Published in final edited form as:

Curr Opin Syst Biol. 2018 October ; 11: 74–81. doi:10.1016/j.coisb.2018.09.001.

Integration of large-scale multi-omic datasets: a protein-centric view

Justin Rendleman¹, Hyungwon Choi^{#2,3}, and Christine Vogel^{#1,#}

¹Center for Genomics and Systems Biology, New York University, Department of Biology, New York, USA

²Department of Medicine, Yong Loo Lin School of Medicine, National University Singapore, Singapore

³Institute of Molecular and Cell Biology, Agency for Science, Technology, and Research, Singapore

These authors contributed equally to this work.

Abstract

Innovative mass spectrometry-based proteomics has enabled routine measurements of protein abundance, localization, interactions, and modifications, covering unique aspects of gene expression regulation and function. It is now time to move from isolated analyses of these datasets toward true integration of proteomics with other data types to gain insights from the interactions and interdependencies of biomolecules. When combined with genomic or transcriptomic data, proteomics expands genome annotation to identify variant or missing genes. Dynamic proteomic measurements can move analysis from predominantly concentration-based framework to that of synthesis and degradation of proteins. Proteomic data from thousands of cancer patients can foster identification of novel pathogenic mutations via detection of protein sequence changes that lead to dysregulated pathways in various tumors. Such comprehensive efforts can exploit the synergy arising from large and complex datasets to advance virtually every field of biology.

Integrating protein knowledge into omics analyses

Rapid technological advances of the past decades, best exemplified by massively parallel sequencing but also by meaningful innovations in mass spectrometry, have enabled routine collection of ‘omic’ scale data, i.e. data that covers a large part if not all of an organism’s genome. Proteomics can now identify close to tens of thousands of genes simultaneously [1,2,3] and assess protein abundance, protein variants, posttranslational modifications, as well as interactions with DNA, RNA, and other proteins (Figure 1). Experimental platforms are continuously added to the repertoire, each permitting the exploration of new areas in

Corresponding author: cvogel@nyu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

biology, e.g. metabolomics [4], lipidomics [5], and techniques charting epigenetic regulation [6] or chromatin structure [7]. However, when conducting such large scale studies it is often the case that each dataset is analyzed in isolation, ignoring the connections between the underlying molecules, thereby rendering the results largely independent of one another. An important next step in systems biology is the integration of these multi-omics data, which is essential to fully understand the complex nature of life.

The isolated analysis approach contrasts biological reality: in general, gene expression is far from solely governed by a single process at a time, but involves a multitude of processes that impact and control one another. A prime example is the cellular response to an accumulation of unfolded proteins in the endoplasmic reticulum (Figure 2), which is critical in the understanding of a variety of human diseases [8,9,10]. The first wave of the mammalian unfolded protein response involves protein phosphorylation of a translation initiation factor which halts translation globally [11,12,13]. However, a subset of genes escape this repression including the key transcription factor ATF4 that triggers a vast transcriptional response [14]. Another important transcription factor is activated through the non-canonical cytosolic splicing of its mRNA by the stress sensor IRE1 [15,16], an endonuclease that can also target unnecessary mRNAs for degradation. Meanwhile, the proteasome degrades irreparably damaged and ubiquitinated proteins [17]. Later in the response, following transcriptional induction, a phosphatase removes the modifications on the translation initiation factor to reactivate protein synthesis. Taken together this stress response provides context for how, over the course of several hours, the cell can systematically employ all levels of gene expression regulation to reestablish homeostasis (Figure 2). We believe this level of complexity in a multilayered regulatory response is not an exception but rather the rule when it comes to how cells react to environmental exposure. Thus, investigation of isolated processes to address any biological question will only capture part of the entire picture, highlighting the critical need for studies that employ data integration.

Truly integrative multi-omics analysis is therefore the next opportunity and challenge of systems biology at the same time. Importantly, this requires the assessment of interactions between data types and their temporal and causal relationships to exploit the synergy arising from them. Here, we review recent examples of such ‘integromics’ studies, with a focus on the burgeoning field of “proteogenomics” that, loosely speaking, maps protein information to data from other platforms, typically genomics, transcriptomics, or epigenomics.

