Editorial

The potential of proteogenomics in oncology

Pamela Pinzani, Francesca Salvianti

Department of Clinical, Experimental and Biomedical Sciences, University of Florence, Florence, Italy

Correspondence to: Pamela Pinzani. Department of Clinical, Experimental and Biomedical Sciences, University of Florence, Viale Pieraccini 6, 50139 Florence, Italy. Email: pamela.pinzani@unifi.it.

Comment on: Mertins P, Mani DR, Ruggles KV, *et al*. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 2016;534:55-62.

Submitted Aug 29, 2016. Accepted for publication Sep 05, 2016.
doi: 10.21037/tcr.2016.10.12
View this article at: http://dx.doi.org/10.21037/tcr.2016.10.12

Great improvements have been achieved during the last 5 years on the detection of somatic mutations and this mainly thanks to the introduction of massive parallel sequencing technologies (NGS). The development of high-throughput methods allows a deeper, easier and faster analysis of the mutational status of tumors and, with no doubts, is going to lead to a new classification of the oncologic diseases that will parallel and permeate the standard anatomo-histological criteria.

NGS is also bringing new potential in the detection of transcripts and the analysis of mRNA expression will take advantage of this new technology as well as of other approaches recently introduced on the market for the reliable and multiplexed analysis of gene expression. NGS has deeply increased the potential of genome characterization generating a huge amount of data by next generation transcriptome sequencing (RNA-Seq) which disclosed many novel transcripts requiring protein-level evidence of expression. Post transcriptional regulation, including variable translation efficiency of mRNA, regulatory actions as well as the dynamic turnover of proteins and their degradation, account for a general lack of correspondence between the variation of mRNA levels and those of the corresponding protein (1).

If on one side DNA and RNA analysis benefits of amplification-based reactions and can be integrated in functional genomic studies, proteins cannot be amplified as well and, since the complexity grows from genome to proteome, integrating the complementary level of complexity requires a proportional increase in enrichment techniques and computing and bioinformatics (2).

On the other side, proteomic analysis intended as a comprehensive, integrative study of proteins and their biological function is very complex for its final goal that is often represented by the complete and quantitative map of the proteome of a species. The technical approach more recently developed is based on tandem mass spectrometry (MS/MS) after protein digestion named shotgun proteomics. It represents a step forward for protein detection and quantification on a large scale, with one main drawback due to the identification of peptides by referring to MS/MS spectra databases. Many peptides in fact are not present in the reference protein sequence databases (Ensembl, RefSeq or UniProtKb). They may contain mutations and may be novel protein-coding loci and alternative isoforms. Despite the existence of some approaches studied to solve the problem (sequence tag-based searching or *de novo* sequencing), proteogenomics offers an integrated view of genomic, transcriptomic and proteomic features in a sample (*Figure 1*) and represents an alternative tool for the identification of novel peptides (3). It was born in 2004 (4) as an improvement of genome analysis and for the characterization of the protein-coding potential. Now this term indicates the identification of novel peptides in protein sequence databases deriving from genomic and transcriptomic data and customized for the detection of predicted novel protein sequences and sequence variants.

Proteogenomics has been applied in multiple studies in humans and in many other model organisms (5-8) although these early works not always represent a coordinated contribution of genomic and proteomic data. Improvements of proteomic methods and the growing number of genomic data produced by NGS lead to a wider application of proteogenomic studies in the strictest sense of the word. A high number of reports on human proteome is focused on the identification of novel peptides and peptide variants

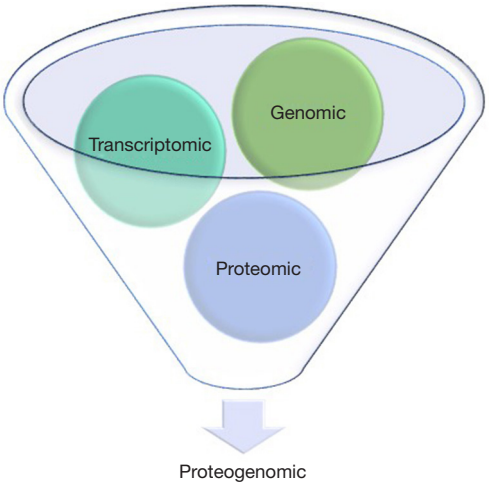

Figure 1 A schematic representation of the proteogenomic integrated approach.

by a proteogenomic approach (9-11). In particular, the most important applications of proteogenomics have been related to the oncology field for the detection of abnormal protein variants in different kind of tumors mainly thanks to the work of the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) (11-13).

