Preface

Plasma and serum are the preferred specimens for non-invasive sampling of normal individuals, at-risk groups, and patients for protein biomarkers discovered and validated to reflect physiological, pathological, and pharmacological phenotypes. These specimens present enormous challenges due to extreme complexity, representing potentially all proteins in the body and their isoforms; at least ten orders of magnitude range in protein concentrations; intra-individual and inter-individual variation from genetics, diet, smoking, hormones, and many other sources; and especially non-standardized methods of sample processing. Furthermore, the inherent limitations of incomplete sampling of peptides by mass spectrometry and high error rates of peptide identifications and protein assignments with various search algorithms and databases lead to low concordance of protein identifications even with repeat analyses of the same sample. These features complicate diagnostic comparisons of specimens.

The Human Proteome Organization (HUPO) has launched several major initiatives to explore the proteomes of liver, brain, and plasma and to generate informatics standards and large-scale antibody production. This book presents the major findings from the pilot phase of the Plasma Proteome Project (PPP). The 17 chapters embrace a combination of collaborative analyses of HUPO PPP reference specimens and several lab-specific projects, both experimental and analytical. The investigators compared PPP reference specimens of human serum and EDTA, heparin, and citrate-anti-coagulated plasma; EDTA-plasma was determined to be the preferred specimen. Together these chapters examine many features of specimen handling, depletion of abundant proteins, fractionation of intact proteins, fractionation of tryptic digest peptides, and analysis of those peptides with various MS/MS instruments. Combinations of technologies gave the most resolution. The subsequent step of matching spectra to peptide sequences with a variety algorithms has numerous, often unspecified parameters. The alignment of peptide sequences with proteins via protein or gene databases likewise is laden with uncertainties and redundancies. Especially for longitudinal and collaborative studies, the periodic issuance of modified versions of the databases creates a moving target for protein identification and annotation, let alone comparison of results from different studies. These challenges are explored in depth. As in the special issue of Proteomics (August 2005) with a total of 28 papers, the authors here provide a revealing snapshot of the output from a variety of proteomics technology platforms across laboratories.

Exploring the Human Plasma Proteome. Edited by Gilbert S. Omenn

ISBN: 3-527-31757-0

The extensive annotations show that present methods already are capable of detecting in plasma large numbers of low-abundance proteins of great biological interest from essentially all cellular compartments. Studies focusing on subproteomes based on glycoprotein enrichment or molecular weight yielded additional findings. As more powerful technologies are applied, we can expect ever more extensive identification, as well as quantitation, of proteins and their isoforms. The high proportion of genes which generate detectable splice isoforms further complicates protein identifications, yet helps to clarify the basis on which humans can have such complex phenotypes with a surprisingly small complement of genes (latest Human Genome Project estimate is about 22,000 protein-encoding genes).

The PPP Core Dataset has 5102 proteins identified with 2 or more peptides, of which 3020 remain after application of our integration algorithm for protein matches which cannot be distinguished with the available peptides. A special feature of the PPP is the set of independent analyses from the raw spectra or peaklists across the multiple laboratories. These independent analyses eliminate the high variability from lab-specific search algorithms, different databases, and investigators' judgments, though each independent analysis has its own peculiar attributes. We also provide comparisons with several published datasets. Meta-analysis of separate studies has similar challenges to those experienced in the integration of datasets from the collaborating PPP laboratories.

Numerous other "cuts" of the data can be made. The primary data are available for such additional analyses at the European Bioinformatics Institute (www.ebi.ac.uk/pride); the University of Michigan (www.bioinformatics.med.umich.edu/hupo/ppp); and the Institute for Systems Biology (www.peptideatlas.org). We are keen to encourage such further analyses. Two examples have already appeared, introducing adjustments for protein length and multiple comparisons testing [1] and enhancing the characterization of the human genome from these proteomics data and gene mapping [2]. This publication presents the foundation for planning the next phases of the Plasma Proteome Project, with Young-Ki Paik, Matthias Mann, and myself as co-chairs. We will:

1. develop standardized operating procedures for specimens, protein and peptide fractionation, and analyses, with attention to replicability of results, to make proteomics practicable for clinical chemistry;
2. select priority PPP proteins for the HUPO Antibody Production Initiative, to generate reagents for biomarker and pathways studies and plasma/organ proteome comparisons;
3. collaborate on informatics, databases, annotations, and error estimation for plasma and serum studies, both HUPO-initiated and published by others;
4. stimulate proteomics technology advances, with special attention to high-resolution/higher-throughput methods and to quantitation of proteins and characterization of modified proteins (primarily glycoproteins and phosphoproteins); and
5. assure paired analyses of plasma and tissue specimens in organ-based and disease-focused proteomics initiatives.

The spirit of collaboration in the Plasma Proteome Project has been splendid. The substantial commitment of so many investigators and sponsors to this pilot phase has been admirable. As a work-in-progress the PPP has generated productive discussions at many scientific meetings. On behalf of the Executive Committee and Technical Committees, I thank everyone involved.

Gilbert S. Omenn
University of Michigan, Ann Arbor
August 2006

1. States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D., Eng, J., Speicher, D. W., Hanash, S. M. *Challenges in deriving high-confidence protein identifications from data gathered by HUPO plasma proteome collaborative study. Nature Biotech* 2006, *24*, 333–338.
2. Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xy, Y., Ulintz, P., Omenn, G. S., States, D. J. *Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biology* 2006, *7:R35*, Published online: http://genomebiology.com/2006/7/4/R35.