Plasma proteome profiling to detect and avoid sample-related biases in biomarker studies

Philipp E. Geyer[1,2], Eugenia Voytik[1], Peter V. Treit[1], Sophia Doll[1,2], Alisa Kleinhempel[3], Lili Niu[2], Johannes B. Müller[1], Jakob Bader[1], Daniel Teupser[3], Lesca M. Holdt[3], Matthias Mann[1,2,*]

Abstract

Plasma and serum are rich sources of information regarding an individual's health state and protein tests inform medical decision making. Despite major investments, few new biomarkers have reached the clinic. Mass spectrometry (MS)-based proteomics now allows highly specific and quantitative read-out of the plasma proteome. Here we employ Plasma Proteome Profiling to define contamination marker panels to assess plasma samples and the likelihood that suggested biomarkers are instead artifacts related to sample handling and processing. We acquire deep reference proteomes of erythrocytes, platelets, plasma and whole blood of 20 individuals (>6000 proteins), and compare serum and plasma proteomes. Based on spike-in experiments we determine contamination-associated proteins, many of which have been reported as biomarker candidates as revealed by a comprehensive literature survey. We provide sample preparation guidelines and an online resource (www.plasmaproteomeprofiling.org) to assess overall sample-related bias in clinical studies and to prevent costly miss-assignment of biomarker candidates.

Introduction

Protein levels determined in blood-based laboratory tests can be useful proxies of diseases. These biomarkers assess normal physiological status, pathogenic processes, or a response to an exposure or intervention (*1*). Proteins and enzymes constitute the largest proportion of laboratory tests, reflecting the importance of the plasma proteome in clinical diagnostics (*2*). Typical protein biomarkers such as the enzymes aspartate aminotransferase (ASAT) and alanine aminotransferase (ALAT) for the diagnosis of liver diseases or cardiac troponins indicating myocardial infarction are used routinely in clinical decision making. Enzymatic activity or antibody-based laboratory tests are performed in high-throughput and at relatively low costs, as the standard of health care. However, specific biomarkers are only available for a very limited number of conditions and most have been introduced decades ago (*3*). There is thus a critical need to make the biomarker discovery process more efficient.

Protein-binder assays quantifying many plasma proteins in parallel have become available (*4, 5*), resulting in large scale biomarker mining efforts (*6-8*). Orthogonal to those technologies, mass spectrometry (MS)-based proteomics has become increasingly powerful in all domains of protein research (*9-11*). MS measures the mass and fragmentation spectra of tryptic peptides derived from the sample with very high accuracy. Because

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
2 Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
3 Institute of Laboratory Medicine, University Hospital, LMU Munich, Munich, Germany
* Corresponding author: Tel: +49 89 8578 2557; E-mail: mmann@biochem.mpg.de

these peptide and fragment masses are unique, proteomics is inherently specific, which can be an advantage over enzyme tests and immunoassays (*12*). Within its` limit of detection, MS-based proteomics can analyze all proteins in a system and is unbiased and hypothesis free in this sense.

The proteomics community has developed guidelines for the development, specificity and potential clinical application of biomarkers. These discuss quality standards and emphasize the importance of selecting cohorts that are appropriate in size, thus ensuring the statistical significance of potential findings (*2, 13-16*). That being said, there are no systematic procedures in place to assess the proteome-wide effects of pre-analytical handling of blood-based samples. Considering that plasma samples are often collected during daily clinical routine and variably processed, sample collection and processing clearly have the potential to negatively influence clinical studies, making it difficult to uncover true biomarkers, meanwhile potentially contributing incorrect ones. Especially in case-control studies, any difference in the collection and processing of samples may result in systematic bias. So far, relatively little attention has been paid to this crucial aspect on a proteome-wide scale (*17-21*).

Recently, we developed 'Plasma Proteome Profiling', an automated MS-based pipeline for high-throughput screening of plasma samples (*22*). In this article, we apply this technology to systematically assess the quality of individual samples and clinical studies with the aim to identify generally applicable contamination marker panels. Blood collection and subsequent errors in preparation are likely sources of plasma contamination. To address this issue, we construct proteomic catalogs of contaminating cell types as well as proteomic changes that may be induced during processing. This results in three panels of contaminating proteins, recommendations for assessing the quality of plasma samples and for consistent sample processing. We develop an online tool for biomarker studies and test the applicability of the panels on a recent investigation on the effects of weight loss on the plasma proteome (*23*). A comprehensive literature review of plasma proteome studies highlights that about half of them potentially suffer from limitations related to sample processing.

Results

Erythrocyte and platelet proteins in the plasma proteome

During the development of our Plasma Proteome Profiling pipeline and its optimization for high-throughput screening of human cohorts (*22*), we repeatedly observed proteins that tended to emerge as groups of statistically significant outliers but appeared to be independent of the particular study. We hypothesized that they reflected sample quality issues. Manual and bioinformatic inspection revealed three classes of origin: erythrocytes, platelets and the blood coagulation system. Consequently, we designed experiments to systematically characterize these main sources of contamination of the plasma proteome.

First, we acquired reference proteomes of erythrocytes and platelets, which are by far the most abundant cellular components (5×10^6 and 3×10^5 cells per µl). We harvested these cellular components from ten healthy females and males to obtain representative erythrocytes, platelets and pure (platelet-free) plasma and further collected platelet-rich plasma and whole blood (Fig. 1A; see Material and Methods). Cell counting confirmed the purity of the samples (Supplemental Table S1). All five blood fractions were separately prepared for each individual by our automated proteomic sample preparation pipeline, followed by liquid chromatography coupled to high resolution mass spectrometry (LC-MS/MS). To create reference proteomes, we generated a very deep library from pooled samples by analyzing extensively pre-fractionated peptides (*24*) (see Materials and Methods). A total of 6130 different proteins were identified from 61,654 sequence-unique peptides

