, methyolome, mRNA, miRNA, proteins, phosphoproteins | A multi-omic analysis of human naive CD4+ T cells | / | Human naive CD4+ T cells | / | p-Site level | / | [105] |
| Proteins, phosphoproteins | Integrating proteomic and phosphoproteomic data for pathway analysis in breast cancer | / | Breast cancer | Integration of proteomics and phosphoproteomics data to perform pathway analysis | p-Protein level | / | [106] |
Table 2. Continued.

| Data type | Title | Algorithm | Type of cancer | Comments | Phospho-data level | Type of integration | Reference |
|-----------|-------|-----------|----------------|----------|--------------------|---------------------|-----------|
| Proteins, phosphoproteins | Integrated analysis of global proteome, phosphoproteome, and glycoproteome enables complementary interpretation of disease-related protein networks | / | Gastric cancer | / | p-Peptide level | / | [107] |
| mRNA, proteins, phosphoproteins | A System-wide Approach to Monitor Responses to Synergistic BRAF and EGFR Inhibition in Colorectal Cancer Cells | / | Colorectal cancer | / | p-Site level | / | [108] |
| CNA, mRNA, proteomics, phosphoproteomics | Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer | WGCNA for subtype clustering and spearman correlation for proteomics and CNA | Ovarian cancer | / | p-Peptide level | / | [66] |

*Epidemiologically associated with cervical and ovarian cancers.*

transcriptomics) and downstream data (proteomics and/or phosphoproteomics) facilitates understanding how post-translational modifications and specific signaling pathways are altered by upstream events such as genetic variants (eQTL), microRNAs, or copy number aberrations. We will now focus on large-scale proteogenomics studies that include phosphoproteomics data and their application methods.

In 2014, Zhang and colleagues[64] performed the first large-scale proteogenomics study analyzing proteomes of 95 colon and rectal tumors previously characterized by TCGA. This analysis demonstrated how proteomics data can enable prioritization of cancer drivers in amplified genomic regions and enable identification of proteomic subtypes, that were only partially overlapping with transcriptomics-based subtypes. These proteomics subtypes also associate differently to clinical outcome. The phosphoproteomic information layer was generated in subsequent cancer proteogenomic studies of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) network. Vasaikar et al. performed a deep molecular characterization of human colon cancer and paired normal adjacent tumors showing that phosphoproteomic integration may correct inaccurate genomics data-based inferences and lead to unexpected discoveries and therapeutic opportunities.[12] This proteogenomic study not only prioritized genomically inferred targets but also provided novel findings. Phosphoproteomics data associated Rb phosphorylation with increased proliferation and decreased apoptosis suggesting this protein known as tumor suppressor also played a role as a putative cancer driver.

Other large-scale proteogenomics studies of the CPTAC network that included phosphoproteomics focused on breast, liver, and ovarian cancer.[64–66] Mertins and colleagues described the integration of genomic and (phospho)proteomics information of 105 genomically annotated breast cancers. This analysis also showed the power of proteogenomics to pinpoint novel oncogenic driver genes. First, mRNA and protein/phosphoprotein abundance correlations for each gene/protein pair were analyzed. Identified proteins were mapped to gene names while phosphosites were aggregated to their corresponding protein identifier by calculating the median log-ratio for all sites arising from the protein. Phosphoproteins collapsed to gene names, were subjected to single sample GSEA (ssGSEA) and resulted pathways were clustered by consensus clustering (same approach was used for proteomics). This multi-omics analysis revealed metabolic functions such as amino acid, sugar, and fatty acid metabolism, particularly enriched among positively correlated genes whereas ribosomal, RNA polymerase, and mRNA splicing functions were negatively correlated. An extra analysis was performed for phopsophotrypticides: outlier kinase analysis that was carried out to compare the kinase at the level of DNA, RNA, and protein expression within each sample. In this breast cancer study,[65] only the phosphoproteome data shed light on a
G-protein-coupled receptor cluster that was not easily identified from transcriptomics data.