In this short review, we will highlight three areas of proteogenomics where the inclusion of protein information has been particularly productive. First, we review recent archetypal applications of proteogenomics, in which proteomics is used to enhance gene annotation by correcting existing predictions to identify previously missed elements. Second, we examine integrative studies that incorporate proteomic and other omic measurements to broaden the understanding of fundamental principles in gene expression regulation. Third, we will review an important clinical application of proteogenomics -- cancer biology -- to demonstrate how this field in particular is poised to move towards truly integrative analysis of the massive data. We will focus on aspects that expand on other excellent reviews published lately that have discussed computational methods and human diseases in more detail [18][19][20][21][22][23][24][25].

Expanding the annotation of genomes

Thirteen years after the complete sequence of the first genome was published [26], sequence-based prediction of genes has yet to capture the full biological complexity. Therefore, to refine annotations many algorithms use external information, for which proteomics is an intuitive data source. The integration of sequence information gained from proteomics with transcriptomics and genomics data is conceptually a straightforward application of proteogenomics, as well as the original intent when the term was coined in 2004 [27].

Such proteogenomic studies have helped to identify novel genes, splice variants, N-terminal extensions, alternative open reading frames, and short proteins in the genomes of both well-characterized and less-studied organisms [1]. Despite the initial excitement, most studies identified only a few hundred new proteins or protein variants: ~100 in the model cell line HEK-293 [28], ~300 in grapevine and monkey [29,30], ~600 in pig [31], and >800 in mosquito [32]. These relatively low numbers may be due to the limited genome coverage that proteomics provides or its bias towards high-abundance proteins, though it may also indicate that the predictive algorithms capture most of the variation, at least within the known feature space.

Perhaps more intriguing are proteogenomic approaches that identify new proteins or protein variants that arise from processes outside of canonical models and expand our ideas of how genes are actually expressed. For example, a study in the parasite *Blastocystis* identified seven new proteins using peptide information at the C-terminal where part of the UAA termination codons are derived by transcript polyadenylation [33]. Other efforts in human cells searched unmatched mass spectra from proteomics experiments against RNA-seq data that was computationally translated into all three possible open reading frames to identify short, missing proteins, identifying a handful of such microproteins [34,35]. Another study used a specific proteomic approach that enriched for the N-terminal ends of proteins that are often missed in both RNA sequencing and proteomics experiments. Examining the plant *Arabidopsis*, integration of Nterminal proteomics with transcript data and information on ribosome-mRNA interactions identified 117 novel translation start sites outside the annotated protein regions [36*].

Additionally, a good amount proteogenomic efforts are aimed at discerning splice variants. As standard RNA-seq practices are insufficient for the delineation of many mRNA isoforms, the incorporation of expressed peptide information is crucial in disambiguating this aspect of gene expression. A recent integrative study in *Arabidopsis* combined short-read RNA sequencing and single molecule long-read sequencing with proteomics and found that 84% of the intron-containing genes were alternatively spliced, primarily through alternative first or last exons, as detected at the RNA level [37*]. The study went further to add another dimension to the analysis: that of differential splicing in cells responding to a stimulus. By incorporating proteomics into the analysis, the authors were able to show that while alternatively spliced isoforms were not frequently observed at the protein level under normal conditions, upon hormone treatment the isoforms that were differentially expressed were

also much more likely to be translated into proteins suggesting a synergy in the cellular response.

Understanding all levels of gene expression regulation

Proteogenomics efforts have also emerged in studies examining gene expression regulation, although these typically investigate a single organism in select conditions. First attempts demonstrate that combining multi-omic platforms can provide additional insights relevant for understanding various disease pathologies, even with simple integration of the data. For example, monitoring cardiac hypertrophy in mice, protein abundance and turnover measurements increased identification of disease-related expression changes by 75% over transcript analysis alone [38]. In a study of osteoarthritis, combining information on DNA methylation with RNA and protein expression identified a set of differentially expressed genes, of which one third were newly implicated in disease progression [39]. Finally, a study of fibroblasts from Down syndrome patients showed that proteomics was capable of detecting changes that arise from an additional chromosome that were not discernible in genomic and transcriptomic data alone [40]. These results identified specific protein complex members that were degraded to dosage-compensate.

