A Human Proteome Detection and Quantitation Project*

N. Leigh Anderson‡§, Norman G. Anderson‡, Terry W. Pearson¶, Christoph H. Borchers‖, Amanda G. Paulovich, Scott D. Patterson‡‡, Michael Gillette§§, Ruedi Aebersold¶¶, and Steven A. Carr§§**

The lack of sensitive, specific, multiplexable assays for most human proteins is the major technical barrier impeding development of candidate biomarkers into clinically useful tests. Recent progress in mass spectrometry-based assays for proteotypic peptides, particularly those with specific affinity peptide enrichment, offers a systematic and economical path to comprehensive quantitative coverage of the human proteome. A complete suite of assays, *e.g.* two peptides from the protein product of each of the ~20,500 human genes (here termed the human Proteome Detection and Quantitation project), would enable rapid and systematic verification of candidate biomarkers and lay a quantitative foundation for subsequent efforts to define the larger universe of splice variants, post-translational modifications, protein-protein interactions, and tissue localization. ***Molecular & Cellular Proteomics 8:883–886, 2009.***

There is growing interest in the idea of a comprehensive Human Proteome Project (1) to exploit and extend the successful effort to sequence the human genome. Major challenges in defining a comprehensive Human Proteome Project (and distinguishing it from the genome effort) are 1) the potentially very large number of proteins with modified forms; 2) the diversity of technology platforms involved in their study; 3) the variety of overlapping biological "units" into which the proteome might be divided for organized conquest; and 4) sensitivity limitations in detecting proteins present in trace amounts. The process of analyzing and discussing these issues may (and ought to) be lengthy, as it addresses core scientific unknowns as well as decisions about the organization and scale of biomedical research in the future. The benefits of taking time to involve the entire biological research community, and especially the medical research segment, in these discussions are substantial.

Progress in systematically measuring proteins, however, need not wait for the conclusion of such discussions. We propose a near-term tactical approach, called the human Proteome Detection and Quantitation (hPDQ)[1] project that will enable measurement of the human proteome in a way that would yield immediately useful results while the strategy for a comprehensive Human Proteome Project is worked out. The hPDQ project is aimed at overcoming present difficulties in answering basic biological questions about the relationship between protein abundance (or concentration) and gene expression, phenotype, disease, and treatment response; *i.e.*, the growing field of protein biomarkers. It is thus focused on the study of biological variation affecting protein expression rather than study of structure and mechanism and in this initial form does not directly address splice variants or most post-translational modifications. It is aimed at providing immediately useful capabilities to the human biology research community, in a way that does not adversely impact funding for individual investigators and does not generate administrative constraints on their ability to set and change courses in the conduct of research. Specifically, the goal of the hPDQ is to enable individual biological researchers to measure defined collections of human proteins in biological samples with 1 ng/ml sensitivity and absolute specificity, at throughput and cost levels that permit the study of meaningfully large biological populations (~500–5,000 samples).

We clearly do not have this capability today. If an investigator defines a set of 20 proteins hypothesized to change in relation to some biological process or event, assays for only a minority (often none!) will typically be available. Further, these assays will lack absolute specificity and will not easily be multiplexed. Current proteomics research platforms are focused mainly on discovery; providing increasingly broad protein sampling surveys, generally at low throughput and high cost. Such approaches generally do not yield an economical or accurate measurement of a defined set of proteins in every

From the ‡The Plasma Proteome Institute, Washington, D. C. 20009, ¶Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia V8W 3P6, Canada, ‖University of Victoria-Genome BC Proteomics Centre, Vancouver Island Technology Park, Victoria, British Columbia V8Z 7X8, Canada, **Fred Hutchinson Cancer Research Center, Seattle, Washington WA 98109-1024, ‡‡Amgen Inc., California 91320-1799, §§The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, and ¶¶Institute of Molecular Systems Biology, HPTE 78, Wolfgang-Pauli-Str. 16, 8093 Zurich, Switzerland

Received, October 17, 2008, and in revised form, January 2, 2009
Published, MCP Papers in Press, January 7, 2009, DOI 10.1074/mcp.R800015-MCP200

[1] The abbreviations used are: hPDQ, human Proteome Detection and Quantitation; MS, mass spectrometry; SISCAPA, stable isotope standards and capture by anti-peptide antibodies; ELISA, enzyme-linked immunosorbent assay.

© 2009 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

sample. There is thus a fundamental barrier to hypothesis testing in quantitative proteomics, where relationships between protein abundance and biology are sought. A particularly important instance of this limitation occurs in the effort to establish useful biomarkers of disease, for diagnosis, for measuring efficacy of treatment, and for monitoring of disease recurrence. This limitation is largely responsible for the research community's failure in recent years to bring forward significant numbers of new proteins as Food and Drug Administration approved diagnostic tests (2). However, if a robust, economical, and widely diffused capability to measure all human proteins existed, the research community would have the collective means to assess the utility of all human proteins as biomarkers in hundreds of diseases and other processes in the most efficient way.

