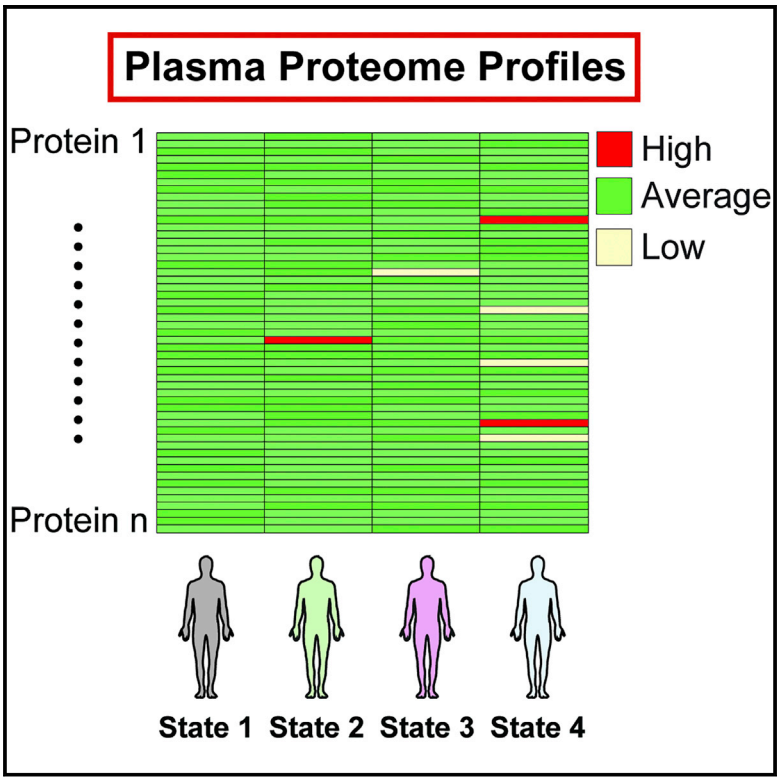
Article

Cell Systems

Plasma Proteome Profiling to Assess Human Health and Disease

Graphical Abstract



Authors

Philipp E. Geyer, Nils A. Kulak, Garwin Pichler, Lesca M. Holdt, Daniel Teupser, Matthias Mann

Correspondence

mmann@biochem.mpg.de

In Brief

A rapid and highly reproducible proteomic workflow delivers a systemic-view proteomic portrait of a person's health state from a single drop of blood.

Highlights

- Automated, highly reproducible, 3-hr proteomic workflow from blood droplet to results
- Plasma protein data reflecting allele differences, metabolic risk, and inflammatory status
- Quantitative 1,000-protein plasma proteome
- The plasma proteome profile as a proteomic portrait of a person's health state

Geyer et al., 2016, Cell Systems 2, 185–195
March 23, 2016
http://dx.doi.org/10.1016/j.cels.2016.02.015

CellPress

Cell Systems
Article

Plasma Proteome Profiling to Assess Human Health and Disease

Philipp E. Geyer,[1,2] Nils A. Kulak,[1] Garwin Pichler,[1] Lesca M. Holdt,[3] Daniel Teupser,[3] and Matthias Mann[1,2,*]
[1]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[3]Institute of Laboratory Medicine, Ludwig-Maximilians University Munich, 80539 Munich, Germany
*Correspondence: mmann@biochem.mpg.de
http://dx.doi.org/10.1016/j.cels.2016.02.015

SUMMARY

Proteins in the circulatory system mirror an individual's physiology. In daily clinical practice, protein levels are generally determined using single-protein immunoassays. High-throughput, quantitative analysis using mass-spectrometry-based proteomics of blood, plasma, and serum would be advantageous but is challenging because of the high dynamic range of protein abundances. Here, we introduce a rapid and robust "plasma proteome profiling" pipeline. This single-run shotgun proteomic workflow does not require protein depletion and enables quantitative analysis of hundreds of plasma proteomes from 1 μl single finger pricks with 20 min gradients. The apolipoprotein family, inflammatory markers such as C-reactive protein, gender-related proteins, and >40 FDA-approved biomarkers are reproducibly quantified (CV <20% with label-free quantification). Furthermore, we functionally interpret a 1,000-protein, quantitative plasma proteome obtained by simple peptide pre-fractionation. Plasma proteome profiling delivers an informative portrait of a person's health state, and we envision its large-scale use in biomedicine.

INTRODUCTION

Blood, plasma, and serum are the predominant samples used for diagnostic analyses in clinical practice and are available in biobanks from thousands of clinical studies (Végvári et al., 2011). The quantitative analysis of individual plasma proteins by immunoassays is used in daily clinical diagnostics. However, immunoassays have inherent limitations with regard to multiplexing, their specificity for protein isoforms, and their incompatibility with hypothesis-free investigations. Mass spectrometry (MS)-based proteomics is a technology that could address all of these limitations and that should be capable of discovering biomarkers in this easily accessible body fluid (Anderson, 2014). However, MS-based plasma proteomics is extremely challenging for a number of reasons, most prominently the extremely large dynamic range of protein abundances (Anderson and Anderson, 2002; Omenn, 2005). There is also a lack of very reproducible, robust, and high-throughput proteomic workflows to identify and verify potential biomarker in large cohorts. As a result, only few novel biomarkers have been established—fewer than 1.5 per year in the 15 years before 2010 (Anderson, 2010)—and this has generally been done by immunoassay-based technologies, such as prostate-specific antigen, one of the best known biomarkers in medicine (Vihko et al., 1978).

Dramatic improvements in the technology of MS-based proteomics over the last few years (Cox and Mann, 2011; Geiger et al., 2010; Muñoz and Heck, 2014) have rekindled an interest in plasma proteomics. Using such technology and combining it with immunodepletion of high- and medium-abundance proteins as well as very extensive peptide fractionation methods, it has now become possible to identify more than 1,000 (Addona et al., 2011; Cao et al., 2012; Paczesny et al., 2010) or even more than 5,000 proteins (Keshishian et al., 2015) in plasma. However, immunodepletion may lead to biases because of cross-reactions of the antibodies used or by proteins bound to carrier proteins such as albumin (Bellei et al., 2011; Tu et al., 2010). Furthermore, extensive pre-fractionation decreases throughput, which is undesirable in clinical practice. Accordingly, the paradigm in biomarker discovery by MS has been to analyze a small number of samples in as much depth as possible, whereas the verification phase was to be done on larger cohorts but with targeted methods and a small number of candidate markers. The final clinical test for a biomarker identified by MS was to be performed with classical immunoassays (Anderson et al., 2009; Surinova et al., 2011). Although this scheme is practical with current technology, it is very laborious and loses much of the promise of systemwide and unbiased investigation of the plasma proteome. Using another approach, Liu et al. (2015) constructed a list of plasma peptide transitions, which they used to interpret the signals in sequential window acquisition of all theoretical MS (SWATH) runs of plasma samples of twins. In this way, the contribution of heritable and environmental changes to the plasma proteome could be distinguished.

In contrast to previous approaches, we here focused on developing a robust and highly streamlined shotgun plasma proteomics workflow. For the MS readout, we used very short liquid chromatography (LC)-MS/MS gradients and recent advances in label-free quantification (Cox et al., 2014). We hypothesized that the resulting "plasma proteome profile" would have a high yield of information about the health state of an individual and that it can be obtained for a large number of clinical samples.

RESULTS

Rapid, Robust, and Highly Reproducible Plasma Proteomic Workflow

Past efforts in shotgun plasma proteomics endeavored to maximize protein identifications, whereas generally less emphasis was placed on quantitative accuracy or throughput. Here we wished to develop a convenient workflow, from sample preparation to data analysis, that can potentially be used in a clinical context. We reasoned that such a workflow should be rapid, optimal for high-throughput, robust, and highly reproducible. Therefore, it should minimize all preparation and analysis steps, while still quantifying clinically interesting proteins accurately. With this in mind, we decided to omit any depletion steps of high-abundance plasma proteins.

Building on the recently described in-StageTip (iST) method (Kulak et al., 2014), we further streamlined the procedure for plasma (Experimental Procedures; Figure 1A). Starting with 1 μl of plasma from a single finger prick, all preparation steps were performed in a single reaction vial. Using ordinary amounts of digestion enzymes, we found that adequate protein digestion had already occurred after 1 hr (protein coefficients of variation [CVs] and tryptic missed cleavage rates were similar to overnight digestion; Table S1). Peptides were then eluted and ready for LC-MS/MS analysis. The entire up-front procedure took less than 2 hr and can readily be performed in a 96-well format and automated in a liquid handling platform, if desired.

Starting with single-run gradient times typical of proteomics experiments, we successively reduced them to determine the maximum information content per unit time. We found that the number of identified proteins decreased very slowly with decreasing time, down to 20 min (only 12 additional identified protein groups in 100 min versus 20 min gradients; Table S1). Below this time, loading and equilibration times become dominant, and therefore we chose 20 min gradients as our standard (33 min between injections, about 50 samples/day). The combination of optimized sample preparation and LC setup allowed for hundreds of plasma proteome analyses, whereas previously clogging of columns was a common occurrence with plasma samples.

We used MaxQuant for quantitative label-free analysis of the LC-MS/MS data (Cox et al., 2014; Cox and Mann, 2008) and for transferring peptide identifications from one LC run to other LC runs in which the peptide was not sequenced (“match between runs”). In combination with a matching library consisting of undepleted plasma of ten different individuals as well as plasma depleted of the 20 highest abundant proteins, this boosted protein identification by 39% (Experimental Procedures; Figures S1A and S1B). Of the 347 protein groups identified in total in the 20 min gradients, 285 were detected in all ten individuals (Figures S1C and S1D). The entire workflow, including the finger-prick procedure and the data analysis, takes less than 3 hr (Figure 1A).

Accuracy of Label-free Quantification of the Plasma Proteome

To investigate the quantitative reproducibility of our workflow (intra-assay variability), we sampled blood by venous puncture from one individual and harvested plasma after centrifugation. We performed the entire workflow 15 separate times on 1 μl aliquots of this stock and correlated protein abundances across the whole measuring range of each of the replicates. The mean R^2 correlation value of the quantified protein signals between individual replicates was excellent at 0.980, with a range of 0.966–0.994 (Figure 1B, excluding keratins; Table S2). We performed 96 blood plasma analyses using multiplexed preparation on a liquid handling robot and short measurement times (5 hr and 51 hr in total, respectively) and achieved a mean R^2 value of 0.97 (Figure S2).

