

# Exploring the Human Plasma Proteome

*Edited by*  
*Gilbert S. Omenn*



**WILEY-  
VCH**

WILEY-VCH Verlag GmbH & Co. KGaA



**Exploring the Human  
Plasma Proteome**

*Edited by  
Gilbert S. Omenn*

## ***Related Titles***

Jungblut, P. R.,  
Hecker, M. (Eds.)

### **Proteomics of Microbial Pathogens**

2006  
ISBN-13: 978-3-527-31759-2  
ISBN-10: 3-527-31759-7

Liebler, D. C., Petricoin, E. F.,  
Liotta, L. A. (Eds.)

### **Proteomics in Cancer Research**

2005  
ISBN-13: 978-0-471-44476-3  
ISBN-10: 0-471-44476-6

Lion, N., Rossier, J. S.,  
Girault, H. (Eds.)

### **Microfluidic Applications in Biology**

**From Technologies to Systems Biology**

2006  
ISBN-13: 978-3-527-31761-5  
ISBN-10: 3-527-31761-9

Sanchez, J.-C., Corthals, G. L., Hoch-  
strasser, D. F. (Eds.)

### **Biomedical Applications of Proteomics**

2004  
ISBN-13: 978-3-527-30807-1  
ISBN-10: 3-527-30807-5

Hamacher, M., Marcus, K., Stühler, K.,  
van Hall, A., Warscheid, B., Meyer, H.E.  
(Eds.)

### **Proteomics in Drug Research**

2006  
ISBN-13: 978-3-527-31226-9  
ISBN-10: 3-527-31226-9

# Exploring the Human Plasma Proteome

*Edited by*  
*Gilbert S. Omenn*



**WILEY-  
VCH**

WILEY-VCH Verlag GmbH & Co. KGaA

#### The Editor

**Prof. Dr. Gilbert S. Omenn**

University of Michigan  
A520 MSRB I Bldg  
1150 West Medical Center  
Dr. Ann Arbor,  
MI 48109-0626  
USA

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

**Library of Congress Card No.:**  
applied for

**British Library Cataloguing-in-Publication Data**  
A catalogue record for this book is available from the British Library.

**Bibliographic information published by  
Die Deutsche Bibliothek**

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <<http://dnb.ddb.de>>.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Printed in the Federal Republic of Germany  
Printed on acid-free paper

**Typesetting** X Con Media AG, Bonn  
**Printing** Strauss GmbH, Mörlenbach  
**Binding** Litges & Dopf Buchbinderei GmbH, Heppenheim

**ISBN:** 3-527-31757-0

**ISBN 13:** 978-3-527-31757-8

## Table of Contents

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>Overview of the HUPO Plasma Proteome Project: Results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database</b>   | <b>1</b> |
|          | <i>Gilbert S. Omenn, David J. States, Marcin Adamski, Thomas W. Blackwell, Rajasree Menon, Henning Hermjakob, Rolf Apweiler, Brian B. Haab, Richard J. Simpson, James S. Eddes, Eugene A. Kapp, Robert L. Moritz, Daniel W. Chan, Alex J. Rai, Arie Admon, Ruedi Aebersold, Jimmy Eng, William S. Hancock, Stanley A. Hefta, Helmut Meyer, Young-Ki Paik, Jong-Shin Yoo, Peipei Ping, Joel Pounds, Joshua Adkins, Xiaohong Qian, Rong Wang, Valerie Wasinger, Chi Yue Wu, Xiaohang Zhao, Rong Zeng, Alexander Archakov, Akira Tsugita, Ilan Beer, Akhilesh Pandey, Michael Pisano, Philip Andrews, Harald Tammen, David W. Speicher and Samir M. Hanash</i> |          |
| 1.1      | Introduction  | 2        |
| 1.2      | PPP reference specimens   | 4        |
| 1.3      | Bioinformatics and technology platforms   | 5        |
| 1.3.1    | Constructing a PPP database for human plasma and serum proteins   | 5        |
| 1.3.2    | Analysis of confidence of protein identifications   | 14       |
| 1.3.3    | Quantitation of protein concentrations  | 15       |
| 1.4      | Comparing the specimens   | 17       |
| 1.4.1    | Choice of specimen and collection and handling variables  | 17       |
| 1.4.2    | Depletion of abundant proteins followed by fractionation of intact proteins   | 19       |
| 1.4.3    | Comparing technology platforms  | 22       |
| 1.4.4    | Alternative search algorithms for peptide and protein identification  | 23       |
| 1.4.5    | Independent analyses of raw spectra or peaklists  | 24       |
| 1.4.6    | Comparisons with published reports  | 25       |
| 1.4.7    | Direct MS (SELDI) analyses  | 27       |
| 1.4.8    | Annotation of the HUPO PPP core dataset(s)  | 27       |
| 1.4.9    | Identification of novel peptides using whole genome ORF search  | 30       |
| 1.4.10   | Identification of microbial proteins in the circulation   | 30       |

1.5 Discussion 31  
 1.6 References 33

**2 Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project 37**

*Marcin Adamski, Thomas Blackwell, Rajasree Menon, Lennart Martens, Henning Hermjakob, Chris Taylor, Gilbert S. Omenn and David J. States*

2.1 Introduction 37  
 2.2 Materials and methods 39  
 2.2.1 Development of the data model 39  
 2.2.1.1 Laboratory 39  
 2.2.1.2 Experimental protocol 39  
 2.2.1.3 Protein identification data set 39  
 2.2.1.4 Peak list 41  
 2.2.1.5 Summary of technologies and resources 41  
 2.2.1.6 MS/MS spectra 41  
 2.2.1.7 SELDI peak list 42  
 2.2.2 Data submission process 42  
 2.2.3 Design of the data repository 42  
 2.2.4 Receipt of the data 43  
 2.3 Inference from peptide level to protein level 44  
 2.4 Summary of contributed data 46  
 2.4.1 Cross-laboratory comparison, confidence of the identifications 49  
 2.5 False-positive identifications 51  
 2.6 Data dissemination 56  
 2.7 Discussion 57  
 2.8 Concluding remarks 58  
 2.9 Computer technologies applied 60  
 2.10 References 61

**3 HUPO Plasma Proteome Project specimen collection and handling: Towards the standardization of parameters for plasma proteome samples 63**

*Alex J. Rai, Craig A. Gelfand, Bruce C. Haywood, David J. Warunek, Jizu Yi, Mark D. Schuchard, Richard J. Mehigh, Steven L. Cockrill, Graham B. I. Scott, Harald Tammen, Peter Schulz-Knappe, David W. Speicher, Frank Vitzthum, Brian B. Haab, Gerard Siest and Daniel W. Chan*

