Cell. Mol. Life Sci. (2015) 72:953–957
DOI 10.1007/s00018-015-1837-y

VISIONS AND REFLECTIONS

Proteogenomics: emergence and promise

Sam Faulkner · Matthew D. Dun · Hubert Hondermarck

Received: 10 November 2014 / Revised: 8 January 2015 / Accepted: 12 January 2015 / Published online: 22 January 2015

Abstract Proteogenomics, or the integration of proteomics with genomics and transcriptomics, is emerging as the next step towards a unified understanding of cellular functions. Looking globally and simultaneously at gene structure, RNA expression, protein synthesis and post-translational modifications have become technically feasible and offer a new perspective to molecular processes. Recent publications have highlighted the value of proteogenomics in oncology for defining the molecular signature of human tumors, and translation to other areas of biomedicine and life sciences is anticipated. This mini-review will discuss recent developments, challenges and perspectives in proteogenomics.

Keywords Omics · Proteomics · Transcriptomics · Genomics · Metabolomics

Introduction

The completion of genome projects, for human and other species, has not only provided a considerable amount of information on DNA and gene structure, but it has also opened the way for the global investigation of gene expression. Building on that foundation, transcriptomics has progressively evolved from the simple analysis of individual transcript levels using microarrays to the simultaneous sequencing of all RNAs expressed in a living entity. As a consequence, the integration of genomics with transcriptomics, also termed functional genomics, has led to a better understanding of the relationship between genotypes and phenotypes, thus significantly impacting the fields of biology and medicine. For instance, genomic and transcriptomic profiles have been established for several pathologies including cancer, and this is starting to impact clinical practice for disease diagnosis, prognosis as well as prediction of risk. However, it has also been realized that the gene and transcript levels could not be regarded as the ultimate window for understanding gene functions and associated phenotypes, as proteins are the functional effectors of gene function. Thus, the necessity of a better integration of functional genomics with analysis at the protein level, proteomics, has progressively emerged. Here, we will discuss the latest developments in integrating the omics, proteogenomics, and the new opportunities it opens in cellular and molecular life sciences.

Proteomics: a step closer to function

It is now clearly established that the abundance of an individual protein cannot be predicted in confidence by the level of the corresponding mRNA. Initial studies in bacteria and yeast [1] have suggested a reasonable correlation of about $\sim$50 % between mRNA and protein levels, but studies in multicellular eukaryotes have revealed a much lower correlation. In humans, global transcriptomic and proteomic analyses have shown that only 30 % of changes in protein levels can be explained by corresponding variations in mRNA [2]. This discrepancy emphasizes the importance of post-transcriptional regulations. Variable translation efficiency of mRNA and regulation by siRNA account at least partially for the difference between mRNA and protein

S. Faulkner · M. D. Dun · H. Hondermarck (✉)
Faculty of Health and Medicine, School of Biomedical Sciences
and Pharmacy and Hunter Medical Research Institute, Life
Science Building, University of Newcastle,
Callaghan NSW 2308, Australia
e-mail: hubert.hondermarck@newcastle.edu.au

levels, but the degradation and dynamic turnover of proteins are also involved. Ranging from a few minutes to several days, the half-life of proteins is controlled by various pathways such as, but not limited to, the ubiquitin and proteasome pathway [3]. Furthermore, the presence or absence of post-translational modifications, such as phosphorylation, glycosylation or ubiquitinylation, has a strong impact on protein stability, adding a further level of complexity. Together, the composition and dynamics of the proteome is not deducible from functional genomics data, and as a consequence proteomics is an indispensable and complementary approach to genomics and transcriptomics.

It is interesting to note that the existence of amino acids and the structure of proteins were established at the beginning of the twentieth century, long before the structure of DNA/RNA was elucidated. However, the genome of many species has already been sequenced but we are still waiting for the definition of their entire proteomes. A first draft of the human proteome has just been delineated [4] but we are still far from a complete description, and the main source of complexity appears to be the organ and tissue compartmentalization of proteins. There are about 200 cell types in the human body, which are organized in tissues and organs, thus creating a variety of proteomes. Therefore, defining the proteomes of human and of other species is going to be challenging and will require a great deal of international effort and coordination.

