



Beyond genomics: biologic insights from the CPTAC proteogenomic analysis of breast cancer

Sandeep Rajput¹, Cynthia X. Ma^{1,2}

¹Section of Medical Oncology, Division of Oncology, Department of Internal Medicine, ²Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO 63110, USA

Correspondence to: Cynthia X. Ma, MD, PhD. Section of Medical Oncology, Division of Oncology, Department of Medicine, Siteman Cancer Center, Washington University School of Medicine, 4515 McKinley Building, Box 8076, St. Louis, MO 63110, USA. Email: cynthiamax@wustl.edu.

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 Sep 14, 2016. Accepted for publication Sep 22, 2016.

doi: 10.21037/tcr.2016.10.95

View this article at: <http://dx.doi.org/10.21037/tcr.2016.10.95>

Global genomic investigations of breast cancers have generated an extensive catalogue of somatic mutations as potential therapeutic targets. However, progress has been slow in differentiating driver mutations from passenger mutations, which hindered the development of therapeutic hypothesis. Proteomic analysis allows an opportunity for functional interpretation of somatic mutations (1). The initial proteomic analysis performed in the Cancer Genome Atlas (TCGA) breast cancer study quantified the expression levels of 171 cancer-related proteins and phosphoproteins by using the antibody based reverse phase protein array (RPPA) platform on 403 tumors, which identified seven proteomic subtypes, including Basal, HER2, Lum A, Lum A/B, that were highly concordant with the mRNA subtypes (2), and Reac I (mostly composed of a subset of mRNA Lum A tumors), and Reac II subtypes (a mixture of mRNA subtypes) that were enriched by proteins such as fibronectin, caveolin 1 and collagen VI, likely produced by the microenvironment and/or cancer-activated fibroblasts, as well as a seventh proteomic subtype “X”, which contained too few cases to analyzed (3). Coordinated genomic and RPPA data in this initial study identified aberrations of a number of key signaling pathways at the protein/phosphoprotein level in association with breast cancer mRNA subtypes (3). However, RPPA interrogation of the cancer proteome was limited by the number of antibodies included in the platform. To provide greater analytical breadth, the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) employed the state of the art mass spectrometry to analyze the global proteome and phosphoproteome for 77 genomically annotated TCGA

breast cancer samples representative of four principal mRNA defined breast cancer intrinsic subtypes (4). This is the first global proteogenomic study in breast cancer and generated important biologic insights that potentially connect somatic mutations to signaling aberrations in breast cancer.

High-resolution mass tandem mass spectrometry (MS/MS) technology has the advantage of quantifying multiple peptides from each protein, compared to the single epitope based antibody detection methodologies such as RPPA, therefore allowing a reliable assessment of protein expression levels (4-6). In addition, MS peptide sequencing has the capacity to detect single amino acid variants, frameshifts, and splice variants, although the coverage with current MS technology was sparse (4,6). Using the MS/MS, the CPTAC breast cancer study identified a total of 15,369 proteins (12,405 genes) and 62,679 phosphosites, with an average of 11,632 proteins and 26,310 phosphosites per tumor, as well as 3,709/90,806 (4.1%) protein coding nonsynonymous single nuclear variants and 672/238,646 (0.28%) RNA splice junction variants (4).

To illustrate the power and utility of MS/MS global proteomic analysis, protein expression levels of the three frequently mutated genes (*TP53*, *PIK3CA*, and *GATA3*), and the three clinical biomarkers [estrogen receptor (ER), estrogen receptor 1 (ESR1), progesterone receptor (PGR), and *ERBB2*] were correlated to mutations, gene copy numbers, RNA-seq and RPPA data (4). Missense mutations in *TP53* were associated with increased protein levels by both MS/MS and RPPA. However, nonsense and frameshift mutations in *TP53* were correlated with

lower p53 protein levels particularly pronounced by MS/MS method. In contrast, *GATA3* frameshift alterations were found to be expressed at both RNA and protein levels, indicating the presence of truncated *GATA3* protein rather than loss of protein expression. As expected, a good Pearson correlation was observed between RNA-seq and MS/MS protein expression levels for *ESR1* ($r=0.74$), *PGR* ($r=0.74$), *ERBB2* ($r=0.84$) and *GATA3* ($r=0.83$). However only modest correlation was observed for *TP53* ($r=0.36$), with lower levels of p53 protein by MS/MS compared to mRNA expression by RNA-seq, especially in luminal tumors. A search of E3 ligases that negatively correlated to p53 protein was therefore performed, which identified UBE3A, a known p53 E3 ligase, as the candidate regulator for post-transcriptional regulation of p53. These studies illustrated the ability of global proteome correlation analysis in confirming suspected regulatory mechanisms and the potential to identify novel regulatory pathways for subsequent investigations.

