Clinical Chemistry 61:9
1126–1128 (2015)

Perspective

Proteomics of Colorectal Cancer in a Genomic Context: First Large-scale Mass Spectrometry–Based Analysis from the Cancer Genome Atlas

Connie R. Jimenez[1*] and Remond J.A. Fijneman[2]

Cancer is a disease caused by DNA alterations that make cells grow in an uncontrolled way. Identifying the alterations in each tumor's complete set of DNA (the genome) and its functionally relevant protein complement (the proteome) is expected to increase our understanding of how such changes interact to drive the disease. This global molecular information will help lay a foundation for improving cancer prevention, early detection, and treatment.

In the past decade, mass spectrometry (MS)[3] technology and (bio)informatics tools have matured to the extent that they can provide high-throughput, comprehensive protein inventories of cells, tissues, and biofluids, even when present at low concentrations *(1)*, paving the way for large-scale profiling of clinical samples. The integration of proteomic and genomic data in which MS/MS data are searched against customized databases of individually matched DNA/RNA sequence data, referred to as proteogenomics, provides a more comprehensive view of the molecular determinants that drive cancer than genomic analysis alone and may help identify the most important targets for cancer detection and intervention.

Six years ago, the first report on the genomic characterization of glioblastoma was published by the Cancer Genome Atlas (TCGA) network. A research team of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) has now assembled an MS-based inventory of protein expression and variants in 95 colon and rectal tumors *(2)* that were previously characterized by the TCGA project. The authors focused on the question of how gene alterations identified in previous analyses of the same samples are expressed at the protein level.

To enable a comprehensive and in-depth analysis of colorectal tumors, the CPTAC team used a label-free, next-generation proteomics approach based on basic and reversed-phase nano-liquid chromatography coupled to high-resolution tandem MS *(2)*. The tandem MS data were analyzed by an innovative bioinformatics pipeline that was developed in-house for stringent identification and quantification of proteins and tumor-specific variants. Spectral counts (the total number of MS/MS spectra acquired for peptides from a given protein) were used for label-free protein quantification. These analyses took almost a year of instrument time. Platform reproducibility was demonstrated using reference quality control samples that were run throughout the analyses. Altogether, the data yielded a proteome portrait of colorectal cancer (CRC) that included >6 299 756 MS/MS spectra, yielding 124 823 identified distinct peptides mapping to 7526 protein groups, with a protein-level false discovery rate of 2.64%. A gene-level assembly of the peptides identified 7211 genes, of which 3899 genes with a protein-level false discovery rate of 0.43% were used to compare relative protein abundance across tumor samples.

Comments on the Main Findings of the Report

The integrated proteogenomic analyses of this study revealed that somatic variants displayed a reduced protein abundance compared to germline variants. More specifically, 796 single amino acid variants were identified across 86 tumors with matched RNA-seq data, among which 108 corresponded to somatic variants reported by TCGA or the Catalogue of Somatic Mutations in Cancer (COSMIC) database. The authors hypothesize that the low identification rate for somatic variants may reflect a reduced translational efficiency or protein stability. Somatic variants mapped to 105 genes, including known

[1] OncoProteomics Laboratory, Department of Medical Oncology, and [2] Tumor Profiling Unit, Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands.
* Address correspondence to this author at: VU University Medical Center, De Boelelaan 1117, Amsterdam, Netherlands 1081HV. E-mail c.jimenez@vumc.nl.
Received December 22, 2014; accepted January 26, 2015.
Previously published online at DOI: 10.1373/clinchem.2014.234807