Cycling processes, such as mitosis, meiosis, or the diurnal rhythm, have been popular targets of proteogenomic approaches investigating gene expression. Analyzing transcript, translation, and protein stability information in mitotic yeast, one study discovers new cycling genes across the different data types [41]. A similar analysis in yeast undergoing meiosis discovered a strong anti-correlation between transcript and protein concentrations for hundreds of genes [42**]. The ribosome footprinting data for these genes revealed that this anti-correlation can be explained by a switch between the transcription of canonical transcripts and 5' extended isoforms that are not efficiently translated. This is coordinated by a single transcription factor which also induces another group of RNAs that are in fact translated efficiently, providing a beautiful link between transcription regulation and translation that allows this transcription factor to simultaneously activate and repress gene expression among distinct sets of genes.

While this example demonstrates how transcription and translation can oppose one another to achieve the desired outcome, examples of concordant regulation also exist. In *Arabidopsis*, the combined analysis of protein and mRNA abundance changes in diurnal regulation identified the synchronization of peaks in transcript levels and translation rates [43]. The authors suggest this result might be “translational coincidence” controlling the most dominant expression changes during the photoperiod.

A few studies have aimed at a more elaborate integrative analysis of RNA- and protein-level regulation of gene expression, moving towards a dynamic assessment of cellular perturbations. An emerging theme in these studies is ability of transcript abundance variation to only partially explain protein abundance variation [44,45,46]. The degree of contribution to gene expression regulation by transcription and translation varies substantially in different biological systems. In a key study in dendritic cells responding to lipopolysaccharide, RNA expression and protein expression were monitored over the course of 12 hours [47]. Using a

mathematical model, the authors show a majority of this response was governed by transcription and RNA degradation, rather than translation and protein degradation.

However, subsequent work has shown that this is not the case for all treatments. In response to a reagent causing accumulation of unfolded proteins in the endoplasmic reticulum, relative contributions of transcription and translation were much more balanced, and protein changes often differed from those of RNA in a non-linear manner [48]. In addition, transcript abundances changed more acutely than protein abundances and in most cases RNA returned to initial states by the end of the 30 hour experiment. Proteins, in contrast, reached substantially different equilibrium points during the time course where they remained throughout. Importantly, both studies moved beyond simple comparisons of protein and mRNA concentrations in steady state conditions, to integrative computational models that assessed dynamic systems.

Finally, proteomics can also generate unique information with respect to protein localization and posttranslational modifications. In a recent study the application of RNA-seq, ribosome footprinting, and proteomics with spatial separation in neurons was used to demonstrate that mRNA localization to cellular projections (neurites) explains half of the differential protein localization in neurites compared to the cell bodies (soma), implying that other processes must account for the remainder [49]. Through the combination of ribosome profiling and multiple proteomics techniques the authors confirm that localized translation does indeed occur within neurites. Another study examined a different type of protein localization: that along the chromosomes. Using information on protein and transcript levels as well as the genome architecture, the authors observed co-regulation of neighboring transcripts that does not transfer into the protein level [50]. Many RNAs appear to be transcribed simply due to stochastic chromatin fluctuations, which is then buffered at the protein level. These findings support an interpretation in which post-transcriptional regulation fine-tunes protein expression levels [46].

Proteogenomics across populations of cancer patients

Another important role for proteogenomics studies is the application to human disease, which is depicted in the large scale profiling of cells and tissues across various populations of cancer patients. The advent of The Cancer Genome Atlas (TCGA)[51,52] has propelled cancer research to be a model of big data collection and integration: to date, TCGA features genomic, epigenomic, transcriptomic, microRNA, and protein array data from more than 11,000 primary tumors [53]. The multi-omic data has enabled comprehensive characterization of tumor tissues and fluids, often with strong emphasis on the somatic and pathogenic germline variants, and improved understanding of tumor diversity and subtyping [54] and the immune landscape [55]. Other studies map the mutation landscape during oncogenesis with respect to the context-depending impact of a mutation on different gene regulatory levels, e.g. epigenome and transcriptome [56], or provided a unique clustering of tumor types and subtypes using multi-omic datasets [57].