The global MS/MS proteomic data also allowed identification of proteins or phosphoproteins abundance in “cis” or “trans” relationship with copy number aberrations (CNAs) (4). About 7,776 CNAs were correlated with mRNA and protein levels while 4,472 CNAs were correlated with phosphoprotein levels. Proteins and phosphoproteins that correlated in cis to CNAs represented a subset of mRNA-CNA pairs correlated in cis. In addition, cancer relevant oncogenes and tumor suppressor genes were more likely to be cis-regulated on both protein and mRNA levels. Trans-effects were found for 68% of CNAs on the mRNA level, 13% on the protein level and 8% on the phosphoprotein level. These correlations were then compared to the functional knockdown data in the Library of Integrated Network-based Cellular Signatures (LINCS) database (<http://www.lincsproject.org/>) to identify candidate driver genes. Ten CNA genes were identified that affected by both CNA gains and losses, among which *ERBB2* was functionally connected only to CNA gain trans-effects. Interestingly, E3 ligase *SKP1* and the ribonucleoprotein export factor *CETN3* on chromosome arm 5q, which is frequent deleted in basal type breast cancer, were found to be potential regulators for the expression of *EGFR* and *SRC* kinase (4).

Clustering and network analyses of the proteomic data revealed striking similarity in the subtypes defined by RNA or by protein when restricted to a set of 35 PAM50 genes, indicating that subtype-defining proteomic features exist.

Unsupervised clustering of global proteome data identified basal-enriched, luminal enriched, and stromal-enriched clusters. The basal-enriched and luminal enriched groups showed a strong overlap with the mRNA-based PAM50 basal-like and luminal subgroups, whereas stromal-enriched proteome represented a mixture of all PAM50 mRNA-based subtypes, similar to the reactive type I subtype defined by RPPA. Subtype specific pathways were identified in tumors of the luminal- and basal- enriched proteome subgroups, with estradiol and *ESR1*-driven protein expression in the luminal-enriched subgroup and *MYC* target gene enrichment involved in cell cycle, checkpoint, and DNA repair pathways such as *AURKA/B*, *ATM*, *ATR*, *CHEK1/2*, and *BRCA1/2*; and for immune response/inflammation, including T-cell, B-cell, and neutrophil signatures. By phosphoproteome, the tumors were clustered into four robustly segregated subgroups, including subgroups 2, 3, and 4, which substantially recapitulated the stromal-, luminal- enriched, and basal/non-basal with *TP53* mutations, respectively, and subgroup 1, a novel subgroup defined by G protein, G-protein-coupled receptor, inositol phosphate metabolism signatures, and ionotropic glutamate signaling. These additional subgrouping by proteomic analysis are hypothesized to be functionally relevant.

The robust quantitative capability of MS also allowed the development of phosphoproteomic signatures of frequently mutated genes such as *PIK3CA* and *TP53* for mutation induced signaling aberrations (4). Phosphoproteomic analysis for upregulated phosphosites in *PIK3CA* mutated breast cancers identified 62 different phosphosites, including the kinases *RPS6KA5* and *EIF2AK4*, that were positively correlated with *PIK3CA* mutation, particularly in the helical domains. Similarly, a total of 56 phosphosites were up-regulated in *TP53* mutant tumors, especially the missense mutations in the DNA-binding regions rather than those with nonsense/frameshift mutations. These studies provided functional readout for these mutations, demonstrating the robustness of global phosphoproteome analysis in the interpretation of functional significances of genomic alterations.

In searching for potential amplified kinases as drug targets, proteogenomic analysis for outlier kinases identified the expected *ERBB2* in HER2-enriched subtype and other subtype specific kinase outliers that exhibited similar gene-amplification-driven proteogenomic patterns to *ERBB2*, including *CDK12* in HER2-enriched tumors, *PAK1* and *TLK2* in luminal breast cancer, and *PRKDC* and *SPEG* in

basal like breast cancer, among others (4).

In conclusion, the CPTAC breast cancer study provided the proof of concept that high quality robust quantitative global proteomic and proteomic and phosphoproteomic analyses could be generated by MS/MS technology to connect somatic mutations to signaling pathways, to narrow down candidate driver genomic events, and to generate therapeutic hypothesis. The preservation of luminal and basal enriched subtypes by proteomic and phosphoproteomic clustering indicated that the mRNA based intrinsic subtypes are captured at the protein level. The additional subtypes identified by proteomic analysis, including the stromal enriched and a novel subtype that was only obvious by phosphoprotein clustering, demonstrated further biological insight gained beyond DNA and RNA analysis. Although the requirement in the quality and quantity of tissues for proteomic analysis present a challenge to incorporate global proteomic approaches in routine clinical care, application of selected biomarkers is possible. Indeed, the well-established clinical markers, including ER, PR and HER2 (*ERBB2*), are protein markers. We recommend incorporation of the proteogenomic investigation in preclinical and clinical trial development of targeted agents so that biomarkers predictive of therapeutic efficacy could be developed. We are optimistic that advances in science and technology will lead to personalized medicine and improved care of cancer patients.

Acknowledgments

Funding: This work was supported by the Susan G. Komen Investigator-Initiated Research Grant (IIR13263475).

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).

Cite this article as: Rajput S, Ma CX. Beyond genomics: biologic insights from the CPTAC proteogenomic analysis of breast cancer. *Transl Cancer Res* 2016;5(Suppl 4):S713-S715. doi: 10.21037/tcr.2016.10.95

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at <http://dx.doi.org/10.21037/tcr.2016.10.95>). 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.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Ellis MJ, Gillette M, Carr SA, et al. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* 2013;3:1108-12.
2. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000;406:747-52.
3. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61-70.
4. Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 2016;534:55-62.
5. Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* 2012;9:555-66.
6. Gillette MA, Carr SA. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nat Methods* 2013;10:28-34.