LIKE ANY GOOD EXPERIMENT, CANCER GENOMIC ANALYSIS GENERATES MORE QUESTIONS THAN ANSWERS

The generation of comprehensive somatic mutation profiles of human tumors through massively parallel sequencing programs, such as The Cancer Genome Atlas (TCGA), has created an unprecedented opportunity for new biologic insights and an expectation that some of these discoveries will be translated into improved clinical outcomes. This promise has already been realized in many excellent examples of precise therapies available for gain-of-function events such as amplifications (HER2-positive breast cancer), point mutations (BRAFV600E-positive melanoma), or translocations (BCR-ABL in chronic myeloid leukemia).

However, the deluge of new data generated by DNA and RNA sequencing has often been more confusing than enlightening. The list of “significantly mutated genes” (SMG) that are recurrent at the level of point mutations or small insertions/deletions is startlingly short in many common epithelial cancers, with the genomic landscape dominated by loss-of-function mutations in tumor suppressor genes rather than the gain-of-function mutations in oncogenic protein kinases that have been so central to recent therapeutic successes. High-grade serous ovarian cancer is an excellent example of this frustrating pattern, as besides TP53 mutation, which was present in almost all cases, the ovarian SMG list was limited to only nine other genes, none of which was an oncogenic kinase and all of which had low recurrence frequencies.

Basal-like breast cancer is very similar to serous ovarian cancer in this regard, but luminal-type breast cancer has a longer list of higher recurrence rate SMGs including an oncogenic lipid kinase (PIK3CA), motivating active exploration of a number of new therapeutic directions (1). More commonly, however, the functions of somatic mutations identified through genomic analysis of common solid malignancies are very poorly understood, particularly in a disease and tissue-specific context, and therefore do not drive therapeutic initiatives. Similarly, though chromosomal amplification and deletion events are thought to be critical driver events, the contributions of each of the often multiple genes within amplicons or deletions are often unclear. Complicating matters further is the recognition that in many instances several genes in one region may function in concert.

The vast majority of mutations are actually not in protein coding space but in regulatory sequences or other noncoding regions that could, nonetheless, be functionally relevant. Furthermore, recurrent clustered mutations in long noncoding RNAs such as MALAT1 are also emerging, but again the associated biochemistry is mysterious. One must therefore inescapably conclude that we still do not understand the biochemical and cell biologic effects of most genomic alterations in cancer revealed by sequencing studies. Nor do we understand the combinatorial effects of many mutations that are present in any single cancer, as these are very hard to model. Our limited understanding of this complex molecular interplay renders even our therapeutic successes
informative, because when proteins harboring significant driver mutations are successfully targeted, therapeutic resistance often develops because of the development of alternative pathways for cell survival and growth. Unbiased discovery of these resistance mechanisms is clearly a priority if we are to translate pathway targeted approaches into curative regimens.

INFORMATIC AND PATHWAY ANALYSES BASED ON DNA AND RNA SEQUENCING DATA GENERATE HYPOTHESES, NOT CONCLUSIONS

Currently the cancer informatics field makes conclusions regarding cancer biology and signaling events by developing databases from a priori understandings of signal transduction or other biochemical studies. The genomic data typically are parsed into networks or “interactomes” and lists of biochemical activities that connect or map into pathways containing over- or underexpressed genes, amplified or deleted genes, or mutant genes. These relationships are then interpreted as the identification of specific biologic processes that drive clinical phenotypes. Though clearly useful, the functional data used to annotate cancer informatics programs nevertheless may not be correct in the biologic context of tumors. As the actual biochemistry is inferred, but not directly observed, the outputs of these programs must be considered hypotheses, not conclusions (2). Thus, biochemical analyses that link genotype to phenotype are critical to the translational success of cancer genomics.

PROTEOMIC TECHNOLOGIES CAN BRIDGE THE GAP BETWEEN GENOTYPE AND PHENOTYPE

The central dogma of biology places proteins and their functions as the direct mediators of phenotypic characteristics. Proteomic analyses therefore offer the means to measure the biochemical impact of cancer-related genomic abnormalities, including expression of variant proteins encoded by mutations, protein changes driven by altered DNA copy number, chromosomal amplification and deletion events, epigenetic silencing, and changes in microRNA expression. Analysis of protein posttranslational modifications, particularly phosphorylation, enables the detection of signaling network adaptations driven by genomic changes.

