Minireview

High-accuracy proteome maps of human body fluids

Alexander Schmidt and Ruedi Aebersold

Address: Institute for Molecular Systems Biology, ETH Zurich, Wolfgang-Pauli-Strasse 16, CH-8093 Zurich, Switzerland.

Correspondence: Ruedi Aebersold. Email: aebersold@imsb.biol.ethz.ch

Published: 28 November 2006

Genome ***Biology*** 2006, **7:**242 (doi:10.1186/gb-2006-7-11-242)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/11/242

Abstract

The proteomes most likely to contain clinically useful disease biomarkers are those of human body fluids. Three recent large-scale proteomic analyses of tears, urine and seminal plasma using the latest mass spectrometric technology will provide useful datasets for biomarker discovery.

Over the past decade, thousands of articles using the term 'proteome' in their title have been published, yet not a single proteome has been comprehensively identified. Each piece of work has typically identified a rather small and biased subset of the proteome under study. With the emergence of methods of quantitative proteomics based on mass spectrometry (MS) [1-4], which have improved both the value of quantitative comparisons and the fraction of the proteome measured, there is an even greater need for comprehensive proteome analyses to use as baseline standards. In studies in which multiple samples are being quantitatively compared, for example in time-course experiments, the whole proteome, or at least a consistent and reproducible subset thereof, needs to be detectable and identifiable in order to avoid an apparent falling-off in the number of different polypeptides measured in successive samples. In addition, the extensive pre-fractionation required to detect low-abundance proteins typically generates ten or more peptide mixtures per sample, each requiring several hours of MS analysis time and creating significant data-analysis overheads. This limits the application of any type of 'shotgun proteomics' approach to high-throughput screening.

Current MS-based proteomic methods sample a limited subset of a proteome in a relatively random manner; this means that neither complete nor reproducibly defined subproteomes are usually analyzed. We have proposed that proteomics research should be divided into two phases - a mapping phase and a scoring phase [5,6]. In the mapping phase, all the proteins and peptides detectable by current technology - ideally all the polypeptides present in a sample - would be confidently identified and the data organized into an easily accessible and searchable database. Initial implementations of such databases include the Global Proteome Machine [7,8] and the Peptide Atlas [9,10]. In the scoring phase, a set of peptides representing the whole proteome, or a consistent subset of particular interest, is identified in the database and measured in samples by one of a number of targeted analytical methods [3,11-13]. The recent publication of three high-quality proteomic analyses of human body fluids - tears, urine and seminal plasma - by Matthias Mann and his colleagues [14-16], along with papers describing large, high-quality datasets of serum [17,18] and yeast [9] proteomes, are significant steps in the mapping phase of this strategy.

Improvements in mass spectrometry

Until recently, the vast majority of proteomic data were collected using ion-trap mass spectrometers, instruments that are extremely robust but have only moderate mass accuracy and resolution. An important consequence of this low mass accuracy is the informatics challenge of assigning peptide sequences to the fragment-ion spectra with high confidence. The recent introduction of mass spectrometers with high mass accuracy has increased the confidence of proteomic results and led to the development of data-collection protocols specifically designed to reduce the likelihood of false sequence assignment [19].

The large-scale analyses of tear-duct fluid, urine and seminal plasma from Mann's group [14-16] were done using the

latest generation of mass spectrometers. First, the complexity of each sample was reduced by protein fractionation, by either one-dimensional gel electrophoresis or reversed-phase chromatography. After tryptic digestion of each fraction, the resulting peptides were analyzed by liquid chromatography followed by tandem mass spectrometry (LC-MS/MS) using two types of high-performance hybrid mass spectrometer - the linear ion trap - Fourier transform mass spectrometer (LTQ-FT) or the linear ion trap-orbitrap mass spectrometer (LTQ-orbitrap).