3.1 Introduction 63  
 3.2 Materials and methods 65  
 3.2.1 HUPO reference sample collection protocol 65  
 3.2.2 Differential peptide display 66  
 3.2.3 Stability studies and SELDI analysis 66  
 3.2.4 SDS-PAGE analysis for stability studies 67



|          |  |           |
|----------|--|-----------|
| 3.2.5    | 2-DE for stability studies   | 67        |
| 3.2.6    | SELDI-TOF analysis for protease inhibitor studies  | 67        |
| 3.2.7    | 2-DE for plasma protease inhibition studies  | 68        |
| 3.2.8    | Tryptic digestion and protein identification for protease inhibition studies   | 69        |
| 3.2.9    | Antibody microarray analysis using two-color rolling circle amplification  | 69        |
| 3.3      | Results  | 69        |
| 3.3.1    | Comparisons of specimen types  | 71        |
| 3.3.1.1  | Analysis of serum  | 71        |
| 3.3.1.2  | Analysis of plasma   | 71        |
| 3.3.2    | Evaluation of storage and handling conditions  | 71        |
| 3.3.3    | Evaluations of the use of protease inhibitors  | 73        |
| 3.3.3.1  | Analysis with SELDI-TOF MS of “time zero” effects of protease inhibitors in plasma   | 73        |
| 3.3.3.2  | Analysis by 2-DE   | 73        |
| 3.3.3.3  | Analysis with antibody arrays  | 76        |
| 3.4      | Discussion   | 77        |
| 3.4.1    | Other pre-analytical variables and control considerations  | 83        |
| 3.4.2    | Reference materials  | 84        |
| 3.5      | Concluding remarks   | 87        |
| 3.6      | References   | 88        |
| <b>4</b> | <b>Immunoassay and antibody microarray analysis of the HUPO Plasma Proteome Project reference specimens: Systematic variation between sample types and calibration of mass spectrometry data</b>   | <b>91</b> |
|          | <i>Brian B. Haab, Bernhard H. Geierstanger, George Michailidis, Frank Vitzthum, Sara Forrester, Ryan Okon, Petri Saviranta, Achim Brinker, Martin Sorette, Lorah Perlee, Shubha Suresh, Garry Drwal, Joshua N. Adkins and Gilbert S. Omenn</i> |           |
| 4.1      | Introduction   | 92        |
| 4.2      | Materials and methods  | 93        |
| 4.2.1    | Reference specimens  | 93        |
| 4.2.2    | DB immunoassays  | 93        |
| 4.2.3    | Antibody arrays at GNF   | 94        |
| 4.2.3.1  | Antibodies, reagents, microarray printing, and platform  | 94        |
| 4.2.3.2  | Microarray layout and processing   | 94        |
| 4.2.3.3  | Array imaging and data analysis  | 95        |
| 4.2.4    | Antibody microarrays at MSI  | 95        |
| 4.2.4.1  | Chip manufacture   | 95        |
| 4.2.4.2  | Rolling circle amplification (RCA) immunoassay   | 96        |
| 4.2.4.3  | Conversion of mean fluorescent intensity to concentration  | 96        |

|          |  |            |
|----------|--|------------|
| 4.2.5    | Antibody microarrays at VARI   | 96         |
| 4.2.5.1  | Fabrication of antibody microarrays  | 96         |
| 4.2.5.2  | Serum labeling   | 97         |
| 4.2.5.3  | Processing of antibody microarrays   | 97         |
| 4.2.5.4  | Analysis   | 97         |
| 4.2.6    | Retrieval and matching of IPI numbers for the analytes   | 97         |
| 4.3      | Results  | 98         |
| 4.3.1    | Antibody-based measurements of the HUPO reference specimens  | 98         |
| 4.3.2    | Systematic variation between the preparation methods of the PPP reference specimens                                  | 100        |
| 4.3.3    | Consistent alterations in specific protein abundances  | 107        |
| 4.3.4    | Linkage of MS data and antibody-based measurements   | 108        |
| 4.4      | Discussion   | 110        |
| 4.5      | References   | 113        |
| <b>5</b> | <b>Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma</b> | <b>115</b> |
|          | <i>Lynn A. Echan, Hsin-Yao Tang, Nadeem Ali-Khan, KiBeom Lee and David W. Speicher</i>                               |            |
| 5.1      | Introduction   | 116        |
| 5.2      | Materials and methods  | 117        |
| 5.2.1    | Serum/plasma collection  | 117        |
| 5.2.2    | MARS   | 118        |
| 5.2.3    | Multiple affinity removal spin cartridge   | 118        |
| 5.2.4    | Microscale solution IEF (MicroSol IEF) (ZOOM™-IEF) fractionation   | 118        |
| 5.2.5    | 2-DE   | 119        |
| 5.2.6    | LC-MS/MS   | 119        |
| 5.3      | Results  | 120        |
| 5.3.1    | Depletion of major proteins to enhance detection of lower abundance proteins   | 120        |
| 5.3.2    | Evaluation of high-abundance protein removal using 2-DE  | 121        |
| 5.3.3    | Specificity of major protein depletion   | 123        |
| 5.3.4    | Impact of Top-6 protein depletion on detection of lower abundance proteins using 2-D gels                            | 125        |
| 5.3.5    | Combining Top-6 protein depletion with microSol IEF prefractionation and narrow pH range gels                        | 125        |
| 5.3.6    | Analysis of Top-6 depleted serum and plasma using protein array pixelation   | 128        |
| 5.4      | Discussion   | 130        |
| 5.5      | References   | 134        |