On average, 284 ± 5 different proteins were quantified (total 313): the large majority in all 15 samples and only 3% uniquely in single LC runs (Figure 1C). We picked six well-characterized plasma proteins across a million-fold abundance range and found that quantification was highly reproducible (Figure 1D). We compared different conditions by the proportion of proteins with CVs less than 20%, because this is a commonly used cutoff for in vitro diagnostic assays (U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Veterinary Medicine, 2001). Notably, 67% of quantified proteins were within the 20% cutoff range, and 30% had CVs below 10% (Figure 1E).

To determine the variability caused by LC-MS/MS analysis alone (analytical variability), peptides from one sample preparation were injected and measured 15 times. This resulted in only slightly better reproducibility (71% with a CV less than 20% and 37% with a CV less than 10%), indicating that up-front sample preparation contributed little to overall quantitative variability (Figure 1F). A notable exception to this trend were certain keratin proteins, which had very small analytical variability but sometimes had a large quantitative difference between repeated analysis of the same sample. This is readily explained by contamination with exogenous keratins during sample preparation. Nevertheless, it is clinically relevant, because we found that plasma proteomes of the same person clustered together much better after excluding keratins and other proteins introduced by sample processing such as hemoglobins (see below).

Intra- and Inter-individual Variability of the Plasma Proteome

The high-throughput of our workflow allowed us to extensively characterize the quantitative variation within and between individuals. To determine inter-individual variability, we performed finger pricks on one person four times a day over 8 consecutive days and analyzed all 32 blood proteomes with less than 24 hr of measuring time. This revealed stability of the plasma proteome over time (55% of proteins below 20% CV; Figure 2A). The proteins with large CVs were the aforementioned keratins, as well as high-abundance erythrocyte-specific proteins. The latter are caused by a slightly different extent of erythrocyte lysis during plasma preparation or by contamination of plasma with erythrocytes during plasma harvesting.

To determine inter-individual variability, we harvested plasma from five female and five male donors in triplicate by finger pricks. The average R^2 value within the technical workflow triplicates was 0.976, excluding keratins and erythrocyte-specific protein groups. For CVs of the technical replicates of all individuals, see Table S3. Of 345 proteins quantified, only a minority was under the CV cutoff. This indicates that overall, the plasma

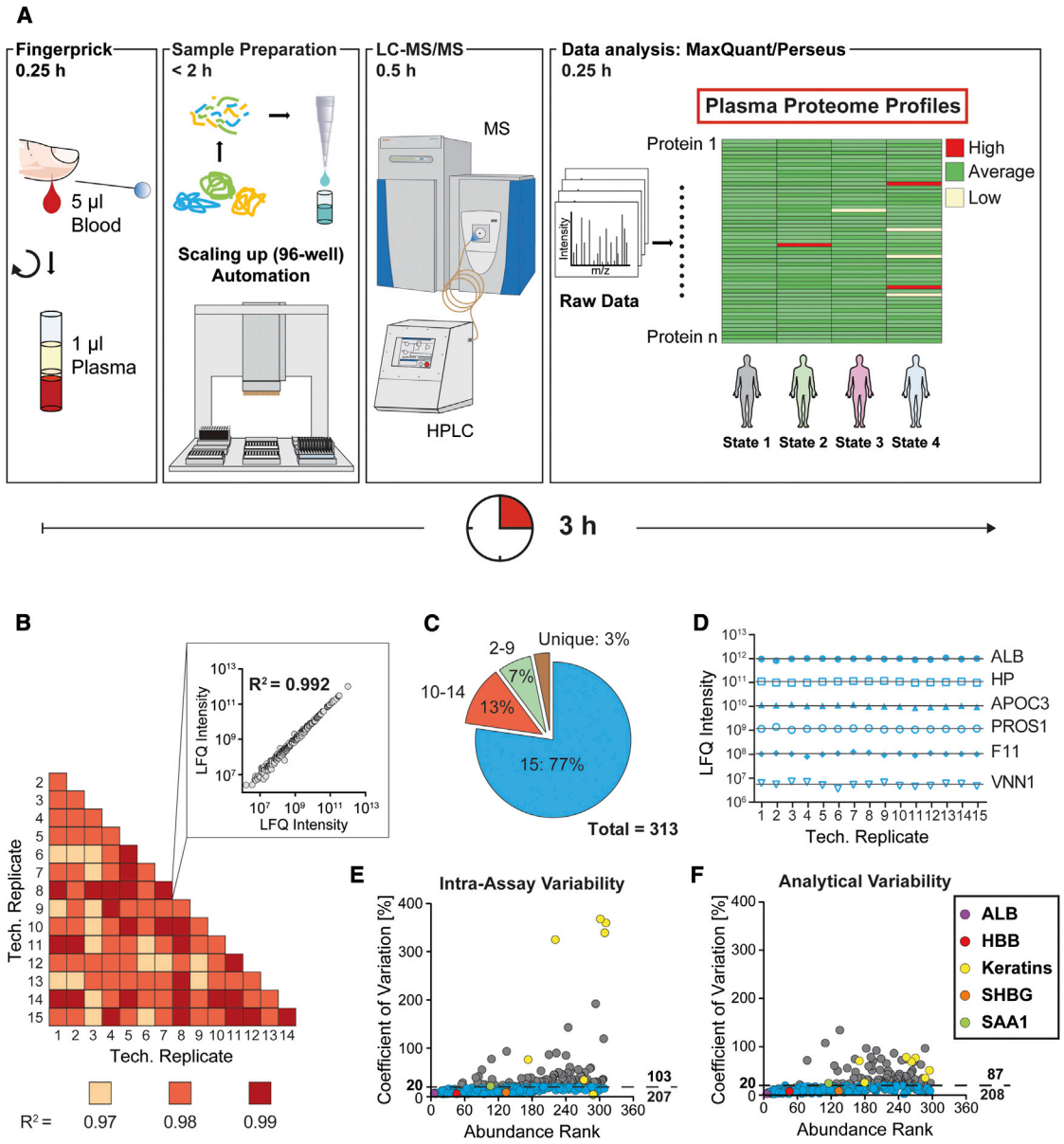


Figure 1. Technological Aspects of Plasma Protein Profiling

(A) Schematic depiction of the workflow. Blood volumes of 5 µl are routinely used to harvest 1 µl of plasma. The workflow is based on the iST protocol and consists of denaturation, reduction, alkylation of cysteines, short 1 hr enzymatic digestion, and purification of peptides. Automation for liquid handling platforms is also indicated. Peptides are separated with optimized short 20 min HPLC gradients and measured online by LC-MS/MS. Data analysis is performed by MaxQuant and Perseus, which deliver information about hundreds of plasma proteins that could reflect an individual's state as symbolized by the plasma proteome profiles.

(B) Color-coded R^2 values for the binary comparison of 15 technical workflow replicates. R^2 values up to 0.994 demonstrate high reproducibility.

(C) Frequency of protein quantification, which was present in all 15 workflow replicates, in 10–14, in 2–9, or only in 1.

(D) Reproducibility of the LFQ intensities of six proteins covering nearly six orders of magnitude for 15 workflow replicates. The line represents the mean values for ALB (serum albumin), HP (haptoglobin), APOC3 (apolipoprotein C-III), PROS1 (vitamin K-dependent protein S), F11 (coagulation factor XI), and VNN1 (pantetheinase).

(E) To determine the intra-assay variability, CVs of all quantified proteins were calculated for the 15 workflow replicates and are plotted according to their abundance. Proteins with CVs < 20% are colored in blue and those with CVs > 20% in gray. HBB, hemoglobin subunit beta. SHBG, sex hormone-binding globulin; SAA1, serum amyloid A-1 protein.

(F) Fifteen repeated injections were used to determine the analytical variability, which includes variability of the LC-MS/MS analysis.

proteome has much higher inter- than intra-individual variability (19% and 55% of proteins within a CV of 20%, respectively; Figure 2B). These general trends have been observed previously (for a recent example, see Liu et al., 2015). Here they suggest that our label-free workflow is well suited to capture the natural or pathological variation of protein levels between individuals.

To directly test this notion, we asked if we could discern systematic differences between the plasma proteomes of women

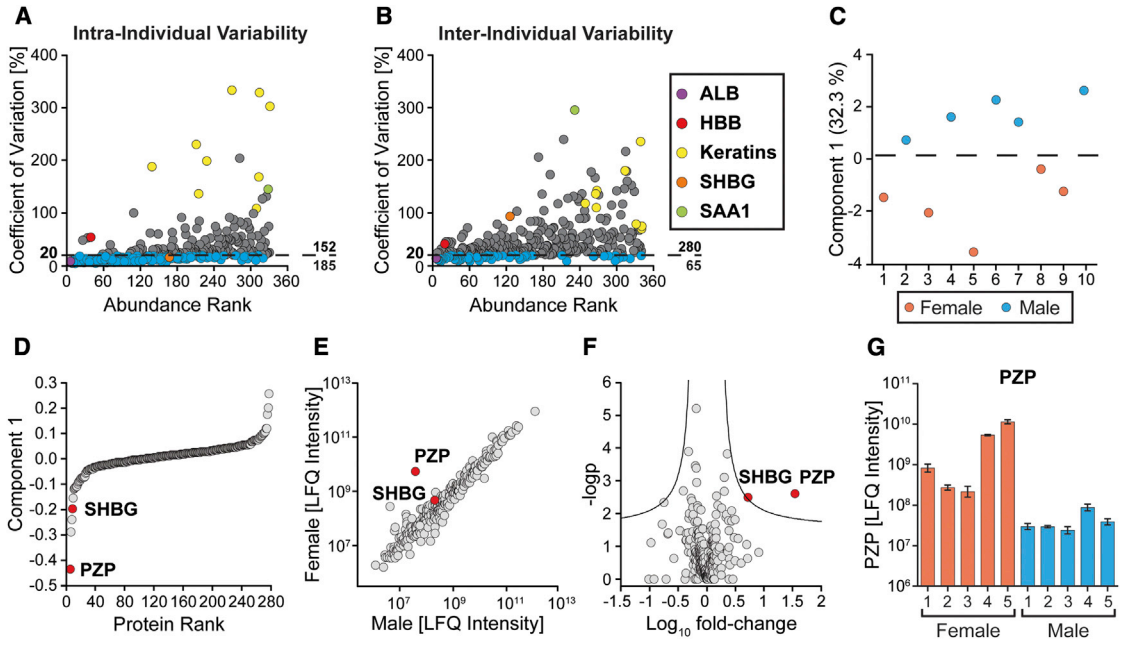


Figure 2. Intra- and Inter-individual Variability of the Plasma Proteome

(A) Intra-individual variability was assessed by finger pricks four times per day for 8 days, and CVs for all quantified proteins were calculated. Proteins with CVs < 20% are colored in blue and those with CVs > 20% in gray.

(B) Inter-individual variation of five women and five men.

(C) Female and male proteomes in one-dimensional PCA.

(D) Proteins and their contribution to the gender separation.