From a methodological standpoint, proteomic analysis of a living entity is made possible because its genome has been sequenced and is accessible for online interrogation [5]. After proteolytic digestion of a protein extract, the resulting protein fragments are analyzed in mass spectrometry and the peptide sequences obtained are used to interrogate genomic databases for identifying proteins. Mining genomics databases and looking for sequence homology that match with mass spectrometry data is the essence of any modern proteomic analysis. Mass spectrometry has progressively become more sensitive and high throughput for protein identification, as illustrated with shotgun proteomics [6], and is also the central tool for protein quantification and comparative analyses, as well as determination of post-translational modifications. It should be emphasized here that bioinformatics is essential to proteomics, just as it is to genomics and transcriptomics [7], and that it is also the cornerstone for integrating data obtained from the different omics.

The emergence of proteogenomics

Two essential facts differentiate genomic and transcriptomic from proteomic approaches. First, proteins cannot be amplified: there is no PCR for proteins and therefore as much protein necessary for identification and quantification has to be purified prior to analysis. Second, at this stage, there are no protein arrays that work efficiently for large-scale analysis. Not only are antibodies not available against all proteins, but also post-translational modifications can alter antibody recognition and affinity. Overall, the methodological approaches used in transcriptomics and proteomics are fundamentally different in principle and as a consequence the integration of these omics has remained a challenge. In this regard, the situation in humans is indicative of the difficulties and challenges. With about 20,000 genes and 10^6 proteins bearing more than 200 types of post-translational modifications, the molecular complexity clearly increases from the genome to the proteome and integrating these complementary levels of complexity requires a proportional increase in fractionation and enrichment techniques [8], as well as in computing and bioinformatics [7].

The concept of proteogenomics (Fig. 1) is that the three levels, DNA and epigenetic regulations, RNA expression,

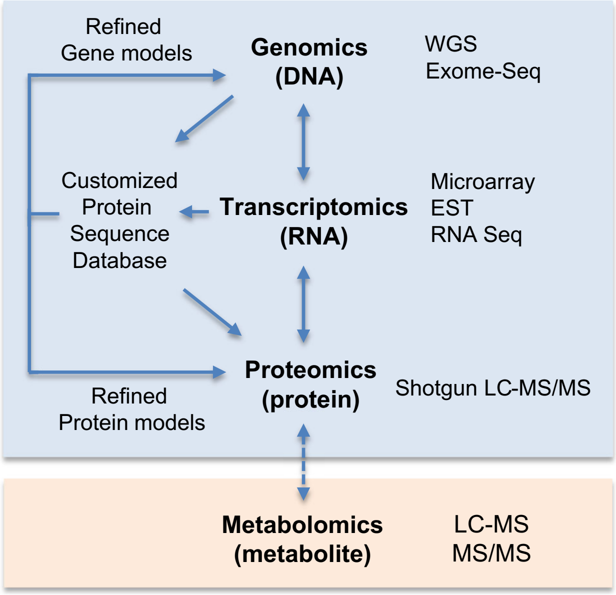

Fig. 1 Concept and methods in proteogenomics. An integrated approach including analyses at the DNA (genomics), RNA (transcriptomics) and protein (proteomics) levels is called proteogenomics and allows assembling the molecular puzzle driving cellular functions. On one hand, genomic and transcriptomic data are used to obtain a customized database of theoretical proteins. On the other hand, proteomic data open a protein window into gene expression data. This reciprocal improvement requires handling of information obtained from the different technologies commonly used in genomics (*WGS* whole genome sequencing, exome sequencing), transcriptomics (microarrays; *EST* expressed sequence tag; RNA sequencing) and proteomics (shotgun *LC–MS/MS* Liquid chromatography tandem mass spectrometry). The analysis of metabolites (metabolomics) and corresponding technologies (*LC–MS* liquid chromatography–mass spectrometry, *MS/MS* tandem mass spectrometry) is the next omics to be integrated. Of note, the methodological proximity between proteomics and metabolomics, both based on the use of mass spectrometry, should facilitate future integration

protein and their post-translational modifications are simultaneously investigated and integrated. Although this is intellectually seducing, until recently it had not been practically implemented. This situation has rapidly changed over the last 2 years and based on methodological convergence and progress in bioinformatics, robust approaches in proteogenomics have been successfully developed [9]. The essential step in proteogenomics is the creation of a customized protein sequence database, derived from genomic data, which can then be utilized to study the model of interest. Because whole genome or exome sequencing and RNA sequencing are now feasible at an accessible cost, it is possible to delineate the entire set of theoretical protein sequences present in particular structures (such as cells or tissues) to be studied. Following this, it is possible to identify proteins using shotgun LC–MS/MS against this customized protein sequence database, instead of using publically available generic databases. As a result, this process delivers increased confidence in protein identification/quantification on one hand, and on the other hand it provides a protein-level validation of gene expression. The overall benefit is a holistic view of the molecular landscape from genes to proteins and a refinement of both protein and gene models. Technical challenges inherent to genomics and proteomics (sensitivity, reproducibility, accuracy) remain relevant in a proteogenomics investigation. In particular, precision and accuracy in both gene and protein sequencing is probably the most crucial issue at this stage, as the existence of errors in annotation can be misleading to subsequent functional and clinical investigations. Although caution has to be applied to avoid potential errors to be amplified during the crossover of genomics and proteomics data, primarily occurring during the constitution of the customized sequence database, proteogenomics ultimately provides a powerful means to correct mistakes in both gene and protein sequences.