Gao et al. demonstrated that phosphoproteome integration into a multi-omics context of proteomics, transcriptomics, and genomics data revealed alterations of metabolic pathways among the most dramatic differences between tumor and non-tumor hepatitis B virus (HBV)-related liver tissues. These results were given by functional analysis done on the subtype classifications based on transcriptomics, proteomics, and phosphoproteomics data. Again, the data streams were analyzed first separately, then significance was estimated of the pathway enrichment scores across the datasets using limma statistics to get a single final score for each pathway. Together with metabolic alterations, proliferation and microenvironmental dysregulation signaling stratified patients into three distinct liver cancer subgroups.\[31\]

Zhang and colleagues presented a proteogenomic integration on 174 high-grade ovarian carcinomas yielding to several insights regarding: i) copy-number alterations that influence the proteome; ii) proteins associated with chromosomal instability; iii) patient stratification by specific protein acylations. Consensus clustering was applied to compare proteomics subtypes to previous subtypes identified by TCGA. Phosphoproteomics data were mainly used to compare pathway activity between short and long survivors. To avoid the issue of ranking pathways based on phosphopeptides unique to a single pathway and phosphopeptides that map to proteins shared between pathways, Zhang and colleagues performed two analyses: one of all proteins associated with specific pathway and a second analysis in which proteins shared between multiple pathways were excluded. Next, they ranked the pathways based on proteomics, transcriptomics, and CNA information. The results of this parallel analysis of multiple data types showed that RhoA signaling, PDGFRB, integrin-like kinases, NOTCH, and Her2 signaling were more active in patients having short survival. The most significant contribution was given by phosphoproteomics data. Together, these studies highlighted the importance of the phosphoproteomics layer for drug target prediction and association with clinical outcome.

In short, large-scale proteogenomics cancer studies represent an outstanding resource of multi-omics data and exemplify the current state-of-the-art in applied integrative methods. In the next section, we review more in detail the tools used in integrative studies when dealing with phosphoproteomics data that have been applied or not to large scale proteogenomics studies.

### 4.2. Data-Driven Discovery of Signaling Network Modules

Phosphoproteomics can inform on the state of intracellular phosphorylation both at the pathways level and at the whole cell level. Cascades of protein phosphorylation induced by protein kinases comprise a core mechanism in the integration and propagation of intracellular signals. Hence, a phosphoproteomics layer is valuable in aiding a data-driven discovery of the activation state of signaling network modules.

OmicsPASS is a tool to search for predictive subnetworks consisting of molecular interactions within and between related omics data types in a supervised analysis setting.\[67\] Based on user-provided network data and phenotype annotations, iOmicsPASS computes a score for each molecular interaction, and applies a modified nearest shrunken centroid algorithm to the scores to select densely connected subnetworks that can accurately predict each phenotypic group. iOmicsPASS analysis of TCGA/CPTAC with phosphoproteomics breast cancer data\[67\] highlighted a new transcriptional regulatory network underlying the basal-like subtype as positive protein markers, a result not seen through analysis of individual omics data.

Another network-based approach is Mixed Network Integration (MiNETi) developed by Santra et al.\[68\] on HeLa cell lines. This algorithm operates in four stages. First, it analyzes protein interactome, phosphoproteomics, and transcriptomics data to reconstruct protein-protein interaction (PPI), kinase-substrate (KS), and transcription factor (TF)-DNA interaction networks. These networks are then linked using prior knowledge from PPI databases to construct integrated networks that can be used to track signals emanating from multiprotein complexes and are transmitted via phosphorylation networks to the nucleus, where they modify the activities of transcriptional regulators to control changes in gene expression. MiNETi is universally applicable to datasets where PPI changes regulate gene transcription via phosphorylation networks. Santra et al. applied this method to analyze HRAS signaling in HeLa cells from different subcellular compartments, showing that changes in HRAS protein interactions at different sites lead to different kinase activation patterns that differentially regulate gene transcription. The integrated networks provide a topologically and functionally resolved view of HRAS signaling. They reveal distinct HRAS functions including the control of cell migration from the endoplasmic reticulum and TP53-dependent cell survival when signaling from the Golgi apparatus.\[68\]