To complete this description of the tumor regulation, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) recently added large-scale proteomics data for breast, colorectal, and

ovarian cancer to TCGA. The new datasets allowed for first large-scale proteogenomic analyses [58**,59,60]. They showed, for example, that unbiased clustering of protein expression and phosphorylation data do not recapitulate the RNA-based subtyping in breast cancer, revealing an orthogonal axis of three clusters [58**]. In addition, the proteomic and phosphoproteomic data revealed functional consequences of whole chromosome-scale amplifications and deletions and pinpointed candidates for druggable kinases, an aim that has been long sought after.

Next steps towards making the sum bigger than the parts

Multi-omics data has enabled scientists to link different biomolecules to improve on genomic annotations, gain a more comprehensive picture of gene expression regulation than any single -omics, and identify molecular switches driving various phenotypes of interest. However, data integration remains challenging as it requires careful choice of the experimental methods and design, including simultaneous controlled data acquisition, and analysis strategies capable of interpreting relationships between the data.

For example, RNA and protein concentration measurements are often casually used as estimates of *transcription* and *translation*, respectively. However, the observed concentrations are the result of both synthesis and degradation -- and to truly estimate these parameters, one has to be disentangled from the other, either through experiment or computational modeling. First tools and studies attempting this disentanglement have emerged, e.g. [47,48,61,62**].

Another example illustrating the many considerations to be made in choosing the appropriate method is the study of translation, for which both sequencing and mass spectrometry based approaches are available (Figure 1). Sequencing-based ribosome footprinting generally has much larger genome coverage than proteomics, provides actual ribosome positions, and does not require metabolic labeling. However, ribosome footprinting typically does not estimate actual translation *rates*, but only - indirectly - translation efficiency via the number of ribosomes bound to an mRNA. In comparison, proteomics-based methods can estimate translation rates directly and even simultaneously assess protein stability, e.g. [63], yet they require metabolic labeling and provide only partial coverage of the genome. Recent work on multiple myeloma cells underlined the importance of proteomics-based measurements of translation: while estimates from ribosome footprinting and pulse-chase proteomics experiments agreed under normal conditions, the correlation broke down when the cells were treated with a proteasome inhibitor [62**]. The proteomics method could detect changes in response to this stress not observed by sequencing.

Proteogenomics has contributed enormously to cancer research, e.g. through CPTAC 3.0 and the International Cancer Proteogenome Consortium. However, major challenges remain. For example, mass spectrometry based peptide coverage of the proteome is still far from being comprehensive: in a recent proteogenomic analysis, mutated peptide sequences with mass spectral evidence verified <10% of the exome variants found in patient-derived-xenografts [64*]. With the continuous improvement of mass spectrometry instrumentation, we should

soon be able to know if this low verification is due to technical limitation or, rather excitingly, indicative of a ‘buffering’ of genomic variation at the protein level.

Other challenges in big multi-omics data acquisition and analysis lie in the inclusion of new ‘dimensions’ such as time. Instead of examination of steady state, we need to study the transition between states to obtain a more realistic picture of the cellular dynamics when striving for homeostasis. New dimensions also lie in metabolomics and lipidomics, as well as in more comprehensive assessments of DNA, RNA, and protein modifications. The mapping between these data is non-trivial and needs to include additional information that can properly account for diverse interactions. First such successful examples have arisen, such as for cancer studies in which alterations in transcriptomes only provide meaningful, disease-relevant interpretations once their impact on protein abundance and protein modifications are taken into account [58**].

Finally, to provide meaningful data interpretation, new statistical methods need not only synthesize information from several sources, but be scalable and adopt standards that streamline the analysis as well. Such standardization is still challenging, as the context is different for every biological or clinical inquiry. The examples described here, while highly interesting, are also specific to individual cases, e.g. they find splice variants in a single organism, discover and quantitate patient specific mutations in one cohort, or dissect regulatory layers in the cell’s response to a specific stimulus. Developments in statistical and visualization methods must enable biological inference at the level of molecular interactions, pathways and networks generated from these multi-omics data [65]. Future work will move towards expanding these studies to - one day in the near future - truly understand the cell as a system.