Interestingly, the overlap of proteins identified from identical samples with different instruments was less than that from repeat analyses in the same instrument, and thus several additional proteins were identified by combining the datasets generated by the two instruments. The difference in peptides identified can be explained by the fact that the two instruments were operated in different cycle modes that correspond to their physical characteristics. The LTQ-FT instrument had a slower peptide-sequencing duty cycle than the LTQ-orbitrap. This was compensated for by the higher mass precision (< 3 ppm) and two consecutive stages of fragmentation (MS^3) that significantly increased confidence in peptide identification [20]. In contrast, the LTQ-orbitrap was set up for higher throughput, providing a larger number of peptide-sequencing cycles per time period with only a slightly lower mass accuracy (< 5 ppm). Using the LTQ-orbitrap, identification of two different peptides was required for confident identification of a protein, whereas the combined MS/MS and MS^3 data of a single peptide identified by the LTQ-FT was considered sufficiently informative to identify a protein with confidence [14-16]. This latter mode of scoring significantly increased the total number of identified proteins, with most of the proteins exclusively detected by the LTQ-FT being identified by a single peptide.

Thus, operating the two instruments in different modes resulted in a reduced number of redundant protein identifications and increased the coverage of the proteome. The rate of false peptide assignments was evaluated by submitting the MS data to a search against a decoy database, in which the protein sequences had been reversed [21], and was found to be very low. The results show that the increased data quality generated by high-performance instruments, compared with the commonly used ion-traps, greatly facilitates the generation of high-confidence datasets.

Proteomics and biomarker discovery

Among samples analyzed by proteomics, blood plasma and other body fluids most clearly illustrate the need for consistent, in-depth and high-throughput analysis, and thus for the implementation of the two-stage proteomic strategy outlined above. Proteomics has raised great expectations for the discovery of biomarkers for improved diagnosis or stratification of a wide range of diseases, including cancers [22]. Blood plasma and other body fluids are expected to be excellent sources of protein biomarkers because they circulate through, or come in contact with, a variety of tissues - with all tissues in the case of plasma. During this contact they are likely to pick up proteins secreted or shed by tissues, a hypothesis that has recently been tested and confirmed [23].

The task of quantitatively analyzing the proteomes of plasma and other body fluids is as daunting as it is attractive, especially if many clinical samples have to be processed in a single study. Human plasma has been termed the most complex human proteome [24] and the large differences in the concentrations of individual proteins, ranging from several milligrams to less than one picogram per milliliter, challenge current MS technology. Another analytical challenge for biomarker discovery arises from the high variability in the concentration and state of modification of some human plasma proteins between different individuals [25]. Therefore, samples from a large number of individuals will have to be analyzed to control for this variability. Despite these limitations, human plasma holds immense diagnostic potential. Recently, several large-scale projects have been initiated, aimed at characterizing the human plasma proteome [9,17]. Although the coverage of the plasma proteome with high-confidence identifications was disappointingly low [18], these publicly available high-confidence datasets provide helpful references for future targeted studies following a proteome-scoring strategy.

As a considerable volume of blood circulates through all organs in humans, it must be expected that proteins secreted or released from a specific tissue or cell type - the proteins that hold the highest potential as biomarkers - will be diluted in plasma to a degree that frequently makes them undetectable with current analytical methods. Interest has, therefore, been focused on the analysis of so-called 'proximal' fluids, which have been in contact with only one or a few tissue types, and for which less dilution of tissue-derived proteins would be expected. Proximal fluids include nipple aspirate, cerebrospinal fluid, bronchial lavage fluid, as well as the urine, seminal plasma and tear fluid that are the subject of the three recent papers from Mann's group [14-16]. These latter studies stand out because the powerful new mass spectrometers have been applied in a consistent manner to all three samples. The results are of excellent quality and have increased the number of proteins identified from the respective samples several fold compared with previous studies, providing unprecedented insight into the complexity of the proteome in these three body fluids. This work, and similar studies that will undoubtedly follow, should provide a rich source of information for the implementation of advanced proteome-scoring strategies.