Proteogenomics breakthrough in oncology

The emergence of proteogenomics is best illustrated with cancer research where major advances have been made. Cancer progression is driven by genomic alterations and instability that result in a series of genomic changes including mutations, methylation, copy number aberrations or translocation [10]. Until recently, most efforts to define molecular changes associated with oncogenesis have been driven by deep genome sequencing under the leadership of the International Cancer Genome Consortium [11] and The Cancer Genome Atlas (TCGA) project [12]. However, as those projects were progressively advanced, it became clear that linking cancer genotypes to phenotypes would also require the definition of cancer proteotypes. In the same time, progress in high-throughput proteomics (shotgun proteomics) has made proteomics a reliable and large-scale approach with capabilities matching those of genomics for the analysis of tumor and blood samples. In this context, the National Cancer Institute (NCI) launched the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [13] in 2011 to accelerate the understanding of the molecular basis of cancer through the application of robust, quantitative, proteomic technologies and workflows [14]. The overarching goal of CPTAC is to improve the ability to diagnose, treat and prevent cancer by defining a proteomic signature to major human cancers.

Significant progress in CPTAC program has recently been illustrated with a major publication [15] describing a proteogenomic characterization of human colon and rectal cancers. This study has identified 5 proteomic subtypes in the TCGA colorectal cancer cohort. Interestingly, chromosome 20q amplicon was associated with the largest global changes at both mRNA and protein levels and proteomic data indicated potential 20q candidate biomarkers and therapeutic targets in colorectal cancer. The results also highlighted that messenger RNA transcript abundance did not reliably predict protein abundance differences between tumors. It is worth noting that tumor classifications established so far are entirely based on genomic and transcriptomic analyses. For instance, in breast cancer, four main classes of tumors (luminal A, luminal B, HER2, triple negative/basal like) have been defined based on gene expression profile and this classification is regularly updated by the addition of subclasses [16] to better match clinical data, particularly in terms of treatment response. The demonstration of a low correlation between mRNA and protein abundance in colorectal cancer [15] raises the limitations of current molecular classifications solely based on gene expression, and it is clear that a refinement at the proteomic level is necessary. Overall, these developments show that proteogenomics has the ability to deliver an additional dimension to the understanding of tumor molecular biology [17] and this could lead to improved diagnostic and therapeutic strategies. In particular, it should be emphasized that traditional genomic and proteomic studies typically use a reference database to model the general population, therefore masking patient specific variation. Perhaps, the most significant potential of proteogenomics for making the leap from the bench to the clinic is to provide a means for personalized analysis, via the constitution of an individual patient-based customized gene/protein sequence database. This clearly represents a significant milestone towards individualized cancer medicine.

In a broader perspective, these recent achievements in cancer proteogenomics are paving the way for similar investigations in other diseases as well as in other fields of biology. Proteogenomics has already been implemented in microbiology [18, 19], plant biology [20, 21] and environmental sciences [22, 23]. These investigations have led to the discovery and revision of genes and proteins involved in basic physiological processes, but also they have helped to remove questionable gene annotation assignments and to confirm the presence of post-translational modifications. Interestingly, some of the issues facing these disciplines, and in particular microbial and plant population heterogeneity, are common with cancer biology. The molecular heterogeneity of tumors and cancer patient population is a limitation to the efficacy of the management of the disease in terms of prediction of risk, diagnosis and treatment. Therefore, the application of proteogenomics across areas of life sciences will presumably result in mutual improvement and should foster its widespread use.

Future directions and conclusion

Although proteogenomics is the latest development in omics, with anticipated practical outcomes in biology and medicine, it is still in infancy stage and future developments are needed before it can become widely used. First, further bioinformatics integration is required to fully exploit the entire spectra of information obtained in genomic, transcriptomic and proteomic investigations. Second, proteogenomics outputs will have to be made accessible to the research community. At the present stage, this is not really the case and more databases, informatics interfaces and dedicated software will have to be developed. Third, it already appears that proteogenomics will not be the ultimate stage of molecular integration from genotypes to phenotypes, as we are already witnessing the birth of the next omics: metabolomics [24]. Defining the metabolites produced by all enzymatic activities will be the next level for a better comprehension of living organisms and this will also have to be integrated with proteogenomics. Thus, more integrative challenges are ahead and the future of biology and medicine is probably taking shape in these efforts.