Acknowledgments

The work was supported by the NIH/NIGMS grant 1R35GM127089–01 (to C.V.) and Singapore Ministry of Education grant MOE2016-T2–1-001(to H.C.).

References

1. Branca RMM, Orre LM, Johansson HJ, Granholm V, Huss M, Pérez-Bercoff Å, Forshed J, Käll L, Lehtiö J: HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods* 2014, 11:59–62. [PubMed: 24240322]
2. Thakur SS, Geiger T, Chatterjee B, Bandilla P, Fröhlich F, Cox J, Mann M: Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol Cell Proteomics* 2011, 10:M110.003699.
3. Shishkova E, Hebert AS, Coon JJ: Now, More Than Ever, Proteomics Needs Better Chromatography. *Cell Syst* 2016, 3:321–324. [PubMed: 27788355]
4. Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, Ogiwara A, Meissen J, Showalter M, Takeuchi K, et al.: Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* 2018, 15:53–56. [PubMed: 29176591]
5. Yang K, Han X: Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. *Trends Biochem Sci* 2016, 41:954–969. [PubMed: 27663237]
6. Lee TA, Bailey-Serres J: Lighting the shadows: methods that expose nuclear and cytoplasmic gene regulatory control. *Curr Opin Biotechnol* 2018, 49:29–34. [PubMed: 28800418]

7. Cafarelli TM, Desbuleux A, Wang Y, Choi SG, De Ridder D, Vidal M: Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Curr Opin Struct Biol* 2017, 44:201–210. [PubMed: 28575754]
8. Zhao L, Ackerman SL: Endoplasmic reticulum stress in health and disease. *Curr Opin Cell Biol* 2006, 18:444–452. [PubMed: 16781856]
9. Morimoto R: Proteostasis: monitoring the health of the proteome in biology, aging and disease. *Alzheimers Dement* 2013, 9:P512.
10. Labbadia J, Morimoto RI: The Biology of Proteostasis in Aging and Disease. *Annu Rev Biochem* 2015, 84:435–464. [PubMed: 25784053]
11. Wek RC, Cavener DR: Translational control and the unfolded protein response. *Antioxid Redox Signal* 2007, 9:2357–2371. [PubMed: 17760508]
12. Harding HP, Zhang Y, Ron D: Protein translation and folding are coupled by an endoplasmicreticulum-resident kinase. *Nature* 1999, 397:271–274. [PubMed: 9930704]
13. Ron D, Walter P: Signal integration in the endoplasmic reticulum unfolded protein response. *Nat Rev Mol Cell Biol* 2007, 8:519–529. [PubMed: 17565364]
14. Vattem KM, Wek RC: Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci U S A* 2004, 101:11269–11274. [PubMed: 15277680]
15. Calton M, Zeng H, Urano F, Till JH, Hubbard SR, Harding HP, Clark SG, Ron D: IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. *Nature* 2002, 415:92–96. [PubMed: 11780124]
16. Hollien J, Lin JH, Li H, Stevens N, Walter P, Weissman JS: Regulated Ire1-dependent decay of messenger RNAs in mammalian cells. *J Cell Biol* 2009, 186:323–331. [PubMed: 19651891]
17. Plemper RK, Wolf DH: Retrograde protein translocation: ERADication of secretory proteins in health and disease. *Trends Biochem Sci* 1999, 24:266–270. [PubMed: 10390615]
18. Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, Fenyö D, Zhang B, Mani DR: Methods, Tools and Current Perspectives in Proteogenomics. *Mol Cell Proteomics* 2017, 16:959–981. [PubMed: 28456751]
19. McAfee A, Foster LJ: Proteogenomics: Recycling Public Data to Improve Genome Annotations. *Methods Enzymol* 2017, 585:217–243. [PubMed: 28109431]
20. Menschaert G, Fenyö D: Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrom Rev* 2015, 36:584–599. [PubMed: 26670565]
21. Rodriguez H, Pennington SR: Revolutionizing Precision Oncology through Collaborative Proteogenomics and Data Sharing. *Cell* 2018, 173:535–539. [PubMed: 29677503]
22. Martens L, Vizcaino JA: A Golden Age for Working with Public Proteomics Data. *Trends Biochem Sci* 2017, 42:333–341. [PubMed: 28118949]
23. Karczewski KJ, Snyder MP: Integrative omics for health and disease. *Nat Rev Genet* 2018, 19:299–310. [PubMed: 29479082]
24. Manzoni C, Kia DA, Vandrovцова J, Hardy J, Wood NW, Lewis PA, Ferrari R: Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform* 2018, 19:286–302. [PubMed: 27881428]
25. Jean Beltran PM, Federspiel JD, Sheng X, Cristea IM: Proteomics and integrative omic approaches for understanding host-pathogen interactions and infectious diseases. *Mol Syst Biol* 2017, 13:922. [PubMed: 28348067]
26. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995, 269:496–512. [PubMed: 7542800]
27. Jaffe JD, Berg HC, Church GM: Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 2004, 4:59–77. [PubMed: 14730672]
28. Lobas AA, Karpov DS, Kopylov AT, Solovyeva EM, Ivanov MV, Ilina IY, Lazarev VN, Kuznetsova KG, Ilgisonis EV, Zgoda VG, et al.: Exome-based proteogenomics of HEK-293 human cell line: Coding genomic variants identified at the level of shotgun proteome. *Proteomics* 2016, 16:1980–1991. [PubMed: 27233776]

