Journal of proteome research

Perspective

pubs.acs.org/jpr

The Biology/Disease-driven Human Proteome Project (B/D-HPP): Enabling Protein Research for the Life Sciences Community

Ruedi Aebersold,[†,‡,§] Gary D. Bader,[§,∥] Aled M. Edwards,[§,⊥,#] Jennifer E. van Eyk,[§,¶] Martin Kussmann,[§,×,□,●] Jun Qin,[§,△,■] and Gilbert S. Omenn[○,◆]

[†]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

[‡]Faculty of Science, University of Zurich, Zurich, Switzerland

[∥]The Donnelly Centre, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada

[⊥]Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada

[#]Division of Cancer Genomics and Proteomics, Ontario Cancer Institute, Toronto M5G 2M9, Canada

[¶]Johns Hopkins Bayview Proteomics Center, Department of Medicine, Division of Cardiology, School of Medicine, Johns Hopkins University, Baltimore, Maryland, United States

[×]Proteomics and Metabonomics Core, Nestlé Institute of Health Sciences, Lausanne, Switzerland

[□]Faculty of Life Sciences, Ecole Polytechnique Fédérale Lausanne (EPFL), Lausanne, Switzerland

[●]Faculty of Science, Aarhus University, Aarhus, Denmark

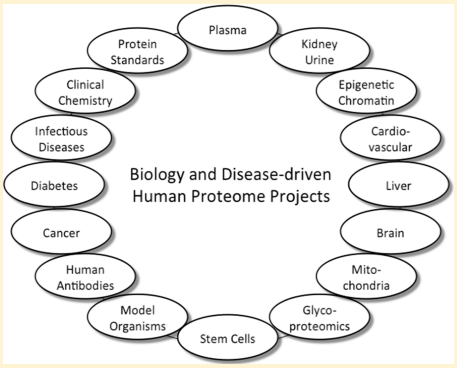
[△]State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, China

[■]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030, United States

[○]Departments of Computational Medicine and Bioinformatics, Internal Medicine, and Human Genetics, and School of Public Health, University of Michigan, Ann Arbor, Michigan 48109-2218, United States; chair, HUPO Human Proteome Project

[◆]Institute for Systems Biology, Seattle, Washington 98101, United States

ABSTRACT: The biology and disease oriented branch of the Human Proteome Project (B/D-HPP) was established by the Human Proteome Organization (HUPO) with the main goal of supporting the broad application of state-of the-art measurements of proteins and proteomes by life scientists studying the molecular mechanisms of biological processes and human disease. This will be accomplished through the generation of research and informational resources that will support the routine and definitive measurement of the process or disease relevant proteins. The B/D-HPP is highly complementary to the C-HPP and will provide datasets and biological characterization useful to the C-HPP teams. In this manuscript we describe the goals, the plans, and the current status of the of the B/D-HPP.



KEYWORDS: *Human proteome project, proteomics, mass spectrometry, affinity reagents, network biology, human disease, biological processes*

PROTEOMICS AND ITS IMPACT ON EXPERIMENTAL BIOLOGY

With the announcement of the first draft of the human genome sequence in 2000, life science research achieved an important milestone. With the nearly completed human genome sequence the genetic basis for the ensemble of molecules that constitute a living human cell and execute all its biochemical processes of life had become available, in principle. However, it is the how, when and where the genetic information is translated into the different classes of biomolecules and the question how these molecules interact that ultimately define a particular phenotype. These questions have remained the subject of intense research, particularly via genomic approaches. Recent progress extensively reported by the ENCODE II program has revealed regulatory roles of genome elements, various RNA molecules, and proteins.[1,2]

It is generally understood that proteins play an essential role in connecting genotype and phenotype. However, the specific mechanisms by which genomic variation is translated into specific (disease) phenotypes remain essentially unknown. For example, in cancer genomics studies, a key objective is to find recurrently mutated or aberrantly expressed genes and their