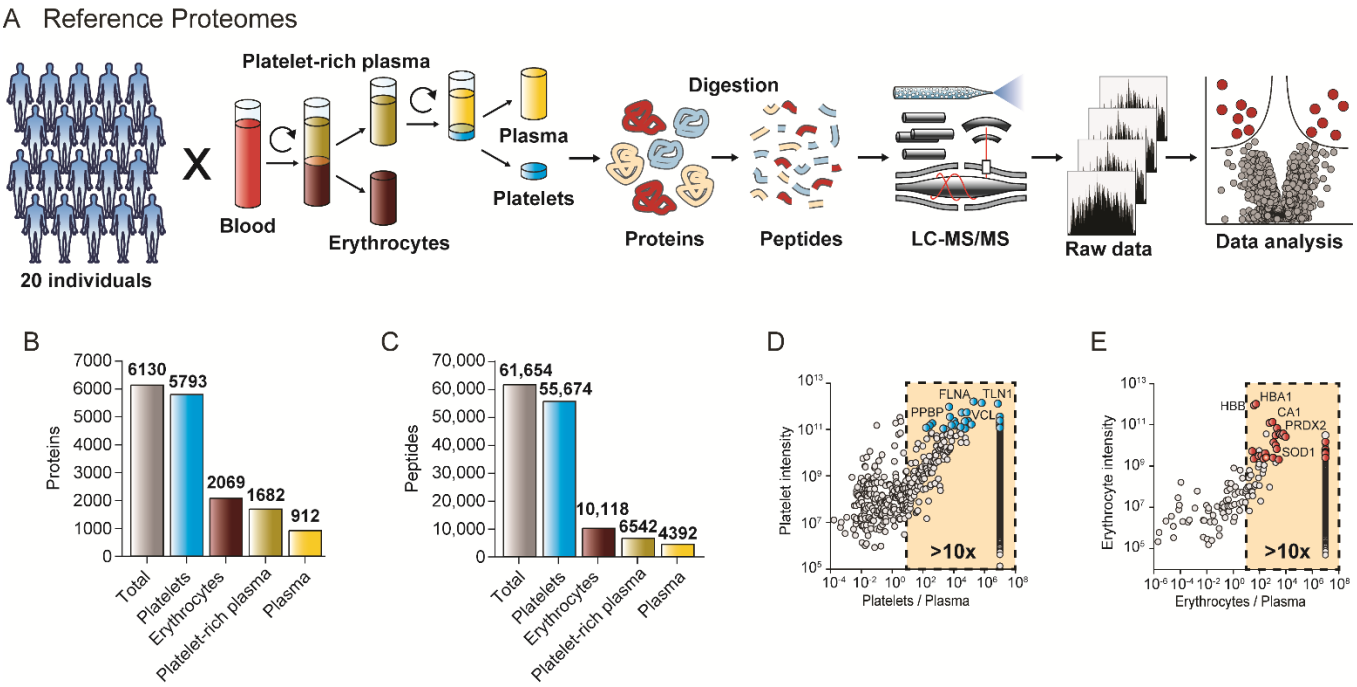


Fig. 1: Identification of blood cell markers. (A) Study outline and proteomics workflow. Erythrocytes, thrombocytes, platelet-rich and platelet-free plasma were harvested from ten healthy female and male individuals by differential centrifugation and successive purification steps. In order to generate reference proteomes for each of the blood compartments, the respective protein samples of the 20 study participates were digested to peptides. (B) Number of proteins and (C) peptides identified for platelets, erythrocytes, platelet-rich and platelet-free plasma. (D) Selection of the most suitable quality marker proteins for platelet contamination (blue dots) and (E) erythrocyte contamination (red dots) based on their abundance, the platelet/erythrocyte to plasma ratio and the coefficient of variation.

(Fig. 1B/C). The platelet proteome was the most extensive (5793 proteins), whereas we detected 2069 proteins in erythrocytes, 1682 in platelet-rich plasma and 912 in platelet-free plasma. The comparison of platelet-rich plasma to platelet-free plasma (84% additional proteins) demonstrates the extent of proteins that can be introduced by platelets.

Next, we investigated purified samples for all 20 study participants individually. The average numbers of identified proteins and peptides were very consistent in all individuals (Supplemental Fig. S1). To construct panels of easily detectable and robust contamination markers, we calculated the average protein intensities and the coefficient of variation (CV) across the study participants. As a prerequisite, we required that the proteins should be substantially more abundant in erythrocytes as well as platelets rather than in plasma. According to these criteria, we selected the 30 most abundant proteins with CVs below 30% and at least a 10-fold higher expression level in the contaminating cell type than in plasma (Fig. 1D/E). NIF3-like protein 1 (NIF3L1), a low abundant erythrocyte specific protein was excluded, because it was inconsistently identified as was the platelet-bound coagulation factor F13A1, whose function makes it an unsuitable platelet marker. The remaining proteins represent our cellular contamination marker panels (Supplemental Table S2). They overlap by just two proteins (actin/ACTB and glyceraldehyde-3-phosphate dehydrogenase/GPDH) and their quantities did not correlate with each other (Supplemental Fig. S2). Thus they are specific and independent indicators for the origin of plasma contamination.

Comparing median expression values of proteins shared between the blood components revealed

that plasma proteins do correlate with whole blood (Pearson correlation coefficient R = 0.43), as expected. In contrast, there was no correlation between the platelet, erythrocyte and plasma proteomes (Supplemental Fig. S2). This indicates that the levels of cellular proteins in plasma are not a constant fraction of those in the cellular proteomes. The platelet panel was enriched in platelet-rich plasma compared to normal (platelet-free) plasma. Both panels are depleted in pure plasma compared to whole blood, however the erythrocyte panel even more strongly decreased as centrifugation removes erythrocytes more efficiently than platelets. A histogram of both panels over the abundance range visualizes their distribution in the different blood compartments (Supplemental Fig. S2). Erythrocytes are ten-fold more abundant and four-fold larger than platelets and indeed the corresponding panel proteins have a 42-fold difference in whole blood. In plasma, however, their ratio was nearly one to one, again pinpointing a more efficient removal of erythrocytes compared to platelets in standard sample preparation. The fact that several proteins of both panels were still detectable in pure plasma indicates a baseline level of contaminants due to imperfect depletion or the life cycle of these cells. The four most abundant erythrocyte proteins, HBA1, HBB, CA1 and HBD were present in pure plasma of almost all individuals, whereas lower abundant proteins were only sporadically identified. In contrast, platelet proteins were quantified over a larger abundance range and some of them were found in every individual.

In addition to the sum of panel protein abundances, we calculated their correlation to the standard reference panel defined by the 20 participants to several hundred plasma samples. A distinct contamination of erythrocyte proteins seems to be a part of the plasma proteome as the erythrocyte panel has in general a relative high correlation between the reference cohort erythrocyte levels and the study plasma samples. In contrast, in many plasma samples there was no correlation detectable between the reference cohort platelet levels and the plasma samples in the study. In practice, a correlation greater than 0.5, indicated that the proteins are present as a result of contamination (Supplemental Fig. S3A,B,C). Note that an apparent contaminant protein could still be applied as a biomarker – however, in this case its abundance value should be different from the pattern in the reference contamination panel.

Spike-in experiments validate the erythrocyte and platelet contamination panels

To determine if the two protein panels correctly quantify contamination in plasma, we generated four pools of erythrocytes and platelets from five study participants at a time. These pools were diluted in nine steps into platelet-free plasma for a total range of 10^7, followed by cell counting and proteomic analysis (Fig. 2A). This resulted in an expected decrease in the cellular proteome ratio to plasma (Fig. 2B, C). All but two of the panel proteins were consistently quantified over the dilution range.

As the protein within each panel have the same origin, we defined a single variable for each cell type by summing their intensities and dividing by the summed intensities of all quantified plasma proteins. This yielded two remarkably robust ‘contamination indices’ that turned out to be linear with respect to the cell numbers determined by cell cytometry (Supplemental Table S3) (R = 0.98 and 0.99, Fig. 2D, E). Spiked-in contaminations of 1:100 could readily be detected, which corresponds to a concentration of 70,000 erythrocytes or 30,000 platelets per µl plasma.