The article "Proteogenomics connects somatic mutations to signaling in breast cancer" by Mertins *et al.* and with the collaboration of the CPTAC, published on June 2016 in *Nature* (14), finds its object in the understanding of how genomic changes drive the proteome and phosphoproteome to generate phenotypic characteristics. It is the second of three important papers on three different tumors (colorectal, breast and ovarian respectively) (11,13,14) previously analyzed by The Cancer Genome Atlas (TCGA). The general aim is integrating proteomic measurements with the genomic yielding a number of insights into the diseases (13).

The authors used shotgun proteomics on breast cancer samples previously submitted to genome analysis and produced, by a proteogenomic approach, the analysis of 105 breast tumor samples composed of a balanced representation of the basal-like, luminal A, luminal B and HER2-enriched subtypes defined on the basis of the PAM50 RT-qPCR test for gene expression-based subtyping. Samples were submitted to high-resolution accurate-mass tandem spectrometry after peptide fractionation and phosphopeptide enrichment followed by isobaric peptide labelling (iTRAQ), a tagging method permitting relative expression measurements of large sets of proteins with a high degree of automation (15). Totally 15,396 proteins (12,405 genes) and 62,679 phosphoproteins were identified with a mean per tumor of 11,632 and 26,310 respectively. Data were filtered in order to select only proteins observed in at least a quarter of the samples (12,553 proteins and 33,239 phosphoproteins) which were quantified and used for the evaluation study. It is worth noting that the technical noise was taken into account as well as the influence on data deriving from the use of two distinct tumor fragments for proteomics and genomic analyses. To confirm the need of error rate estimation methods, the results of this analysis implied the exclusion of 28 over the initial 105 tested samples mainly due to protein degradation. Here comes the importance of the preanalytical phase especially for samples undergoing complex analytical processes at the level of molecules prone to degradation and deregulation (as proteins and RNA). The number of identified peptides found by MS/MS corresponding to identified genome and transcriptomic variants was low (corresponding to 4.1% of the total DNA and RNA variants detected).

An advantage of the MS/MS approach is the analysis of multiple peptides for each protein that allows the determination of the overall level of the target protein. This represents one of the advantages over antibody-based protein expression analysis, which detects proteins based on a single epitope.

The analysis of three frequently mutated genes in breast cancer (TP53, PIK3CA and GATA3) and three clinical biomarkers (ER, PGR and ERBB2) identified correlations between protein levels and mutations.

The results are dense of correlations and findings. Among all the reported relationships, the association of TP53 nonsense and frameshift mutations with a decrease of TP53 protein levels demonstrated a reduction in the expression of the mutated proteins. On the other hand, other mutations (such as C-terminal frameshift alterations) in different genes (i.e., GATA3) did not result in decreased protein expression suggesting protein expression despite truncation. For the third gene PIK3CA, no consistent effect of somatic mutations on protein expression could be evidenced.

Overall, the authors report a good correlation between RNA-Seq and MS/MS protein expression levels, even if in particular subgroups of the case study lower protein levels compared to corresponding mRNAs might indicate posttranscriptional regulatory mechanisms such as proteasomal degradation.

In support of the usefulness of global proteome correlation analysis and with the indication of the need of

further investigations, the authors compared copy number alterations (CNAs), RNA and protein levels for some genes of interest and interestingly found some negative correlations between CNAs and RNA and protein levels.

Correlation analysis between CNA-mRNA, CNA-protein and CNA-phosphoprotein paired measurements for 7,776 genes reported significant positive results in 64%, 31% and 20% of genes respectively. Moreover, the study led to the conclusion that CNA events with tumor promoting outcome more likely lead to cis-regulatory effects on both the protein and the mRNA, whereas CNA events with no documented role on tumorogenesis are more likely to be neutralized at the protein level than at the mRNA level. "Hot spots" of significant trans-effects have been evidenced as well.