Special Issue: Chromosome-centric Human Proteome Project

Received: December 7, 2012
Published: December 21, 2012

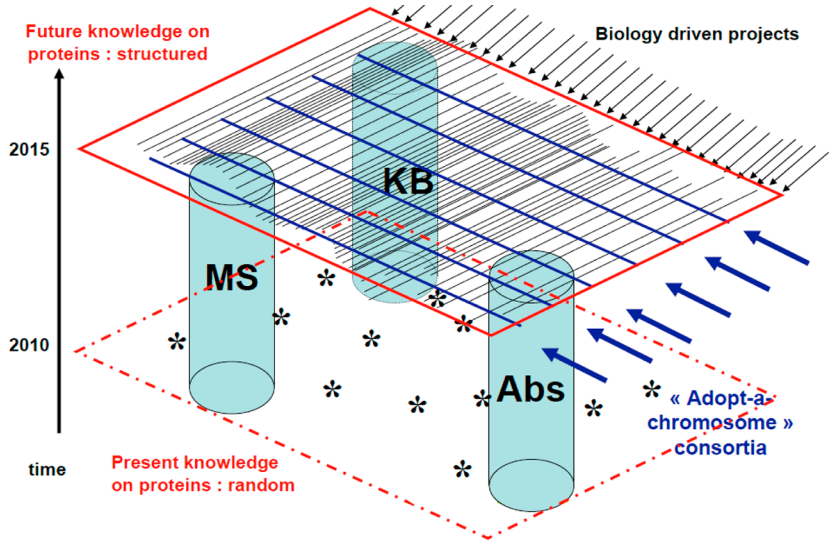


Figure 1. Schematic representation of the HUPO Human Proteome Project. The three pillars represent: knowledge base (KB), mass spectrometry (MS), and protein capture (antibodies, Abs). The biology-driven project are now captured in the B/D-HPP, and the adopt-a-chromosome consortia are the chromosome-specific C-HPP teams. (Figure reproduced from Legrain et al.[10])

corresponding proteins, which are thus implicated in malignancy and could be targeted by drugs or other agents. Although useful for relatively simple diseases that involve only a few genes, it is becoming increasingly clear that this approach largely fails for the many diseases that result from complex interactions between multiple genes and proteins. In these cases, many different and often rare combinations of mutations among multiple genes in a pathway may lead to the same phenotype, making it difficult to identify potential causes by observing recurrence within a single gene. Further, as increasing numbers genes and alterations are identified, multiple testing corrections reduce the power of the approach. These problems are solved, in part, by interpreting genomics data in the context of prior knowledge about pathways and interactions among genes and their products, which can explain the many complex combinations of alterations in the context of functional modules and processes. For instance, a seemingly complex pattern of alterations may simply affect a single pathway. Therefore, it can be anticipated that proteomic studies will increasingly focus on both the quantification of individual proteins, for example, the enzymes that are responsible for the synthesis of all classes of biomolecules, including transcripts, microRNAs, lipids, glycans, metabolites and proteins themselves, and their interactions within pathways and networks.[3,4]

The genome project also catalyzed the emergence of high-throughput biology whereby molecules of a particular type expressed by specific cells—DNA and histone sequences, transcripts, proteins, lipids and metabolites—are completely enumerated and, if possible, quantified. Proteomics represents high(er)-throughput biology specifically applied to proteins; most proteomic studies have been and still are based on mass spectrometry. Over the last two decades enormous progress has been made in the technologies that measure protein abundances in cells and the number of peptides and proteins that can be identified and quantified has been steadily increasing. Despite these impressive advances, the use and impact of mass spectrometric technologies and their resulting impact in experimental biology have lagged significantly compared to genomics. Without question, this "knowledge translation" gap is slowing the pace by which gene function(s) are determined and genomic discoveries are translated into therapies.

There are four reasons for this discrepancy, both conceptual and technical. First, mass spectrometry technologies are perceived by many scientists as highly complex and the equipment expensive and in constant transformation. While the same is true for genomics and high-throughput sequencing, the nonlinear nature of proteomic analyses and the multiple workflows and modules utilized render proteomics distinctively more complex. Taken together, mass spectrometry tools are inaccessible to many scientists. Indeed, the literature shows that, even within the field of proteomics, a sizable fraction of the high-impact results is generated by a relatively small number of laboratories. Highly confident identification of 7000 to 10000 nonredundant proteins from mass spectrometric analysis of cultured cell lines has become feasible,[5−7] yet many studies still report only 100 or fewer highly abundant proteins. Second, the output of high-throughput mass spectrometry studies carried out in a discovery or screening mode interfaces poorly with the hypothesis-driven research method that remains the mainstay of life science research. It is not obvious how the process of iterative cycles of hypothesis generation and testing benefit from proteomic data sets beyond the discovery stage of an initial screen. Third, the recognition has emerged that cataloguing of proteins in a sample, or according to the coding gene locations on specific chromosomes, a main objective of the C-HPP, is necessary but not sufficient for biological understanding. The physical and functional interactions in the context of dynamic molecular networks in which proteins participate are as important as the structure and function of the individual proteins.[8] The view that biological processes should be considered as dynamic networks of interacting molecules and that changes in the structure or topology of such networks determines phenotypes is the basis for the emerging field of systems biology.[3] Fourth, technical limitations of mass spectrometry technologies with respect to data comprehensiveness and reproducibility of peptide identifications and protein matches have limited the potential to compare proteomic data sets across studies and laboratories. The overall result is that specialized mass spectrometry research