|          |   |
|----------|---|
| <b>6</b> | <b>A novel four-dimensional strategy combining protein and peptide separation methods enables detection of low-abundance proteins in human plasma and serum proteomes</b> 135 |
|          | <i>Hsin-Yao Tang, Nadeem Ali-Khan, Lynn A. Echan, Natasha Levenkova, John J. Rux and David W. Speicher</i>  |
| 6.1      | Introduction 135  |
| 6.2      | Materials and methods 138   |
| 6.2.1    | Materials 138   |
| 6.2.2    | Top six protein depletion 138   |
| 6.2.3    | MicroSol-IEF fractionation 139  |
| 6.2.4    | Protein array pixelation 139  |
| 6.2.5    | LC-ESI-MS/MS methods 140  |
| 6.2.6    | Data analysis 140   |
| 6.3      | Results and discussion 141  |
| 6.3.1    | Protein array pixelation strategy 141   |
| 6.3.2    | Optimization of protein array pixelation 143  |
| 6.3.3    | Total analysis time for protein array pixelation of human plasma proteome 146   |
| 6.3.4    | Systematic protein array pixelation of the human plasma proteome 147  |
| 6.3.5    | Systematic protein array pixelation of the human serum proteome 150   |
| 6.3.6    | Analyses of human plasma and serum proteomes using HUPO filter criteria 153   |
| 6.4      | Concluding remarks 157  |
| 6.5      | References 157  |
| <b>7</b> | <b>A study of glycoproteins in human serum and plasma reference standards (HUPO) using multilectin affinity chromatography coupled with RPLC-MS/MS</b> 159                    |
|          | <i>Ziping Yang, William S. Hancock, Tori Richmond Chew and Leo Bonilla</i>  |
| 7.1      | Introduction 159  |
| 7.2      | Materials and methods 160   |
| 7.2.1    | Materials 160   |
| 7.2.2    | Isolating glycoproteins using multilectin affinity columns 161  |
| 7.2.3    | Analysis of glycoproteins on LC-LCQ MS 161  |
| 7.2.4    | Analysis of glycoproteins on LC-LTQ MS 162  |
| 7.2.5    | Protein database search 162   |
| 7.3      | Results and discussion 162  |
| 7.3.1    | Protein IDs from the plasma and serum samples 162   |
| 7.3.2    | Comparison between serum and plasma glycoproteomes 179  |
| 7.3.3    | Comparison of the glycoproteins present in the samples collected from three ethnic groups 179   |

|          |  |            |
|----------|--|------------|
| 7.4      | Concluding remarks   | 182        |
| 7.5      | References   | 183        |
| <b>8</b> | <b>Evaluation of prefractionation methods as a preparatory step for multidimensional based chromatography of serum proteins</b>  | <b>185</b> |
|          | <i>Eilon Barnea, Raya Sorkin, Tamar Ziv, Ilan Beer and Arie Admon</i>  |            |
| 8.1      | Introduction   | 185        |
| 8.1.1    | The HUPO Plasma Proteome Project (PPP) goals and the serum as a complex sample   | 185        |
| 8.1.2    | The scope of this manuscript   | 187        |
| 8.2      | Materials and methods  | 187        |
| 8.2.1    | Depletion from serum albumin and antibodies  | 187        |
| 8.2.2    | MudPIT and mass segmentation   | 187        |
| 8.2.3    | Protein separation by SDS-PAGE   | 188        |
| 8.2.4    | SCX separation of intact proteins followed by MudPIT   | 188        |
| 8.2.5    | Liquid-phase IEF followed by MudPIT  | 188        |
| 8.2.6    | Capillary RP-LC-MS/MS  | 189        |
| 8.2.7    | MS data processing and peptide/protein identifications   | 189        |
| 8.3      | Results  | 189        |
| 8.3.1    | Comparisons between the prefractionation methods   | 190        |
| 8.3.2    | Identification of different protein subsets  | 191        |
| 8.3.3    | Proteins identified by only one prefractionation method  | 193        |
| 8.3.4    | Different methods resulted in diverse peptide coverage   | 193        |
| 8.4      | Discussion   | 196        |
| 8.4.1    | Giving every peptide a chance  | 196        |
| 8.4.2    | How to identify more of the marginal proteins  | 197        |
| 8.4.3    | Clustering and comparing raw data  | 197        |
| 8.4.4    | High throughput and ruggedness versus high sensitivity   | 197        |
| 8.4.5    | The cost effectiveness of the different methods  | 198        |
| 8.5      | Concluding remarks   | 198        |
| 8.6      | References   | 199        |
| <b>9</b> | <b>Efficient prefractionation of low-abundance proteins in human plasma and construction of a two-dimensional map</b>  | <b>201</b> |
|          | <i>Sang Yun Cho, Eun-Young Lee, Joon Seok Lee, Hye-Young Kim, Jae Myun Park, Min-Seok Kwon, Young-Kew Park, Hyoung-Joo Lee., Min-Jung Kang, Jin Young Kim, Jong Shin Yoo, Sung Jin Park, Jin Won Cho, Hyon-Suk Kim and Young-Ki Paik</i> |            |
| 9.1      | Introduction   | 202        |
| 9.2      | Materials and methods  | 203        |
| 9.2.1    | Plasma sample preparation  | 203        |
| 9.2.2    | Depletion of major abundance proteins with an immunoaffinity column  | 203        |

|           |  |            |
|-----------|--|------------|
| 9.2.3     | 2-DE   | 204        |
| 9.2.4     | Identification of proteins by MS   | 204        |
| 9.2.5     | Fractionation of the plasma samples by FFE   | 204        |
| 9.2.6     | LC-MS/MS   | 205        |
| 9.2.7     | Bioinformatics   | 206        |
| 9.3       | Results and discussion   | 206        |
| 9.3.1     | 2-DE map of human plasma devoid of high-abundance proteins   | 206        |
| 9.3.2     | Expression of different anticoagulant-treated plasma   | 214        |
| 9.3.3     | FFE/1-DE/nanoLC-MS/MS and 2-DE/MALDI-TOF   | 215        |
| 9.4       | Concluding remarks   | 219        |
| 9.5       | References   | 219        |
| <b>10</b> | <b>Comparison of alternative analytical techniques for the characterisation of the human serum proteome in HUPO Plasma Proteome Project</b>  | <b>221</b> |
|           | <i>Xiaohai Li, Yan Gong, Ying Wang, Songfeng Wu, Yun Cai, Ping He, Zhuang Lu, Wantao Ying, Yangjun Zhang, Liyan Jiao, Hongzhi He, Zisen Zhang, Fuchu He, Xiaohang Zhao and Xiaohong Qian</i> |            |
| 10.1      | Introduction   | 222        |
| 10.2      | Materials and methods  | 223        |
| 10.2.1    | Materials  | 223        |
| 10.2.2    | Human serum samples  | 223        |
| 10.2.3    | Integrated strategy for characterising analytical approaches   | 223        |
| 10.2.4    | Depletion of the highly abundant serum proteins by MARS  | 224        |
| 10.2.5    | Desalting and concentrating the flow-through fractions by centrifugal ultrafiltration  | 224        |
| 10.2.6    | Fractionation of depleted serum samples by anion-exchange HPLC   | 225        |
| 10.2.7    | Protein fractionation by 2-D HPLC with nonporous RP-HPLC   | 225        |
| 10.2.8    | The 2-DE strategy for the analysis of serum proteins   | 226        |
| 10.2.8.1  | 2-DE   | 226        |
| 10.2.8.2  | In-gel digestion <i>via</i> automated workstation  | 227        |
| 10.2.8.3  | Protein spot identification by MALDI-TOF-MS/MS   | 227        |
| 10.2.9    | Shotgun strategy for the analysis of serum proteins  | 228        |
| 10.2.9.1  | Trypsin digestion of serum proteins  | 228        |
| 10.2.9.2  | Protein identification by micro2-D LC-ESI-MS/MS  | 228        |
| 10.2.9.3  | Data processing  | 229        |
| 10.2.10   | Protein fractionation strategy for the analysis of serum proteins  | 229        |
| 10.2.10.1 | 2-D LC fractionation of serum proteins   | 229        |
| 10.2.10.2 | Digestion of the 2-D LC separated fractions  | 229        |
| 10.2.10.3 | 1-D microRP-HPLC-ESI-MS/MS identification of digested serum proteins   | 230        |
| 10.2.11   | Offline shotgun strategy for the analysis of serum proteins  | 230        |