Contamination panel for blood coagulation

In addition to contamination due to cellular constituents, partial and variable coagulation could contribute to systematic bias in biomarker studies. Indeed, we had found coagulation-related proteins to be connected to sample handling from finger pricks while developing our Plasma Proteomics

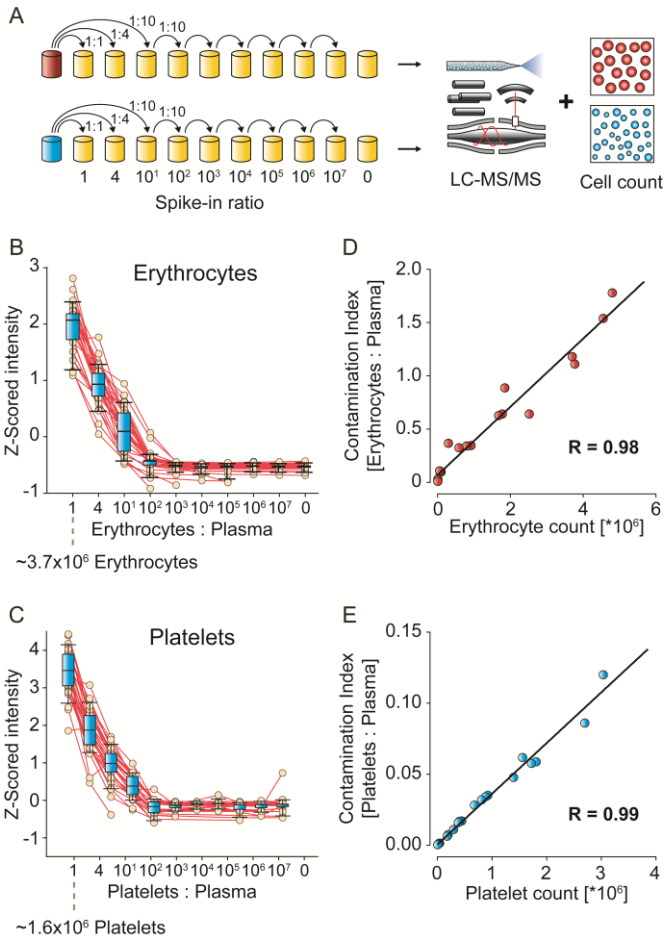


Fig. 2: Spike-in of erythrocyte and platelet fractions into pure plasma. (A) Dilution and analysis scheme. (B) Z-scored protein intensities of the 29 contamination markers of erythrocytes and (C) of the 29 platelet contamination markers as a function of their spike-in proportion to plasma. Whiskers indicate 10-90 percentiles and horizontal lines denote the mean. (D) Correlation of erythrocyte count to the 'contamination index' for the erythrocyte marker panel. (E) Correlation of platelet count to contamination index for the platelet marker panel.

Pipeline (*22*). In clinical practice, an anticoagulant is pre-added to commercially available containers so that it is combined with blood upon withdrawal. Prompt shaking mixes the anticoagulant with the blood, yielding pure plasma after centrifugation (Fig. 3A). Any delay in adding or mixing could cause partial coagulation – in the most extreme case of missing anti-coagulant and waiting for 30 min, one would obtain serum instead of plasma.

To generate a panel for assessing blood coagulation, we systematically compared 72 plasma vs. 72 serum samples (4 individuals, 18 aliquots). From a total of 2099 quantified proteins, 299 were significantly altered (Fig. 3B). The most significantly decreased proteins after clotting were typical constituents of the coagulation cascade such as fibrinogen chains alpha (FGA), beta (FGB), gamma (FGG) ($p < 10^{-130}$, > 40-fold), whereas the platelet-associated coagulation factor F13A1 and antithrombin-III (SERPINC1) decreased by more than half. Interestingly, the strongest elevated proteins in serum were all connected to platelets: platelet basic protein (PPBP), platelet glycoprotein Ib alpha chain (GP1BA), thrombospondin 1 (THBS1) and platelet glycoprotein V (GP5) ($p < 10^{-10}$; 2 to 5-fold increase). In total, 208 proteins increased and 91 decreased due to coagulation. The former set of proteins were also quantitatively enriched within the higher abundant proteins in the platelet proteome ($p < 10^{-5}$; median rank 699 of 3150 proteins), indicating coagulation-induced activation of platelets.

To define a robust panel of contamination markers assaying the extent of coagulation, we first selected the 30 most significantly altered proteins between serum and plasma. Although not among the top 30, we added the platelet factor 4 variant 1 (PF4v1) ($p < 10^{-11}$, 2.2-fold up in serum), because it was an excellent indicator of coagulation in our studies and has already been reported in the context of pre-analytical variation (*21*).

In contrast to the erythrocyte and platelet panels, proteins of the coagulation panel increase or decrease due to blood clotting and the fold-changes vary strongly between them. Because fold-changes are greatest for the decreasing proteins, we calculated the coagulation contamination ratio only from them (sum of all plasma proteins divided by sum of plasma-elevated coagulation proteins). This ratio was very robust when comparing serum and plasma, clearly separating them with median contamination ratios of 9 and 120 for these distinct sample types (Fig. 3C). Of the coagulation marker

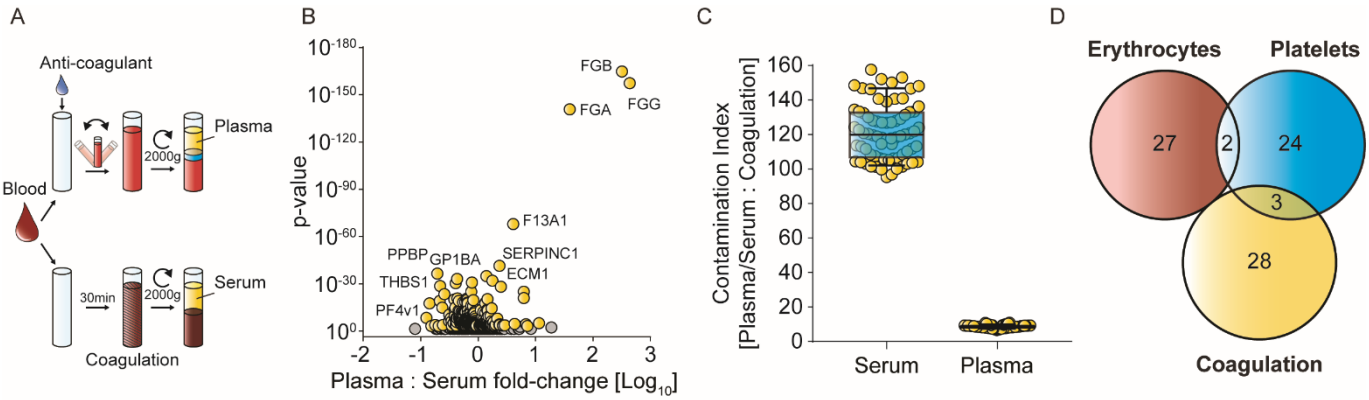


Fig. 3: Contamination marker panel for blood coagulation. (A) Preparation of plasma and serum samples. EDTA was used as anti-coagulation agent and incubation and centrifugation values are indicated. (B) Volcano plot comparing 72 plasma vs. 72 serum proteomes. Proteins highlighted in yellow were chosen according to their p-value as markers for coagulation. Only the plasma enriched proteins (compared to serum) were used in the calculation of the coagulation contamination index. (C) Ratio of the summed intensities of all plasma or serum proteins to the sum of the plasma-enriched panel proteins are plotted for all samples. Whiskers indicate the 10-90 percentile. (D) Overlap of the three quality marker panels.

panel, only F13A1, PPBP and THBS1 were in common with the platelet panel and none with the erythrocyte panels (Fig. 3D). The low overlap observed for the three quality marker panels should make them highly specific tools to elucidate the presence and origin of sample related bias.

