Proteomics Is Not an Island: Multi-omics Integration Is the Key to Understanding Biological Systems*

Bing Zhang‡§** and Bernhard Kuster¶‖**

Proteins serve as a critical link between genotype and phenotype. Proteomic profiles reflect cellular responses to genomic, epigenomic, and environmental alterations, and they in turn shape these responses. Adapting a quote from the English poet John Donne, "proteomics is not an island, entire of itself." Integrating proteomics data with other types of data from genomics, epigenomics, transcriptomics, metabolomics, image-omics, and phenomics generates the bigger picture and thus holds great potential for revealing novel biology and transforming clinical practice. Realizing this potential requires computational methods and tools that can drive effective multi-omics data integration. This special issue brings together a series of articles describing novel computational methods and tools, biological applications, and perspectives on multi-omics integration with the aim to raise awareness in the field for this ever-growing need.

We start with a minireview from Vogel and colleagues (1) on the current state and limitations of multi-omics data integration, termed integromics by the authors. After a brief summary on the well-reviewed topic of integrating proteomic and transcriptomic measurements, they detail on the integration of proteomics with other types of omics data including translatome profiling, genomic alterations, post-translational modifications, polysome profiling, cellular thermal shift assays (CETSA)[1], and metabolomics and lipid measurements. Next, they present a high-level summary of computational tools and techniques for omics data integration, including visualization tools that provide a holistic view of merged data and interpretation tools that harmonize biological information across heterogeneous platforms. The authors note that inter-omics causal and regulatory relationships have been largely overlooked, but such relationships should be recognized and emphasized in statistical modeling. Finally, they provide a step-by-step guide to and considerations on productive integromics to exploit the synergy of multi-omics data.

The first group of articles describes new methods and tools for studying associations between proteomics data and other types of omics data, including somatic mutations, somatic copy number alterations (CNA), DNA methylation, mRNA expression, protein phosphorylation, and morphological features. Chen *et al.* (2) present an updated version of the cancer proteome atlas (TCPA), which includes reverse-phase protein arrays (RPPA)-based proteomic data for ~8000 patient samples across 32 cancer types through The Cancer Genome Atlas (TCGA) project. The updated version introduces a new module called "TCGA Pan-cancer Analysis," which provides comprehensive protein-centric analyses that integrate RPPA data with other TCGA data across cancer types. This new module allows examining the correlation between protein expression and somatic mutations, assessing the predictive power of somatic copy-number alterations, DNA methylation and mRNA on protein expression, inferring regulatory effects of miRNAs on protein expression, constructing coexpression networks of proteins and pathways, and identifying clinically relevant protein markers. The updated TCPA provides a comprehensive resource for cancer researchers to test their protein-centric hypotheses using multi-omics data from a broad range of cancer types.

Arshad *et al.* (3) integrated MS-based proteomic and phosphoproteomic data from breast and ovarian tumors to study the relationship between kinase activity, substrate specificity, and phosphorylation. They found that phosphorylation levels were largely unrelated to the protein abundance of the cognate protein or the phosphorylation of other sites on the same protein. Abundance of the kinases and phosphorylation of kinases on their annotated activating or inhibiting sites did not seem to correlate well with their activity, as assessed by phosphorylation of known substrates. However, focusing on highly correlated kinases and phosphosites provides a reasonable approach for identifying novel substrates for some kinases.

From the ‡Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas; §Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas; ¶Chair of Proteomics and Bioanalytics, Technische Universitat Munchen, Freising, Germany; ‖Bavarian Biomolecular Mass Spectrometry Center, Technische Universitat Munchen, Freising, Germany

Received July 22, 2019
Published, MCP Papers in Press, July 22, 2019, DOI 10.1074/mcp.E119.001693

Zhan *et al.* (4) systematically examined the relationships between RNA-Seq and proteomic data and morphological features of breast cancer. Their integrative data mining inferred four biological processes associated with various interpretable morphological features. The biological processes were related to cancer hallmarks and included metabolism, cell cycle, immune response, and extracellular matrix development. The morphological features were related to area, density, and shapes of epithelial cells, fibroblasts, and lymphocytes. Interestingly, unfavorable prognostic morphological features were linked to large cell nuclei or large distances to neighboring cells, which were highly associated with metabolic or extracellular matrix related processes, whereas favorable prognostic morphological feature tended to be small distances to neighboring cells, which were highly correlated with metabolic or immune related processes.