OmicsIntegrator is a software package that takes a variety of omics data as input and identifies putative underlying molecular pathways.\[69\] The approach applies advanced network optimization algorithms to thousands of molecular interactions to find subnetworks that best explain the data. These subnetworks connect changes observed in gene expression, protein abundance, or other global assays to proteins that may not have been measured in the mass spectrometer due to stochastic sampling, inherent bias, or noise in measurement. The strength of the approach is that it reveals unannotated molecular pathways that would not be detectable by searching pathway databases. The software comprises two individual classifiers, Garnet and Random Forest,\[70,71\] that can be run together or independently. The tool successfully identified heterogeneity in medulloblastoma subgroups,\[16\] samples collected from children’s hospitals in the United States, highlighting MYC phosphorylation as a prognosis predictor for medulloblastoma subgroup patients. In Sychev et al.\[72\] a Steiner forest algorithm from the OmicsIntegrator package was used to identify altered cellular pathways in Kaposi's sarcoma associated herpesvirus demonstrating that such pathways were not found when analyzing single biology data stream.

Similarity network fusion (SNF) developed by Wang et al.\[73\] and applied to genomics data of medulloblastoma samples from Advanced Genomics International Consortium. SNF first constructs a sample similarity network for each of the data types and then iteratively integrates these networks using a novel network fusion method. Working in the sample network space allows SNF to avoid dealing with different scale, collection bias, and noise in different data types. Integrating data
in a non-linear fashion allows SNF to take advantage of the common as well as complementary information in different data types. Cavalli et al. demonstrated that the algorithm was able to find intertumoral heterogeneity within medulloblastoma subgroups using DNA-methylation and gene expression data.\(^{[74]}\)

Another method is the prize-collecting Steiner tree problem (PCST).\(^{[75]}\) This mathematical model was successfully applied on GBM data by Huang and colleagues\(^{[76]}\) to find a network of interactions that link phosphorylation events and differentially transcribed genes. Basically the approach searches for a tree structure in the interactome. The tree structure can be seen as a group of connected nodes (e.g., genes, proteins, miRNAs) and the interactome is usually given a priori using public databases such as STRING,\(^{[77]}\) MiNT,\(^{[78]}\) and IntACT.\(^{[79]}\) As a result, the PCST approach searches for a tree structure in the interactome that connects as many of the experimental data as it can, giving a score (probability). Briefly, in a large gene–gene or protein–protein interaction network, PCST attempts to find sub-networks where many genes are differentially expressed. This method was also applied to address the challenge of data integration in a study conducted by Balbin et al.\(^{[80]}\) where transcriptome, proteome, and phosphoproteome were profiled in a panel of non-small cell lung cancer (NSCLC) cell lines to reconstruct a targetable network associated to KRAS. They first defined an “abundance-score” combining the -omics data to nominate differentially abundant proteins and then used the PCST algorithm to identify functional sub-networks. Finally, LCK was validated as critical gene for cell proliferation in KRAS-dependent NSCLCs and nominated as possible druggable target. Despite the application of this method to multiple cancer datasets, this approach has a limitation of using already pre-defined interaction datasets. While the Steiner tree algorithm was one of the most adopted method in proteogenomic studies integrating PTMs with other -omics, it was always applied to peptide-level phosphoproteomics data, thereby losing important site-specific PTM information.

### 4.3. Correlation Analysis

Phosphoproteomics provides a regular data matrix that can be incorporated into existing statistical, integrative methods. Multiple-block orthogonal projections to latent structures (OnPLS)\(^{[81]}\) is a projection method that simultaneously models multiple data matrices, reducing feature space without relying on a priori biological knowledge and was successfully applied to transcriptionomics, proteomics, and metabolomics data showing multi-level oxidative stress response in early onset gastric cancer.\(^{[15]}\)