Application of the contamination marker panels to a biomarker study

The above defined marker panels can assess sample-related issues at three levels: the quality of each sample in a clinical cohort, potential systematic bias in the entire study and the likelihood that individual biomarker candidates belong to the contaminant proteomes.

We recently investigated changes in the plasma proteome upon weight loss (*22, 23*). Briefly, caloric restriction in 52 individuals for two months was followed by weight maintenance for a year. Plasma Proteome Profiling of seven longitudinal samples revealed significant changes in the profile of apolipoproteins, a decrease in inflammatory proteins and markers correlating with insulin sensitivity. Given that protein abundance changes of less than 20% were often highly significant, we expected that overall sample quality was high, making this study suitable for testing the practical applicability of the contamination marker panels.

First, we assessed the quality of each sample separately by calculating the three contamination indices and plotting their distribution in the total of 318 measurements. For each index, we initially defined potentially contaminated samples as those with a value more than two standard deviations above the mean (red lines in Fig. 4A). This flagged 12 samples, six with platelet contamination, one with increased erythrocyte levels, and five with signs of partial coagulation. Resolving the three quality marker panels to the levels of individual proteins resulted in almost perfectly parallel trajectories (Supplemental Fig. S4A-C). Accordingly, the correlations to the reference contamination panels were substantial ($R > 0.77$). Overall, the variation of the contamination indices was highest for the platelets also visible by a contamination index difference (max/min ratio) of a factor 182 between the least and the most contaminated sample, followed by erythrocytes (max/min 23) and lowest for coagulation (max/min 5). The platelet proteins talin-1 (TLN1), myosin-9 (MYH9) and alpha-actinin-1 (ACTN1) had the largest variations, all with maximal changes greater

than 5000-fold. Catalase (CAT), carbonic anhydrase 1 and 2 (CA1, CA2) from the erythrocyte index varied maximally by more than 500-fold. The three fibrinogens in the coagulation panel changed by up to 20-fold, indicating that only partial coagulation events took place (Fig. 4A).

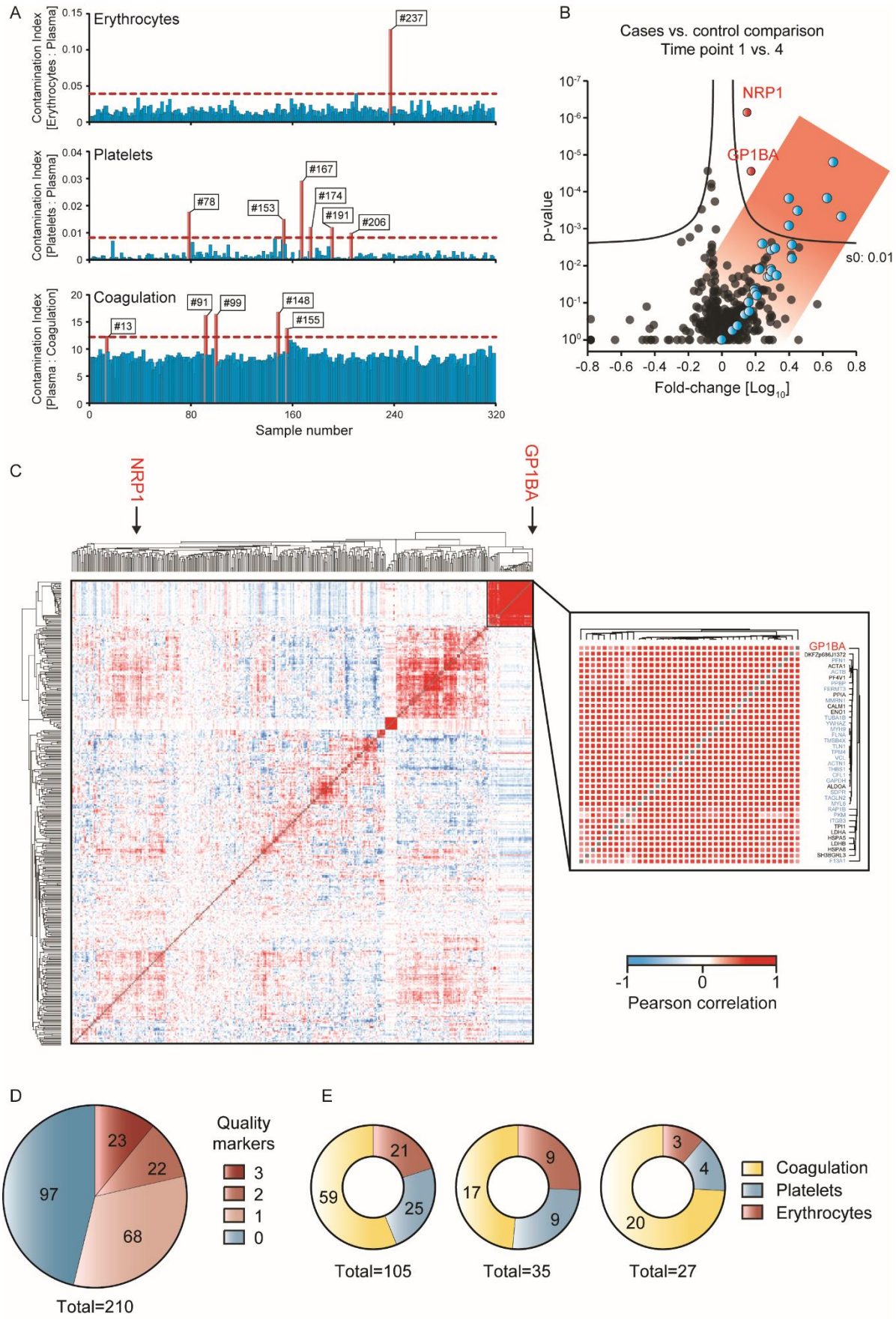
Note that evaluating individual sample quality based on the standard deviation of all samples, as done here, has the benefit of being independent of the specific proteomic method used to measure protein amounts. However, this requires that most samples have low levels of contamination, so that outliers of the statistical distribution are clearly apparent. If this is not the case, we propose using general, study-independent cut-off values to differentiate between samples of high and poor quality in such studies.