(E) Direct comparison of female subject 4 (F4) and male subject 5 (M5) depicts the extreme difference of PZP between women and men against the background of all other quantified proteins.

(F) Volcano plot of female against male proteomes (x axis, fold change of females to males serving as t test difference; y axis, p value). The black curves show the threshold for statistical significance, where we used a false discovery rate of 5% and an S0 of 0.8.

(G) LFQ intensities for PZP in all ten individuals.

and men, a question that to our knowledge has not been addressed by shotgun proteomics before. Indeed, one-dimensional principal-component analysis (PCA) was already sufficient for complete separation (Figure 2C). Inspection of the drivers of the PCA separation revealed that several of them are known to be regulated by estrogen (Figure 2D) (Christensen et al., 1989; Ottosson et al., 1981; Sand et al., 1985). Direct comparison of the plasma proteome profiles of a woman and a man shows that pregnancy zone protein (PZP) and sex hormone-binding globulin (SHBG) are of high absolute abundance in the plasma proteome of women and can be as high as 1% of human serum albumin (Figure 2E). This suggests a functional role in plasma, and indeed, SHBG binds estrogen, whereas PZP traps proteases (Figure 2F). On average, PZP levels were 33-fold higher in women compared with men. Furthermore, two women had 10- to 100-fold higher levels than the other three, likely because of highly elevated levels of estrogen (Figure 2G).

Rapid Assessment of Sample Quality by Plasma Proteomes

A frequently discussed issue in plasma proteomics as well as in clinical laboratory medicine is the potentially deleterious effects of inconsistent sample handling, such as variable time between blood taking and workup. We reasoned that our rapid and highly reproducible workflow might also allow the determination of protein markers of sample quality.

In clinical practice, a certain degree of hemolysis is not uncommon. Starting from our observation that high-abundance erythrocyte-specific proteins often showed high variability (Figures 2A, 2B, and 3A), we deliberately spiked in increasing amounts of erythrocyte lysates to pure plasma. We obtained a proportionate increase of erythrocyte-specific proteins, specifically, hemoglobin subunits alpha, beta, delta, and carbonic anhydrase 1. Notably, these proteins increased linearly (R^2 = 0.99), and even an admixture of 1 in 10,000 could easily be spotted (Figure 3B). This demonstrates that plasma proteome profiling readily indicates even small amounts of cellular contamination, in which case the values of pertinent proteins could be disregarded or corrected. The importance of this analysis step is illustrated by triplicate plasma proteome analysis, in which the samples from individual donors clustered together much more tightly in PCA when keratins and prominent red blood cell proteins were removed (Figure S3).

The blood coagulation system is primed for clotting in case of injury and wound repair. Although serum is harvested by inducing coagulation, harvesting of plasma requires addition of appropriate amounts of anticoagulants. Our plasma proteome profile contained many proteins with a function in the coagulation cascade, and we next evaluated the coordinate behavior

of these proteins as a quality control for appropriate plasma preparation. Plasma from each of the fingers of one individual was processed (Experimental Procedures). The levels of fibrinogen alpha (FGA), fibrinogen beta (FBA), and fibrinogen gamma (FGG) were lower in two of the samples. In addition, platelet basic protein (PPBP) and platelet factor 4 variant (PF4V1), which are released from activated blood platelets, were increased only in these same samples (Figure 3C), suggesting that partial coagulation had occurred. To test this hypothesis, we collected plasma and serum from two individuals and carried out sample preparation in triplicates. Indeed, levels of FGA, FGB, and FGG were much lower and levels of PPBP and PF4V1 much higher in serum compared with plasma (Figure 3D).

These observations prompted us to investigate coagulation and erythrocyte status in optimally prepared plasma. For this purpose, we obtained reference samples from a blood bank, which had gone through an extremely rigorous sample collection procedure (Experimental Procedures). They had very low and constant levels of red blood cell-specific proteins, and none had evidence of partial clotting. Although our plasma samples were also virtually coagulation free, this is in our experience not always the case with samples obtained from clinical studies (Figure S4).

Quantification of Clinically Interesting Markers in Short Gradients

Apolipoproteins are functional blood proteins involved in lipid homeostasis. They therefore reflect an individual's metabolic status, and some of them are classical markers of cardiovascular risk and metabolic disorders such as diabetes (Jenkins et al., 2014; Jensen et al., 2014). We quantified 15 apolipoproteins at each of 32 different time points in one individual. Apolipoprotein-a (LPA) had the strongest variation (CV = 20%), whereas APOB had the lowest (CV = 6%). The distribution of LPA levels in the population is skewed toward zero, with most individuals having low LPA levels but some (~20%) having higher levels. The successful quantification of the apolipoproteins in 32 plasma proteomes demonstrates the feasibility of a longitudinal measurement of risk factors known to be associated with an individual's propensity for certain diseases (Figure 3E).

Some of the apolipoproteins have allelic variants occurring with high frequency in populations that can easily be detected by MS (Krastins et al., 2013; Martínez-Morillo et al., 2014). The apolipoprotein allele APOE4 in the homozygous form is the largest known risk factor for late-onset Alzheimer's disease with a 10-fold higher risk compared with the homozygous APOE3 form (Tanzi, 2012). APOE4 has an arginine at position 112 instead of a cysteine residue in APOE2 and APOE3. In the 20 min LC-MS/MS data, we were able to clearly distinguish between the peptides LGADMEDVR (APOE4) and LGADMEDVCGR (APOE2, APOE3). In our group of ten individuals, two had one APOE4 allele (Figures 3F and 3G). The second allele was either an APOE3 or the APOE2 allele.

Serum amyloid A-1 protein (SAA1) and C-reactive protein (CRP) are acute phase proteins that are routinely measured in the clinic. Both are correlated with inflammatory states, and chronic elevation is strongly associated with increased risk for future cardiovascular events (Hua et al., 2009; Wilson et al., 2008). We found that their expression levels varied up to 1,000-fold among the ten individuals, and in a correlated manner (R^2 = 0.6; Figures 3H and 3I). In the plasma proteome with the highest levels of SAA1 and CRP, these are by far the largest differences to the plasma proteomes of the other healthy individuals, and this is presumably caused by recovery from a common cold (Figure 3J).

Next, we asked if our rapid proteome profiles contained information on any further known biomarkers. We scanned the raw data of the 15 technical workflow replicates to calculate the CVs for Food and Drug Administration (FDA)-cleared or FDA-approved biomarkers, as listed Anderson (2010). In total, 49 FDA-approved biomarkers were present in this data set (46 quantified in all 15 workflow replicates); 41 of them had CVs of less than 20%, and 28 had CVs even less than 10% (Figure 3K). When dividing these FDA-approved biomarkers into different classes (Anderson, 2010), 45 fell into "act in plasma," 2 into "tissue leakage," and 1 into "receptor ligand," and 1 was lysozyme, which had not been assigned to any category. The 20 min gradients already covered 45 of a total of 54 proteins among the "act in plasma" biomarkers. Interestingly, 42 of them were among the 180 highest abundance proteins, whereas the next 133 proteins contained only 7 known biomarkers (Figure 4A; Table S2).

Plasma Protein Epitope Signature Tags as Internal Standards for Protein Quantification

In clinical applications, quantification is almost always performed with internal standards. To add this capability to our fast workflow, we investigated the use of stable isotope labeling of amino acids in cell culture (SILAC)-protein epitope signature tags (PrESTs), which are recombinant expressed stable isotope-labeled protein fragments. This approach has the advantage that it controls for digestion efficiency, alkylation rate, and other workflow aspects and that a "master mix" of dozens of proteins of interest can be readily prepared and quantified (Edfors et al., 2014; Zeiler et al., 2012). We used APOA1, APOA4, APOB, APOE, and SHBG to construct a master mix for quantification of multiple plasma proteins in short gradients. Samples from ten individuals were prepared in triplicate and measured (Figure S5, Table S4). This resulted in low CVs for these proteins (APOA1 = 2.3%, APOA4 = 3.8%, APOB = 5.3%, APOE = 3.8%, and SHBG = 14.7%). Optimized targeted methods applied to peptides resulting from the PrESTs could improve these CVs even further.

A Quantitative Proteome of 1,000 Plasma Proteins

The above experiments highlight the value of quantifying hundreds of proteins in a very short analysis time. To obtain estimates of abundances for a deeper plasma proteome, we used a combination of peptide pre-fractionation, a matching library consisting of depleted plasma, and 100 min high-performance LC (HPLC) gradients. With 16 hr of measurement time, we identified 1,040 proteins in non-depleted plasma, of which 965 had label-free protein quantification (LFQ) values. Although MS signals for these proteins span more than six orders of magnitude, the majority of them were confined to a 100-fold abundance range (Figure 4B). The deep proteome data can be assessed in Table S5 and in the MaxQB database (Schaab et al., 2012), which also displays the mass spectrometric evidence and MS/MS transitions for all identified peptides.

Unexpectedly, the deep plasma proteome contained only 14 additional FDA-approved biomarkers compared with the 49 already found in the 20 min gradients. Nine of them were