Data clustering plays an important role in integrative -omics analysis, for example, the re-classification of a sample based on its molecular features. iCluster\(^{[82]}\) is a penalized latent regression model for joint modeling of multiple types of -omics data. The result is a unique cluster assignment, which includes all -omics data. iCluster achieves dimension reduction of the data and is demonstrated to identify potentially novel subtypes of breast cancer and lung cancer.\(^{[83,84]}\) Mo et al. developed iClusterPlus,\(^{[85]}\) which is an extension of the iCluster algorithm and incorporates a diversity of data types, including binary mutation status, mutational category copy number states (gain/normal/loss), read count sequencing data, and continuous data, and adopts tailored modeling strategies such as logistic regression, multi-logit regression, Poisson regression, and standard linear regression. The iClusterPlus algorithm has been widely used in subtyping prediction. A recently published example of multi-omics data integration with phosphoproteomics data is the early-onset gastric cancer study from Mun et al.\(^{[15]}\) that provided new insights in stratification and cancer-subtype biology from 80 patients. The method for subtyping is summarized in Figure 2 adapted from publication. First, subtypes were obtained for each individual omics data type (transcriptomics, proteomics, phosphoproteomics, N-glycoproteomics) using orthogonal non-negative matrix factorization (ONMF). The input for ONMF was based on the median absolute deviation (MAD) value using the top 10% or 20% of MADs (depending on the data type). This analysis led to the identification of two RNA subtypes (RNA1-2), four proteomics subtypes (Prot1-4), three phosphoproteomics subtypes (Phos1-3), and three N-glycoproteomics subtypes (Gly1-3). The same input data used for clustering of the individual types of data was used for iClusterPlus (Figure 2A). The resulting clustering identified four different subtypes clearly showing that proteomics and phosphoproteomics data can further subdivide subtypes 1 and 2 where RNA data cannot. Two subtypes were associated with best and worst survival (Figure 2B) and the association of these subtypes to the immune and invasion-related pathways were mainly identified by phosphorylation (circled P) and glycosylation data (circled G) (Figure 2C). The authors concluded that the identification of new subtypes in early-onset gastric cancer was not possible without the proteomics and phosphoproteomics information. Particularly, it would be inefficient to elucidate up-regulation of signaling pathways in the immune and invasive (RHOA) subtypes. Taken together, Mun and colleagues showed how a simple and ready-to-use tool such as iClusterPlus can integrate several data types and extrapolate important features for subtype biology, particularly relevant for patient stratification into drug-sensitive versus resistance. The latter was computed by predicting drug sensitivity based on phosphorylation of ARID1A, CDH1 (immune-subtype), and RHOA (invasive-subtype). However, this publication presents some limitations such as the lack of an in-depth study down to the phosphosite level and a functional validation of the data.

Recently, a new algorithm named iProFun based on multiple linear regression was presented by Song et al.\(^{[79]}\) to facilitate screening for cancer drivers and identify proteogenomics functional traits perturbed by DNA copy number alterations (CNAs) and methylations. This tool takes as input the association statistics from associating CNAs and methylations of genes to each type of cis-molecular trait, aiming to detect the joint associations of DNA variations and molecular traits in various association patterns. Of particular interest are the genes with “cascading effects” on all cis molecular traits of interest and the genes whose functional regulations are unique at global/phospho protein levels. Downstream enrichment analysis is also embedded into this pipeline. Song et al. demonstrated the feasibility of the algorithm analyzing 570 tumors of ovarian cancer from the TCGA and CPTAC project. The algorithm was able to identify CNAs (e.g., AKT1) driving perturbations on mRNA/proteins levels.
4.4. Other Opportunities Suitable for Phosphoproteomics Integration

A number of multi-omics integration methods have not been applied yet to phosphoproteomics data. Here, we highlight the most widely used and promising approaches (listed in Table 3).

A large number of publications showed the application of weighted gene correlation network analysis (WGCNA)\(^{[86]}\) as a great approach to find intramodular hub genes as potential biomarkers or drug targets (depending on the research question). By selecting intramodular hubs in consensus modules, WGCNA can also be applied to integrative data mining. Besides its extensive use in integrative biology, there are no publications so far related to the use of WGCNA and PTM datasets. It would be of a great interest to see the application of WGCNA to phosphoproteomics data and at phosphosite level.