To assess potential systematic bias for groups of samples such as cases and controls or different time points, we applied a volcano plot. This revealed that the statistically significantly upregulated proteins of time point 4 of our study mainly due to the platelet panel (Fig. 4B). With this information in hand, we contacted our collaboration partners, who tracked down the platelet contamination to a switch of the blood taking equipment due to low supplies.

In practice, such sample issues will occasionally happen in a clinical study, and our contamination marker panels would allow elimination of the affected samples. However, if contaminating proteins can reliably be distinguished from relevant biomarker candidates, the data could still be used. In our example, six of the eight significant outliers were from the platelet panel, and the other two proteins – GP1BA and NRP1 – could still be of interest. To investigate this further, we inspected the global correlation map of all proteins, time points and participants (*25*). In this hierarchical clustering analysis, proteins that are co-regulated have a high correlation to each other and appear in groups, visualized as red patches (Fig. 4C). Here, the platelet cluster was the second largest one with 38 proteins (R = 0.69). All quantified platelet panel proteins were in this cluster, as was GP1BA, flagging them as likely contaminants (Fig. 4C and inset). Interestingly, NRP1, a receptor involved in angiogenesis did not group with the platelet proteins, suggesting a potential biological role. This is supported by the fact that NRP1, was significantly regulated over all time points compared to the baseline, in contrast to the platelet cluster proteins.

The other two quality marker panels are also readily apparent in the global correlation map. Ten members of the erythrocyte panel cluster tightly as do the three fibrinogen chains (Supplemental Fig. S5). However, in this study the fibrinogens group with proteins involved in low-grade inflammation, reduction of which was one of the main findings of our study (Supplemental Fig. S5). In contrast, the coagulation marker PF4v1, which is also a highly abundant protein in platelets, clustered in the platelet group in this analysis, indicating that it varied as a result of sample preparation.

To make the above described analysis readily available, we created an online platform at www.plasmaproteomeprofiling.org. It provides a toolbox for the interactive assessment of the quality of plasma proteomics data. Lists of protein abundances can be uploaded by drag-and-drop and the system automatically generates the three contamination index values as shown in Figure 4A. If the user indicates cases and controls, the dataset will be analyzed for systematic bias as visualized in a volcano plot (Fig. 4B). The global correlation map is also displayed with the clusters of the contamination panels (Fig. 4C). The website is designed in the Dash data visualization framework, which allows further interactive analysis of the data (see Materials and Methods). Potential biomarker candidates in the volcano plot can be selected and displayed in the global correlation map to check if the protein falls into or near one of the



contamination marker clusters.

Revisiting results of published biomarker studies

Having examined one study in detail, we set out to survey the extent to which probably contamination proteins are reported as biomarker candidates in the literature. To this end, we performed a comprehensive PubMed search requiring the terms `proteomics`, `proteome`, `plasma OR serum`, `biomarker` and `mass spectrometry` spanning the time frame from 2002 to April 2018. We excluded review papers, purely technological publications without biomarker candidates, animal studies and publications without proteins as qualitative or quantitative variables. From the resulting 210 publications, we manually extracted the lists of the biomarker candidates that were reported as 'significantly altered proteins' by the authors. Gene and protein names were mapped to the corresponding protein identifiers in our reference panels and analyzed for their frequencies.

Remarkably, 113 studies (54%) reported at least one potential contamination marker as a biomarker candidate or as a statistically significant association (Fig. 4D). As the total contamination panel consists of 84 proteins and the median number of candidates per clinical study was seven, a certain overlap is not entirely unexpected. However, the candidates in question almost always were near the top of most abundant proteins of the contamination panels, making it highly likely that they are indeed contaminants. Furthermore, while an individual protein could still be a genuine biomarker candidate, the fact that 22 studies (11%) reported two of them, and a further 23 studies (11%) three or more, again makes contamination the likely explanation.

The majority of these studies reported proteins as potential biomarkers or as significant outliers of the coagulation panel, followed by the erythrocyte and platelet panels (Fig. 4E). The most frequent one was clusterin (CLU; 27 times), followed by the fibrinogen (alpha, beta and gamma; 22, 10, 15 times), prothrombin (F2; 17 times), kininogen (KNG1; 15 times), antithrombin-III (SERPINC1, 13 times) and platelet basic protein (PPBP, 10 times). It is worth noting that proteins related to erythrocyte leakage may falsely be taken to indicate activation of oxidative pathways. For example, the hemoglobin subunits (e.g. HBA1, HBB, HBD, listed 1, 6 and 1 times), carbonic anhydrases (CA1, CA2, 6 and 6 times), fructose-bisphosphate aldolase (ALDOA, 5 times), peroxiredoxin 2 (PRDX2, 3 times) and superoxide dismutase (SOD1; 2 times) are annotated with keywords linked to oxidation. To illustrate this, a recent publication connected plasma proteome alterations in Type 1 diabetes to oxidative stress. This may be a spurious link because the reported proteins were mostly members of the erythrocyte contamination panel (*26*). Although platelet panel proteins are not prominent in the biomarker literature yet, we expect that they – along with lower abundant erythrocyte specific proteins – will play an increasing role as technological progress enables higher plasma proteome coverage. We caution that platelet proteins already found in the biomarker literature such as PPBP, THBS1 and PF4 are often linked to coagulation events.

Fig. 4: Quality marker panels in a weight loss study and literature study. (A) Assessment of individual sample quality with respect to the three contamination indices using the online tool at www.plasmaproteomeprofiling.org. Samples with indices that are more than two standard deviations from the mean (horizontal red lines) are flagged as potentially contaminated (red bars and sample numbers). (B) Volcano plot of the proteome comparison of time point 1 vs 4. Proteins of the platelet panel are highlighted in blue and two additional significantly regulated proteins in red. (C) Global correlation map on the left with an inset of the platelet cluster on the right. The two significant outliers of the volcano plot in (B) are marked in red. Platelet panel proteins are highlighted in blue in the inset. Red patches in the global correlation map indicate positive and blue patches negative correlations. (D) Literature analysis of 210 publications using MS-based plasma proteomics to identify new biomarkers. The number of quality markers reported as biomarker candidates in these studies is indicated. (E) Distribution of the reported quality markers according to the three types of likely contaminations. The distribution is shown across studies that report one, two or three proteins of the same contamination panel.