dx.doi.org/10.1021/pr301151m | J. Proteome Res. 2013, 12, 23−27

groups generate increasingly large high-quality data sets, while the vast majority of life science researchers whose research depends on the analysis of a small set of proteins still rely on methods developed decades ago, exemplified by the Western blot and ELISA assays.

The general aim of the Biology/Disease branch of the Human Protein Project (B/D-HPP) initiated by the Human Proteome Organization (HUPO) is to explore the impact that proteomic technologies, exemplified by mass spectrometry, can have when applied to a focused area of biology, in collaboration with experts in that particular field of biology. As reflected in this special issue of the journal, the complementary chromosome-centric branch of the HUPO Human Proteome Project (C-HPP) aims to identify every human protein and map its sequence to the gene coding for it.[9] The combined HPP will create an in-depth database of expressed isoforms, PTMs and proteins. The B/D-HPP will apply this information to understanding the molecular networks that underlie molecular mechanisms that govern health and human disease. It is the organization of the proteome into functional modules determining the cellular, organellar and organ complexity and diversity that drives this initiative.

THE BEGINNING: B/D-HPP ORIGINS AND RATIONALE

For several years, at least since the 2008 HUPO World Congress in Amsterdam, the protein community represented by the HUPO has been discussing the need for an HPP. At the 2010 HUPO meeting in Sydney the members present agreed to a general outline for the HPP (Figure 1).

It was agreed that the overarching goal of the HPP is to enable a large number of biologists, clinician-investigators, and other researchers whose research depends on the analysis of proteins or proteomes to carry out state-of-the-art proteomic projects and to generate high quality results. It was envisaged that this would be accomplished by providing to these researchers three types of information, each deemed critical for the analysis of the proteome, represented as the three pillars in Figure 1: knowledge base, mass spectrometry, and protein capture. The knowledge base compiles, curates, and organizes the information gained about each human protein. This pillar started with the UniProtKB database (www.uniprot.org), supplemented by neXtProt (www.nextprot.org), PRIDE (www.ebi.ac.uk/pride), PeptideAtlas (www.peptideatlas.org), and GPMdb (gpmdb.thegpm.org). The second pillar is the sum of the technology platforms and analytical methods that enable reliable and reproducible identification and quantification of peptides and proteins in cell and tissue extracts using mass spectrometry. The resource comprises validated reference fragment ion spectra for the (enzymatic cleavage-derived) peptides of each human protein. These spectra, in turn, are intended to support and facilitate proteomic studies, for example, as a reference for the peptides and proteins already detected in various biofluids or organs[11] and as the basis for development of validated assays for targeted proteomics by selected reaction monitoring (SRM). The initial instance of this pillar is the PeptideAtlas database at the Institute of Systems Biology[12] and associated database instances, including SRMAtlas[13] (http://www.srmatlas.org) and PASSEL[14] (http://www.peptideatlas.org/passel/). The third pillar is a resource of antibodies or other protein-specific binding and capture reagents for each human protein. These reagents support the detection and quantification of human proteins in biological specimens via a multitude of affinity reagent based assays. The initial instance of this pillar is the Human Protein Atlas database (http://www.proteinatlas.org), with information about tissue distribution and subcellular localization of protein expression.[15]

The resolution to launch a HPP within this framework received broad support in the proteomics community worldwide and catalyzed efforts to initiate its pilot phase. Consortia, broadly formed along national efforts, organized themselves to divide the human proteome into well-defined segments and to characterize the proteins within, thus providing content for the three pillars. As in the early days of the genome-sequencing project, each human chromosome was allocated to a different consortium, and the effort was termed C-HPP. The current state of these efforts is described in this volume.