Multi-factor analysis (MFA) is devoted to the simultaneous exploration of multiple data tables (omics data types) where the same individuals are described by several tables of variables. The aims of MFA are similar to those of PCA. Application of MFA to single-cell analysis is done with MOFA (multi-omic factor analysis), a statistical method for integrating multiple modalities of -omics data in an unsupervised fashion.\(^{[87]}\) Importantly, MOFA disentangles to what extent each factor is unique to a single data modality or is manifested in multiple modalities, thereby revealing shared axes of variation between the different -omics layers. Once trained, the model output can be used for a range of downstream analyses, including visualization, clustering, and classification of samples in the low-dimensional space(s) spanned by the factors, as well as the automated annotation of factors using (gene set) enrichment analysis, the identification of outlier samples, and the imputation of missing values.

Meng et al. also provided an extension of common gene set enrichment analysis (GSEA) to multi-omics data called MOGSA.\(^{[88]}\) This method does not require filtering data to the intersection of features (gene IDs); therefore, all molecular features, including
Table 3. Methods and tools available for general data integration.

| Name | Computational platform | Description | Availability | Number of data types | Type of integration | Reference |
|------|------------------------|-------------|--------------|----------------------|---------------------|-----------|
| WGCNA | R | Integrate 2 datatypes based on module’s construction | https://cran.r-project.org/web/packages/WGCNA/index.html | 2 | Correlation analysis | [86] |
| MixOmics | R | Offers a wide range of multivariate methods for the exploration and integration of biological datasets with a particular focus on variable selection | https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html | Unlimited | / | [110] |
| Multiple Dataset integration | Matlab | MDI can integrate information from a wide range of different datasets and data types simultaneously (including the ability to model time series data explicitly using Gaussian processes) | https://warwick.ac.uk/fac/cross_fac/zeeman_institute/zeeman_research/software/ | Unlimited | Matrix factorization | [92] |
| OnPLS | Python | OnPLS extracts a minimal number of globally predictive components that exhibit maximal covariance and correlation | https://github.com/tomlof | Unlimited | Correlation analysis | [111] |
| MCIA | R | Identifies co-relationships between multiple datasets | https://bioconductor.org/packages/release/bioc/html/omicade4.html | Unlimited | Matrix factorization | [112] |
| Mogsa | R | Gene set analysis on multiple omics data | https://bioconductor.org/packages/release/bioc/html/mogsa.html | Unlimited | Matrix factorization | [88] |
| iCluster | R | Joint Gaussian latent variable models | http://www.mskcc.org/mskcc/html/85130.cfm | Unlimited | Matrix factorization | [82] |
| MFA | R | Combine multiple data sets and to add formalized knowledge | https://cran.r-project.org/web/packages/FactoMineR/index.html | Unlimited | Matrix factorization | [87] |
| MOFA | R / Python | Method for discovering the principal sources of variation in multi-omics data sets | https://github.com/bioFAM/MOFA | Unlimited | Matrix factorization | [87] |
| MO network based | R | Network-based analysis of omics with MO optimization | mimornics@ibc.cnr.it | Unlimited | Network based | [90] |
| OmicsIntegrator | Web / Python | Package designed to integrate proteomic data, gene expression data and/or epigenetic data using a protein-protein interaction network | http://fraenkel-nsf.csbi.mit.edu/omicsintegrator/ or https://github.com/fraenkel-lab/OmicsIntegrator | 3 | Network based | [69] |
| MiNETi | Matlab / Java / R | Integration of multi-omics datasets that reconstructs and integrates interaction data | https://www.sciencedirect.com/science/article/pii/S2211124719302098?via%3Dihub#mmc8 | Unlimited | Network based | [68] |
| LinkFinder | web | Integration and visualization of two -omics datasets at time | http://www.linkedomics.org/admin.php | 2 | Correlation analysis | [97] |
| iProFun | R | Integrative analysis tool to screen for proteogenomic functional traits perturbed by DNA copy number alterations (CNAs) and DNA methylations | https://github.com/songxiaoyu/iProFun | 3 | Linear regression | [78] |

those that lack annotation may be included in the analysis. The publication provides a use case of MOGSA on NCI-60 cell lines panel with transcriptome and proteome data, but can be applied to more than two -omics data types.