29. Chapman B, Bellgard M: Plant Proteogenomics: Improvements to the Grapevine Genome Annotation. *Proteomics* 2017, 17.
 30. Proffitt JM, Glenn J, Cesnik AJ, Jadhav A, Shortreed MR, Smith LM, Kavanagh K, Cox LA, Olivier M: Proteomics in non-human primates: utilizing RNA-Seq data to improve protein identification by mass spectrometry in vervet monkeys. *BMC Genomics* 2017, 18:877. [PubMed: 29132314]
 31. Marx H, Hahne H, Ulbrich SE, Schnieke A, Rottmann O, Frishman D, Kuster B: Annotation of the Domestic Pig Genome by Quantitative Proteogenomics. *J Proteome Res* 2017, 16:2887–2898. [PubMed: 28625053]
 32. Prasad TSK, Mohanty AK, Kumar M, Sreenivasamurthy SK, Dey G, Nirujogi RS, Pinto SM, Madugundu AK, Patil AH, Advani J, et al.: Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. *Genome Res* 2017, 27:133–144. [PubMed: 28003436]
 33. Armengaud J, Pible O, Gaillard J-C, Cian A, Gantois N, Tan KSW, Chabé M, Viscogliosi E: Proteogenomic Insights into the Intestinal Parasite *Blastocystis* sp. Subtype 4 Isolate WR1. *Proteomics* 2017, 17.
 34. Lee S-E, Song J, Bösl K, Müller AC, Vitko D, Bennett KL, Superti-Furga G, Pandey A, Kandasamy RK, Kim M-S: Proteogenomic Analysis to Identify Missing Proteins from Haploid Cell Lines. *Proteomics* 2018, 18:e1700386. [PubMed: 29474001]
 35. Ma J, Saghatelian A, Shokhirev MN: The influence of transcript assembly on the proteogenomics discovery of microproteins. *PLoS One* 2018, 13:e0194518. [PubMed: 29584760]
 - 36*. Willems P, Ndah E, Jonckheere V, Stael S, Sticker A, Martens L, Van Breusegem F, Gevaert K, Van Damme P: N-terminal Proteomics Assisted Profiling of the Unexplored Translation Initiation Landscape in *Arabidopsis thaliana*. *Mol Cell Proteomics* 2017, 16:1064–1080.
- Proteogenomics exploited sequence and N-terminal proteomics data for the identification of novel translation initiation sites. The analysis identified 117 N-terminal extensions by peptide evidence; 23 of these were further supported by ribosome profiling data
- . [PubMed: 28432195]
- 37*. Zhu F-Y, Chen M-X, Ye N-H, Shi L, Ma K-L, Yang J-F, Cao Y-Y, Zhang Y, Yoshida T, Fernie AR, et al.: Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in *Arabidopsis* seedlings. *Plant J* 2017, 91:518–533.
- Using single molecule long-read sequencing, the authors reveal that >80% of intron containing genes in *Arabidopsis* express multiple splice variants under normal conditions, although the variants are rarely detected at the protein level. However, after treatment with abscisic acid, differentially expressed isoforms are more likely observed at the protein level suggesting possible synergy between isoform transcript expression and translation
- . [PubMed: 28407323]
38. Lau E, Cao Q, Lam MPY, Wang J, Ng DCM, Bleakley BJ, Lee JM, Liem DA, Wang D, Hermjakob H, et al.: Integrated omics dissection of proteome dynamics during cardiac remodeling. *Nat Commun* 2018, 9.
 39. Steinberg J, Ritchie GRS, Roumeliotis TI, Jayasuriya RL, Clark MJ, Brooks RA, Binch ALA, Shah KM, Coyle R, Pardo M, et al.: Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis. *Sci Rep* 2017, 7:8935. [PubMed: 28827734]
 40. Liu Y, Borel C, Li L, Müller T, Williams EG, Germain P-L, Buljan M, Sajic T, Boersema PJ, Shao W, et al.: Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. *Nat Commun* 2017, 8:1212. [PubMed: 29089484]
 41. Aviner R, Shenoy A, Elroy-Stein O, Geiger T: Uncovering Hidden Layers of Cell Cycle Regulation through Integrative Multi-omic Analysis. *PLoS Genet* 2015, 11:e1005554. [PubMed: 26439921]