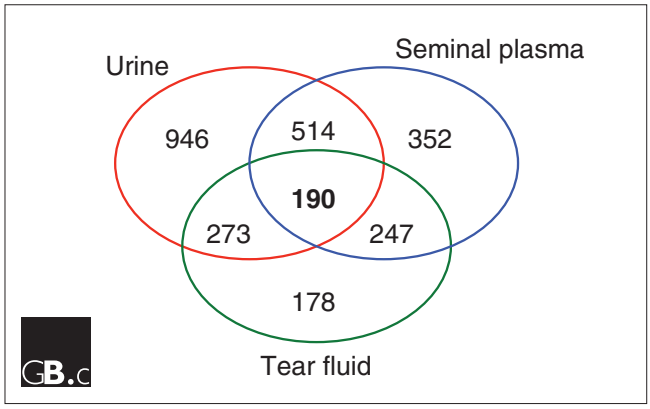

Figure 1
The numbers of proteins identified in urine, seminal plasma and tear fluid. All overlaps of proteins (two-way and three-way) are shown for all three datasets: urine (red), seminal plasma (blue) and tear fluid (green). Numbers represent the number of shared proteins in the respective overlapping and non-overlapping areas.

A comparison of body-fluid proteomes

In spite of considerable effort and the application of state-of-the art MS (as in [14-16]), none of the proteomes analyzed so far can be considered to be completely mapped. Nevertheless, the extensive data collected enable interesting comparisons to be made that will guide the use of the datasets for biomarker discovery. The proteins identified from the different body fluids by Mann's group [14-16] were compared with each other and with a high-quality reference list of peptide sequences already observed by MS in human plasma [14]. The overlaps between the individual studies are shown in Figure 1. Interestingly, more than half of the proteins identified in seminal plasma and in tear fluid were also identified in the urine dataset. The combined dataset contains the impressive number of 2,130 unique protein hits, but only 190 proteins were found in all three studies. The urine proteome was analyzed most extensively; it contained the highest number of exclusive proteins and therefore represents the richest resource for biomarker discovery of the three body fluids discussed here.

A comparison between the urine dataset and the latest version (February 2006) of the public human plasma Peptide Atlas database [9] showed that about two-thirds of the urine proteins had already been detected in human plasma using MS. As expected, most proteins exclusively found in urine have very low concentrations in plasma (215 ng/ml to 11 pg/ml) [26] and were therefore more difficult to identify in this body fluid. For instance, the widely used protein biomarker prostate-specific antigen (PSA; Swiss-Prot accession number: P07288) was not included in the large human plasma dataset, but could be unambiguously detected in urine and in seminal plasma. Proteins exclusively identified in urine include corticotropin-lipotropin (a marker for pituitary tumors; Swiss-Prot: P01189), kallikrein II (a marker for ovarian cancer; Swiss-Prot: Q9UBX7), prostate secretory protein PSP94 (Swiss-Prot: P08118), prostate acid phosphatase (Swiss-Prot: P15309) and pancreatic secretory trypsin inhibitor (TATI, Swiss-Prot: P00995). All these are already in use as clinical markers or are being evaluated as biomarkers for prostate or pancreatic diseases [26].

Looking to the future

The high number of proteins identified in urine, seminal plasma and tear fluid suggests that differences in protein concentrations in these samples are significantly less than in plasma, making these body fluids easier to analyze by MS. Although some proteins exclusively detected in urine were not identified in plasma by MS-based methods, they were detected in plasma by sensitive antibody-based approaches. This underlines the fact that biomarkers discovered in other body fluids can also be screened for in plasma [27]. The major limitation of proximal fluid proteomes over that of plasma is their lack of comprehensiveness, which restricts their biomarker potential to particular diseases. In addition, the limited dynamic range of current MS methods, even those as advanced as the ones used by Mann and colleagues [14-16], suggests that this proteome coverage is still incomplete. New methods will have to be developed to expand the detectable protein concentration range and increase sample throughput.

Nevertheless, the protein datasets provided by Mann and colleagues [14-16] significantly expand the proteome coverage of urine, seminal plasma and tear fluid, and represent very useful high-quality references for future proteome studies, including targeted LC-MS/MS approaches. The datasets represent an important step towards the implementation of two-stage proteomic strategies in biomarker discovery.

Acknowledgements

Our work is supported in part with federal funds from the National Heart, Lung, and Blood Institute, NIH, under contract N01-HV-28179 (to R.A.), by the Swiss National Science Foundation and ETH Zurich.