10.2.11.1 Offline SCX for first-dimension chromatographic separation of peptides 230

10.2.11.2 1-D capillary RP-HPLC/microESI-IT-MS/MS analysis for the SCX-separated peptide fractions 231

10.2.12 Optimised nanoRP-HPLC-nanoESI IT-MS/MS for the reanalysis of offline SCX-separated peptides (offline-nanospray strategy) 231

10.3 Integrated analysis of the whole data sets 231

10.3.1 Protein grouping analysis 231

10.3.2 Sequence clustering 232

10.4 Results and discussion 233

10.4.1 Depletion of the highly abundant serum proteins 233

10.4.2 The 2-DE strategy for the analysis of serum proteins 233

10.4.3 2-D HPLC fractionation for the analysis of serum proteins 234

10.4.4 Shotgun strategy for the analysis of serum proteins with online SCX 237

10.4.5 Shotgun strategy for the analysis of serum proteins with offline SCX 237

10.4.6 Offline SCX shotgun-nanospray strategy for the analysis of serum proteins 239

10.4.7 Comparison of the five strategies for the analysis of the human serum proteome 241

10.5 Concluding remarks 246

10.6 References 246

**11 A proteomic study of the HUPO Plasma Proteome Project's pilot samples using an accurate mass and time tag strategy 249**

*Joshua N. Adkins, Matthew E. Monroe, Kenneth J. Auberry, Yufeng Shen, Jon M. Jacobs, David G. Camp II, Frank Vitzthum, Karin D. Rodland, Richard, C. Zangar, Richard D. Smith and Joel G. Pounds*

11.1 Introduction 250

11.2 Materials and methods 251

11.2.1 Human blood serum and plasma 251

11.2.2 Depletion of Igs and trypsin digestion 252

11.2.3 Peptide cleanup 252

11.2.4 Capillary RP-LC 253

11.2.5 IT-MS 254

11.2.6 SEQUEST identification of peptides 254

11.2.7 Putative mass and time tag database from SEQUEST results 254

11.2.8 FT-ICR-MS 255

11.2.9 cLC-FT-ICR MS data analysis 255

11.2.10 OmniViz cluster and visual analysis 257

11.3 Results 257

11.3.1 PuMT tag database 257

|           |  |            |
|-----------|--|------------|
| 11.3.2    | Summary of peptide/protein identifications by AMT tags   | 258        |
| 11.3.3    | Protein concentration estimates from ion current   | 260        |
| 11.3.4    | Global protein analysis  | 261        |
| 11.4      | Discussion   | 264        |
| 11.4.1    | Application of FT-ICR MS as a proteomic technology bridge  | 264        |
| 11.4.2    | Confidence in any MS-based proteomic approach  | 266        |
| 11.4.3    | Peptide/protein redundancy   | 267        |
| 11.4.4    | Identification sensitivity versus specificity  | 267        |
| 11.4.5    | Throughput and differential analysis   | 269        |
| 11.5      | References   | 270        |
| <b>12</b> | <b>Analysis of Human Proteome Organization Plasma Proteome Project (HUPO PPP) reference specimens using surface enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry: Multi-institution correlation of spectra and identification of biomarkers</b> | <b>273</b> |
|           | <i>Alex J. Rai, Paul M. Stemmer, Zhen Zhang, Bao-ling Adam, William T. Morgan, Rebecca E. Caffrey, Vladimir N. Podust, Manisha Patel, Lih-yin Lim, Natalia V. Shipulina, Daniel W. Chan, O. John Semmes and Hon-chiu Eastwood Leung</i>  |            |
| 12.1      | Introduction   | 273        |
| 12.2      | Materials and methods  | 275        |
| 12.2.1    | Sample preparation   | 275        |
| 12.2.2    | Sample preprocessing   | 275        |
| 12.2.3    | Target (CM10) chip preparation and sample incubation   | 275        |
| 12.2.4    | Scanning protocol  | 276        |
| 12.2.5    | Data processing  | 276        |
| 12.2.6    | Bioinformatics analysis of data and correlation coefficient matrix   | 276        |
| 12.2.7    | Protein purification, SDS-PAGE analysis, and extraction of proteins  | 276        |
| 12.2.8    | Peptide mass fingerprinting (PMF)  | 277        |
| 12.2.9    | MS/MS analysis   | 277        |
| 12.2.10   | Western blot analysis  | 277        |
| 12.3      | Results  | 278        |
| 12.4      | Discussion   | 283        |
| 12.5      | References   | 286        |
| <b>13</b> | <b>An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis</b>  | <b>289</b> |
|           | <i>Eugene A. Kapp, Frédéric Schütz, Lisa M. Connolly, John A. Chakel, Jose E. Meza, Christine A. Miller, David Fenyo, Jimmy K. Eng, Joshua N. Adkins, Gilbert S. Omenn and Richard J. Simpson</i>  |            |
| 13.1      | Introduction   | 289        |
| 13.1.1    | Heuristic algorithms   | 291        |