The C-HPP effort will provide invaluable information, but it is not yet clear that this form of the HPP project will enable effective knowledge translation to the larger community. Thus, as a complementary strategy, at the 2010 and 2011 HUPO World Congress meetings in Sydney and Geneva, the proteomics community adopted a project to carry out state-of-the-art proteomic measurements that specifically address biological problems, within more focused areas of biology or disease and in collaboration with expert biologists. This was termed B/D-HPP, the subject of this document.

THE START: B/D-HPP GOALS AND IMPLEMENTATION

Quantitative bibliometric analysis of research activity on human genes and proteins reveals that most publications are focused on a relatively small set of human proteins.[16] Strikingly, despite the advent of the human genome project and genomic and proteomic technologies, the pattern of research activity is changing little. There is also a marked and often temporal correlation between the availability of high-quality research tools for a given protein and the research activity on that protein.[17] There are three important conclusions from these analyses. First, the vast majority of human proteins remain essentially unexplored. Second, neither knowledge of the human genome sequence nor powerful proteomic techniques have fundamentally affected the fraction of the proteome that is most intensely studied, predominantly by hypothesis-driven research. In other words, the most studied proteins remained the favorite. There was no broadening due to the discovery and exploration of the many "new" proteins being discovered and correlated to changes in a wide number of circumstances. Third, the pattern of protein-related research can be influenced by the availability of research tools and methods. Scientists will explore relatively understudied proteins if provided with research tools and specific assays that are readily available, that is, that are reproducible, accurate and easily quantified.

The B/D-HPP aims to explore whether access to reliable and reproducible measurement techniques will facilitate research on a much larger fraction of the proteome by a broader range of scientists than those representing the core proteomics community. These goals will be pursued, along the HPP three-pillar strategy, by providing: (1) specific and validated mass spectrometric ways to quantify the abundance of all proteins within a focus area of biology, (2) specific and selective protein affinity reagents for each of these proteins, and (3) the knowledge how to use these assays and reagents, in a format easily accessible for the scientific community. In short, the B/D-HPP attempts to provide a spectrum of research tools for

dx.doi.org/10.1021/pr301151m | J. Proteome Res. 2013, 12, 23−27

studying the proteins of relevance in a specific area of biology or specific disease. Since these resources generated will support, in principle, the reproducible quantification of any protein in a suitable sample, biological processes or disease mechanisms are expected to be increasingly analyzed in the full complexity of the living cell.

Toward this end, the B/D-HPP defined the following specific goals:

1) Select biological focus areas. One or more areas of biology, or disease areas, should be selected as a focus. The area(s) might derive from self-organizing groups of scientists who study specific biological processes or diseases and who have great interest in supporting the application of proteomics technologies to their field.
2) Generate a target list. Once the areas of biology are defined, the area experts should deliver a list of proteins known to be, or hypothesized to be, relevant for that field. For example, a biological theme on cellular signaling might comprise all human protein kinases.
3) Define and generate relevant assays and reagents for these targets. The collaborative team of biologists and proteomics experts should then define the assays and reagents of most relevance to the problem, and generate a scientific plan to generate and characterize those assays and reagents.
4) Disseminate the knowledge and reagents. The impact of the B/D-HPP will depend on community uptake, and thus the B/D-HPP proposes that all mass spectrometry and affinity based assays and reagents would be made publicly accessible, with no restriction on use.

This process is exemplified by the recently published study by Hüttenhain et al. on cancer-associated proteins.[18] An available list of more than 1000 cancer-related proteins compiled from the literature by Leigh Anderson and colleagues[19] was used as the starting point for the development and refinement of selected reaction monitoring (SRM) assays for these human proteins. These SRM assays were then used to determine the detectability of the target proteins in two types of samples, human plasma and urine; 182 proteins were detected in immuno-depleted plasma, spanning 5 orders of magnitude and reaching below a concentration of 10 ng/mL. The narrower concentration range of proteins in urine compared to plasma allowed the detection of 408 proteins. Through public access to all the assay reagents and the spectral library via SRMAtlas[13] and PASSEL,[14] researchers are now able to target cancer-associated proteins contained in the assay library that are of interest to their respective project in any sample, using the detectability information in plasma and urine as a guidance.