Also included in the first group are two articles from Wang and colleagues describing statistical methods for integrative multi-omics relationship modeling. The first article (5) describes iProFun, an algorithm to discover *cis*-associations between CNAs and DNA methylations and molecular quantitative traits including mRNA, protein, and phosphoprotein abundance. Instead of analyzing each molecular trait separately, their approach allows joint modeling of multi-omics data, which leads to enhanced power for detecting significant *cis*-associations shared across different omics data types and higher accuracy in inferring *cis*-associations unique to certain type(s) of molecular trait(s). The second article (6) describes ProMAP, a computational algorithm that systematically characterizes both *cis*- and *trans*- regulatory relationships between CNAs and proteins. A multivariate regression framework is used to model the multiple-to-multiple regulatory relationships between CNA and proteins. Further, statistical regularization is used to facilitate the detection of master genetic regulators, which affect the activities of many proteins and often play important roles in genetic regulatory networks. ProMAP also includes a linear mixed effects model to account for the batch structure and explicitly incorporate the abundance-dependent-missing-data mechanism of proteomic data.

The second group of the special issue includes four articles reporting novel pathway analysis methods and tools with a focus on addressing specific needs for interpreting multi-omics data. Pathway or gene set analysis is indispensable in the functional interpretation of omics data. Applying such analysis to data generated from a single omics experiment has been well standardized, but its application to multi-omics remains challenging. Savage *et al.* (7) present two graph algorithms implemented in an R package named Sumer for condensing and consolidating gene set analysis results from multiple experiments, as in multi-omics studies. Sumer uses a weighted set cover algorithm to reduce redundancy of gene sets identified in a single experiment and then uses affinity propagation to consolidate similar gene sets identified from multiple experiments into clusters and to automatically determine the most representative gene set for each cluster. Case studies demonstrate that Sumer can greatly facilitate the interpretation of gene set analysis results from multi-omics and other types of integrative studies.

Rather than post-processing gene set analysis results from individual omics data types, Meng *et al.* (8) introduce a new computation method named multi-omics gene set analysis (MOGSA), which integrates multiple types of omics data from the same set of samples to perform an integrated multivariate single sample gene set analysis. MOGSA first learns a low dimensional representation that capture the most prominent correlated structure among different datasets using multiple factor analysis, which is a multiple table extension of principal component analysis. Next, it calculates an integrated gene-set score from the most informative features in each data type. Case studies show that integrating multiple types of omics data increase the power to discover subtle changes in gene-sets and reduces the impact of unreliable information in any single data type.

Along the same line, Liu *et al.* (9) applied independent component analysis, another dimension reduction technique to the study of human breast cancer transcriptomic and proteomic data. Applying this unsupervised feature extraction method to both transcriptomic and proteomic data constructs signatures that can be linked to known biological processes and pathways. Moreover, constructed transcriptomic and proteomic signatures can be associated by their respective correlation with patient clinical features. Thus, the method allows an unbiased discovery of phenotype-related biological processes or pathways.

The last article in this group concerns metaproteomics in microbiome studies. Easterly *et al.* (10) present metaQuantome, a comprehensive software suite that enables fully quantitative differential abundance analysis of the functional and taxonomic profile of a metaproteome. It is amenable to label-free proteomics data quantified using peptide-level MS1 intensity values as well as spectral counting. A unique feature of metaQuantome is the integration of taxonomic annotation and functional annotation to perform a multi-faceted analysis of a metaproteomics dataset, enabling users to determine microbe-specific contributions to the functional profile, or the profile of microbes contributing to a specific functional protein class.

Whereas the first two groups of articles focus on the analysis of proteome profiling data in the context of multi-omics studies, the two articles in the third group are devoted to protein-protein interactions, another important dimension of

[1] The abbreviations used are: CETSA, cellular thermal shift assays; CAN, copy number alterations; TCGA, The Cancer Genome Atlas; TCPA, The Cancer Protein Atlas; RPPA, reverse-phase protein array; MOGSA, multi-omic gene set analysis; HD, Huntington's disease; Htt, huntingtin.

protein function. Federspiel *et al.* (11) used a multi-omics approach to uncover the function of histone deacetylase 4 (Hdac4) in the context of Huntington's disease (HD) progression. The authors characterized the interactomes of endogenous Hdac4 in the brains of HD mouse models with wild type and mutant forms of the huntingtin (Htt) gene at both presymptomatic and symptomatic ages. Further integration of whole proteome and transcriptome datasets from these HD mouse models revealed how mutant Htt and age affect protein abundance and gene expression patterns, and how these synergize with functional enrichments observed in the Hdac4 interactomes. Thus, the integrative analysis provided new knowledge of the molecular underpinning of HD phenotypes, which may be useful for designing therapeutic interventions.