|           |   |            |
|-----------|---|------------|
| 13.1.2    | Probabilistic algorithms  | 292        |
| 13.2      | Materials and methods   | 292        |
| 13.2.1    | HUPO-PPP reference specimens  | 292        |
| 13.2.2    | Sample preparation and MS analysis  | 293        |
| 13.2.3    | Protein sequence databases  | 293        |
| 13.2.4    | MS/MS database search strategy  | 293        |
| 13.2.4.1  | SEQUEST and MASCOT workflow performed by the JPSL research group  | 294        |
| 13.2.4.2  | SEQUEST and PeptideProphet workflow performed by the ISB research group   | 294        |
| 13.2.4.3  | Spectrum Mill workflow performed by the Agilent group   | 295        |
| 13.2.4.4  | Sonar and X!Tandem workflow performed by David Fenyo  | 295        |
| 13.2.5    | Web interface for data validation, integration, and cross annotation  | 295        |
| 13.2.6    | ROC curve generation  | 297        |
| 13.3      | Results and discussion  | 298        |
| 13.3.1    | Comparison of MS/MS search algorithms   | 299        |
| 13.3.1.1  | Sensitivity and concordance between MS/MS search algorithms   | 299        |
| 13.3.1.2  | Specificity and discriminatory power of the primary score statistic for the different MS/MS search algorithms: Distribution of scores and ROC plots               | 301        |
| 13.3.1.3  | Calculation of score thresholds based on specified FP identification error rates  | 304        |
| 13.3.1.4  | Benchmarking of the different MS/MS search algorithms at 1% FP error rate   | 310        |
| 13.3.1.5  | Effect of database size and search strategy   | 311        |
| 13.3.1.6  | Utility of reversed sequence searches   | 311        |
| 13.3.1.7  | Consensus scoring between MS/MS search algorithms   | 312        |
| 13.4      | Concluding remarks  | 313        |
| 13.5      | References  | 314        |
| <b>14</b> | <b>Human Plasma PeptideAtlas</b>  | <b>317</b> |
|           | <i>Eric W. Deutsch, Jimmy K. Eng, Hui Zhang, Nichole L. King, Alexey I. Nesvizhskii, Biaoyang Lin, Hookeun Lee, Eugene C. Yi, Reto Ossola and Ruedi Aebersold</i> |            |
| 14.1      | References  | 322        |
| <b>15</b> | <b>Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories</b>  | <b>323</b> |
|           | <i>Lennart Martens, Alexey I. Nesvizhskii, Henning Hermjakob, Marcin Adamski, Gilbert S. Omenn, Joël Vandekerckhove and Kris Gevaert</i>                          |            |
| 15.1      | References  | 328        |



|           |   |            |
|-----------|---|------------|
| <b>16</b> | <b>A functional annotation of subproteomes in human plasma</b>  | <b>329</b> |
|           | <i>Peipei Ping, Thomas M. Vondriska, Chad J. Creighton, TKB Gandhi, Ziping Yang, Rajasree Menon, Min-Seok Kwon, Sang Yun Cho, Garry Drwal, Markus Kellmann, Suraj Peri, Shubha Suresh, Mads Gronborg, Henrik Molina, Raghohama Chaerkady, B. Rekha, Arun S. Shet, Robert E. Gerszten, Haifeng Wu,, Mark Raftery, Valerie Wasinger, Peter Schulz-Knappe, Samir M. Hanash, Young-ki Paik, William S. Hancock, David J. States, Gilbert S. Omenn and Akhilesh Pandey</i> |            |
| 16.1      | Introduction  | 330        |
| 16.2      | Materials and methods   | 330        |
| 16.2.1    | Coagulation pathway and protein interaction network analysis  | 331        |
| 16.2.2    | Gene ontology annotations   | 331        |
| 16.2.3    | Analysis of MS-derived data for identification of proteolytic events and post-translational modifications   | 331        |
| 16.3      | Results and discussion  | 331        |
| 16.3.1    | Bioinformatic analyses of the functional subproteomes   | 332        |
| 16.3.1.1  | An interaction map of human plasma proteins   | 332        |
| 16.3.1.2  | Gene Ontology annotation of protein function  | 334        |
| 16.3.2    | Proteins involved in the blood coagulation pathway  | 335        |
| 16.3.3    | Proteins potentially derived from mononuclear phagocytes  | 337        |
| 16.3.4    | Proteins involved in inflammation   | 338        |
| 16.3.5    | Analyzing the peptide subproteome of human plasma   | 339        |
| 16.3.6    | Liver related plasma proteins   | 339        |
| 16.3.7    | Cardiovascular system related plasma proteins   | 341        |
| 16.3.8    | Glycoproteins   | 342        |
| 16.3.9    | DNA-binding proteins  | 342        |
| 16.3.9.1  | Histones  | 343        |
| 16.3.9.2  | Helicases   | 344        |
| 16.3.9.3  | Zinc finger proteins  | 345        |
| 16.3.10   | Annotation through reanalysis of mass spectrometry data   | 345        |
| 16.3.10.1 | Cleavage of signal peptides and transmembrane domains   | 346        |
| 16.3.10.2 | Identification of PTMs  | 347        |
| 16.4      | Concluding remarks  | 348        |
| 16.5      | References  | 349        |
| <b>17</b> | <b>Cardiovascular-related proteins identified in human plasma by the HUPO Plasma Proteome Project Pilot Phase</b>   | <b>353</b> |
|           | <i>Beniam T. Berhane, Chenggong Zong, David A. Liem, Aaron Huang, Steven Le, Ricky D. Edmondson, Richard C. Jones, Xin Qiao, Julian P. Whitelegge, Peipei Ping and Thomas M. Vondriska</i>  |            |
| 17.1      | Introduction  | 353        |
| 17.1.1    | HUPO Plasma Proteome Project pilot phase  | 354        |

|        |  |     |
|--------|--|-----|
| 17.1.2 | Need for novel insights into cardiovascular disease        | 354 |
| 17.2   | Materials and methods                                      | 355 |
| 17.3   | Groups of cardiovascular-related proteins                  | 356 |
| 17.3.1 | Markers of inflammation and CVD                            | 356 |
| 17.3.2 | Vascular and coagulation proteins                          | 357 |
| 17.3.3 | Signaling proteins   | 359 |
| 17.3.4 | Growth- and differentiation-associated proteins            | 360 |
| 17.3.5 | Cytoskeletal proteins                                      | 360 |
| 17.3.6 | Transcription factors                                      | 361 |
| 17.3.7 | Channel and receptor proteins                              | 363 |
| 17.3.8 | Heart failure- and remodeling-related proteins             | 364 |
| 17.4   | Functional analyses and implications                       | 365 |
| 17.4.1 | Organ specific cardiovascular-related proteins in plasma   | 365 |
| 17.4.2 | Novel cardiovascular-related proteins identified in plasma | 366 |
| 17.5   | Methodology considerations                                 | 368 |
| 17.6   | Conclusions and future directions                          | 368 |
| 17.7   | References   | 370 |

## 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 of 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.

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 sub-proteomes 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](http://www.ebi.ac.uk/pride)); the University of Michigan ([www.bioinformatics.med.umich.edu/hupo/ppp](http://www.bioinformatics.med.umich.edu/hupo/ppp)); and the Institute for Systems Biology ([www.peptideatlas.org](http://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>.