[3] Nonstandard abbreviations: MS, mass spectrometry; TCGA, the Cancer Genome Atlas; CPTAC, Clinical Proteomic Tumor Analysis Consortium; CRC, colorectal cancer; COSMIC, Catalogue of Somatic Mutations in Cancer; ALDH2, aldehyde dehydrogenase 2; HSD17B4, hydroxysteroid (17-beta) dehydrogenase 4; PARP1, poly(ADP-ribose) polymerase 1; P4HB, prolyl 4-hydroxylase, beta polypeptide; TST, thiosulfate sulfurtransferase (rhodanese); GAK, cyclin G–associated kinase; SLC25A24, solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 24; SUPT16G, suppressor of Ty 16 homolog (*S. cerevisiae*); HNF4A, hepatocyte nuclear factor 4, α; TOMM34, translocase of outer mitochondrial membrane 34; SRC, Src protooncogene, nonreceptor tyrosine kinase.

cancer genes such as *KRAS* (Kirsten rat sarcoma viral oncogene homolog),[4] *CTNNB1* [catenin (cadherin-associated protein), beta 1, 88 kDa], *SF3B1* (splicing factor 3b, subunit 1, 155 kDa), *ALDH2* [aldehyde dehydrogenase 2 family (mitochondrial)], and *FH* (fumarate hydratase) and 14 targets of FDA-approved drugs or drugs in clinical trials, such as aldehyde dehydrogenase 2 (ALDH2), hydroxysteroid (17-beta) dehydrogenase 4 (HSD17B4), poly(ADP-ribose) polymerase 1 (PARP1), prolyl 4-hydroxylase, beta polypeptide (P4HB), thiosulfate sulfurtransferase (rhodanese) (TST), cyclin G–associated kinase (GAK), solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 24 (SLC25A24), and suppressor of Ty 16 homolog (*S. cerevisiae*) (SUPT16H).

Furthermore, the authors found that DNA copy number alterations showed strong *cis*- and *trans*-effects on mRNA abundance, but that relatively few of these extended to the protein level. However, proteomics identified a few amplifications that had dramatic effects on protein concentrations and that may prioritize targets for diagnosis or therapeutic interventions. Examples on the chromosome 20q amplicon include hepatocyte nuclear factor 4, α (HNF4A), translocase of outer mitochondrial membrane 34 (TOMM34), and Src protooncogene, nonreceptor tyrosine kinase (SRC).

The relationship between mRNA transcript levels and protein abundance has long been debated. The CRC proteogenomics effort allowed, for the first time, a global analysis of mRNA transcript–protein correlations in a large cohort of tumors. Although 89% of mRNA-protein pairs across all samples showed a positive correlation, this correlation was significant for only 32% of the pairs. These data indicate that mRNA transcript abundance cannot be used to reliably predict protein abundance. In itself, this discordance is not unexpected, as many regulatory controls lie between RNA and protein expression. At the same time, the current methods used to correlate mRNA to protein concentrations may need to be revisited. One of the outcomes of the recently published first MS-based draft of the human proteome is that the protein/mRNA ratio is remarkably conserved between tissues for any given protein *(3)*. As such, the rate of translation appears to be a fundamental characteristic of each individual gene transcript, with major influence on the actual protein abundance. Therefore, the accuracy of predicting protein abundance based on mRNA levels may improve by taking transcript-specific translation rates into account.

One of the key challenges in the field of oncology is to classify genetically diverse tumors with similar phenotypic appearance into distinct cancer subtypes with clinically similar behavior and outcome. Genome-wide characterization of tumors, which has been technically feasible for RNA profiling for more than a decade, has revealed multiple RNA signatures for classification of CRC into subtypes with a relatively good or bad prognosis. Considering the limited correlation between mRNA and protein concentrations, one interesting analysis that could now be performed is the classification of tumors based on proteomics data. Unsupervised cluster analysis uncovered 5 CRC proteomic subtypes. Compared to the DNA-based (epi)genomic TCGA classification of these tumors, almost all hypermutated and microsatellite instable tumors were included in 2 of the proteomic subtypes, subtypes B and C. Because subtype B lacked *TP53* (tumor protein 53) mutations and exhibited loss of chromosome 18q, 2 DNA alterations known to be highly prevalent in CRC, the proteomic data suggest that these subtypes developed through different biological mechanisms. Interestingly, whereas microsatellite instable tumors often have a relatively good prognosis, the proteomic subtype C appeared to be enriched in molecular features associated with poor clinical outcome based on comparison with 2 RNA-based classification systems and Gene Ontology enrichment analysis. Unfortunately, the TCGA CRC series lacks sufficient clinical follow-up information and statistical power to directly determine clinical outcome for each of the proteomic CRC subtypes.