By taking into account functional knockdown data from the Library of Integrated Network-based Cellular Signature (LINCS) database, the researchers identified 10 candidate driver genes (affected by CNA gain or loss) with CNA that are direct driver of trans-effects. An example of gene functionally connected only to CNA gain trans effect is ERBB2, while E3 ligase SKP1 and the ribonucleoprotein export factor CETN3 were indicated as potential regulators affecting the expression of tyrosine kinase and therapeutic target EGFR.

Although analytical weakness and source of errors can be identified for the proposed proteomic analysis and despite the use of different tissue sections of the same tumor for RNA-seq and protein analyses, very similar subtype-defining features could be observed using the PAM50 breast cancer classification scheme and RNA and proteome clustering.

On the basis of the phosphoproteome profiling the authors developed a signaling pathway-based classification articulated into four robustly segregated groups in which subgroups 2, 3 and 4 substantially recapitulated the stromal, the luminal enriched and the basal-enriched proteomic subgroups respectively, while subgroup 1 is novel.

The authors focused on PIK3CA and TP53-mutated tumors since mutations in these genes are the most frequently encountered in breast cancer.

In order to verify the expected activation of the PIK3CA signaling cascade in tumors showing PIK3CA missense mutations, phosphosites markers were selected leading to the identification of activated PIK3CA mutation signature in 58% of PIK3CA mutated tumors. The PIK3CA mutant phosphoproteome signature was activated in all tumors harboring helical domain mutations. Analogously, the TP53 mutant phosphosignature was enhanced in tumors in which mutations occurred almost exclusively in the DNA-binding region.

Since many protein kinases are drug targets, the authors focused on this group whose expression is at least 1.5 interquartile ranges higher than the median. The ERBB2 locus showed the strongest effect of increased phosphoprotein levels associated with gene-amplification-driven RNA and protein overexpression. Other amplicon-associated highly phosphorylated kinases were identified, including CDK12, PAK1, PTK2, RIPK2 and TLK2.

The manuscript reports the results of a huge amount of experimental work (the authors themselves declare that 10 months of instrument time was required to analyze just over 100 patient samples) and a complex, well-structured statistical and computational analysis of proteogenomic data.

Through the results, the reader becomes aware of many of the strengths of mass spectrometry-based proteomics for cancer discovery, but also of some of the limitations inherent in proteolytic peptide sequencing.

First of all, the method requires almost one milligram of protein extract per sample, mainly devoted to phosphoproteomics despite the enrichment procedure, which revealed its relative inefficiency.

With regard to somatic mutation detection, the most striking result concerns the discrepancy between the number of variants detected at the DNA and RNA levels and those identified by the analysis of the proteome. Just a low percentage of all the mutations could be detected on proteins evidencing possible analytical problems, which contribute to the low detection rate of variants at the protein level, such as the optimal length of tryptic peptides, the difficulties in the detection of small peptides as well as of the very low abundance proteins by MS/MS. Other more general technical problems can be cited affecting also the more abundant proteins, like digestion efficiency and the size and hydrophobicity of the peptides.

Excluding the presence of a protein in a given sample is, by this methodological approach, a difficult objective to reach. Absence calls can derive from the analysis of protein whose concentration is below the limit of detection or, on the contrary, as in the case of isobaric mass tag labeling, it can be always possible to get signals for every target peptide and, to distinguish true signals from the biological noise, normalization by appropriate cut-offs has to be assessed.

The high amount of sample required for the analysis and the low throughput achieved by the proposed technology limit the evaluation of patient's samples on extended

case studies. Moreover, it is now unfeasible to evaluate proteomic characteristics in different fractions of a tumor to evidence intra-tumor heterogeneity due to the high protein input need. For this reason, a technological development is required to increase the instruments throughput as well as the introduction of new reagents with greater efficiency to allow multiplexing and improve sensitivity.

Anyway, proteogenomics represents the link between cell genotype and the cellular function and phenotype and it constitutes a new important tool to the refinement of the oncological research now mainly based on genomic and transcriptomic evaluations (16). In perspective, the new contribute of the proteogenomic analysis could improve the diagnostic, treatment and prevention of cancer. In view of the translational potential of this new "omic" and looking to the contribute it can bring to the clinic, the major advantage we can expect is related to the identification of a personalized analysis of tumors by the creation of a gene/protein sequence database specific for each single patient. The proteomic approach allows assessing whether DNA or RNA modifications are translated to the protein level identifying new therapeutic targets and new diagnostic or prognostic biomarkers. Also in the analyzed paper, in fact, the main object of the study is the identification of druggable kinases beyond HER2 to enable hypotheses for new inhibitors development.