## List of Contributors

***Dr. Gilbert S. Omenn***

Internal Medicine,  
University of Michigan, MSRB 1,  
1150 W. Medical Center Dr.  
Ann Arbor,  
MI 48109-0656, USA

***Dr. David J. States***

University of Michigan,  
2017 Palmer Commons,  
100 Washtenaw Avenue,  
Ann Arbor,  
MI 48109-2218, USA

***Dr. Alex J. Rai***

Assistant Professor and Director  
of General Chemistry,  
The Johns Hopkins University  
School of Medicine,  
Department of Pathology,  
600 N. Wolfe St.,  
Meyer B-121, Baltimore,  
MD 21287-7065, USA

***Brian B. Haab***

Ph.D.,  
The Van Andel Research Institute,  
333 Bostwick NE,  
Grand Rapids,  
MI 49503, USA

***Dr. David W. Speicher***

The Wistar Institute,  
3601 Spruce St. Rm. 151,  
Philadelphia,  
PA 19104, USA

***Dr. David W. Speicher***

The Wistar Institute,  
3601 Spruce Street,  
Philadelphia,  
PA 19104, USA

***Professor William S. Hancock***

Barnett Institute and Department  
of Chemistry and Chemical Biology,  
Northeastern University,  
Boston,  
MA 02115, USA

***Professor Arie Admon***

Department of Biology,  
Technion,  
Haifa 3200, Israel

***Professor Young-Ki Paik***

Yonsei Proteome Research Center  
and Biomedical Proteome  
Research Center,  
Yonsei University,  
134 Shinchon-dong,  
Sudaemoon-ku,  
Seoul 120-749, Korea

***Dr. Xiaohong Qian***

Ph.D.,  
Beijing Institute of Radiation  
Medicine,  
27 Taiping Road,  
Beijing 100850,  
China

***Dr. Joel G. Pounds***

Biological Sciences Division,  
Pacific Northwest National  
Laboratory,  
Box 999 MSIN: P7-58,  
Richland,  
WA 99352, USA

***Dr. Alex J. Rai***

Assistant Professor and Director  
of General Chemistry,  
Department of Pathology,  
Division of Clinical Chemistry,  
Johns Hopkins University  
School of Medicine,  
600 N. Wolfe St., Meyer B-121,  
Baltimore,  
MD 21287-7065, USA

***Professor Richard J. Simpson***

Joint ProteomicS Laboratory,  
Ludwig Institute for Cancer  
Research,  
P.O. Box 2008,  
Royal Melbourne Hospital,  
Parkville, Victoria 3050,  
Australia

***Dr. Eric W. Deutsch***

Institute for Systems Biology,  
1441 N 34th Street,  
Seattle,  
WA 98103, USA

***Dr. Lennart Martens***

Department of Biochemistry,  
Faculty of Medicine and Health  
Sciences,  
Ghent University,  
A. Baertsoenkaai 3,  
B-9000 Ghent, Belgium

***Dr. Akhilesh Pandey***

McKusick-Nathans Institute of  
Genetic Medicine,  
733 N. Broadway, BRB 569,  
Johns Hopkins University,  
Baltimore,  
MD 21205, USA

***Thomas M. Vondriska***

Cardiovascular Research  
Laboratories,  
Departments of Physiology  
and Medicine,  
Division of Cardiology,  
David Geffen School of Medicine  
at UCLA,  
Room 1619, MRL Building,  
Los Angeles,  
CA 90095, USA



## 1

## Overview of the HUPO Plasma Proteome Project: Results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database\*

*Gilbert S. Omenn, David J. States, Marcin Adamski, Thomas W. Blackwell, Rajasree Menon, Henning Hermjakob, Rolf Apweiler, Brian B. Haab, Richard J. Simpson, James S. Eddes, Eugene A. Kapp, Robert L. Moritz, Daniel W. Chan, Alex J. Rai, Arie Admon, Ruedi Aebersold, Jimmy Eng, William S. Hancock, Stanley A. Hefta, Helmut Meyer, Young-Ki Paik, Jong-Shin Yoo, Peipei Ping, Joel Pounds, Joshua Adkins, Xiaohong Qian, Rong Wang, Valerie Wasinger, Chi Yue Wu, Xiaohang Zhao, Rong Zeng, Alexander Archakov, Akira Tsugita, Ilan Beer, Akhilesh Pandey, Michael Pisano, Philip Andrews, Harald Tammen, David W. Speicher and Samir M. Hanash*

HUPO initiated the Plasma Proteome Project (PPP) in 2002. Its pilot phase has (1) evaluated advantages and limitations of many depletion, fractionation, and MS technology platforms; (2) compared PPP reference specimens of human serum and EDTA, heparin, and citrate-anti-coagulated plasma; and (3) created a publicly-available knowledge base ([www.bioinformatics.med.umich.edu/hupo/ppp](http://www.bioinformatics.med.umich.edu/hupo/ppp); [www.ebi.ac.uk/pride](http://www.ebi.ac.uk/pride)). Thirty-five participating laboratories in 13 countries submitted datasets. Working groups addressed (a) specimen stability and protein concentrations; (b) protein identifications from 18 MS/MS datasets; (c) independent analyses from raw MS-MS spectra; (d) search engine performance, subproteome analyses, and biological insights; (e) antibody arrays; and (f) direct MS/SELDI analyses. MS-MS datasets had 15 710 different International Protein Index (IPI) protein IDs; our integration algorithm applied to multiple matches of peptide sequences yielded 9504 IPI proteins identified with one or more peptides and 3020 proteins identified with two or more peptides (the Core Dataset). These proteins have been characterized with Gene Ontology, InterPro, Novartis Atlas, OMIM, and immunoassay-based concentration determinations. The database permits examination of many other subsets, such as 1274 proteins identified with three or more peptides. Reverse protein to DNA matching identified proteins for 118 previously unidentified ORFs.

\* Originally published in *Proteomics* 2005, 13, 3226–3245

We recommend use of plasma instead of serum, with EDTA (or citrate) for anti-coagulation. To improve resolution, sensitivity and reproducibility of peptide identifications and protein matches, we recommend combinations of depletion, fractionation, and MS/MS technologies, with explicit criteria for evaluation of spectra, use of search algorithms, and integration of homologous protein matches.

This Special Issue of *PROTEOMICS* presents papers integral to the collaborative analysis plus many reports of supplementary work on various aspects of the PPP workplan. These PPP results on complexity, dynamic range, incomplete sampling, false-positive matches, and integration of diverse datasets for plasma and serum proteins lay a foundation for development and validation of circulating protein biomarkers in health and disease.