Table 1: Practical considerations to minimize systematic bias

General instructions
• Avoid pooling of samples
• Use plasma or serum exclusively, not a combination
Sample collection
• Standardize blood collection and pre-analytical procedures (preferably same person collecting blood, centrifuge, sampling container, storage temperature and time)
• Centrifuge blood to generate plasma immediately
• Centrifuge according to manufacturer's instruction
• Harvest plasma immediately after centrifugation
• Harvest the plasma starting from the top of the container, and pool it before aliquotting
• Discard the last 500 µl of plasma to avoid contamination with platelets or use a second centrifugation step to generate platelet-poor plasma
• Freeze study samples immediately after harvesting
Principal assessment of study sample quality
• When working with a new batch of samples from collaborators: run at least ten test samples of each study group by mass spectrometry
• Use quality marker panels to check for any indication of contamination
Main study
• Continuously assess quality during the project to detect and avoid systematic bias (pre-analytics, mass-spectrometric analyses)
• Overall quality: Report the number of contaminated samples
• Systematic bias: Report potential systematic bias
• Check if biomarker candidates are contained in the quality marker panels
• Identification of several quality markers as biomarker candidates may be indicative of a study vector
• If a quality marker is among the biomarker candidates, thorough validation required

Recommendations for future proteomics studies

Based on our experience with the above defined three quality marker panels (Supplemental Table S2) and analysis of thousands of plasma proteomes, we devised a general guideline for minimizing and detecting biases related to sample taking and processing (Table 1). The centrifugation device itself plays the largest role in causing platelet contamination and the blood taking equipment must be kept the same throughout a study. We recommend to not collect the lowest layer of the plasma above the platelet bed after centrifugation as this can strongly contribute to contamination (Supplemental Fig. S6). Furthermore, any delay from centrifugation to plasma harvest has a high potential to induce platelet protein contamination. The above factors mainly influence the platelet rather than the erythrocyte contamination index, indicating that proteins from the platelet proteome are the most likely cause of erroneous assignment of biomarker candidates.

Discussion

Blood plasma remains the predominant biological matrix to assess health and disease in clinical settings. Around the world, every day hundreds of thousands of samples are analyzed to determine the levels of individual proteins. Likewise, blood plasma is directly or indirectly assessed in most clinical trials. Protein levels in plasma can readily be affected by cellular contamination or handling-related issues and in clinical practice this is partially addressed by simple tests such as those for hemoglobin contamination. However, these tests are not systematic or quantitative and they can only be used to exclude clearly contaminated samples.

Because of its high specificity and unbiased nature, MS-based proteomics is ideally suited to characterize the quality of blood plasma and it requires less than one µl of material. So far, research on sample quality involving MS has mainly been restricted to the stability of internal standards in targeted assays and has rarely addressed overall sample quality (*14, 17, 18*). To our knowledge, there has been no systematic effort to examine the common sources of plasma proteome contamination using proteomics. Employing this technology on a systematic basis, we found that platelets, erythrocytes and coagulation are by far the most important causes of plasma contamination. We acquired very deep reference proteomes for these cell types and blood compartments, which we provide to the community to evaluate the possible origin of proteins emerging

from biomarker studies. We defined three panels of about 30 proteins each that can serve as contamination indices (Supplemental Table S2). Using the example of a longitudinal Plasma Proteome Profiling study of weight loss and our online resource, we illustrated how the contamination indices can flag individual, suspect samples and systematic biases. Furthermore, correlation analysis reveals whether or not potential biomarkers emerging from a given study are likely to be associated with contamination-related proteome changes instead. Conversely this procedure can 'rescue' genuine biomarker candidates that are part of the contamination proteomes.

The clinical potential of the plasma proteome has long been realized and is also emphasized by the fact that more than 50 FDA-approved biomarkers can be quantified even in relatively shallow proteomics measurements of plasma (*22*). If there are as many new biomarkers among the less abundant proteins, there should be a diagnostic treasure trove still to be discovered (*2*). Millions of plasma samples are stored in biobanks worldwide, representing an immense untapped resource that could be analyzed by MS-based proteomics or large-scale affinity based methods. Despite initial enthusiasm and community efforts such as the Human Proteome Organization's plasma proteomics initiative (*27, 28*), few if any new protein biomarkers have entered the clinic in recent decades. This is probably at least partially due to technological limitations to characterize the vast dynamic range of the plasma proteome, which in turn has led to underpowered study designs (*2*). While many of these challenges are already being addressed, we suspected that problems with sample quality represent another important reason for the paucity of new biomarkers and even more seriously, for incorrect biomarkers being used. Examining our own data as well as the scientific literature, we here show that sample quality issues indeed have an impact on reported results. Nearly half of the reviewed studies reported at least one potential biomarker that is in our contamination panels, and many had two or more, making sample contamination very likely. While coagulation related issues are currently most prominent, increasing depth of plasma proteome coverage may replace platelet contamination as the most important source of error in the future. A corollary of the very large abundance variation of proteins introduced by contamination is that it should further discourage pooling of samples. While this increases throughput, even a single contaminated sample can readily skew an entire batch.

Systematic bias introduced by imperfect sample handling or processing may lead to reporting incorrect biomarkers. Conversely, randomly distributed samples with poor quality will diminish overall statistical quality and may obscure true biomarker candidates.

We here provide general considerations for minimizing sample-related issues, ranging from immediate harvest of the plasma after centrifugation to discarding the lowest layer of plasma to avoid recontamination with platelets (Table 1). These recommendations update and extend general good laboratory practices as well as HUPO guidelines (*20, 27*). We also advocate that plasma samples are quality-checked by MS-based proteomics, at least for a representative subset. This is especially important for clinical studies but also for targeted, single analyte measurements, which by their nature are blind to the overall composition of the sample. Although it would be possible to determine contamination indices by multiplexed affinity-based methods, we recommend MS for this purpose because of its very high specificity and its unbiased nature. Furthermore, the proteomics depth needed to assess contamination is easily achievable even rapid and economical measurements.

The concepts and methods put forward in this study could readily be adapted to other body fluids such as urine, saliva or cerebrospinal fluid. This would require developing the appropriate contamination indices. Furthermore, the three contamination categories are the largest but not the only ones. For instance, we imagine that similar experiments can be performed to gauge the effect of storage duration and temperature on the plasma proteome as it influences MS-based proteomics.

In conclusion, sample-related quality issues are clearly a concern for biomarker studies. However, we show here that they can be addressed rigorously and comprehensively by MS-based proteomics. As this technology continues to improve in throughput, depth and robustness, we envision that it will be employed in routine clinical practice. Biomarker panels instead of single markers will be measured by MS-based proteomics as this takes advantage of its inherently multiplexed nature and allows the characterization of clinical conditions more comprehensively. These biomarker panels could routinely be extended with contamination panels as introduced here, helping to establish biomarker guided decisions in a wide variety of clinically important areas.

Acknowledgments

We thank all members of the Proteomics and Signal Transduction Group for help and discussions and in particular Igor Paron, Christian Deiml and Gaby Sowa for technical assistance, Mario Oroshi for assistance with the online resource, Nicolai Jacob Wewer Albrechtsen and Nils A. Kulak for discussion and Jürgen Cox for bioinformatics tools.

Funding

The work carried out in this project was partially supported by the Max Planck Society for the Advancement of Science, the European Union's Horizon 2020 research and innovation program with the MSmed project (no. 686547), by the Novo Nordisk Foundation (NNF15CC0001) and the BMBF grant German Biobank Alliance (BMBF 01EY1711C).