B/D-HPP PRESENT STATUS

At the 2012 HUPO World Congress in Boston, nine B/D-HPP workshops were held in which scientists discussed the formation of working groups focused on specific biological processes and disease areas. These initial discussions have continued after the congress and several additional working groups are being constituted. The following B/D-HPP teams are emerging:

- Diabetes (Jean-Charles Sanchez, Geneva, Switzerland; Peter Bergsten, Uppsala, Sweden; and Martin Kussmann, Lausanne, Switzerland; cochairs). The group held a workshop Nov seventh, 2012 in which they specified 1300 proteins that are of central interest and relevance to the diabetes research community. The group already has attracted external funding (http://cms.unige.ch/medecine/hdpp.info/node/1).
- Cancer Proteomics (Juan Pablo Albar, Madrid, Spain; and Hui Zhang, Baltimore, MD; cochairs). The group is compiling lists of cancer-associated proteins to extend the available assay libraries.
- Mitochondria (Andrea Urbani, Rome, Italy). This is a bridge project between B/D-HPP and C-HPP (the "mitochondrial chromosome"). Mitochondrial proteins may be encoded in both mitochondrial DNA and nuclear DNA.
- Infectious Diseases (Ileana Cristea, Princeton, NJ). This project will feature co-analysis of host and micro-organism proteomes.
- Epigenetics/chromatin-associated proteins (Structural Genomics Consortium, Alec Edwards, chair, Toronto, Canada). Proteins related to chromatin and epigenetics comprise a set of ~500 that are relevant to biology and drug discovery.
- Pre-existing HUPO organ-based initiatives have expressed a desire to join and become embedded in the B/D-HPP. They will benefit from stimulation to pursue fresh directions as laid out by the B/D-HPP. These include the Plasma, Liver, Brain, Kidney-Urine, Stem Cells, Cardiovascular, and Model Organisms Proteome Projects, and the Human Antibody, Glycoproteomics, and Protein Standards Initiatives (see Figure 2).

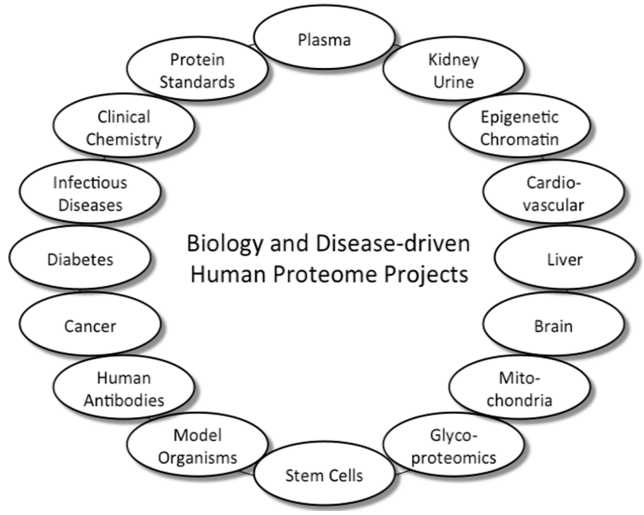


Figure 2. Initial components of the B/D-HPP.

MOVING FORWARD: BD HPP PLANS FOR THE NEXT YEAR

Over the next year, the B/D-HPP program will focus on launching several pilot projects. The most advanced working groups may generate mass spectrometric or affinity reagent based sets for a significant fraction of their initial Target List by the end of 2013; this information will be made available to the three pillar databases of the HPP. We will encourage collaborations that bridge protein capture and mass spectrometry, as in the use of PrEST epitope tags with mass

dx.doi.org/10.1021/pr301151m | J. Proteome Res. 2013, 12, 23−27

spectrometry by Zeiler et al.[20] A key success metric will be the deployment of the B/D-HPP output be used by scientists unaffiliated with the project. It is expected that the working groups and additional scientists will provide updates on their progress at the 2013 HUPO World Congress in Yokohama.

As pilot projects are launched, working groups and the B/D-HPP will work with area experts to identify funding mechanisms to support B/D-HPP projects. The community-enabling aims of the B/D-HPP, as well as the commitment to place the information into the public domain without restriction, should make the projects very attractive for public and private sector funding.

CONCLUSIONS

The B/D-HPP is undertaking the ambitious goal of fundamentally transforming biological research by enabling the reliable detection and quantification of every human protein by the general life science research community. The impact of such an advance cannot be overestimated because both the scope of proteins accessible to experimentation and the range of scientists with access to state-of-the-art proteomic measurements would vastly increase.