## 1.1

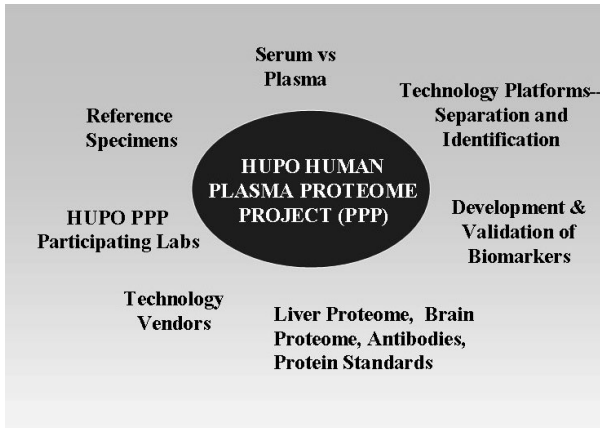
### Introduction

A comprehensive, systematic characterization of circulating proteins in health and disease will greatly facilitate development of biomarkers for prevention, diagnosis, and therapy of cancers and other diseases [1]. Proteomics technologies now permit extensive fractionation of proteins in complex specimens, analysis of peptides by MS, and matching of peptide sequences to protein “hits” through gene and protein databases generated directly and indirectly from the sequencing of the human genome [2, 3], as well as other methods for identifying proteins.

The HUPO, formed in 2001, aims to accelerate the development of the field of proteomics and to stimulate and organize international collaborations in research and education [4]. HUPO has launched major initiatives focused on the plasma, liver, and brain proteomes, proteomics standards and databases, and large-scale antibody production. The plasma proteome is linked with these other initiatives (see Fig. 1).

The long-term scientific goals of the HUPO Plasma Proteome Project (PPP) are (1) comprehensive analysis of the protein constituents of human plasma and serum; (2) identification of biological sources of variation within individuals over time due to physiology (age, sex, menstrual cycle, exercise, stress), pathology (various diseases, special cohorts), and treatments (common medications); and (3) determination of the extent of variation across individuals within populations and across populations due to genetic, nutritional and other factors. The pilot phase aims to (1) compare advantages and limitations of many technology platforms; (2) contrast reference specimens of human plasma (EDTA, heparin, or citrate-anticoagulated) and serum in terms of numbers of proteins identified and any interferences with various technology platforms; and (3) create a global, open-source knowledge base/data repository.

The collaborative nature of this Project permitted exploration of many variables and adoption during the study phase of emerging technologies. Planning proceeded expeditiously from the organizing meeting of HUPO in Bethesda in



**Fig. 1** Schema showing relationship of HUPO Plasma Proteome Project (PPP) to other HUPO initiatives and components of the PPP.

April 2002, to the first PPP meeting in Ann Arbor in September 2002, the expression of interest by numerous investigators at the 1st HUPO World Congress on Proteomics in Versailles in November 2002, and then the PPP Workshop for Technical Committees and participating laboratories in Bethesda in July 2003 to launch the pilot phase. PPP reference specimens were prepared and distributed, beginning in September 2003, and first data were submitted, analyzed, and presented at a workshop at the 2nd HUPO World Congress in Montreal in November 2003. An intensive 4 day Jamboree Workshop was organized for Ann Arbor in June 2004, at which numerous work groups pursued cross-laboratory analyses and proposed further work. Investigators were advised to adopt more stringent criteria for high confidence peptide and protein identifications, and a commitment was made to collect raw spectra from the 18 laboratories that had submitted MS/MS or FT-ICR/MS datasets for independent analyses by three different groups. The datasets were moving targets, as some, but not all, labs submitted expanded or updated analyses, and about 15 laboratories completed “special projects” stimulated by HUPO PPP with a competition for small grants following the Montreal workshop.

The PPP provided participating laboratories with 1.0 mL of reference specimens of serum and plasma by three different methods of anticoagulation for plasma (EDTA, citrate, heparin) from specific donor pools. Investigators utilized their established and emerging technologies for fractionation and analysis of proteins. Investigators were encouraged to “push the limits” of their methods to detect and identify low abundance proteins. Comparisons of findings across laboratories provide a special opportunity for confirmation of protein identifications. Results were submitted to centralized bioinformatics functions at the University of Michigan and the European Bioinformatics Institute to create an integrated data repository from which PPP and other investigators could initiate further analyses and annotations. The approaches and core results have been presented at the US HUPO inaugural meeting in March 2005, the HUPO World Congress in Munich in August 2005, and at other meetings.

Here we present a comprehensive account of the major findings from the pilot phase of the Human Plasma Proteome Project, including the many associated special projects.

## 1.2

### PPP reference specimens

The primary specimens were sets of four reference specimens prepared under the direction of the HUPO PPP Specimens Committee by BD Diagnostics for each of three ethnic groups: Caucasian-American (B1), African-American (B2), and Asian-American (B3). Each pool consisted of 400 mL of blood each from one male and one post-menopausal female healthy, fasting donor, collected into 10 mL tubes in a prescribed sequence (see Supplementary Protocol) after informed consent. Very large pools were rejected as requiring too prolonged specimen handling and processing unlike the collection of individual specimens; even a protocol for two males and two females proved to require more than the 2 h limit we set. Equal numbers of tubes and aliquots were generated with appropriate concentrations of K<sub>2</sub>-EDTA, lithium heparin, or sodium citrate for plasma or permitted to clot at room temperature for 30 min to yield serum (with micronized silica as clot activator). The additives were dry-sprayed on the inner walls of the tubes, except for 1.0 mL of 0.105 M buffered sodium citrate, which gave a final ratio of 9:1 for blood to citrate in a 10 mL final volume, causing an 11% dilution of the blood. No protease inhibitor cocktails were used. This procedure required 2 h, mostly at 2 to 6°C. After centrifugation, volumes from the male and female donors in each donor pair for each specimen type were pooled and then aliquoted into numerous 250 µL portions in vials which were frozen and stored at -70°C. The centrifugation conditions with citrate consistently produced platelet-poor plasma (platelet count <10<sup>3</sup>/µL). Aliquots tested negative for HIV, HBV, HCV, HTLV-1, and syphilis. We supplied four × 250 µL aliquots for each of the four plasma/serum specimens in each set. These vials were shipped on dry ice *via* courier in early May 2003 (and later to additional laboratories which petitioned to join the project, some of which could no longer be supplied the B1 set). No reshipping was permitted.