Author contributions

PEG designed, performed and interpreted the MS-based proteomic analysis of patient plasma and wrote the paper and generated the Figs. PVT and EV wrote the manuscript and performed together with LN, SD, JBM, DH, MS, AK, JBM and JB experiments and generated article text. DT and LH designed experiments, drafted practical considerations for sample preparation and worked on the article text. EV designed and established the interactive online resource. MM designed and interpreted the MS-based proteomic analysis of plasma, supervised and guided the project and wrote the manuscript.

Competing interests:

The authors declare no competing interests.

Literature

1. FDA-NIH:Biomarker-Working-Group. (Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health (US), Maryland, 2016).
2. P. E. Geyer, L. M. Holdt, D. Teupser, M. Mann, Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol* **13**, 942 (2017).
3. N. L. Anderson, A. S. Ptolemy, N. Rifai, The riddle of protein diagnostics: future bleak or bright? *Clin Chem* **59**, 194-197 (2013).
4. L. Gold *et al.*, Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).
5. E. Assarsson *et al.*, Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One* **9**, e95192 (2014).
6. B. B. Sun *et al.*, Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).
7. P. Ganz *et al.*, Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA* **315**, 2532-2541 (2016).

8. C. Herder *et al.*, A Systemic Inflammatory Signature Reflecting Cross Talk Between Innate and Adaptive Immunity Is Associated With Incident Polyneuropathy: KORA F4/FF4 Study. *Diabetes* **67**, 2434-2442 (2018).
9. R. Aebersold, M. Mann, Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).
10. R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355 (2016).
11. J. Munoz, A. J. Heck, From the human genome to the human proteome. *Angew Chem Int Ed Engl* **53**, 10864-10866 (2014).
12. D. Wild, *The immunoassay handbook : theory and applications of ligand binding, ELISA, and related techniques*. (Elsevier, Oxford ; Waltham, MA, ed. 4th, 2013), pp. xxi, 1013 p.
13. H. Mischak *et al.*, Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med* **2**, 46ps42 (2010).
14. A. N. Hoofnagle *et al.*, Recommendations for the Generation, Quantification, Storage, and Handling of Peptides Used for Mass Spectrometry-Based Assays. *Clin Chem* **62**, 48-69 (2016).
15. S. Surinova *et al.*, On the development of plasma protein biomarkers. *J Proteome Res* **10**, 5-16 (2011).
16. S. J. Skates *et al.*, Statistical design for biospecimen cohort size in proteomics-based biomarker discovery and verification studies. *J Proteome Res* **12**, 5383-5394 (2013).
17. A. S. Schrohl *et al.*, Banking of biological fluids for studies of disease-associated protein biomarkers. *Mol Cell Proteomics* **7**, 2061-2066 (2008).
18. M. E. Hassis *et al.*, Evaluating the effects of preanalytical variables on the stability of the human plasma proteome. *Anal Biochem* **478**, 14-22 (2015).
19. U. Qundos *et al.*, Profiling post-centrifugation delay of serum and plasma with antibody bead arrays. *J Proteomics* **95**, 46-54 (2013).
20. A. J. Rai *et al.*, HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics* **5**, 3262-3277 (2005).
21. J. F. Timms *et al.*, Preanalytic influence of sample handling on SELDI-TOF serum protein profiles. *Clin Chem* **53**, 645-656 (2007).
22. P. E. Geyer *et al.*, Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* **2**, 185-195 (2016).
23. P. E. Geyer *et al.*, Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol Syst Biol* **12**, 901 (2016).
24. N. A. Kulak, P. E. Geyer, M. M., Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Molecular Cellular Proteomics; Manuscript in Press*, (2017).
25. N. J. W. Albrechtsen *et al.*, Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis
markers after Roux-en-Y gastric bypass surgery. *Cell Syst*, (2018).
26. C. W. Liu *et al.*, Temporal expression profiling of plasma proteins reveals oxidative stress in early stages of Type 1 Diabetes progression. *J Proteomics* **172**, 100-110 (2018).
27. G. S. Omenn *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* **5**, 3226-3245 (2005).
28. J. M. Schwenk *et al.*, The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J Proteome Res* **16**, 4299-4310 (2017).
29. N. A. Kulak, G. Pichler, I. Paron, N. Nagaraj, M. Mann, Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* **11**, 319-324 (2014).
30. R. A. Scheltema, M. Mann, SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J Proteome Res* **11**, 3458-3466 (2012).
31. F. Meier, P. E. Geyer, S. Virreira Winter, J. Cox, M. Mann, BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods (2018); doi:10.1038/s41592-018-0003-5*, (2018).
32. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).

33. J. Cox *et al.*, Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805 (2011).
34. N. Nagaraj *et al.*, 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 (2012).
35. J. Cox *et al.*, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526 (2014).
36. S. Tyanova, Temu, T., Sinitcyn, P., Carlson, A., Hein, M., Geiger, T., Mann, M., Cox, J., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, (2016).

Materials and Methods

Samples for defining the three contamination marker panels

Whole blood was harvested by venipuncture of ten females and ten males into commercial EDTA containing sampling containers. All participants gave written informed consent for their participation in the Munich Study on Biomarker Reference Values (MyRef), which is registered under the local ethic number 11-16. The blood was centrifuged at 200 g for 10 min and both the pellet and the supernatant were kept for further processing steps. The bottom layer of 500 µl plasma was discarded to avoid contamination of the platelet-rich plasma fraction with erythrocytes. The pellet was centrifuged at 2000 g for 15 min and the top layer containing plasma, the buffy coat and 1 ml of erythrocytes were discarded. After adding 4 ml PBS containing 1.6 mg/ml EDTA, the suspension was centrifuged at 2000 g for 15 min and the supernatant was discarded together with 500 µl of the top layer of the erythrocytes. This step was repeated and the pure erythrocyte fraction was harvested. We centrifuged the supernatant from the first centrifugation step containing plasma and platelets a second time at 200 g for 10 min and harvested the supernatant, which constitutes the platelet-rich plasma. This step was repeated and we collected the supernatant and the platelet after centrifugation at 2000 g for 15 min. The supernatant was centrifuged a second time at 2000 g for 15 min to harvest platelet-free plasma by sampling only top layer of the supernatant, but discarding the bottom layer of 500 µl. The platelets were washed twice by adding 4 ml PBS containing 1.6 mg/ml EDTA and centrifugation at 2000 g for 15 min. The supernatant was discarded and the pure platelet fraction was harvested.

For the serum and plasma comparison, blood samples from two females and two males were split into 18 samples each and serum and plasma were harvested after centrifugation at 2000 g for 15 min.

To evaluate the platelet contamination in different layers of plasma after centrifugation, blood was collected in two different 9 ml S-Monovette EDTA containing sampling containers (Sarstedt). The blood of one container was transferred to a 15 ml centrifugation tube without separation gel. Both container were centrifuged at 2000 g for 15 min. Plasma was harvested in nine volume fractions starting from the top layer in 500 µl steps to the top of the buffy coat. The buffy coat itself was not touched and a small amount of plasma (~200 µl) remained on top.