Protein-protein interactions in protein complexes have been linked to protein level attenuation of copy number variation in cancer samples. To better understand the interaction mediated control of protein abundance, Sousa *et al.* (12) performed a multi-omics study that combines genomics, proteomics, and phosphoproteomics data from hundreds of cancer samples. The authors find that up to 42% of the 8124 proteins analyzed showed evidence of post-transcriptional attenuation. Over 500 protein-protein interactions showed indirect protein abundance control through interaction, and some interactions were further controlled by phosphorylation. Interestingly, further integration of structural data showed that the fraction of interface residues of a protein is a strong determinant of attenuation. These results suggest that protein complex formation is an important factor in post-transcriptional control, likely via a high degradation rate of unassembled subunits.

The last group of articles in this special issue is devoted to proteogenomics-driven interpretation of MS/MS data. Li *et al.* (13) describe a software pipeline including two algorithms for identifying peptides and proteins from metaproteomics data using protein databases derived from matching metagenomic and metatranscriptomic data. The first algorithm Graph2Pro retains and uses uncertainties of metagenome assembly for reference-based MS/MS data analysis, whereas the second algorithm Var2Pep considers the variations found in metagenomic/metatranscriptomic sequencing reads that are not retained in the assemblies (contigs). The pipeline doubled the spectra identification rate compared with conventional contig- or read-based approaches in two microbiome data sets. The study highlights the importance of considering assembly uncertainties and genomic variants in metaproteomic data interpretation.

Verbruggen *et al.* (14) present an updated version of Proteoformer, a pipeline for processing ribosome profiling data to generate protein sequence database, which can be used to improve MS/MS data interpretation. Major updates in Proteoformer 2.0 include improved Ribo-Seq data quality assessment, read preprocessing and alignment, transcript calling, and proteoform calling using multiple methods. A protein sequence database is generated by combining proteoforms identified by different methods. Application of Proteoformer 2.0 to matching Ribo-Seq and MS/MS data from two cell lines led to the identification and validation of different categories of new proteoforms, including translation products of up- and downstream open reading frames, 5′ and 3′ extended and truncated proteoforms, single amino acid variants, splice variants, and translation products of so-called noncoding regions.

Despite the power of proteogenomics-driven MS/MS data integration in genome and metagenome annotation, peptide-spectral matches (PSMs) supporting novel proteoforms are prone to false positive identifications and usually require careful (and often manual) evaluation. Brademan *et al.* (15) present IPSA, a web-based spectrum annotator that visualizes and characterizes peptide tandem mass spectra. ISPA can visualize peptides collected using a wide variety of experimental and instrumental configurations. Single spectra can be analyzed through provided web forms, whereas data for multiple peptide spectral matches can be uploaded using the Proteomics Standards Initiative file formats mzTab, mzIdentML, and mzML, and annotated spectra are customizable via a selection of interactive features and can be exported as editable scalable vector graphics to aid in the production of publication-quality figures. IPSA may therefore become very useful for inspection of PSMs supporting novel proteoforms.

We end the special issue with a perspective article from Payne and colleagues (16) on reproducibility and transparency of data analysis. Reproducibility has recently gained significant attention in biomedical research. Efforts have been made to make meta data and raw omics data publicly available and early steps of omics data analysis transparent and reproducible. However, downstream data integration and interpretation, which are essential to support scientific conclusions from multi-omics studies, are typically not openly shared. The authors propose one possible solution, which requires documenting the entire data analysis, including posting code for analysis and the resulted figures to an open version control software repository like GitHub, posting data tables used in the analysis in the same repository or in a password-free download if they are too large, and listing the URL to specific scripts in the repository in figure legends and methods sections in the manuscripts. Gladly, some of these have already been implemented in a few exemplary publications highlighted by the authors in the article.

In summary, multi-omics integration is an exciting and fast-growing field of research, and proteomics plays a central role in such integration. We hope this special issue offers useful resources and ideas, stimulates the development of new methods, and encourages more researchers to leverage multi-omics data to pursue a deeper and more complete understanding of biological systems.

Acknowledgments—We thank all authors and reviewers for their contributions to this special issue. We thank the Molecular and Cellular Proteomics Editors and staff, especially Emily Huff, for their support for the special issue.

* B.Z.'s work relevant to this editorial was supported by grant U24CA210954 from the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium, by grant CPRIT RR160027 from the Cancer Prevention & Research Institutes of Texas (CPRIT), and by funding from the McNair Medical Institute at The Robert and Janice McNair Foundation. B.Z. is a CPRIT Scholar in Cancer Research and a McNair Scholar. B.K. is a Carl von Linde Senior Fellow of the Institute for Advanced Study (IAS) of the Technical University of Munich.