Proteomic technologies, which are based primarily on mass spectrometry (MS), have fundamentally advanced our knowledge of biochemistry and cell biology, including protein dynamics, multiprotein complexes, and signaling networks (3), and have enabled systematic development of targeted assays for precise quantification of proteins in tissues and biofluids (4). MS analytic platforms are capable of providing both global profiles of protein expression and posttranslational modification, as well as precise, targeted quantification of proteins and their modified and variant forms.

A complement to MS-based proteomics is the reverse-phase protein array (RPPA) platform, which uses antibodies to probe printed arrays of tissue lysates and has been used recently in TCGA tumor analyses (5). The key advantage of RPPA is a small sample requirement and an ability to efficiently probe cancer-relevant signaling pathways. However, RPPA requires highly specific antibodies to reliably detect targeted analytes in unfraccionated lysates. The number of phosphosite-specific antibodies remains very limited, and few antibodies can distinguish between members of closely related protein families that may, nonetheless, have very different biologic effects, such as AKT1, 2, and 3. Despite these constraints, RPPA analyses have shown utility and can be used to inform more in-depth pathway characterization by targeted MS approaches.

CPTAC AND CLINICAL PROTEOMIC TECHNOLOGIES

The NCI launched the Clinical Proteomic Technology Assessment for Cancer in 2006, which was designed to evaluate the performance of proteomic technology platforms for both global profiling and targeted quantitative analysis in tissues and biofluids. Major contributions of this first program were:

1. Demonstration of the intralaboratory reproducibility of unbiased, data-dependent proteomic platforms for biologic discovery, together with generation of reference materials and performance metrics for system assessment (6).

2. Confirmation of the ability of targeted protein quantification by multiple reaction monitoring (MRM) to achieve reproducible, precise quantification of protein levels in tissues and biofluids (7) across multiple laboratories throughout the CPTAC network. This group also developed Skyline, a widely used and community-supported open-source software platform for MRM assay design and data analysis on all commercial instrument platforms (8).

The renewal of the Clinical Proteomic Technology Assessment for Cancer program as the Clinical Proteomic Tumor Analysis Consortium (CPTAC) shifted the focus from technology assessment to integrated cancer genomics and proteomics. The CPTAC centers are applying standardized proteome analysis platforms to analyze tumor tissues from TCGA, as well as unique cell and xenograft models and other tissue collections, all of which are accompanied by rich genomic datasets. The current CPTAC program is thus a broad, proof-of-concept initiative to bring high-throughput proteomic technologies into the cancer genomics enterprise.

Five multi-institution Proteome Characterization Centers (PCC) comprise CPTAC: (i) Broad Institute of MIT and Harvard/Fred Hutchinson Cancer Research Center/Massachusetts General Hospital; (ii) Johns Hopkins University/Memorial-Sloan Kettering/Stanford University/University of Chicago; (iii) Pacific Northwest National Laboratory/Massachusetts Institute of Technology/Oregon Health and Science University/The University of Texas MD Anderson Cancer Center; (iv) Vanderbilt University/Massachusetts Institute of Technology; (v) Washington University/University of North Carolina. Each team combines required capabilities in MS and related proteomic technologies, bioinformatics and biostatistics, cancer biology, and clinical/translational cancer research.

A CPTAC-Data Coordinating Center (CPTAC-DCC) provides a central repository for mass spectrometry raw data
and metadata from biospecimens analyzed by the CPTAC PCCs. These data are available to the research community through the CPTAC Data Portal, which is accessed from the NCI Office of Cancer Clinical Proteomics Research website (http://proteomics.cancer.gov/). In addition to the original (“raw”) MS data files, users can access peptide and protein assemblies, posttranslational modification maps, and quantitative assay results processed through a standardized data analysis pipeline. The CPTAC-DCC incorporates standards (metadata, protein names, file formats) established by the Human Proteome Organization, the Human Genome Organization, and genomics resources to facilitate vertical integration of multiple data types to better define the molecular features of cancer.