This approach might reach a higher degree of personalization in cancer treatment and therapy and could be extended to other diseases and other biological fields. Further development is needed on the bioinformatics to reach a complete integration among genomic, transcriptomic and proteomic data and on the technical side to overcome the methodological challenges by improving the main analytical features of accuracy, sensitivity and reproducibility to avoid crucial issues causing misleading errors in the evaluation of the cell functional characteristics and the deriving clinical investigations. At present proteogenomics probably represents the first step towards an integrated view of the (tumor) cell biology that will open new opportunities in cellular and molecular life sciences.

The proteogenomic studies proposed up to now are focused on tissue samples in order to delineate the relationships among the three main molecular targets (DNA, RNA and proteins) and to evidence the most reliable ones to be investigated in view of a clinical application. Nonetheless the heterogeneous cell composition of the tissue samples and the consequent averaging procedure in the evaluation of the results deriving from bulk analysis can be the source of incoherent results (as already highlighted for other technologies like qPCR) being impossible to evaluate the contribution of different cell types and cell compartments. Proteomic analysis on samples with homogenous cellular composition or even single-cell proteogenomics might help in elucidating this issue when suitable technical tools will be developed to allow the necessary sensitivity at the proteomic level. Proof-of-principles have been published already on this topic (17), but mass spectrometry-based single cell proteomics has not been achieved yet despite tremendous interest in the field (18). Microfluidics coupled to mass spectrometry seems to represent the best solution to reach this goal.

As far as the clinical oncologic research is concerned, the proteogenomic analysis of the liquid biopsy of a tumor would efficiently solve most of the problems related to the therapy identification and monitoring required for precision medicine with a minimally invasive procedure for the patients.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned and reviewed by the Section Editor Zi-Guo Yang, MD (Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Breast Center, Peking University Cancer Hospital & Institute, Beijing, China).

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tcr.2016.10.12). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 2012;13:227-32.
2. Faulkner S, Dun MD, Hondermarck H. Proteogenomics: emergence and promise. Cell Mol Life Sci 2015;72:953-7.
3. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11:1114-25.
4. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 2004;4:59-77.
5. Lasonder E, Ishihama Y, Andersen JS, et al. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature 2002;419:537-42.
6. Merrihew GE, Davis C, Ewing B, et al. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res 2008;18:1660-9.
7. Brunner E, Ahrens CH, Mohanty S, et al. A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol 2007;25:576-83.
8. Baerenfaller K, Grossmann J, Grobei MA, et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science 2008;320:938-41.
9. Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. Nature 2014;509:575-81.
10. Wilhelm M, Schlegl J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. Nature 2014;509:582-7.
11. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature 2014;513:382-7.
12. Whiteaker JR, Halusa GN, Hoofnagle AN, et al. Using the CPTAC Assay Portal to Identify and Implement Highly Characterized Targeted Proteomics Assays. Methods Mol Biol 2016;1410:223-36.
13. Zhang H, Liu T, Zhang Z, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell 2016;166:755-65.
14. Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 2016;534:55-62.
15. Multiplex Protein Quantitation using iTRAQ™ Reagents – 8plex. Available online: http://tools.thermofisher.com/content/sfs/manuals/cms_049786.pdf
16. Dimitrakopoulos L, Prassas I, Diamandis EP, et al. Proteogenomics: Opportunities and Caveats. Clin Chem 2016;62:551-7.
17. Rubakhin SS, Sweedler JV. Characterizing peptides in individual mammalian cells using mass spectrometry. Nat Protoc 2007;2:1987-97.
18. Wang D, Bodovitz S. Single cell analysis: the new frontier in 'omics'. Trends Biotechnol 2010;28:281-90.

Cite this article as: Pinzani P, Salvianti F. The potential of proteogenomics in oncology. Transl Cancer Res 2016;5(Suppl 4):S708-S712. doi: 10.21037/tcr.2016.10.12