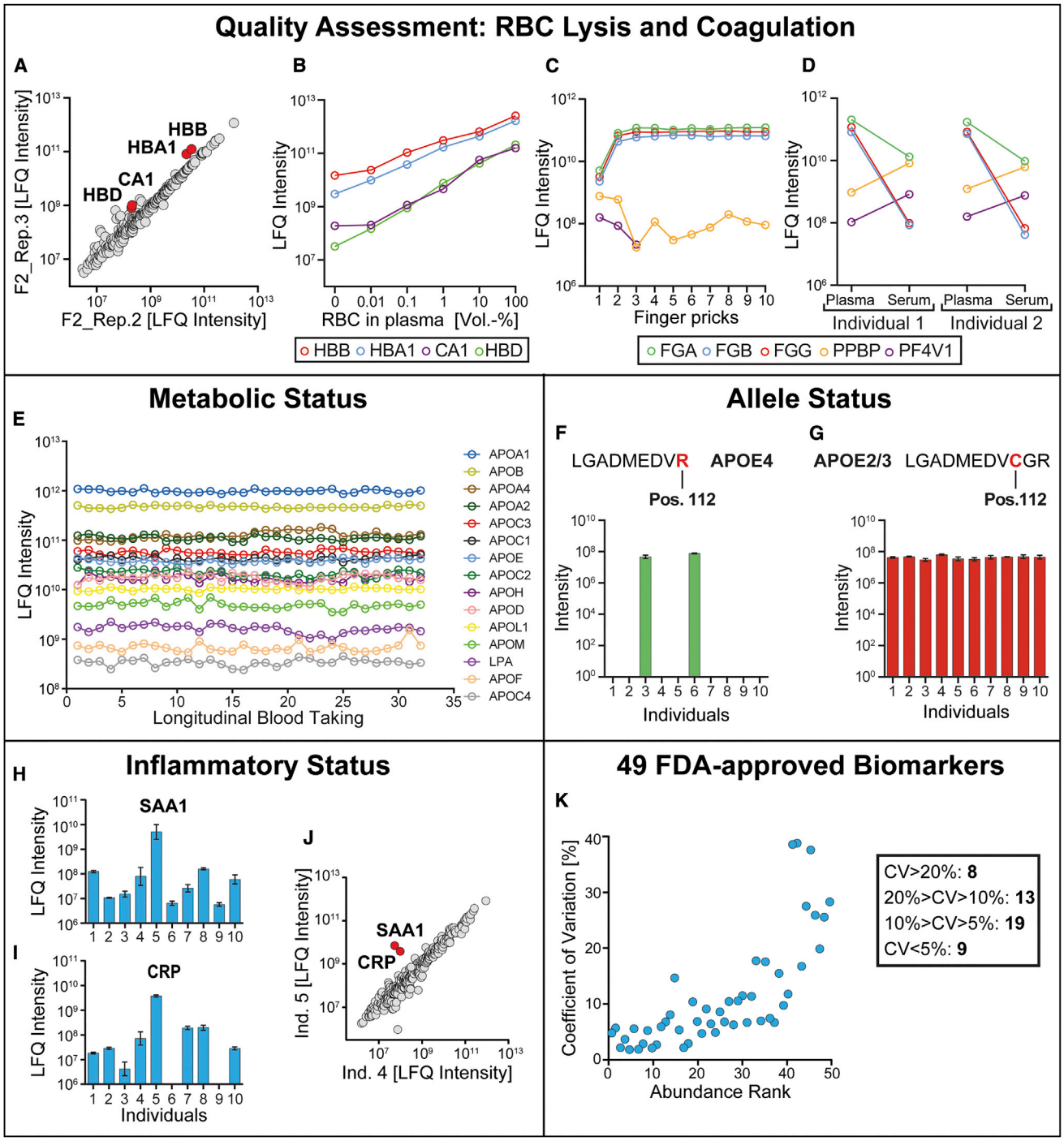


Figure 3. Quantification of Clinically Interesting Proteins

(A) Scatterplot of repeated finger pricks of one individual (replicate 2 versus replicate 3) showing that erythrocyte-specific proteins were elevated as a group. HBA1, hemoglobin subunit alpha; HBB, hemoglobin subunit beta; HBD, hemoglobin subunit; CA1, carbonic anhydrase 1.

(B) Spike-in of erythrocytes into plasma resulting in an increase of these proteins.

(C) Blood was processed from ten different fingers of one individual after finger pricking, and LFQ intensities of FGA, FGB, FGG, PPBP, and PF4V1 are plotted. In samples 1 and 2, fibrinogens are decreased, whereas platelet-specific proteins are increased.

(D) FGA, FGB, and FGG levels are decreased, and PPBP as well as PF4V1 levels are elevated in serum compared with plasma in two individuals.

(E) Fifteen apolipoproteins were quantified without any missing value after longitudinal collection of 32 plasma samples of one individual (four finger pricks per day over 8 days).

(F) The peptide LGADMEDVR is specific for the APOE4 allele and was present and quantified in two of ten individuals.

(G) Presence of at least one APOE2 or APOE3 allele in all ten individuals.

(legend continued on next page)

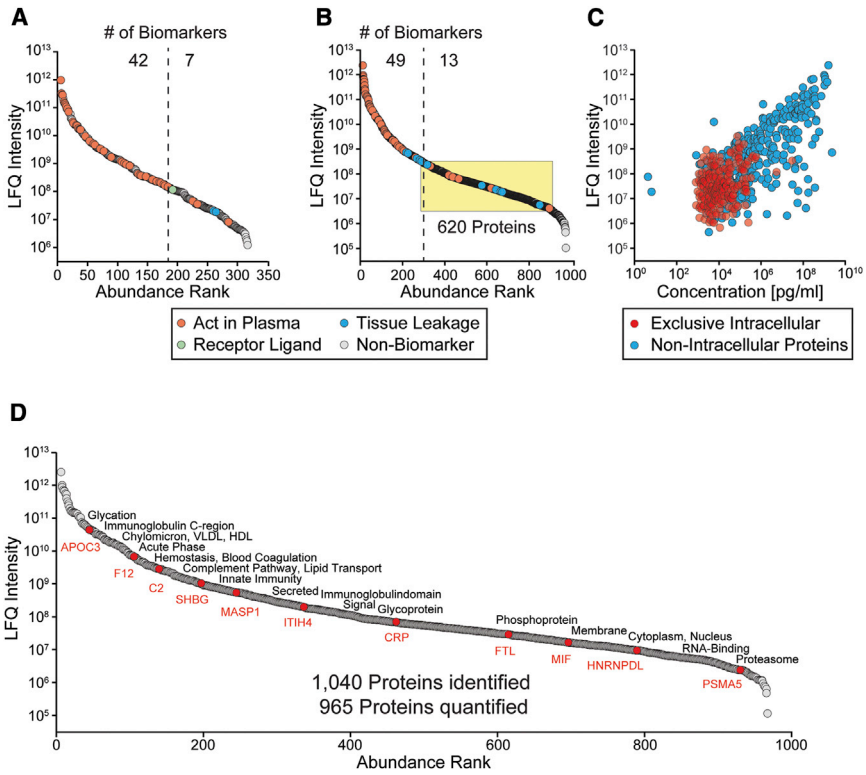


Figure 4. FDA-Approved Biomarker and Deep Plasma Proteomics

(A) Distribution of proteins quantified in the 20 min gradients of the 15 workflow replicates. FDA-approved biomarkers are color coded. The dashed line separate the regions densely populated and sparsely populated by biomarkers.

(B) LFQ intensities of 965 proteins quantified after separating peptides into eight fractions. The yellow rectangle encloses a 100-fold range, which contains the majority of the measured plasma proteome.

(C) Correlation of LFQ intensities of the deep plasma proteome data set and absolute concentrations from the Plasma Proteome Database.

(D) UniProtKB keyword annotations and their enrichment along the whole abundance range as determined by 1D enrichment (see main text). Exemplary proteins contributing to keywords are highlighted in red. APOC3, apolipoprotein C-III; F12, coagulation factor XII; C2, complement C2; MASP1, Mannan-binding lectin serine protease 1; ITIH4, inter-alpha-trypsin inhibitor heavy chain H4; HNRNPDL, heterogeneous nuclear ribonucleoprotein D-like; PSMA5, proteasome subunit alpha type 5.

classified as ''tissue leakage,'' only 4 as ''act in plasma,'' and 1 was not assigned to any category. A depth of 450 plasma proteins would be sufficient to cover 87% of the FDA-approved biomarkers present in our deep data set according to Anderson (2010).

The fact that we did not use protein depletion allowed us to investigate the quantitative nature of the plasma proteome. Bioinformatics analysis revealed that 457 of all quantified proteins had an extracellular annotation and 651 an intracellular one, with 221 overlapping proteins. Interestingly, the 430 proteins with exclusive intracellular annotation, which also have an independent abundance estimate in the Plasma Proteome Database (Nanjappa et al., 2014), are almost completely excluded from the top three order of magnitudes of protein abundance (Figure 4C). These proteins are likely of tissue origin and have been released by normal tissue damage, without necessarily having a function in blood. In contrast to the deep proteome, these intracellular proteins were largely absent in the 20 min measurements (25 of 313 proteins; Table S2). Reassuringly, 91% of proteins identified in plasma by us had also been identified in at least one of the studies collected in the PeptideAtlas repository (Farrah et al., 2014). In the absence of MS-derived quantitation of a deep, non-depleted proteome, we turned to the Plasma Proteome Database, which lists absolute concentrations of 597 of the proteins that are also quantified in our data set. Although these concentrations derive from the literature from a wide variety of individuals, health states, and quantification methods, we found a reasonable correlation, with an R^2 value of 0.53. This analysis also confirmed that we had quantified many proteins of clinical interest in the lower abundance range, such as the plasma protein ferritin (FTL) (12 ng/ml), which is widely used to diagnose dysregulation of iron homeostasis, or the cytokine macrophage migration inhibitory factor (MIF) (10 ng/ml). A total of 183 proteins in our data set have reported concentrations below 10 ng/ml.

To bioinformatically analyze the functional nature of the plasma proteome, we used the ''1D annotation'' algorithm in the MaxQuant software (Cox and Mann, 2008, 2012), to assign UniProtKB keywords to distinct abundance ranges. This resulted in 58 statistically significant features (Table S6). Classical characteristics of the plasma proteome were typically located in the high-abundance range. These include ''glycation,'' ''immunoglobulin,'' and ''chylomicrons,'' which are expected because functional plasma proteins are typically glycated, a large proportion of functional plasma proteins are antibodies, and apolipoproteins are the structural components of chylomicrons. In the low-abundance tail of the distribution, we found highly abundant intracellular complexes such as the proteasome as well as RNA-binding and processing proteins. ''Phosphoprotein'' was situated close to the middle of the distribution, and above ''membrane,'' ''cytoplasm,'' and ''nucleus,'' presumably because most intracellular proteins have by now been shown to be phosphorylatable, in addition to some of the extracellular ones. The

(H) Variation of the acute phase protein SAA1 in ten individuals.

(I) Variation of the inflammatory marker CRP in the same ten individuals.

(J) Direct comparison of two individuals to visualize the magnitude of SAA1 and CRP in the background of the other quantified plasma proteins.

(K) The CVs of 49 FDA-approved biomarkers from 15 workflow replicates as a function of protein abundance rank.

mean of functionally important Gene Ontology annotated biological processes, such as "protein-lipid complex assembly," "sterol and cholesterol transport," "acute phase response," and "regulation of coagulation" processes all scored in the upper third of the distribution, highlighting that these functions are overwhelmingly carried out by high-abundance plasma proteins. The above analysis can also be used to infer the likely function or lack thereof of a protein found at a certain concentration in normal plasma. For instance, the hormone-binding protein SHBG is in the upper range of the plasma proteome, which correlates well with its carrier function for an abundant circulating hormone (Ottosson et al., 1981) (Figure 4D).

DISCUSSION

Using state-of-the-art shotgun proteomics technology, in particular the recently described iST preparation (Kulak et al., 2014), Orbitrap instrumentation with very high sequencing speed (Kelstrup et al., 2014; Scheltema et al., 2014), and advances in label-free quantification (Cox et al., 2014), we here developed a streamlined and robust workflow for shotgun plasma proteomics. Sample preparation steps are minimized without loss of performance, and the procedure can be performed in 96-well format by a liquid handling platform. In this way, hundreds of plasma proteomes can be processed and sample preparation is not a limiting step for plasma proteomics in our workflow. We found that even extremely short measurements of 20 min still allowed the identification of more than 300 proteins, which was aided by a reference data set and the "match between runs" functionality. Accuracy and precision of the label-free workflow were excellent with intra-assay correlation of about $R^2 = 0.98$ and CVs smaller than 20% for the majority of quantified proteins.