We believe that there are few technological impediments to this objective. The base technology for targeted mass spectrometry quantification of an individual protein is well established, including software programs for assay development,[21,22] robust data acquisition strategies and statistically sound data analysis tools[23,24] and assay repositories.[13,14] Exciting new technical developments, particularly in the field of data independent acquisition (DIA)[25] mass spectrometry also promise to vastly extend the range of proteins that can be targeted in a single analysis. The methods to generate high affinity, renewable antibodies for any proteins are now also established.

What are the main hurdles? Although there are remaining technological challenges that remain, the hurdles are largely organizational. If methods are developed, will the scientific community immediately use them? If not, how will uptake by the community be accelerated? How should the project be funded and managed? Can one achieve global agreement to make all the information and reagents available without restriction? Will instrument manufacturers produce robust, simple to operate mass spectrometers for use by the nonexpert community? We believe that these organizational hurdles will be overcome by joint efforts of the community of life scientists, funding agencies and the private sector (e.g., instrument manufacturers, reagent providers). The benefits for life science research are well worth the effort.

AUTHOR INFORMATION

Corresponding Author

*E-mail: aebersold@imsb.biol.ethz.ch.

Notes

The authors declare no competing financial interest.

[§]Member of Executive committee of HUPO B/D-HPP project.

ACKNOWLEDGMENTS

We acknowledge the support of and the numerous discussions with HUPO members during the development of the presented vision. We thank Dr. Ruth Hüttenhain for her help in the preparation of this document.

REFERENCES

(1) Gerstein, M. B.; et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **2012**, *489* (7414), 91–100.

(2) Dunham, I.; et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489* (7414), 57–74.

(3) Bensimon, A.; Heck, A. J. R.; Aebersold, R. Mass spectrometry–based proteomics and network biology. *Annu. Rev. Biochem.* **2012**, *81* (1), 379–405.

(4) Hood, L. E.; et al. New and improved proteomics technologies for understanding complex biological systems: Addressing a grand challenge in the life sciences. *Proteomics* **2012**, *12* (18), 2773–2783.

(5) Nagaraj, N.; et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **2011**, 7.

(6) Beck, M.; et al. The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **2011**, *7*, 549.

(7) Munoz, J.; et al. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol. Syst. Biol.* **2011**, *7*, 550.

(8) Vidal, M.; et al. The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. *Clin. Proteomics* **2012**, *9* (1), 6.

(9) Berglund, L.; et al. A gene centric Human Protein Atlas for expression profiles based on antibodies. *Mol. Cell. Proteomics* **2008**, *7* (10), 2019–2027.

(10) Legrain, P.; et al. The Human Proteome Project: Current state and future direction. *Mol. Cell. Proteomics* **2011**, DOI: 10.1074/mcp.M111.009993-1.

(11) Farrah, T.; et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* **2011**, *10* (9), M110.006353.

(12) Deutsch, E.; et al. Human Plasma PeptideAtlas. *Proteomics* **2005**, *5* (13), 3497–3500.

(13) Picotti, P.; et al. A database of mass spectrometric assays for the yeast proteome. *Nat. Methods* **2008**, *5* (11), 913–914.

(14) Farrah, T.; et al. PASSEL: The PeptideAtlas SRM Experiment Library. *Proteomics* **2012**, *12*, 1170–1175.

(15) Uhlen, M.; et al. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28* (12), 1248–1250.

(16) Edwards, A. M.; et al. Too many roads not taken. *Nature* **2011**, *470* (7333), 163–165.

(17) Isserlin, R., et al. Preprint at http://arxiv.org/abs/1102.0448v2, 2011.

(18) Hüttenhain, R.; et al. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. *Sci. Transl. Med.* **2012**, *4* (142), 142ra94–142ra94.

(19) Polanski, M.; Anderson, N. L. A list of candidate cancer biomarkers for targeted proteomics. *Biomarker Insights* **2007**, *1*, 1–48.

(20) Zeiler, M.; et al. A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol. Cell. Proteomics* **2012**, *11* (3), O111.009613.

(21) MacLean, B.; et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26* (7), 966–968.

(22) Brusniak, M. Y.; et al. ATAQS: A computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry. *BMC Bioinform.* **2011**, *12*, 78.

(23) Reiter, L.; et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **2011**, *8*, 430–435.

(24) Chang, C.-Y.; et al. Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell. Proteomics* **2012**, *11* (4), M111.014662.

(25) Gillet, L. C.; et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **2012**, *11* (6), O111.016717.