A different approach is given by joint and individual variation explained (JIVE) that partitions the variations of the multi-omics data into the sum of three components: 1) a low rank approximation accounting for common variations among -omics data from all platforms, 2) low rank approximations for platform-specific structural variations, and 3) residual noise. JIVE has been extended to joint and individual clustering analysis (JIC) to simultaneously conduct joint and omics-specific clustering on gene expression and miRNA levels on TCGA breast cancer.[89]

Network-based multi-objective (MO) optimization is another method to generate networks of biological components that incorporate multi-omics information.[90] The MO optimization procedure drives the identification of networks that are enriched according to several statistical estimators. The analysis can be applied to different types of -omics and biological interactions. The authors found...
protein networks that participate in the establishment of the increased basal differentiation observed in breast tumors of BRCA1-mutation carriers. Additionally, they carried out a network-based comparison among several -omic datasets using transcriptomic data from two types of breast tumors and the corresponding epithelial cells. They found a protein network that shows a strong and coherent (the same direction) differential expression when comparing each tumor with its respective epithelial tissue.

Finally, Bayesian-methods are particularly well suited for integrating different biological data sources, offering a natural setting for integrative analysis including the possibility of incorporating prior information. In contrast, non-Bayesian approaches typically assess different data types in a sequential way or are based on measures of the association between (two) data types rather than on a joint analysis. Multiple dataset integration (MDI) is a Bayesian method for unsupervised integrative modeling of multiple datasets. MDI can integrate information from a wide range of different datasets and data types simultaneously (including the ability to model time series data explicitly using Gaussian processes). Each dataset is modeled using a Dirichlet-multinomial allocation mixture model, with dependencies between these models captured through parameters that describe the agreement among the datasets.

4.5. Open Source Software for Multi-Omics Visualization and Integration

There are open source software platforms that are commonly used for network or pathways visualization. As mentioned before, Cytoscape is the most common used tool where dedicated app such as “Omics Visualizer” is relatively easy-to-use for multi-omics data visualization. Recently, the Cytoscape community has released “omics analysis collection.” This app is a collection of already published apps that enables the user to generate a base network from STRING, WikiPathways, and Agilent Literature Search and subsequently projects -omics data on top of them. Afterward, the user is able to perform cluster analysis and functional enrichment while preferred network and heatmap can be visualized as well.

Another extendable pathway analysis toolbox is PathVisio. Unlike Cytoscape, PathVisio allows users to visualize -omics data directly on WikiPathways graphs. Moreover, the user can add/remove specific genes/proteins/metabolites and visualize the specific expression. This tool is intended only for -omics visualization and can be downloaded from https://pathvisio.github.io.

Several databases are emerging as data collection, analysis, and visualization. Databases such as cBioPortal, TCPAv3.0, TCGA and CPTAC collect multi-omics data previously characterized by TCGA and CPTAC and provide the ability to visualize/analyze those data together with your own data in a comprehensive way. The web portals are easily accessible via their websites (https://www.cibioportal.org, https://tcpaportal.org, http://www.linkedomics.org, and https://www.omicsdi.org). CBioPortal is the most used website in terms of genomic data integration. The user can upload their own data and run basic analysis like gene–gene correlation, survival analysis, heatmap visualization of different genomics data, and boxplots related to the expression of a certain gene with specific mutation. While cBioPortal, TCPA, and LinkedOmics relies on TCGA and CPTAC sample characterization, OmicsDI is a database storing all data available from several species, tissues, and diseases. The user can look for a specific disease and select the type of data to analyze/download. In addition, cBioPortal provides the option to analyze your own data by installing the instance available on GitHub (https://github.com/cBioPortal/cbioportal). Nevertheless, none of these visualization tools do justice to phosphoproteomics data, as they treat the phosphoproteomics layer as phosphoprotein expression instead of focusing on multiple phosphosites of the protein. At the moment, “Omics Visualizer” is the only Cytoscape plugin available that shows phosphoproteomics data at the phosphosite level together with others -omics data. Here, network connections are based on the STRING database and not yet customizable for different types of input such as kinase-substrate relations.