CRC Proteome, a Starting Point for Biomarker Development: Implications for Personalized Medicine

The comprehensive data that have been generated by TCGA's network and CPTAC members are freely available through a data portal for broader use by the cancer research community. Other large-scale proteomics efforts of note are the Human Protein Atlas (http://www.proteinatlas.org/) *(4)*, which uses antibody staining to catalog the location of proteins within cells and tissues, and the Human Proteome Project, a consortium effort that mainly uses MS to characterize the role of each protein in human biology and disease. The resources generated by 2 recent studies on the draft human proteome reported in *Nature* this year—Human Proteome Map (http://www.humanproteomemap.org) and ProteomicsDB *(3, 5)* (https://www.proteomicsdb.org)—are also available to researchers across the biomedical sciences.

Protein biomarkers are well suited for the development of clinical applications, in particular antibody-based assays for diagnosis, prognosis, therapy prediction, and disease monitoring. Public access to the TCGA

[4] Human genes: *KRAS,* Kirsten rat sarcoma viral oncogene homolog; *CTNNB1,* catenin (cadherin-associated protein), beta 1, 88 kDa; *SF3B1,* splicing factor 3b, subunit 1, 155 kDa; *ALDH2,* aldehyde dehydrogenase 2 family (mitochondrial); *FH,* fumarate hydratase; *TP53,* tumor protein 53.

CRC proteome and exploration using an interactive gene network browser (http://www.netgestalt.org/crc/main.html) now allows researchers to examine expression of their favorite protein biomarker candidates across a series of tumors, in relation to somatic DNA alterations and mRNA expression levels. Visualization of protein staining of normal and cancer tissues can easily be obtained in silico by visiting the Human Protein Atlas *(4)*, which will indicate whether these proteins are expressed by neoplastic epithelial cells or nonneoplastic cells in the (tumor) stroma or have been accumulated in the extracellular matrix. As such, public resources of large-scale genome-wide molecular profiling initiatives provide an invaluable source of information to accelerate biomarker discovery and validation.

The CPTAC team has cataloged the CRC proteome from 95 genomically characterized human tumor samples using shotgun MS. They integrated their proteomics results with genomics data generated for the same tumors. This integrated approach offers new insights into colorectal tumor biology and important information on potential drivers of CRC. This large-scale, in-depth CRC proteome provides a resource for design of targeted MS-based assays to enable high-throughput multiplexed biomarker validation. In addition, the most discriminatory proteins may be followed up with routine antibody-based applications. This study lays the foundation for future validation studies and experimentation to propel advances in personalized medicine.

The use of MS-based proteomics to generate the proteomic landscape of a large set of colorectal tumors, including tumor-specific variants, is an impressive accomplishment. Integrated proteogenomic analysis, as reported in this study, can provide functional context to interpret genomic abnormalities in terms of cancer biology. Nonetheless, the clinical implications need to be explored in larger series of clinically well-annotated samples. We anticipate that this data set will fuel insight into CRC heterogeneity and subtypes and provide drug targets and biomarkers for future clinical applications in CRC.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

References

1. Pham TV, Piersma SR, Oudgenoeg G, Jimenez CR. Label-free mass spectrometry-based proteomics for biomarker discovery and validation. Expert Rev Mol Diagn 2012;12:343–59.
2. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. NCI CPTAC. Proteogenomic characterization of human colon and rectal cancer. Nature 2014;513:382–7.
3. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, et al. Mass-spectrometry-based draft of the human proteome. Nature 2014;509:582–7.
4. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 2010;28:1248–50.
5. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature 2014;509:575–81.