References

1. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003, **422:**198-207.
2. Flory MR, Griffin TJ, Martin D, Aebersold R: **Advances in quantitative proteomics using stable isotope tags.** *Trends Biotechnol* 2002, **20(12 Suppl):**S23-S29.
3. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** *Nat Biotechnol* 1999, **17:**994-999.
4. Ong SE, Mann M: **Mass spectrometry-based proteomics turns quantitative.** *Nat Chem Biol* 2005, **1:**252-262.
5. Aebersold R: **Constellations in a cellular universe.** *Nature* 2003, **422:**115-116.
6. Kuster B, Schirle M, Mallick P, Aebersold R: **Scoring proteomes with proteotypic peptide probes.** *Nat Rev Mol Cell Biol* 2005, **6:** 577-583.
7. Beavis RC: **Using the global proteome machine for protein identification.** *Methods Mol Biol* 2006, **328:**217-228.

8. **The Global Proteome Machine** [http://www.thegpm.org]
9. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R: **The PeptideAtlas project.** *Nucleic Acids Res* 2006, **34Database issue:**D655-D658.
10. **Peptide Atlas** [http://www.peptideatlas.org]
11. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, *et al.*: **Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.** *Genome Biol* 2005, **6:**R9.
12. Anderson L, Hunter CL: **Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins.** *Mol Cell Proteomics* 2006, **5:**573-588.
13. Domon B, Aebersold R: **Mass spectrometry and protein analysis.** *Science* 2006, **312:**212-217.
14. de Souza GA, Godoy LM, Mann M: **Identification of 491 proteins in the tear fluid proteome reveals a large number of proteases and protease inhibitors.** *Genome Biol* 2006, **7:**R72.
15. Adachi J, Kumar C, Zhang Y, Olsen JV, Mann M: **The human urinary proteome contains more than 1500 proteins including a large proportion of membrane proteins.** *Genome Biol* 2006, **7:**R80.
16. Pilch B, Mann M: **Large-scale and high-confidence proteomic analysis of human seminal plasma**. *Genome Biol* 2006, **7:**R40.
17. Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, *et al.*: **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.** *Proteomics* 2005, **5:**3226-3245.
18. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM: **Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study.** *Nat Biotechnol* 2006, **24:**333-338.
19. Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP: **Optimization and use of peptide mass measurement accuracy in shotgun proteomics.** *Mol Cell Proteomics* 2006, **5:**1326-1337.
20. Olsen JV, Mann M: **Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation.** *Proc Natl Acad Sci USA* 2004, **101:**13417-13422.
21. Moore RE, Young MK, Lee TD: **Qscore: an algorithm for evaluating SEQUEST database search results.** *J Am Soc Mass Spectrom* 2002, **13:**378-386.
22. Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, Radich J, Anderson G, Hartwell L: **The case for early detection.** *Nat Rev Cancer* 2003, **3:**243-252.
23. Zhang H, Liu AY, Loriaux P, Wollscheid B, Zhou Y, Watts JD, Aebersold R: **Mass spectrometric detection of tissue proteins in plasma.** *Mol Cell Proteomics* 2006, doi: 10.1074/mcp.M600255-MCP200.
24. Anderson NL, Polanski M, Pieper R, Gatlin T, Tirumalai RS, Conrads TP, Veenstra TD, Adkins JN, Pounds JG, Fagan R, *et al.*: **The human plasma proteome: a nonredundant list developed by combination of four separate sources.** *Mol Cell Proteomics* 2004, **3:**311-326.
25. Nedelkov D, Kiernan UA, Niederkofler EE, Tubbs KA, Nelson RW: **Investigating diversity in human plasma proteins.** *Proc Natl Acad Sci USA* 2005, **102:**10852-10857.
26. Polanski M, Anderson L: **A list of candidate cancer biomarkers for targeted proteomics.** *Biomarker Insights* 2006, **1:**1-48.
27. Rosty C, Christa L, Kuzdzal S, Baldwin WM, Zahurak ML, Carnot F, Chan DW, Canto M, Lillemoe KD, Cameron JL, *et al.*: **Identification of hepatocarcinoma-intestine-pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology.** *Cancer Res* 2002, **62:**1868-1875.