5. Concluding Remarks and Outlook

Multi-omics analysis has yielded detailed insights into the heterogeneity and complexity of cancers. The main strength of -omics data integration is the capability to track the information flow from upstream (epigenomics, genomics) to downstream effects (proteomics, phosphoproteomics, metabolomics). Recent analyses highlighted the importance of phosphoproteomics in understanding tumor complexity, subtypes, and therapeutic vulnerabilities. We have reviewed state-of-the-art methods for analysis of kinase activities, the latest large-scale proteogenomics studies with phosphoproteomics data, and bioinformatics methods where phosphoproteomics data can be treated as another -omics layer. We have highlighted the importance of multi-omics data visualization and data portals to enable the incorporation of expert knowledge in data analysis.

Integration of phosphoproteomics including site-specific peptide information is important. The INKA method for analysis of kinase activities illustrates the power of bringing to the equation in-depth knowledge about phosphosites within kinase activation loops. We expect that research in this direction will benefit from advances in mass spectrometry-based methods for more comprehensive site-specific phosphorylation profiling and the ability to integrate all available phosphorylation datasets for the analysis of individual patients at hand. Developing new tools for -omics data integration remains a constant evolving field. The joint efforts of mathematicians, physicists, bioinformaticians, and biologists will allow cancer researchers to obtain more and more detailed knowledge of tumor complexity, biology, and therapeutic vulnerabilities. We expect these tools will become available for use in the clinic, to enable patient stratification and personalized medicine.

Acknowledgements

The research reported in this publication was supported by the KWF Dutch Cancer Society (Project 2003887, the Netherlands), funded by the KWF grant Dutch Cancer Society (#10212) (C.J.).

Conflict of Interest

The authors declare no conflict of interest.
Giulia Mantini is a Ph.D. student in computational biology at the Amsterdam UMC–Cancer Center. Her Ph.D. project involves the identification of drug targets and diagnostic markers in solid and liquid tumor biopsies with a focus on cancer proteomics and phosphoproteomics, data integration, systems biology, and machine learning. She holds a master’s degree in bioinformatics from Alma Mater Studiorum Bologna, Italy, while investigating a new methodology for massive parallel sequencing and barcoding in a joint project with the SciLifeLab–Stockholm in the Molecular Diagnostic Group. She is currently in the last year of her Ph.D.
**Thang Pham** graduated from the RMIT University, Australia, in computer science in 1998 (B.Sc.) and in 2005 from the University of Amsterdam, The Netherlands, in computer vision (Ph.D.). He is a research scientist at the OncoProteomics Laboratory, Amsterdam University Medical Center. His research interests are in mass spectrometry-based proteomics data processing, statistical analysis, and application of artificial intelligence techniques in the field.

**Sander Piersma** trained as a biochemist and enzymologist at the universities of Amsterdam and Delft. At Wageningen University, he was introduced to protein mass spectrometry as a post-doc. As project leader at the Institute of Atomic and Molecular Physics in Amsterdam, he developed biological tissue imaging mass spectrometry methods. In 2006, he joined the OncoProteomics Lab of Amsterdam University Medical Center as a research scientist and is responsible for mass spectrometry. His research interests include post-translational modifications and label-free proteomics. Sander enjoys developing innovative mass spectrometry-based (phospho) proteomics methods and data analysis strategies to improve (early) diagnostics and precision medicine for patients.

**Connie Jimenez** is full professor of translational oncoproteomics, group leader of the OncoProteomics Laboratory and Director of Proteomics Core Facility, all at Amsterdam University Medical Center, Amsterdam, The Netherlands. She has been working on biological and biomedical applications of mass spectrometry and proteomics since her Ph.D. studies in the 90s. Her lab’s mission is to apply innovative mass spectrometry-based (phospho)proteomics and data analysis to improve (early) diagnostics and treatment of cancer and neurodegenerate disease. She has a special interest in multi-omics data integration to obtain a systems biology insight into multi-factorial complex diseases.