- 42**. Cheng Z, Otto GM, Powers EN, Keskin A, Mertins P, Carr SA, Jovanovic M, Brar GA: Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. *Cell* 2018, 172:910–923.e16.
- Comparing RNA and protein abundances during meiosis in yeast, ~8% of all genes measured (N=380) display anti-correlation between the transcript and protein level. Incorporating ribosome profiling data, the authors then show that this anti-correlation is due to a switch between canonical transcription start sites that result in a translated mRNA and a 5' terminal extended isoform that is inefficiently translated. A single transcription factor is responsible for this switch among this set of genes. The same transcription factor can also induce a separate set of genes thereby acting both an activator and a repressor
- . [PubMed: 29474919]
43. Seaton D, Graf A, Baerenfaller K, Stitt M, Millar A, Gruissem W: Photoperiodic control of the Arabidopsis proteome reveals a translational coincidence mechanism. 2017, 10.1101/182071.
44. Liu Y, Aebersold R: The interdependence of transcript and protein abundance: new data--new complexities. *Mol Syst Biol* 2016, 12:856. [PubMed: 26792872]
45. McManus J, Cheng Z, Vogel C: Next-generation analysis of gene expression regulation-comparing the roles of synthesis and degradation. *Mol Biosyst* 2015, 11:2680–2689. [PubMed: 26259698]
46. Vogel C, Marcotte EM: Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* 2012, 13:227–232. [PubMed: 22411467]
47. Jovanovic M, Rooney MS, Mertins P, Przybylski D, Chevrier N, Satija R, Rodriguez EH, Fields AP, Schwartz S, Raychowdhury R, et al.: Dynamic profiling of the protein life cycle in response to pathogens. *Science* 2015, 347:1259038–1259038. [PubMed: 25745177]
48. Cheng Z, Teo G, Krueger S, Rock TM, Koh HWL, Choi H, Vogel C: Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol Syst Biol* 2016, 12:855.
49. Zappulo A, van den Bruck D, Ciolli Mattioli C, Franke V, Imami K, McShane E, Moreno-Estelles M, Calviello L, Filipchuk A, Peguero-Sanchez E, et al.: RNA localization is a key determinant of neurite-enriched proteome. *Nat Commun* 2017, 8:583. [PubMed: 28928394]
50. Kustatscher G, Grabowski P, Rappsilber J: Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol Syst Biol* 2017, 13:937.
51. Blum A, Wang P, Zenklusen JC: SnapShot: TCGA-Analyzed Tumors. *Cell* 2018, 173:530. [PubMed: 29625059]
52. The TCGA Legacy. *Cell* 2018, 173:281–282. [PubMed: 29625044]
53. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al.: An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 2018, 173:400–416.e11. [PubMed: 29625055]
54. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al.: The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015, 21:1350–1356. [PubMed: 26457759]
55. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al.: The Immune Landscape of Cancer. *Immunity* 2018, 48:812–830.e14. [PubMed: 29628290]
56. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang K-L, Tokheim C, et al.: Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* 2018, 173:305–320.e10. [PubMed: 29625049]
57. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al.: Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* 2018, 173:291–304.e6. [PubMed: 29625048]
- 58**. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, et al.: Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 2016, 534:55–62.