** To whom correspondence should be addressed: E-mail: bing.zhang@bcm.edu or kuster@tum.de.

Author contributions: B.Z. and B.K. wrote the paper.

REFERENCES

1. Vitrinel, B., Koh, H. W., Kar, F. M., Maity, S., Rendleman, J., Choi, H., and Vogel, C. (2019) Exploiting inter-data relationships in next-generation proteomics analysis. *Mol. Cell. Proteomics* **14,** S5–S14
2. Chen, M. M., Li, J., Wang, Y., Akbani, R., Lu, Y., Mills, G. B., and Liang, H. (2019) TCPA v3.0: An Integrative Platform to Explore the Pan-cancer Analysis of Functional Proteomic Data. *Mol. Cell. Proteomics* **14,** S15–S25
3. Arshad, O. A., Danna, V., Petyuk, V. A., Piehowski, P. D., Liu, T., Rodland, K., and McDermott, J. E. (2019) An integrative analysis of tumor proteomic and phosphoproteomic profiles to examine the relationships between kinase activity and phosphorylation. *Mol. Cell. Proteomics* **14,** S26–S36
4. Zhan, X., Cheng, J., Huang, J., Han, Z., Helm, B., Liu, X., Zhang, J., Wang, T., Ni, D., and Huang, K. (2019) Correlation analysis of histopathology and proteogenomics data for breast cancer. *Mol. Cell. Proteomics* **14,** S37–S51
5. Song, X., Ji, J., Gleason, K. J., Yang, F., Martignetti, J. A., Chen, L. S., and Wang, P. (2019) Insights into impact of DNA copy number alteration and methylation on the proteogenomic landscape in human ovarian cancer via a multi-omics integrative analysis. *Mol. Cell. Proteomics* **14,** S52–S65
6. Ma, W., Chen, L. S., Özbek, U., Han, S. W., Lin, C., Paulovich, A. G., Zhong, H., and Wang, P. (2019) Integrative proteo-genomic analysis to construct CNA-protein regulatory map in breast and ovarian tumors. *Mol. Cell. Proteomics* **14,** S66–S81
7. Savage, S. R., Shi, Z., Liao, Y., and Zhang, B. (May 29, 2019) Graph algorithms for condensing and consolidating gene set analysis results. *Mol. Cell. Proteomics* **14,** S141–S152
8. Meng, C., Basunia, A., Peters, B., Gholami, A. M., Kuster, B., and Culhane, A. C. (2019) MOGSA: integrative single sample gene-set analysis of multiple omics data. *Mol. Cell. Proteomics* **14,** S153–S168
9. Liu, W., Payne, S. H., Ma, S., and Fenyö, D. (2019) Extracting pathway-level signatures from proteogenomic data in breast cancer using independent component analysis. *Mol. Cell. Proteomics* **14,** S169–S182
10. Easterly, C. W., Sajulga, R., Mehta, S., Johnson, J., Kumar, P., Hubler, S., Mesuere, B., Rudney, J., Griffin, T. J., and Jagtap, P. D. (June 24, 2019) metaQuantome: An integrated, quantitative metaproteomics approach reveals connections between taxonomy and protein function in complex microbiomes. *Mol. Cell. Proteomics* **14,** S82–S91
11. Federspiel, J. D., Greco, T. M., Lum, K. K., and Cristea, I. M. (2019) Hdac4 interactions in Huntington's Disease viewed through the prism of multi-omics. *Mol. Cell. Proteomics* **14,** S92–S113
12. Sousa, A., Gonçalves, E., Mirauta, B., Ochoa, D., Stegle, O., and Beltrao, P. (2019) Multi-omics characterization of interaction-mediated control of human protein abundance levels. *Mol. Cell. Proteomics* **14,** S114–S125
13. Li, S., Tang, H., and Ye, Y. (May 29, 2019) A meta-proteogenomic approach to peptide identification incorporating assembly uncertainty and genomic variation. *Mol. Cell. Proteomics* **14,** S183–S192
14. Verbruggen, S., Ndah, E., Criekinge, W. V., Gessulat, S., Kuster, B., Wilhelm, M., Van Damme, P., and Menschaer, G. (May 6, 2019) PROTEOFORMER 2.0: further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. *Mol. Cell. Proteomics* **14,** S126–S140
15. Brademan, D. R., Riley, N. M., Kwiecien, N. W., and Coon, J. J. (May 14, 2019) Interactive peptide spectral annotator: a versatile web-based tool for proteomic applications. *Mol. Cell. Proteomics* **14,** S193–S201
16. Petyuk, V. A., Gatto, L., and Payne, S. H. (July 4, 2019) Reproducibility and transparency by design. *Mol. Cell. Proteomics* **14,** S202–S204