Starting from only a finger prick of blood, the entire workflow, including database search and label-free quantification, can be performed in less than 3 hr. Previous plasma proteome studies typically started from milliliter amounts of blood, used depletion, and extensive pre-fractionation and therefore required days for completion (Cao et al., 2012; Keshishian et al., 2015; Liu et al., 2015; Such-Sanmartín et al., 2014). The ability to use small sample amounts makes blood testing much less invasive, improves cost-efficiency, and is clinically attractive in many situations, including the testing of infants as well as elderly patients (Bai et al., 2013). Likewise, fast response time is frequently important such as in the case of myocardial infarction. Our procedure uses a very short digestion time (60 min), which could be reduced by further optimization, so that the entire procedure could conceivably be performed in less than 1 hr.

Our very short LC-MS/MS runs contain nearly 50 proteins that are already subject to FDA-approved diagnostic tests, whereas the deep proteome only added few additional ones. Furthermore, the proportion of functional plasma proteins was very high, in contrast to the lower abundance range, which was dominated by tissue-derived "leakage proteins." Nevertheless, the deeper proteome still contained many proteins of known clinical significance, and it is interesting to speculate whether the relative paucity of approved biomarkers in this range is due to the greater difficulties associated with studying these proteins. Even in the short analysis runs, the lower half of the distribution has not yet been associated with specific patient states. We suggest that mixtures of recombinant isotope-labeled protein fragments, so-called SILAC-PrESTs (Edfors et al., 2014; Zeiler et al., 2012), could routinely be added to plasma samples. This would enable very high accuracy in absolute quantification for the discovery and validation of such biomarkers at high-throughput.

Further throughput improvements can be achieved with chemical labeling strategies, for instance with isobaric chemical tags such as TMT (Thompson et al., 2003). For 10-plex encoding, this could increase throughput to hundreds of patients per day. This compares favorably with metabolomics studies, which are already performed at large scale in plasma cohorts (Suhre et al., 2011), while providing equally useful and complementary information. Alternatively, TMT could be used to label patient samples before peptide pre-fractionation. This should result in deep proteome coverage, while keeping effective MS measurement time reasonably short at 1–2 hr per patient and compatible with large-scale studies.

The proteins characterized in our short workflow already contain a plethora of useful information. For example, it was easy to distinguish the gender of the donor and to obtain some risk-associated genotype information. The spectrum of apolipoproteins, as well as inflammatory markers, was excellently quantified, reflecting the cardiovascular and metabolic health state. Unexpectedly, the global nature of shotgun proteomics supplies us with valuable information about sample quality, which is not tested in routine clinical practice but can influence test results and medical decisions.

As mentioned above, the current strategy in plasma biomarker discovery by MS-based proteomics involves a narrowing down and widening strategy: a small number of patients and controls are analyzed in great depth with unbiased and relatively low-throughput methods. Resulting potential biomarkers are then envisioned to be validated with targeted MS-based methods or classical immunoassays in much larger cohorts (Anderson, 2014; Keshishian et al., 2015; Surinova et al., 2011).

Here we suggest an additional strategy, which we term "plasma proteome profiling." It consists of the measurement of large numbers of plasma proteomes at the greatest possible depth with streamlined and high-throughput technologies as described in this paper. This allows us to retain one of the basic attractions of unbiased, systemwide methodologies, namely, that associations do not have to be predefined but emerge naturally from "big data mining." Although our current work is only a first step in this direction, we believe that rapid development in the underlying technology will make this strategy more and more attractive. Given the low resource requirements, large cohorts could be investigated in the future, and one can even envision individuals routinely and repeatedly have their plasma proteome profile recorded. These high-dimensional profiles could indicate current disease risk as well as efficacy of lifestyle changes or pharmacological interventions and thereby contribute to individual and public health.

EXPERIMENTAL PROCEDURES

Tryptophan Fluorescence Emission Assay for Protein Quantification

Protein concentrations were determined after solubilizing of samples in 8 M urea by tryptophan fluorescence emission at 350 nm using an excitation wavelength of 295 nm. Tryptophan at a concentration of 0.1 μg/μl in 8 M urea was

used to establish a standard calibration curve (0–4 μl). From this, we estimated that 0.1 μg/μl tryptophan is equivalent to the emission of 7 μg/μl of human protein extract, assuming that tryptophan on average accounts for 1.3% of the human protein amino acid composition.

Blood Collection from Finger Pricks and Venous Blood Sampling

Blood was taken by lancets (Vitrex Sterilance Lite II) to obtain small quantities of capillary blood, and 5 μl of blood was transferred by a pipette into a pipette-tip-based centrifugal devices containing 0.56 μl 106 mM trisodium citrate (end concentration 10.6 mM trisodium citrate, as commonly used in blood collection tubes). The pipette-tip-based centrifugal device was made by melting the end of a pipette tip to seal it. When larger amounts of plasma were needed, blood was taken by venipuncture using a commercially available winged infusion set into collection tubes containing sodium citrate. The blood was centrifuged for 15 min at 2,000 × *g*, and plasma was harvested. Blood was sampled from healthy donors, who provided written informed consent, with prior approval of the ethics committee of the Max Planck Society.

Plasma taken by venipuncture was used to determine analytical and intra-assay variability, because in this case, larger amounts of plasma (15 μl) were needed.

Plasma for intra-individual variability was taken from one person by four finger pricks (at 6 a.m., 9 a.m., 12 p.m., and 3 p.m.) per day for 8 days. To determine the inter-individual variability, blood was taken by finger pricking of ten different individuals in triplicate (five women and five men), and samples were randomized within the gender groups. Furthermore, blood was taken from all ten fingers of one individual to one individual plasma and by venipuncture from two individuals for the comparison of plasma and serum.

Highly reliable plasma samples ($Plasma^{Ref}$ Panels) were obtained from the blood bank Blutspendedienst des Bayerischen Roten Kreuzes.

High-Abundance Protein Depletion for Building a Matching Library

A combination of two immunodepletion kits was used for optimal removal of the 20 highest abundance plasma proteins with the purpose of establishing a peptide library for matching between runs (Nagaraj et al., 2012). First we used the Agilent Multiple Affinity Removal Spin Cartridge for removal of the top six high-abundance proteins (albumin, IgG, IgA, antitrypsin, transferrin, and haptoglobin), followed by ProteoPrep20 Plasma Immunodepletion Kit for the 20 highest abundance proteins from human plasma (Albumin, IgG, IgA, IgM, IgD, transferrin, fibrinogen, α2-macroglobulin, α1-antitrypsin, haptoglobin, α1-acid glycoprotein, ceruloplasmin, apolipoprotein A-I, apolipoprotein A-II, apolipoprotein B, complement C1q, complement C3, complement C4, plasminogen, and prealbumin). Both depletion steps were carried out according to the manufacturer's instructions. The depleted plasma was digested and measured as described below. Raw data of the depleted plasma of one individual and undepleted plasma of ten different individuals served as a "library" for matching between runs for the 20 min gradients.

Sample Preparation: Protein Digestion and iST Purification

Sample preparation was performed as described previously (Kulak et al., 2014) with optimization for blood plasma as follows: 24 μl of SDC reduction and alkylation buffer (Kulak et al., 2014) were added to 1 μl of blood plasma. The mixture was boiled for 10 min to denature proteins. After cooling down to room temperature, the proteolytic enzymes LysC and trypsin were added in a 1:100 ratio (micrograms of enzyme to micrograms of protein). Digestion was performed at 37°C for 1 hr. Peptides were acidified to a final concentration of 0.1% trifluoro-acetic acid (TFA) for SDB-RPS binding, and 20 μg was loaded on two 14-gauge StageTip plugs. Ethylacetate/1% TFA (125 μl) was added, and the StageTips were centrifuged using an in-house-made StageTip centrifuge (a centrifuge with identical specifications is available from Sonation) for up to 2,000 × *g*. After washing the StageTips using two wash steps of 100 μl ethylacetate/1% TFA and one of 100 μl ddH_2O/0.2% TFA consecutively, purified peptides were eluted by 60 μl of elution buffer (80% acetonitrile, 19% ddH_2O, 1% ammonia) into auto sampler vials. The collected material was completely dried using a SpeedVac centrifuge at 45°C (Eppendorf, Concentrator plus). Peptides were suspended in buffer A* (2% acetonitrile, 0.1% TFA) and afterward sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510).

For the deep plasma data set, 20 μg purified and digested plasma peptides were fractionated using basic reversed-phase pre-fractionation. The peptides were loaded onto a reversed-phase C18 column (1.9 μm Reprosil-Pur C18 beads; Dr. Maisch) and were eluted using an EASY-nLC 1000 system (Thermo Fisher Scientific). A gradient was generated by using a dual-buffer system with buffer A (ddH_2O) and buffer B (ddH_2O, 80% ACN) adjusted to pH 10 with ammonium hydroxide. Peptides were separated and eluted from 5% B to 40% B in 50 min followed by a linear increase to 60% B in 10 min. The gradient was followed by a 12 min washout with 60%–95% B. We concatenated the 46 collected fractions into 8 fractions (concatenation scheme: 1 + 9 + 17 + 25, 2 + 10 + 18 + 26, etc.). A total of 1 μg of each concatenated fraction was loaded and measured by LC-MS/MS as described below.

Plasma samples from two individuals were dispensed into a 96-well plate (48 samples for each individual), and the complete sample preparation, with the exception of the centrifugation steps, was performed on an Agilent Bravo liquid handling platform.

Design, recombinant expression, purification and quantification of plasma PrESTs was as described in (Zeiler et al., 2012). Plasma PrESTs of the proteins APOA1, APOA4, APOB, APOE, and SHBG were combined in a master mix, which was added together with the SDC reduction and alkylation buffer to the blood plasma. The subsequent steps for sample preparation workflow are described above.

Ultra-High-Pressure LC and MS

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1000 ultra-high-pressure system (Thermo Fisher Scientific) coupled via a nano-electrospray ion source (Thermo Fisher Scientific) to a Q Exactive HF Orbitrap (Thermo Fisher Scientific) (Scheltema et al., 2014). Purified peptides were separated on 40 cm HPLC-columns (internal diameter 75 μm; in-house packed into the tip with ReproSil-Pur C18-AQ 1.9 μm resin; Dr. Maisch). For each LC-MS/MS analysis, about 1 μg peptides were used for 20 min runs and for each fraction of the deep plasma data set.