The Chinese Academy of Medical Sciences (CAMS) used a variant of the BD protocol to generate similar reference serum and plasma specimens, as described by Li *et al.* [5] and He *et al.* [6]. Pools were prepared after review by the CAMS Ethics Committee and informed consent by ten male and ten female donors in Beijing. Donors were fasting and avoided taking medicines or drinking alcohol for the 12 h before sampling. A subsequent pooling of 20 mL from each of the male and female serum or plasma specimens created the C1-CAMS PPP reference specimens which were sent to the 15 laboratories requesting these specimens after storage at -80°C. They were shipped on dry ice using the same courier in September 2003. C1-CAMS specimens were centrifuged originally, and then again upon thawing, at 4°C [6].

Finally, the UK National Institute of Biological Standards and Control (NIBSC) made available to the PPP their lyophilized citrated plasma standard prepared for hemostasis and thrombosis studies from a pool of 25 donors [1].

A standard questionnaire was sent to all laboratories expressing interest. Of 55 laboratories that originally committed to participate, 41 received the BD B1 specimens, 27 the B2 and B3 specimens, 15 the CAMS specimens, and 45 the NIBSC specimens. Laboratories varied on how many of the specimens they actually analyzed.

### 1.3

#### Bioinformatics and technology platforms

As intended, laboratories used a wide variety of methods, including multiple LC-MS/MS instruments, MALDI-MS, and FT-ICR-MS; depletion of abundant proteins; fractionation of intact proteins on 2-D gels or with LC or IEF methods; protein enrichment or labeling methods; immunoassays or antibody arrays; and direct (SELDI) MS. They also varied on choice of search algorithm and database, and criteria for declaring high or lower confidence identification of peptide sequences and matching proteins (Tab. 1). In general, the numbers of proteins reported individually by the labs do not have the integration feature which was applied to the whole PPP dataset. In several cases, much more extensive analyses were reported. Thus, many of the individual papers in this special issue have additional protein identifications not included in the project-wide dataset(s).

#### 1.3.1

##### Constructing a PPP database for human plasma and serum proteins

Data management for this project included guidance and protocols for data collection, then centralized integration, analysis, and dissemination of findings worldwide *via* a communications infrastructure. As described in great detail by Adamski *et al.* [7, 8], key challenges were integration of heterogeneous datasets, reduction of redundant information to minimal identification sets, and data annotation. Multiple factors had to be balanced, including when to “freeze” on a particular release of the ever-changing database selected for the PPP and how to deal with “lower confidence” peptide identifications. Freezing of the database was essential to conduct extensive comparisons of complex datasets and annotations of the dataset as a whole. However, it complicates the work of linking findings of the current study to evolving knowledge of the human genome and its annotation. Many of the entries in the protein sequence database(s) available at the initiation of the project or even the analytical phase were revised, replaced, or withdrawn over the course of the project, and continue to be revised. Our policies and practices anticipated the guidelines issued recently by Carr *et al.* [9], as documented by Adamski *et al.* [7].

The 18 participating laboratories using MS/MS or FT-ICR-MS submitted a total of 42 306 protein identifications using various search engines and databases to handle spectra and generate peptide sequence lists from the specimens analyzed.

Tab. 1 Protein identifications by lab, by specimen, and by methods

| Lab ID | Specimen | Depletion | Protein separation    | Reduction/alkylation | Peptide separation | Mass spectrum                    | Search software | 3020 High confidence | 3020 Lower confidence | Single peptide |
|--------|----------|-----------|-----------------------|----------------------|--------------------|----------------------------------|-----------------|----------------------|-----------------------|----------------|
| 1      | b1-cit   | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 61                   | 39                    | 12             |
| 1      | b1-edta  | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 35                   | 30                    | 14             |
| 1      | b1-hep   | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 50                   | 38                    | 13             |
| 1      | b1-serum | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 21                   | 6                     | 5              |
| 1      | b2-cit   | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 57                   | 37                    | 12             |
| 1      | b2-hep   | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 58                   | 30                    | 12             |
| 1      | b2-serum | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 59                   | 31                    | 12             |
| 1      | b3-serum | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_decasp                 | PepMiner        | 17                   | 6                     | 7              |
| 2      | b1-cit   | none      | cho affinity          | iam                  | scx/rp             | esi-ms/ms_qtof                   | SEQUEST         | 165                  | 79                    | 94             |
| 2      | b1-serum | none      | cho affinity          | iam                  | scx/rp             | esi-ms/ms_qtof                   | SEQUEST         | 136                  | 48                    | 38             |
| 2      | nibsc    | none      | cho affinity          | iam                  | scx/rp             | esi-ms/ms_qtof                   | SEQUEST         | 171                  | 121                   | 85             |
| 11     | b1-cit   | none      | cho affinity          | iam                  | rp                 | esi-ms/ms_decasp                 | SEQUEST         | 59                   | 4                     | 9              |
| 11     | b1-edta  | none      | cho affinity          | iam                  | rp                 | esi-ms/ms_decasp                 | SEQUEST         | 64                   | 6                     | 4              |
| 11     | b1-hep   | none      | cho affinity          | iam                  | rp                 | esi-ms/ms_decasp                 | SEQUEST         | 62                   | 9                     | 15             |
| 11     | b1-serum | none      | cho affinity          | iam                  | rp                 | esi-ms/ms_decasp                 | SEQUEST         | 64                   | 3                     | 16             |
| 12     | b1-cit   | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_deca                   | SEQUEST         | 111                  | 0                     | 113            |
| 12     | b1-edta  | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_deca                   | SEQUEST         | 111                  | 0                     | 101            |
| 12     | b1-hep   | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_deca                   | SEQUEST         | 127                  | 0                     | 130            |
| 12     | b1-serum | aig       | none                  | iam                  | rp/scx/rp          | esi-ms/ms_deca                   | SEQUEST         | 123                  | 0                     | 111            |
| 17     | b1-serum | aig       | 1s sds                | iam                  | rp                 | esi-ms/ms_lcq                    | SEQUEST         | 50                   | 19                    | 7              |
| 21     | b1-cit   | top6      | rotofor-ief/rp/1d-sds | iam                  | rp                 | esi-ms/ms_qtof                   | MASCOT          | 40                   | 0                     | 1              |
| 21     | b1-cit   | top6      | rotofor-ief/rp/1d-sds | none                 | none               | mal-di-ms/ms <sup>ab</sup> i4700 | MASCOT          | 51                   | 0                     | 3              |
| 21     | b1-cit   | top6      | rotofor-ief/rp/1d-sds | none                 | rp                 | esi-ms/ms_qtof                   | MASCOT          | 39                   | 0                     | 1              |
| 21     | b1-edta  | top6      | rotofor-ief/rp/1d-sds | iam                  | rp                 | esi-ms/ms_qtof                   | MASCOT          | 40                   | 0                     | 1              |