The need for new or improved biomarkers in many areas of healthcare has become critical. Early detection of cancer, coupled with surgical intervention, has the potential to radically improve survival (3), provided early markers exist and can be found. Without good biomarkers, degenerative diseases such as Alzheimer and chronic obstructive pulmonary disease (COPD) are difficult to detect early enough to benefit from the potential therapies. Clinical development of new drugs increasingly depends on identification of biomarkers for pharmacodynamic assessment of drug action to help guide dose and schedule, and predictive biomarkers for selection of patients who will benefit from therapy (4). Companion diagnostics are the currency of personalized medicine and represent those predictive or response biomarkers that are linked to specific therapeutics, substantially increasing their clinical value. Surrogate biomarkers (those biomarkers that substitute for a clinical outcome or response) are the most difficult to discover and to verify because of the long timeframe required but can radically shorten appropriate clinical trials. The impact of a vigorous increase in clinical biomarkers could thus be enormous, both in terms of patient well being and financial viability of healthcare systems worldwide.

Protein measurements are also likely to play an important role in assessing the quality of material stored in large clinical sample collections (Biobanks). Much discussion has occurred recently regarding the value of banked samples because of unknown degrees of protein degradation occurring during acquisition, processing, and storage. This matter is of acute concern in the case of serum, where coagulation initiates a plethora of proteolytic cleavage events. The hPDQ may provide the opportunity to determine the value of each sample through the development of prototypic peptides tracking the stability of labile proteins.

An attractive technology for achieving the objective of hPDQ is quantitative mass spectrometry, the sensitivity, and specificity of which are well established in the measurement of small molecules (5, 6) and peptides (7, 8). To achieve comprehensive quantitation of proteins, given the immense variability in their physical properties, these larger molecules are digested to component peptides using an enzyme such as trypsin, and protein amount is measured using proteotypic peptides (9, 10) as specific stoichiometric surrogates. Multiple peptides from a target protein provide independent confirmation of this stoichiometry (equivalent to having multiple enzyme-linked immunosorbent assays with different antibody pairs), serving to control for the possibility of incomplete digestion or subsequent losses. Accurate calibration is achieved by spiking digested samples with known quantities of synthetic stable-isotope labeled peptides as internal standards (11, 12). The sensitivity of this approach for multiplexed analysis of proteins in plasma has been extended from the microgram (13) to nanogram/ml levels by depletion of abundant proteins and limited peptide fractionation prior to analysis (14) or by capture of the subset of glycopeptides (15). Sensitivity and throughput of peptide MS measurements can be further increased to levels required in hPDQ by specific enrichment of the target peptides using anti-peptide antibodies. This method, called SISCAPA (for "stable isotope standards and capture by anti-peptide antibodies") (16) or iMALDI (for immuno-MALDI) (17), combines the enhanced sensitivity of immunoassays with the specificity of mass spectrometry, while maintaining multiplexing capability. For these reasons we emphasize SISCAPA and iMALDI in this hPDQ proposal, although proteins in the 100 ng/ml or higher concentration are readily accessible by targeted MS in plasma without antibody enrichment. Combining these elements results in a measurement system, with the potential to measure 10–100 selected proteins at ng/ml levels in small ($\sim$10 μl) samples of human plasma in a single short analytical run. Sensitivity can be further increased through the use of larger samples and/or advances in MS sensitivity. In comparison to the conventional ELISA approach, MS-based SISCAPA assays are less expensive to develop (one antibody instead of a carefully matched pair), easier to multiplex (off-target interactions being less likely with peptides than proteins), and provide absolute structural specificity (by reading the masses of multiple specific peptide fragments). This improved specificity solves a major problem plaguing clinical immunoassays for proteins such as thyroglobulin (18) and has led to the development of first clinical SISCAPA assay (19). In addition, since the mass spectrometer functions as a "second antibody" that identifies the captured peptides, the anti-peptide antibody used for peptide enrichment need not have perfect specificity. This greatly reduces the cost of affinity reagents, currently a limiting factor in developing ELISA assays for large numbers of protein analytes.