Peptides were loaded in buffer A (0.1% v/v formic acid) and eluted with a linear 15 min gradient of 10%–50% of buffer B (0.1% v/v formic acid, 60% v/v acetonitrile), followed by a 5 min 98% wash at a flow rate of 450 nl/min. Column temperature was kept at 60°C by a Peltier element-containing, in-house-developed oven, and parameters were monitored in real time by the SprayQC software (Scheltema and Mann, 2012). MS data were acquired with a Top5 data-dependent MS/MS scan method (topN method). Target values for the full scan MS spectra were 3×10^6 charges in the 300–1,650 m/z range, with a maximum injection time of 25 ms and a resolution of 60,000 at m/z 400. A 1.5 m/z isolation window and a fixed first mass of 100 m/z was used for MS/MS scans. Fragmentation of precursor ions was performed by higher energy C-trap dissociation with a normalized collision energy of 27 eV. MS/MS ssans were performed at a resolution of 15,000 at m/z 200 with an ion target value of 5×10^4 and a maximum injection time of 25 ms. Dynamic exclusion was set to 15 s to avoid repeated sequencing of identical peptides.

Data Analysis

MS raw files were analyzed by MaxQuant software version 1.5.2.10 (Cox and Mann, 2008), and peptide lists were searched against the human Uniprot FASTA database (version June 2014) and a common contaminants database by the Andromeda search engine (Cox et al., 2011) with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidations as variable modifications. The false discovery rate was set to 0.01 for both proteins and peptides with a minimum length of seven amino acids and was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin as protease, and a maximum of two missed cleavages were allowed in the database search. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation of 20 ppm. Matching between runs was performed with depleted plasma and undepleted plasma of ten different individuals serving as a library. Proteins matching to the reversed database were filtered out. LFQ was performed with a minimum ratio count of 1 (Cox et al., 2014).

Bioinformatics Analysis

All bioinformatics analyses were performed with the Perseus software of the MaxQuant computational platform (Cox and Mann, 2008). Absolute quantification of protein abundances was computed using peptide label-free

quantification values, sequence length, and molecular weight (Cox et al., 2014). For enrichment analysis, a false discovery rate of <0.02 after Benjamini-Hochberg correction was used.

Statistical Analysis

Reproducibility was analyzed by calculating R^2 values for direct comparison of the LFQ intensities of any two LC-MS/MS runs. CV values were calculated on the basis of LFQ intensities. To determine the analytical and intra-assay variability, we used 15 raw data files, for intra-individual variation 32 files, and for inter-individual variation 30 files, and triplicates for each individual were combined before determining the CV.

ACCESSION NUMBERS

The accession number for the raw and processed data reported in this paper is PRIDE proteomeXchange: PXD002854.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.02.015.

AUTHOR CONTRIBUTIONS

Conceptualization, M.M., P.E.G., and N.A.K.; Methodology, P.E.G., N.A.K., G.P., L.M.H., and D.T.; Validation, P.E.G., N.A.K., and G.P.; Formal Analysis, P.E.G.; Investigation, P.E.G.; Writing – Original Draft: M.M., P.E.G., N.A.K., and G.P.; Supervision, M.M., L.M.H., and D.T.; Project Administration, M.M.; Funding Acquisition, M.M.

ACKNOWLEDGMENTS

We thank all members of the Proteomics and Signal Transduction Group for help and discussions and in particular Igor Paron, Korbinian Mayr, Gaby Sowa for MS technical assistance, Jürgen Cox for bioinformatic tools, and Niklas Grassl and Sean Humphrey for fruitful discussions. Nils Kulak and Garwin Pichler received an m^4 award from the Bio^M Munich Biotech Cluster funded by the Bavarian government. The work carried out in this study was partially supported by the Max Planck Society for the Advancement of Science and by the Novo Nordisk Foundation (grant NNF15CC0001).

Received: October 12, 2015
Revised: January 19, 2016
Accepted: February 24, 2016
Published: March 23, 2016

REFERENCES

Addona, T.A., Shi, X., Keshishian, H., Mani, D.R., Burgess, M., Gillette, M.A., Clauser, K.R., Shen, D., Lewis, G.D., Farrell, L.A., et al. (2011). A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. Nat. Biotechnol. *29*, 635–643.

Anderson, N.L. (2010). The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clin. Chem. *56*, 177–185.

Anderson, L. (2014). Six decades searching for meaning in the proteome. J. Proteomics *107*, 24–30.

Anderson, N.L., and Anderson, N.G. (2002). The human plasma proteome: history, character, and diagnostic prospects. Mol. Cell. Proteomics *1*, 845–867.

Anderson, N.L., Anderson, N.G., Pearson, T.W., Borchers, C.H., Paulovich, A.G., Patterson, S.D., Gillette, M., Aebersold, R., and Carr, S.A. (2009). A human proteome detection and quantitation project. Mol. Cell. Proteomics *8*, 883–886.

Bai, J.P., Barrett, J.S., Burckart, G.J., Meibohm, B., Sachs, H.C., and Yao, L. (2013). Strategic biomarkers for drug development in treating rare diseases and diseases in neonates and infants. AAPS J. *15*, 447–454.

Bellei, E., Bergamini, S., Monari, E., Fantoni, L.I., Cuoghi, A., Ozben, T., and Tomasi, A. (2011). High-abundance proteins depletion for serum proteomic analysis: concomitant removal of non-targeted proteins. Amino Acids *40*, 145–156.

Cao, Z., Tang, H.Y., Wang, H., Liu, Q., and Speicher, D.W. (2012). Systematic comparison of fractionation methods for in-depth analysis of plasma proteomes. J. Proteome Res. *11*, 3090–3100.

Christensen, U., Simonsen, M., Harrit, N., and Sottrup-Jensen, L. (1989). Pregnancy zone protein, a proteinase-binding macroglobulin. Interactions with proteinases and methylamine. Biochemistry *28*, 9324–9331.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Cox, J., and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. Annu. Rev. Biochem. *80*, 273–299.

Cox, J., and Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics *13* (*Suppl 16*), S12.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. *10*, 1794–1805.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. Proteomics *13*, 2513–2526.

Edfors, F., Boström, T., Forsström, B., Zeiler, M., Johansson, H., Lundberg, E., Hober, S., Lehtiö, J., Mann, M., and Uhlen, M. (2014). Immunoproteomics using polyclonal antibodies and stable isotope-labeled affinity-purified recombinant proteins. Mol. Cell. Proteomics *13*, 1611–1624.

Farrah, T., Deutsch, E.W., Omenn, G.S., Sun, Z., Watts, J.D., Yamamoto, T., Shteynberg, D., Harris, M.M., and Moritz, R.L. (2014). State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. J. Proteome Res. *13*, 60–75.

Geiger, T., Cox, J., and Mann, M. (2010). Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol. Cell. Proteomics *9*, 2252–2261.

Hua, S., Song, C., Geczy, C.L., Freedman, S.B., and Witting, P.K. (2009). A role for acute-phase serum amyloid A and high-density lipoprotein in oxidative stress, endothelial dysfunction and atherosclerosis. Redox Rep. *14*, 187–196.

Jenkins, A.J., Toth, P.P., and Lyons, T.J., eds. (2014). Lipoproteins in Diabetes Mellitus (Humana Press).

Jensen, M.K., Bertoia, M.L., Cahill, L.E., Agarwal, I., Rimm, E.B., and Mukamal, K.J. (2014). Novel metabolic biomarkers of cardiovascular disease. Nat. Rev. Endocrinol. *10*, 659–672.

Kelstrup, C.D., Jersie-Christensen, R.R., Batth, T.S., Arrey, T.N., Kuehn, A., Kellmann, M., and Olsen, J.V. (2014). Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. J. Proteome Res. *13*, 6187–6195.

Keshishian, H., Burgess, M.W., Gillette, M.A., Mertins, P., Clauser, K.R., Mani, D.R., Kuhn, E.W., Farrell, L.A., Gerszten, R.E., and Carr, S.A. (2015). Multiplexed, quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury. Mol. Cell. Proteomics *14*, 2375–2393.

Krastins, B., Prakash, A., Sarracino, D.A., Nedelkov, D., Niederkofler, E.E., Kiernan, U.A., Nelson, R., Vogelsang, M.S., Vadali, G., Garces, A., et al. (2013). Rapid development of sensitive, high-throughput, quantitative and highly selective mass spectrometric targeted immunoassays for clinically important proteins in human plasma and serum. Clin. Biochem. *46*, 399–410.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods *11*, 319–324.

Liu, Y., Buil, A., Collins, B.C., Gillet, L.C., Blum, L.C., Cheng, L.Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T.D., et al. (2015). Quantitative variability of 342 plasma proteins in a human twin population. Mol. Syst. Biol. *11*, 786.

Martínez-Morillo, E., Nielsen, H.M., Batruch, I., Drabovich, A.P., Begcevic, I., Lopez, M.F., Minthon, L., Bu, G., Mattsson, N., Portelius, E., et al. (2014). Assessment of peptide chemical modifications on the development of an accurate and precise multiplex selected reaction monitoring assay for apolipoprotein e isoforms. J. Proteome Res. *13*, 1077–1087.

Muñoz, J., and Heck, A.J. (2014). From the human genome to the human proteome. Angew. Chem. Int. Ed. Engl. *53*, 10864–10866.

Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol. Cell. Proteomics *11*, M111.013722.

Nanjappa, V., Thomas, J.K., Marimuthu, A., Muthusamy, B., Radhakrishnan, A., Sharma, R., Ahmad Khan, A., Balakrishnan, L., Sahasrabuddhe, N.A., Kumar, S., et al. (2014). Plasma Proteome Database as a resource for proteomics research: 2014 update. Nucleic Acids Res. *42*, D959–D965.

Omenn, G.S. (2005). Exploring the Human Plasma Proteome, *Volume 5* (John Wiley).

Ottosson, U.B., Damber, J.E., Damber, M.G., Selstam, G., Solheim, F., Stigbrand, T., Södergård, R., and von Schoultz, B. (1981). Effects of sex hormone binding globulin capacity and pregnancy zone protein of treatment with combinations of ethinyl-oestradiol and norethisterone. Maturitas *3*, 295–300.

Paczesny, S., Braun, T.M., Levine, J.E., Hogan, J., Crawford, J., Coffing, B., Olsen, S., Choi, S.W., Wang, H., Faca, V., et al. (2010). Elafin is a biomarker of graft-versus-host disease of the skin. Sci. Transl. Med. *2*, 13ra2.

Sand, O., Folkersen, J., Westergaard, J.G., and Sottrup-Jensen, L. (1985). Characterization of human pregnancy zone protein. Comparison with human alpha 2-macroglobulin. J. Biol. Chem. *260*, 15723–15735.