- Quantitative proteomics of 77 high-quality breast cancer samples elucidates the impacts of various somatic mutations on protein expression, building on the previous subtype clustering using only RNA expression data. The further addition of phosphoproteomics data reveals a new G-protein-coupled-receptor cluster that would not otherwise have been identified
- . [PubMed: 27251275]
59. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al.: Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014, 513:382–387. [PubMed: 25043054]
 60. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou J-Y, Petyuk VA, Chen L, Ray D, et al.: Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* 2016, 166:755–765. [PubMed: 27372738]
 61. Teo G, Bin Zhang Y, Vogel C, Choi H: PECAplus: statistical analysis of time-dependent regulatory changes in dynamic single-omics and dual-omics experiments. *NPJ Syst Biol Appl* 2018, 4:3. [PubMed: 29263799]
 - 62*. Liu T-Y, Huang HH, Wheeler D, Xu Y, Wells JA, Song YS, Wiita AP: Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. *Cell Syst* 2017, 4:636–644.e9.
- Ribosome profiling and pulsed-SILAC estimates of protein synthesis in multiple myeloma cells correlate well in static conditions. However, in dynamic conditions after treatment with the protease inhibitor bortezomib, ribosome footprints and pSILAC data diverge as the proteome is remodelled. Pulsed-SILAC measurements more accurately reflect alterations in translation rates
- . [PubMed: 28578850]
63. Savitski MM, Zinn N, Faelth-Savitski M, Poeckel D, Gade S, Becher I, Muelbaier M, Wagner AJ, Strohmmer K, Werner T, et al.: Multiplexed Proteome Dynamics Profiling Reveals Mechanisms Controlling Protein Homeostasis. *Cell* 2018, 173:260–274.e25. [PubMed: 29551266]
 - 64*. Ruggles KV, Tang Z, Wang X, Grover H, Askenazi M, Teubl J, Cao S, McLellan MD, Clauser KR, Tabb DL, et al.: An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol Cell Proteomics* 2016, 15:1060–1071.
- Proteogenomic analysis of breast-cancer-patient-derived xenografts revealed that only about 10% of somatic mutations noted in DNA or RNA sequencing resulted in a translated mutant peptide. These variant peptides are differentially expressed when comparing between basal-like and luminal suggesting differences in translation and/or degradation rates
- . [PubMed: 26631509]
65. Huang S, Chaudhary K, Garmire LX: More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front Genet* 2017, 8:84. [PubMed: 28670325]

Figure 1. Advanced technologies produce omics data around the Central Dogma of Biology
 In proteogenomics work, mass spectrometry based proteomics data is typically integrated with sequencing based DNA and RNA data, providing insights into different aspects of the different steps of gene expression regulation (blue arrows and text). Both sequencing and proteomics have many different methods that assess, at large scale, a specific aspect of the Central Dogma of Biology. The figure exemplifies these methods (black arrows, beige boxes). Some processes such as translation, can be assessed by both sequencing and proteomics.

Figure 2. The Unfolded Protein Response exemplifies gene expression regulation at multiple levels

Many biological processes are regulated at multiple levels in a dynamic, highly coordinated fashion. Many regulators have unknown targets. The figure illustrates this complexity at the example of the Unfolded Protein Response (UPR). The UPR pathways shown here are much simplified to illustrate the different processes involved (blue), key regulators (black italics), and organelles (grey italics). The goal of integrated large-scale work is to identify new targets for these regulators that expand and interconnect this regulatory network further.