Achieving the hPDQ goal by this approach would require that four resources be generally available. 1) A comprehensive database of proteotypic (protein-unique) peptides for each of the 21,500 human proteins (20), coupled with experimental or computational data identifying the best peptides for MS measurement and associated optimized MS instrument parameters. 2) At least two synthetic proteotypic peptides, la-

beled with stable isotope(s) and available in accurately quantitated aliquots, for use as internal measurement standards for quantitation of each protein. Such peptides are readily available today through custom order, at rapidly declining prices. 3) Anti-peptide antibodies specific for the same two proteotypic peptides per target protein, capable of binding the peptides with dissociation constants $<$ 1e-9 (the level required in theory and practice to enrich low-abundance peptides from complex sample digests). Such antibodies are now being made for a variety of targets, and a robust production pipeline is being developed. Monoclonal antibodies would be preferred, despite their higher development cost, to establish a stable reagent supply, especially for those targets that prove useful as biomarkers. 4) Robust and affordable instrument platforms for quantitative analysis of small (amol to fmol) amounts of tryptic peptides and for sample preparation. Existing triple-quadrupole mass spectrometers (with a current worldwide installed base of more than 6,000 instruments) coupled with nanoflow ($\sim$300–600 nl/min) liquid chromatography systems can meet this requirement and are undergoing rapid improvement with declining cost. MALDI platforms may provide similar capabilities at even higher throughput.

We estimate that an initial pilot phase targeting 2,000 proteins selected for biomarker potential could be completed in two years at a cost of less than $50 million through funding of existing academic and commercial resources in a distributed network. In the following five years, the remaining 18,500 proteins could be targeted for $250 million, making use of anticipated technical improvements, particularly in the strategies for generating suitable high affinity monoclonal antibodies (21) in large numbers at low cost (22).

Although the natural mechanism for providing the hPDQ database (resource 1 above) is through an academic collaboration, perhaps modeled on the successful Global Protein Machine (23) and Peptide Atlas (24) databases, the other resources would benefit from commercial distribution by experienced providers of instruments and reagents. The required instrument platforms (4 above) serve existing markets, and their further development is unlikely to require additional funding for hPDQ applications. However, business economics does not presently justify the expense of developing well characterized antibodies and peptides for quantitation of proteins that are not already recognized as pivotal in biological research (*i.e.* precisely those in need of the attention of the research community). Hence a substantial portion of the required funding for the proposed approach for such antibody and peptide reagents will be needed from government and philanthropic sources. A significant advantage of such diversified support would be the leverage it would provide in retaining in the public domain the identities of the selected peptides, their parameters and basic measurement protocols.

The value of a general protein measurement capability for research is very substantial, but the proposed effort would not solve several larger issues that must await definition of a broader human proteome program. For example, the hPDQ project does not address the basic process of *de novo* proteome-wide discovery; the comprehensive exploration of splice forms, post-translational modifications, active fragments of preproteins or genetic variants (although once known, most of these can be targeted by the methods used here); interactions among proteins or with other molecules; or spatial arrangement of proteins in organs and tissues. Each of these areas would benefit from the resources proposed in hPDQ, but will likely require separate, coordinated large-scale efforts that are likely to identify additional sets of biomarkers. Thus although a complete suite of targeted assays is only a first step toward the complete human proteome, we feel that its fundamental importance for progress in biomarker research and its value as a foundation for protein quantitation justifies consideration as an initial step.

In the beginning of the study of protein diagnostics, investigators at the Behring Institute discovered many of the well known plasma proteins and made associated specific antibodies and antibody-based quantitative tests available to the research community worldwide, spurring the initial round of plasma biomarker research. The application of monoclonal antibodies sparked additional discoveries through close coupling of protein "discovery" with simple quantitative monoclonal antibody-based assays - this "shortcut" to clinical measurement allowed investigators to publish more than 1,000 papers referring to the ovarian cancer marker CA125 (measured by ELISA) before the sequence of the protein was finally identified in 2001 (25). The broader proteomics technologies (beginning with the two-dimensional electrophoresis technology that formed the basis of the Human Protein Index Project (26) formulated by two of us almost 30 years ago, and extending to modern shotgun-style MS-based approaches) have radically expanded the universe of observable proteins. However, quantitative specific assay capabilities have not kept pace with this expansion, leading to the current gap between biomarker proteomics and clinical biomarker output. It is now time to address this gap and realize the benefits of a clinically accessible human proteome. Effective translation of basic research into tangible medical benefit requires it.

* This work was supported, in whole or in part, by National Institutes of Health Grant 1U24 CA126476-02 (to N. L. A., T. W. P., C. H. B., A. G. P., and S. A. C.) from the National Cancer Institute as part of the NCI Clinical Proteomic Technologies Initiative.

§ To whom correspondence should be addressed: Tel.: 301-728-1451; Fax: 202-234-9175; E-mail: leighanderson@plasmaproteome.org.