Schaab, C., Geiger, T., Stoehr, G., Cox, J., and Mann, M. (2012). Analysis of high accuracy, quantitative proteomics data in the MaxQB database. Mol. Cell. Proteomics *11*, M111.014068.

Scheltema, R.A., and Mann, M. (2012). SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. J. Proteome Res. *11*, 3458–3466.

Scheltema, R.A., Hauschild, J.P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., and Mann, M. (2014). The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. Mol. Cell. Proteomics *13*, 3698–3708.

Such-Sanmartín, G., Ventura-Espejo, E., and Jensen, O.N. (2014). Depletion of abundant plasma proteins by poly(N-isopropylacrylamide-acrylic acid) hydrogel particles. Anal. Chem. *86*, 1543–1550.

Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E., et al.; CARDIoGRAM (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature *477*, 54–60.

Surinova, S., Schiess, R., Hüttenhain, R., Cerciello, F., Wollscheid, B., and Aebersold, R. (2011). On the development of plasma protein biomarkers. J. Proteome Res. *10*, 5–16.

Tanzi, R.E. (2012). The genetics of Alzheimer disease. Cold Spring Harb. Perspect. Med. *2*, 2.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A.K., and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. *75*, 1895–1904.

Tu, C., Rudnick, P.A., Martinez, M.Y., Cheek, K.L., Stein, S.E., Slebos, R.J., and Liebler, D.C. (2010). Depletion of abundant plasma proteins and limitations of plasma proteomics. J. Proteome Res. *9*, 4982–4991.

U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Veterinary Medicine (2001). Guidance for Industry, Bioanalytical Method Validation (U.S. Department of Health and Human Services).

Végvári, A., Welinder, C., Lindberg, H., Fehniger, T.E., and Marko-Varga, G. (2011). Biobank resources for future patient care: developments, principles and concepts. J. Clin. Bioinforma. *1*, 24.

Vihko, P., Sajanti, E., Jänne, O., Peltonen, L., and Vihko, R. (1978). Serum prostate-specific acid phosphatase: development and validation of a specific radioimmunoassay. Clin. Chem. *24*, 1915–1919.

Wilson, P.W., Pencina, M., Jacques, P., Selhub, J., D'Agostino, R., Sr., and O'Donnell, C.J. (2008). C-reactive protein and reclassification of cardiovascular risk in the Framingham Heart Study. Circ Cardiovasc Qual Outcomes *1*, 92–97.

Zeiler, M., Straube, W.L., Lundberg, E., Uhlen, M., and Mann, M. (2012). A protein epitope signature tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. Mol. Cell. Proteomics MCP *11*, O111.009613.

Cell Systems, Volume *2*

Supplemental Information

Plasma Proteome Profiling

to Assess Human Health and Disease

Philipp E. Geyer, Nils A. Kulak, Garwin Pichler, Lesca M. Holdt, Daniel Teupser, and Matthias Mann

Supplemental Figures

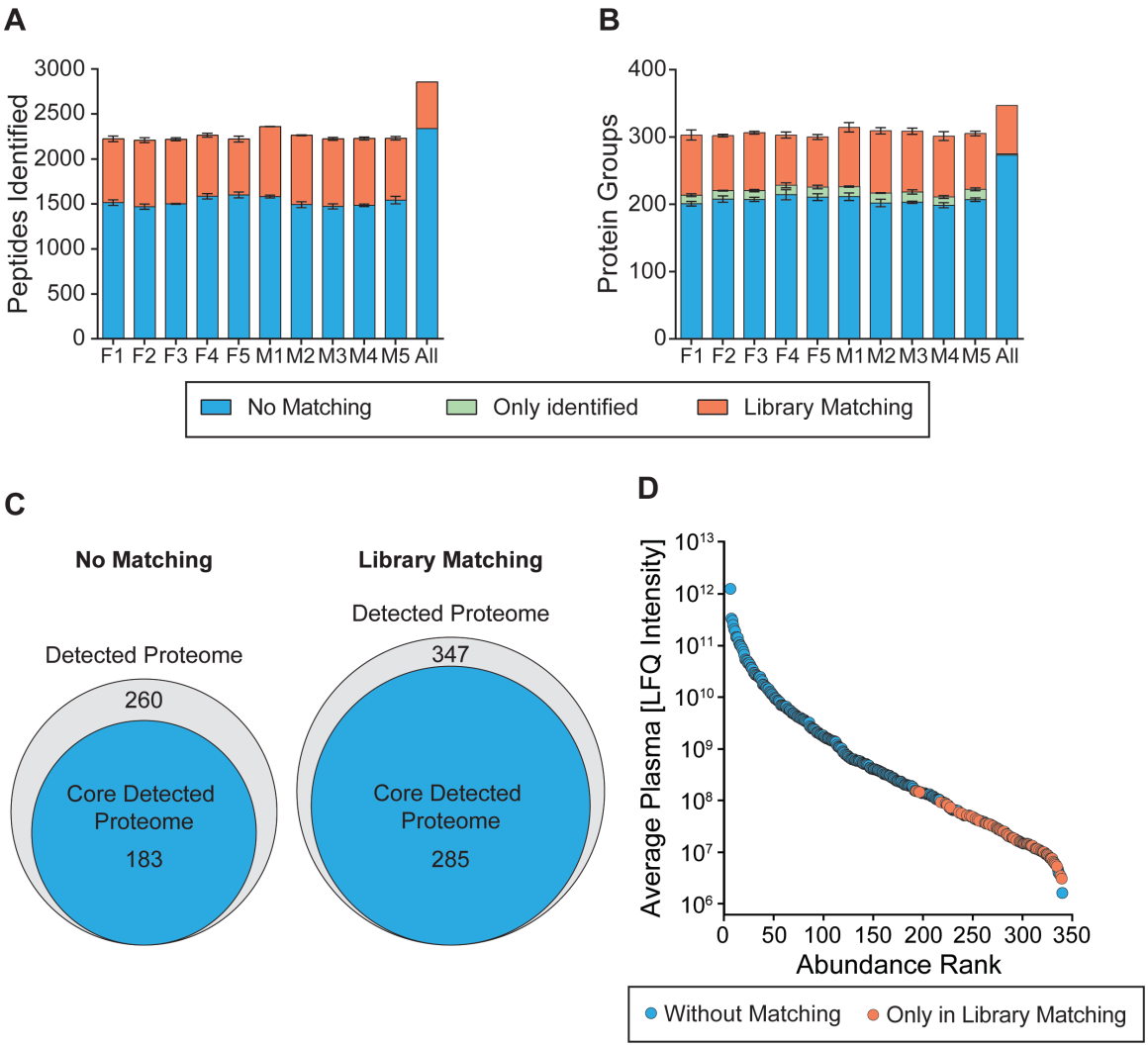


Figure S1. Related to Figure 1. Gain of using a library for match between runs.

A Number of peptides identified in ten different individuals with and without employing match between runs.

B Number of proteins that were identified and quantified with and without the gain of a matching library. Proteins that were only identified, but not quantified are indicated in green.

C 285 out of 347 proteins were identified in all ten individuals, reflecting the core detected proteome in this dataset.

D The additional proteins after library matching are all present in the lower concentration rage and are shown in orange. Blue dots represent proteins that were also present in the analysis without matching.

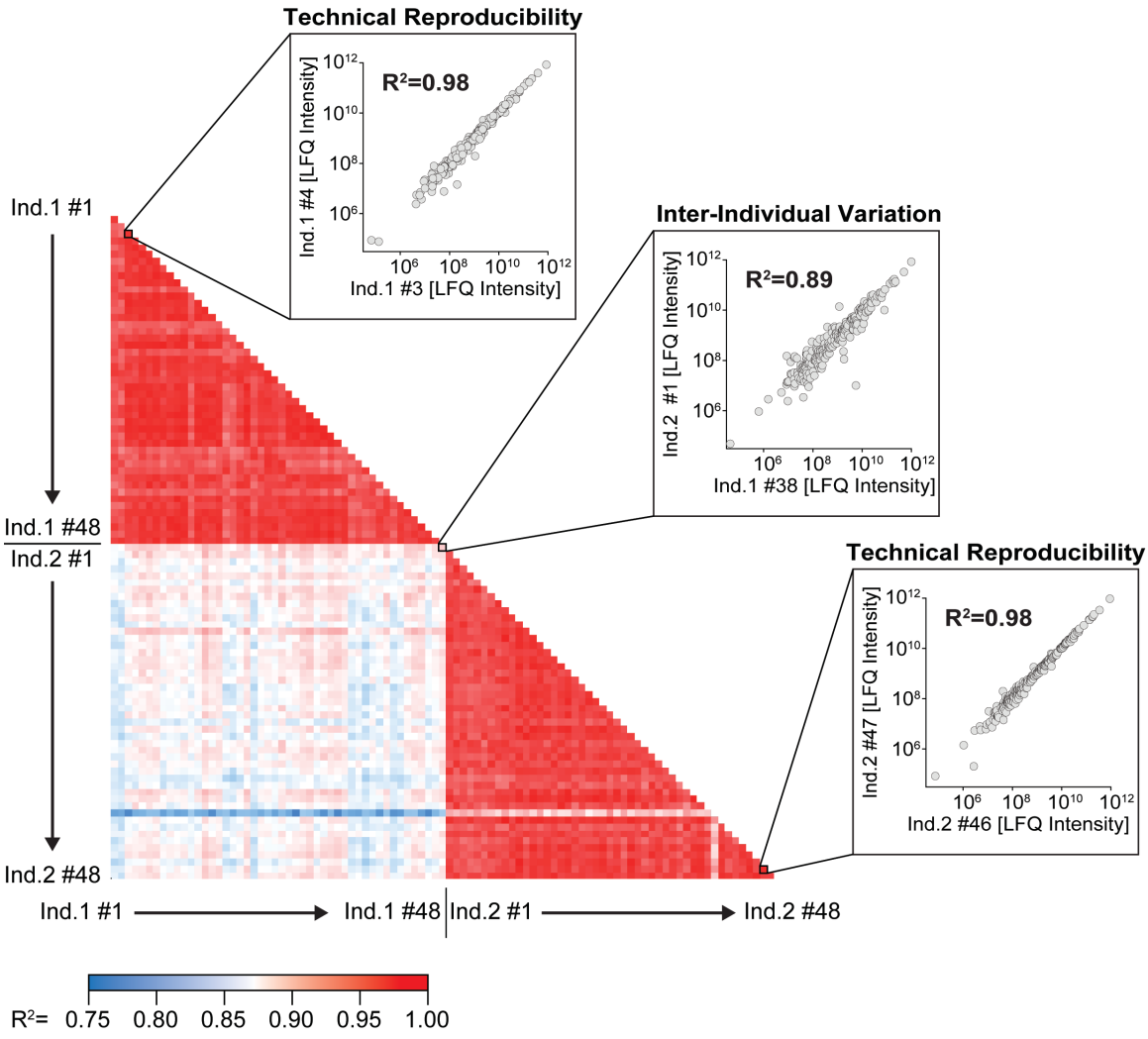


Figure S2. Related to Figure 1. Reproducibility of an automated preparation of 96 plasma samples.