Acknowledgments This work was supported by the University of Newcastle Australia and Hunter Cancer Research Alliance.

Conflict of interest The authors declare no conflict of interest.

References

1. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19(3):1720–1730
2. Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 13(4):227–232. doi:10.1038/nrg3185
3. Claydon AJ, Beynon R (2012) Proteome dynamics: revisiting turnover with a global perspective. Mol Cell Proteomics 11(12):1551–1565. doi:10.1074/mcp.O112.022186
4. Kim MS, Pinto SM et al (2014) A draft map of the human proteome. Nature 509(7502):575–581. doi:10.1038/nature13302
5. Tyers M, Mann M (2003) From genomics to proteomics. Nature 422(6928):193–197. doi:10.1038/nature01510
6. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR 3rd (2013) Protein analysis by shotgun/bottom-up proteomics. Chem Rev 113(4):2343–2394. doi:10.1021/cr3003533
7. Schneider MV, Orchard S (2011) Omics technologies, data and bioinformatics principles. Methods Mol Biol 719:3–30. doi:10.1007/978-1-61779-027-0_1
8. Altelaar AF, Heck AJ (2012) Trends in ultrasensitive proteomics. Curr Opin Chem Biol 16(1–2):206–213. doi:10.1016/j.cbpa.2011.12.011
9. Nesvizhskii AI (2014) Proteogenomics: concepts, applications and computational strategies. Nat Methods 11(11):1114–1125. doi:10.1038/nmeth.3144
10. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674. doi:10.1016/j.cell.2011.02.013
11. International Cancer Genome C, Hudson TJ et al (2010) International network of cancer genome projects. Nature 464(7291):993–998. doi:10.1038/nature08987
12. Cancer Genome Atlas Research N, Weinstein JN et al (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45(10):1113–1120. doi:10.1038/ng.2764
13. Ellis MJ, Gillette M et al (2013) Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. Cancer Discov 3(10):1108–1112. doi:10.1158/2159-8290.CD-13-0219
14. Rivers RC, Kinsinger C et al (2014) Linking cancer genome to proteome: NCI's investment into proteogenomics. Proteomics. doi:10.1002/pmic.201400193
15. Zhang B, Wang J et al (2014) Proteogenomic characterization of human colon and rectal cancer. Nature 513(7518):382–387. doi:10.1038/nature13438
16. Dawson SJ, Rueda OM, Aparicio S, Caldas C (2013) A new genome-driven integrated classification of breast cancer and its implications. EMBO J 32(5):617–628. doi:10.1038/emboj.2013.19
17. Alfaro JA, Sinha A, Kislinger T, Boutros PC (2014) Onco-proteogenomics: cancer proteomics joins forces with genomics. Nat Methods 11(11):1107–1113. doi:10.1038/nmeth.3138
18. Chapman B, Bellgard M (2014) High-throughput parallel proteogenomics: a bacterial case study. Proteomics 14(23–24):2780–2789. doi:10.1002/pmic.201400185
19. Kucharova V, Wiker HG (2014) Proteogenomics in microbiology: taking the right turn at the junction of genomics and proteomics. Proteomics 14(23–24):2360–2675. doi:10.1002/pmic.201400168
20. Castellana NE, Shen Z et al (2014) An automated proteogenomic method uses mass spectrometry to reveal novel genes in *Zea mays*. Mol Cell Proteomics 13(1):157–167. doi:10.1074/mcp.M113.031260
21. Chapman B, Castellana N et al (2013) Plant proteogenomics: from protein extraction to improved gene predictions. Methods Mol Biol 1002:267–294. doi:10.1007/978-1-62703-360-2_21
22. Trapp J, Armengaud J, Salvador A, Chaumot A, Geffard O (2014) Next-generation proteomics: towards customized biomarkers for environmental biomonitoring. Environ Sci Technol 48(23):13560–13572. doi:10.1021/es501673s

23. Armengaud J, Hartmann EM, Bland C (2013) Proteogenomics for environmental microbiology. Proteomics 13(18–19):2731–2742. doi:10.1002/pmic.201200576

24. Patti GJ, Yanes O, Siuzdak G (2012) Innovation: metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 13(4):263–269. doi:10.1038/nrm3314