PROTEOMIC TECHNOLOGY PLATFORMS IN CPTAC

The CPTAC PCCs use untargeted proteomics (also referred to as “discovery proteomics” or “shotgun proteomics”) for unbiased, global profiling of both protein expression and posttranslational modifications. Quantitative comparisons in data-dependent MS analyses are made by tagging peptides with isotope-encoded mass tags (e.g., iTRAQ reagents; ref. 9) or by a label-free, spectral counting approach (10). The iTRAQ-based approach is also applied for global phosphoproteomic profiling (9) and glycoprotein profiling, and can be extended to N-acetylated and ubiquitylated protein inventories. Shotgun proteomics analyses use a new generation of tandem hybrid Orbitrap and time-of-flight mass analyzers, which provide high-resolution MS and tandem MS (MS/MS) analysis of peptides at high scan rates (10–100 Hz), yielding more than 10 MS/MS spectra per second (11). These instruments, when combined with intelligent sample processing and fractionation, enable both deep inventory and quantitative comparisons of complex peptide mixtures from cell, tissue, and biofluid proteomes. For example, recent studies in the Broad and Pacific Northwest National laboratories have yielded more than 11,000 distinct proteins and more than 25,000 phosphorylated peptides from individual iTRAQ-labeled tumor samples. The number of distinct proteins identified and quantified in these experiments now rivals the number of transcripts observed in microarray experiments, while delivering valuable additional information on PTMs hidden to genomic analyses.

The complementary technology to global profiling is targeted analysis by MRM (also termed selected reaction monitoring or “SRM”; refs. 3, 4), which enables systematic development of quantitative protein assays through measurements of specific peptides in proteolytic digests. MRM technology overcomes a fundamental limitation of immunochemical methods—the availability of specific antibodies. Moreover, MRM enables selective quantification of variant/ mutant or posttranslationally modified sequences, both or which are difficult to achieve with antibodies.

MRM assays typically use triple quadrupole instruments, but similar assays are beginning to be deployed on new, hybrid quadrupole–time-of-flight and quadrupole-orbitrap instruments that have higher resolution and mass precision than triple quadrupole instruments (12). A key strength of MRM assays, and one that further distinguishes them from antibody-based approaches, is the ability to monitor multiple peptides in a single analysis, thus providing for multiplexed quantification of as many as a hundred proteins in a single assay. For example, phosphorytrosine capture combined with MRM enabled quantitative and dynamics assessment of growth factor signaling networks (13) and effects of oncogenic mutations on tyrosine kinase networks (14, 15). MRM assays can be applied to simultaneously quantify wild-type and variant protein forms (16), and the multiplexing capability supports simultaneous monitoring of multiple components in signaling pathways (17). Combining immunoaffinity enrichment of specific proteins and peptides with MRM enables systematic deployment of highly sensitive, specific assays (18). Similar sensitivities can be attained without immunoaffinity enrichment by combining another stage of high-performance liquid chromatography separation with depletion of major serum proteins (e.g., limit of detection of 50 pg/ml for prostate-specific antigen; ref. 19).

INTEGRATING GENOMIC AND PROTEOMIC DATA TO UNDERSTAND BREAST, COLON, AND OVARIAN CANCER

Proteomic technologies are analogous to genomic technologies in providing broad inventories of diverse sequences and modified or variant forms, together with quantitative data that indicate expression and dynamics. Proteomics thus can be placed immediately downstream of genomics in a unified analysis scheme that describes the translation of genomic characteristics to functions and phenotypes (Fig. 1). This creates important synergies in proteogenomic analysis. RNA sequencing data from individual tissue specimens can be used to generate customized databases for the identification of sequence variants at the protein level. Comparison of the identified sequence variants with database for single-nucleotide polymorphisms (dbSNP) germ-line sequencing and cancer mutation databases distinguishes polymorphisms from somatic mutations and identifies peptides that arise from the unique genetic background of each patient. With appropriate alignment tools, variant peptides can indicate expression of novel gene structures associated with cancer-specific events. Determination of which DNA or RNA sequence level variants are expressed as proteins provides a basis for prioritizing mutations for further study of their contributions to cancer phenotypes.

Integrated proteogenomics enables key insights into other genomic abnormalities. For example, quantitative comparisons at the copy number, RNASeq, and protein expression levels can identify the gene drivers in focally amplified chromosomal segments. An obvious benefit of proteomic analysis, particularly in profiling phosphoproteins, is the direct measurement of the status of cancer signaling pathways (13). However, global protein expression and posttranslational modification profiling can also reveal unanticipated network and pathway adaptations to mutations, amplification events, and deletion events. Combining global proteome expression profiles with miRNA and mRNA data indicates mechanisms by which tumor-specific alterations in miRNA expression control gene expression (20).

The CPTAC PCCs initially conducted a series of due diligence studies characterizing analysis platform performance,
optimizing sample preparation methods, establishing suitable controls, and scrutinizing the effects of ischemia on the stability of protein expression and phosphorylation. Having showed the suitability of platforms, methodologies, and samples, the next major objective of the CPTAC PCCs is the analysis of approximately one hundred genome-annotated samples each from breast, ovarian, and colon/rectal cancer by global shotgun proteomic, phosphoproteomic, and glycoproteomic profiling and targeted MRM analyses. The goals of these analyses are:

1. To generate rich proteomic datasets for these genome-annotated tumors with state-of-the-art analysis platforms, together with quality control datasets from analyses of a common performance standard.