Plasma of two individuals was distributed on a 96 well plate (48 plasma samples for each individual) and samples were prepared on a liquid handling platform. The figure shows 4,560 binary comparisons between the samples with color-coded R^2 values. For illustration, two correlations of technical replicates and one of the two different individuals are zoomed and displayed in the insets.

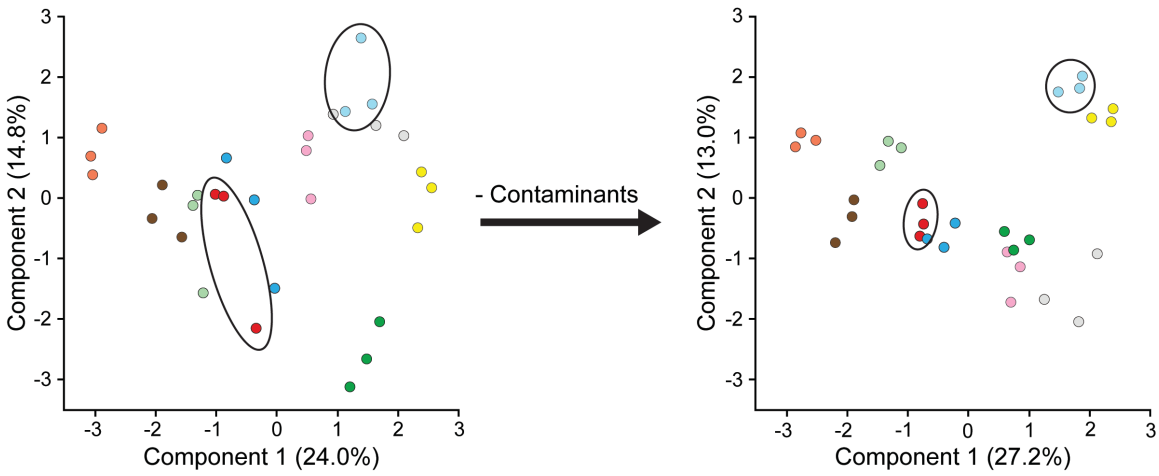
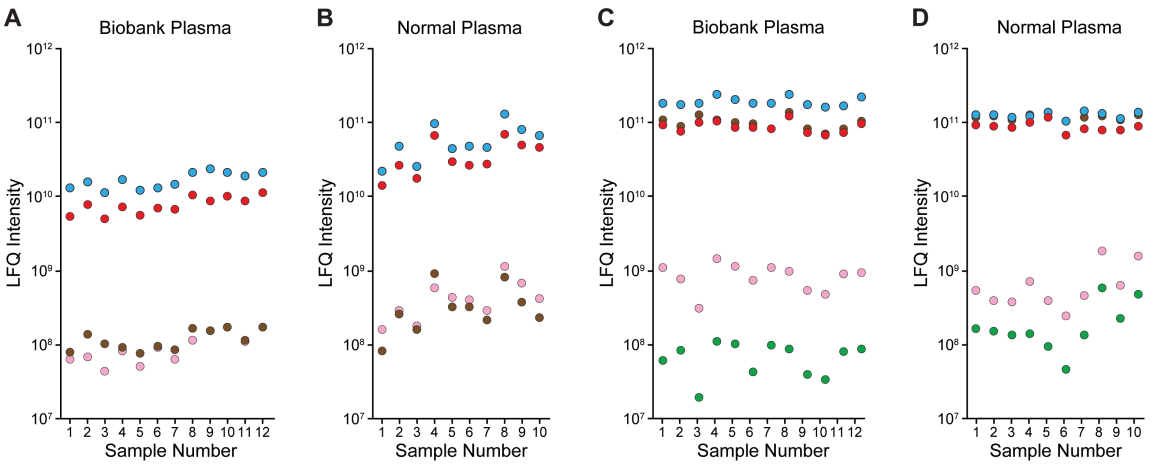


Figure S3. Related to Figure 2. Effect of removal of specific contaminants.

Removal of typical contaminants like keratins and high abundant erythrocyte specific proteins from the analysis results in a stronger clustering of workflow triplicates of ten individuals in a two-dimensional PCA. The grey circles exemplify the stronger clustering for two individuals measured in triplicates.



	HBB	HBA1	HBD	CA1
Biobank CV [%]	25	26	31	47
Normal CV [%]	55	53	75	63

	FGA	FGB	FGG	PPBP	PF4V1
Biobank CV [%]	14	17	20	37	42
Normal CV [%]	10	16	6	76	80

Figure S4. Related to Figure 3. Sample quality marker of a plasma reference panel.

A Variations of markers of erythrocyte lysis within twelve high quality reference plasma samples from a blood biobanks compared to randomly chosen normal plasma samples from ten individuals, in which variation is much higher. HBA, HBB, HBD: Hemoglobin subunit alpha, beta, delta; CA1: Carbonic anhydrase 1.

B Protein marker for coagulations in plasma samples from a biobank compared to normal plasma. FGA, FGB, FGG: Fibrinogen alpha, beta, gamma chain; PPBP: Platelet basic protein; PF4V1: Platelet factor 4 variant.

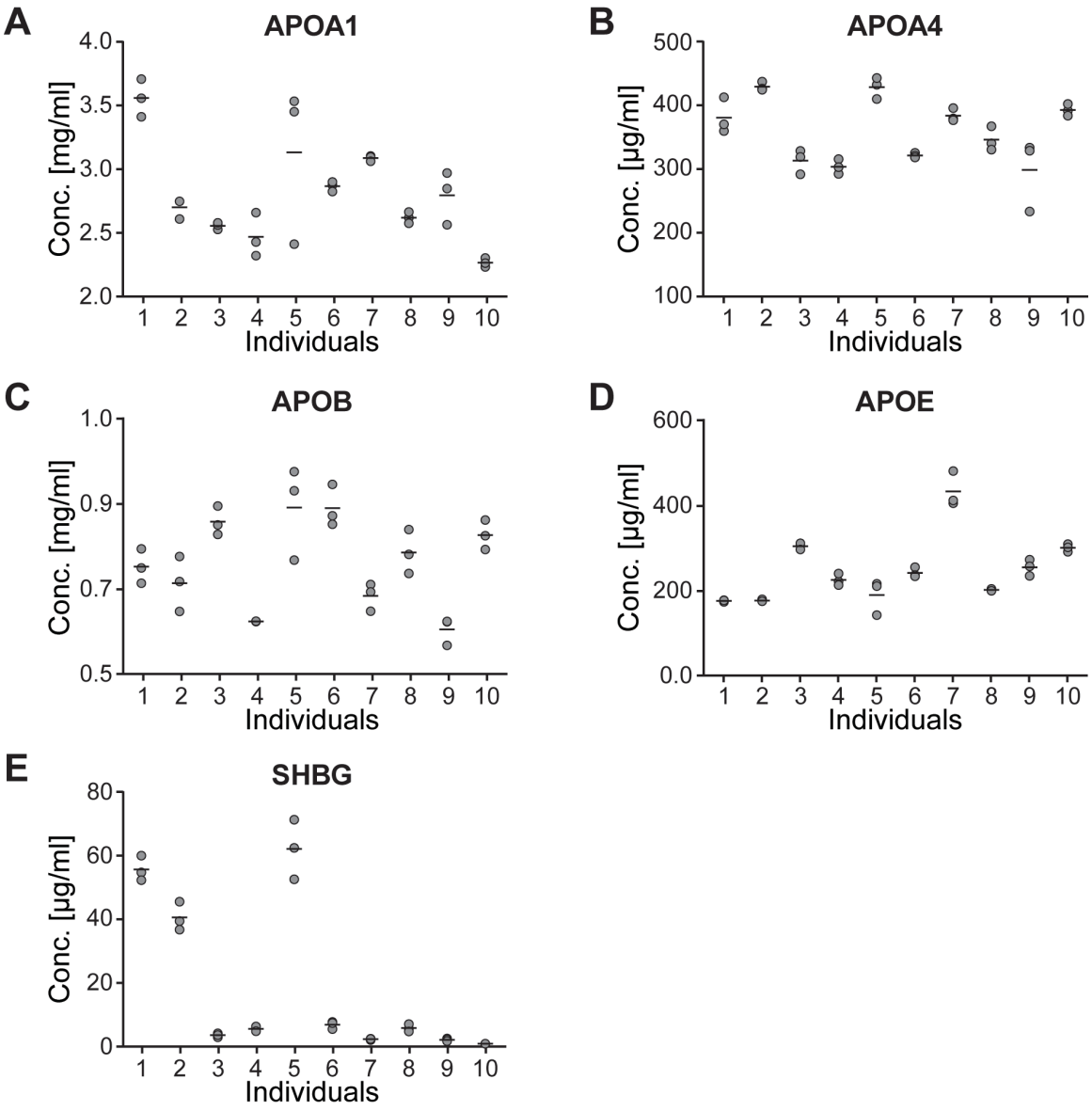


Figure S5. Related to Table S4. Plasma PrESTs as internal standards for protein quantification

A Ratios of heavy labeled PrESTs to light endogenous apolipoprotein A1 (APOA1) for ten individuals in triplicates.

B Ratios of heavy labeled PrESTs to light endogenous apolipoprotein A4 (APOA4) for ten individuals in triplicates.

C Ratios of heavy labeled PrESTs to light endogenous apolipoprotein B (APOB) for ten individuals in triplicates.

D Ratios of heavy labeled PrESTs to light endogenous apolipoprotein E (APOE) for ten individuals in triplicates.

E Ratios of heavy labeled PrESTs to light endogenous sex hormone-binding globulin (SHBG) for ten individuals in triplicates. Individuals 1-5 are women and 6-10 are male.

Supplemental Table Legends

Table S1. Related to Figure 1. Comparison of 1h versus overnight digestions and 20 min versus 100 min gradients.

Table S2. Related to Figure 1. Reproducibility of protein quantification.

Table S3. Related to Figure 2. CVs of the technical replicates of all individuals.

Table S4. Related to Figure S5. CVs and median concentrations of plasma PrESTs.

Table S5. Related to Figure 4. Deep proteome data.

Table S6. Related to Figure 4. Statistically significant features identified by '1D annotation' of plasma proteome.