REFERENCES

1. Editorial (2008) The big ome. *Nature* **452,** 913–914
2. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
3. Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., Radich, J., Anderson, G., and Hartwell, L. (2003) Early detection: the case for early detection. *Nat. Rev. Cancer* **3,** 243–252

4. Severino, M. E., Dubose, R. F., and Patterson, S. D. (2006) A strategic view on the use of pharmacodynamic biomarkers in early clinical drug development. *IDrugs* **9,** 849–853
5. Kostiainen, R., Kotiaho, T., Kuuranne, T., and Auriola, S. (2003) Liquid chromatography/atmospheric pressure ionization-mass spectrometry in drug metabolism studies. *J. Mass Spectrom.* **38,** 357–372
6. Streit, F., Armstrong, V. W., and Oellerich, M. (2002) Rapid liquid chromatography-tandem mass spectrometry routine method for simultaneous determination of sirolimus, everolimus, tacrolimus, and cyclosporin a in whole blood. *Clin. Chem.* **48,** 955–958
7. Tuthill, C. W., Rudolph, A., Li, Y., Tan, B., Fitzgerald, T. J., Beck, S. R., and Li, Y. X. (2000) Quantitative analysis of thymosin alpha1 in human serum by LC-MS/MS. *AAPS PharmSciTech* **1,** E11
8. Desiderio, D. M., Yamada, S., Tanzer, F. S., Horton, J., and Trimble, J. (1981) High-performance liquid chromatographic and field desorption mass spectrometric measurement of picomole amounts of endogenous neuropeptides in biologic tissue. *J. Chromatogr.* **217,** 437–452
9. Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6,** 577–583
10. Craig, R., Cortens, J. P., and Beavis, R. C. (2005) The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **19,** 1844–1850
11. Barr, J. R., Maggio, V. L., Patterson, D. G., Jr., Cooper, G. R., Henderson, L. O., Turner, W. E., Smith, S. J., Hannon, W. H., Needham, L. L., and Sampson, E. J. (1996) Isotope dilution-mass spectrometric quantification of specific proteins: model application with apolipoprotein a-i. *Clin. Chem.* **42,** 1676–1682
12. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 6940–6945
13. Anderson, L., and Hunter, C. L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics* **5,** 573–588
14. Keshishian, H., Addona, T., Burgess, M., Kuhn, E., and Carr, S. A. (2007) Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell. Proteomics* **6,** 2212–2229
15. Stahl-Zeng, J., Lange, V., Ossola, R., Eckhardt, K., Krek, W., Aebersold, R., and Domon, B. (2007) High sensitivity detection of plasma proteins by multiple reaction monitoring of n-glycosites. *Mol. Cell. Proteomics* **6,** 1809–1817
16. Anderson, N. L., Anderson, N. G., Haines, L. R., Hardie, D. B., Olafson, R. W., and Pearson, T. W. (2004) Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (siscapa). *J. Proteome Res.* **3,** 235–244
17. Jiang, J., Parker, C. E., Hoadley, K. A., Perou, C. M., Boysen, G., and Borchers, C. H. (2007) Development of an immuno tandem mass spectrometry (iMALDI) assay for EGFR diagnosis. *Proteomics Clin. Appl.* **1,** 1651–1659
18. Hoofnagle, A. N., and Wener, M. H. (2006) Serum thyroglobulin: a model of immunoassay imperfection. *Clin. Lab. Int.* **8,** 12–14
19. Hoofnagle, A. N., Becker, J. O., Wener, M. H., and Heinecke, J. W. (2008) Quantification of thyroglobulin, a low-abundance serum protein, by immunoaffinity peptide enrichment and tandem mass spectrometry. *Clin. Chem*.
20. Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 19428–19433
21. Pope, M. E., Soste, M. V., Eyford, B. A., Anderson, N. L., and Pearson, T. W. (2009) Anti-peptide antibody screening: Selection of high affinity monoclonal reagents by a refined surface plasmon resonance technique. *J. Immunol. Methods* 10.1016/j.jim.2008.11.004
22. De Masi, F., Chiarella, P., Wilhelm, H., Massimi, M., Bullard, B., Ansorge, W., and Sawyer, A. (2005) High throughput production of mouse monoclonal antibodies using antigen microarrays. *Proteomics* **5,** 4070–4081
23. Craig, R., Cortens, J. P., and Beavis, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3,** 1234–1242
24. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) Peptideatlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9,** 429–434
25. Lloyd, K. O., Yin, B. W., and Kudryashov, V. (1997) Isolation and characterization of ovarian cancer antigen ca 125 using a new monoclonal antibody (vk-8): identification as a mucin-type molecule. *Int. J. Cancer* **71,** 842–850
26. Anderson, N. G., and Anderson, L. (1982) The human protein index. *Clin. Chem.* **28,** 739–748