2. To identify differential protein expression characteristics that distinguish tumors of known biologic subtypes (e.g., microsatellite instability in colon tumors or basal and luminal breast tumors) and potentially provide a new molecular taxonomy that can be compared with DNA- and RNA-based classification approaches.

3. To identify variant protein sequences corresponding to somatic mutations and to evaluate the relationship between mutation frequency and variant protein expression.

4. To determine how copy number variation translates into protein expression differences.

5. To evaluate the impact of genomic features on the status of signaling networks through direct analysis of phosphoprotein intermediates.

6. To derive preliminary associations with clinical characteristics, such as platinum resistance in ovarian cancer.

These analyses will comprise a strong proof-of-concept for the implementation of proteomic technology in tumor tissue analysis, but clearly are only a first step, as discoveries based on proteomics will require validation in independent cohorts.

Finally, additional studies in more precise clinical contexts will begin. For example, in collaboration with the National Surgical Adjuvant Breast and Bowel Project, jointly funded by the Breast Cancer Research foundation and CPTAC, we will accrue rapidly frozen samples from patients with HER2-positive breast cancer, both before and within 48 to 72 hours of receiving a dose of neoadjuvant treatment with paclitaxel and trastuzumab. The objective is to determine whether the acute proteomic response to chemotherapeutic stress is predictive of treatment outcome. Similar studies are under consideration for colon and ovarian cancers. Other studies in CPTAC will focus on cancers sampled longitudinally during therapy, using the unique lens of proteomics to detect molecular adaptations that may be silent at the genome level. A baseline sample anchors quantitative comparisons of acute drug responses against the longer-term effects of the treatment. These studies will leverage high-precision MRM assays, which may enable detection of important indicator proteins in the limiting protein amounts available from core needle biopsies.

**CONCLUSION**

The CPTAC program is deploying rapidly maturing proteomic technologies for analysis of fully genomically characterized tumors, to conduct large-scale comprehensive gene-to-protein integration in cancer for the first time. The combined use of genomic and proteomic data produces strong analytic synergy; more importantly, proteomic analyses can enable direct detection of the impact of genomic alterations. The emphasis on analysis of human tumors, rather than experimental model systems, is deliberate, minimizing the distance from data to clinical impact. An important lesson from the TCGA program is the need to plan in advance for validation studies by carefully accruing samples through precisely designed protocols suitable to address central unresolved questions in clinical oncology. Effective planning must engage the talents and resources of the wider clinical and...
scientific community. Thus, the fundamental purpose of this article is a call for collaboration between clinical investigators, basic scientists, and computational biologists. All of the data from CPTAC studies will be made available to the research community, so that the value of this investment made by the NCI in cancer proteomics will continue to grow.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors’ Contributions

Conception and design: M.J. Ellis, M. Gillette, A.G. Paulovich, K.D. Rodland, D.C. Liebler
Development of methodology: M.J. Ellis, R.D. Smith
Writing, review, and/or revision of the manuscript: M.J. Ellis, M. Gillette, S.A. Carr, A.G. Paulovich, K.D. Rodland, R.K. Townsend, C. Kinsinger, M. Mesri, H. Rodriguez, D.C. Liebler
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.J. Ellis, R.D. Smith, C. Kinsinger, M. Mesri
Figure design and execution: M. Gillette

Acknowledgments

CPTAC funding is provided to the following institutions: Broad Institute, Fred Hutchinson Cancer Research Center, Cambridge, MA (U24CA160034), Johns Hopkins University, Baltimore, MD (U24CA160036), Pacific Northwest National Laboratory, Richland, WA (U24CA160019), Vanderbilt University, Nashville, TN (U24CA159988), Washington University, St. Louis, MO, and University of North Carolina, Chapel Hill, NC (U24CA160035).

Grant Support

This work was supported by grants U24CA160034 (Broad Institute, Fred Hutchinson Cancer Research Center; to M. Gillette, S.A. Carr, and A.G. Paulovich), U24CA160036 (Johns Hopkins University; to D. Chan), U24CA160019 (Pacific Northwest National Laboratory; to R.D. Smith and K.K. Rodland), U24CA159988 (Vanderbilt University; to D.C. Liebler), and U24CA160035 (Washington University, St. Louis; University of North Carolina, Chapel Hill; to M.J. Ellis, X. Chen, and R.R. Townsend).

Published online October 11, 2013.

REFERENCES