High abundant protein depletion for building a matching library

We created a matching library and applied a consecutive depletion strategy, in which the top 6 and top 14 most abundant plasma proteins were depleted by using a combination of two immunodepletion kits, as described in ref. (*22*). Briefly, the Agilent Multiple Affinity Removal Spin Cartridge was used for the depletion of the top six highest abundant proteins (albumin, IgG, IgA, antitrypsin, transferrin, haptoglobin), followed by Seppro Human 14 Sigma immunodepletion for the 14 highest abundant proteins (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, prealbumin).

Following depletion, we fractionated our samples using the high-pH reversed-phase "Spider fractionator" into 24 fractions as described previously (*24*).

Sample preparation: Protein digestion and in-StageTip purification

Sample preparation was carried out according to our Plasma Proteome Profiling pipeline as described in (*22, 23*) with an automated set-up on an Agilent Bravo liquid handling platform. In brief, plasma samples were diluted 1:10 with $_{dd}H_2O$ and 10 µl of the sample were mixed with 10 µl PreOmics lysis buffer (P.O. 00001, PreOmics GmbH) for reduction of disulfide bridges, cysteine alkylation and protein denaturation at 95°C for 10 min (*29*). Trypsin and LysC were added to the mixture after a 5 min cooling step at room temperature, at a ratio of 1:100 micrograms of enzyme to micrograms of protein. Digestion was performed at 37 °C for 1 h. An amount of 20 µg of peptides was loaded on two 14-gauge StageTip plugs, followed by consecutive purification steps according to the PreOmics iST protocol (www.preomics.com). The StageTips were centrifuged using an in-house 3D-printed StageTip centrifugal device at 1500 g. The collected material was completely dried using a SpeedVac centrifuge at 60 °C (Eppendorf, Concentrator plus). Peptides were suspended in buffer A* (2% acetonitrile (v/v), 0.1% formic acid (v/v)) and sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510). Pools for each of the five sample types (whole blood, erythrocytes, platelets, plasma and platelet-free plasma) were generated from the 20 individuals and prepared according to the procedure above. The peptides were fractionated using the high-pH reversed-phase "Spider fractionator" into 24 fractions as described previously to generate deep proteomes (*24*).

Ultra-high pressure liquid chromatography and mass spectrometry

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1000 or 1200 ultra-high pressure system (Thermo Fisher Scientific), which was coupled to a Q Exactive HF Orbitrap (Thermo Fisher Scientific) using a nano-electrospray ion source (Thermo Fisher Scientific). Purified peptides were separated on 40 cm HPLC-columns (ID: 75 µm; in-house packed into the tip with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH)). For each LC-MS/MS analysis about 0.5 µg peptides were used for 45 min runs and for each fraction of the deep plasma data set.

Peptides were loaded in buffer A (0.1% formic acid, 5% DMSO (v/v)) and eluted with a linear 35 min gradient of 3-30% of buffer B (0.1% formic acid, 5% DMSO, 80% (v/v) acetonitrile), followed stepwise by a 7 min increase to 75% of buffer B and a 1 min increase to 98% of buffer B, followed by a 2 min wash of 98% buffer B at a flow rate of 450 nl/min. Column temperature was kept at 60 °C by an in-house-developed oven containing an Peltier element, and parameters were monitored in real time by the SprayQC software (*30*). MS data was acquired with a Top15 data-dependent MS/MS scan method for the construction of the library and BoxCar scans (*31*) for the study samples. Target values for the full scan MS spectra were 3×10^6 charges in the 300-1650 m/z range with a maximum injection time of 55 ms and a resolution of 60,000 at m/z 200. Fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 27 eV. MS/MS scans were performed at a resolution of 30,000 at m/z 200 with an ion target value of 1×10^5 and a maximum injection time of 120 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of identical peptides.

Data analysis

MS raw files were analyzed by MaxQuant software, version 1.5.6.8, (*32*) and peptide lists were searched against the human Uniprot FASTA database. A contaminant database generated by the Andromeda search engine (*33*) was configured with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. We

set the false discovery rate (FDR) to 0.01 for protein and peptide levels with a minimum length of 7 amino acids for peptides and the FDR was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin and LysC as proteases. A maximum of two missed cleavages were allowed. Peptide identification was performed with an initial precursor mass deviation up to 7 ppm and a fragment mass deviation of 20 ppm. The 'match between run algorithm' in the MaxQuant quantification (*34*) was enabled after constructing a matching library consistent of depleted and all the undepleted plasma samples. All proteins and peptides matching to the reversed database were filtered out. Label-free protein quantitation (LFQ) was performed with a minimum ratio count of 2 (*35*).

Bioinformatics analysis

All bioinformatics analyses were performed with the Perseus software of the MaxQuant computational platform (*32, 36*). For the global correlation analysis, proteins were filtered for at least 50% valid values in the weight loss study and the hierarchical clustering was performed using Euclidean distance. The weight loss study contained in total 28 proteins of the platelet panel, but after sorting for 50% valid values only 24 were left and all of them clustered in the platelet panel.

Online platform for automated analysis of clinical studies

Our online portal is equipped with a user-friendly graphical interface that supports the most common web browsers, such as Google Chrome, Firefox and Internet Explorer. For the front-end development, a Dash framework was used (version 0.27.0), which consists of a Flask server (1.0.2) that communicates with front-end React.js components using JSON, or JavaScript Object Notation, packets (a minimal, readable format for structuring data) over HTTP, or Hypertext Transfer Protocol, requests that work as request-response protocols between a client and server. Taking advantage of the full power of Cascading Style Sheets (CSS), every graphical element was customized: the sizing, the positioning, the colors, and the fonts.

The platform takes the results of the MS data processed by the popular MaxQuant software (*32*) integrated with the Andromeda search engine (*33*), which is a so called ProteinGroups table (to be extended to other formats). During the data uploading, the input file is verified through a combination of preliminary tests. To build a complex data structure using general Python libraries, such as NumPy, Pandas and SciPy, we used three panels of markers for platelet contamination, erythrocyte contamination and coagulation events in plasma samples, respectively, to identify samples affected by quality issues. Samples having at least 50% 'valid values' (i.e. those with quantification results), are preprocessed by cleaning the data and prepare them for the subsequent visualization step.

Data and materials availability: The MS-based proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository and are available via ProteomeXchange with identifier PXD011749.

Supplemental Figures

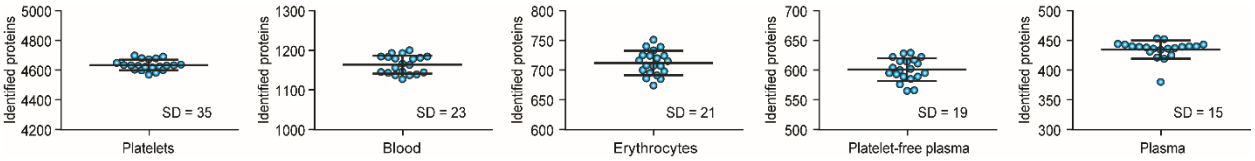


Fig. S1: Number of identified proteins per sample type for the 20 study participants. The number of identified proteins is shown for all individuals. The whiskers indicate the standard deviation (SD) and the mean is also indicated.

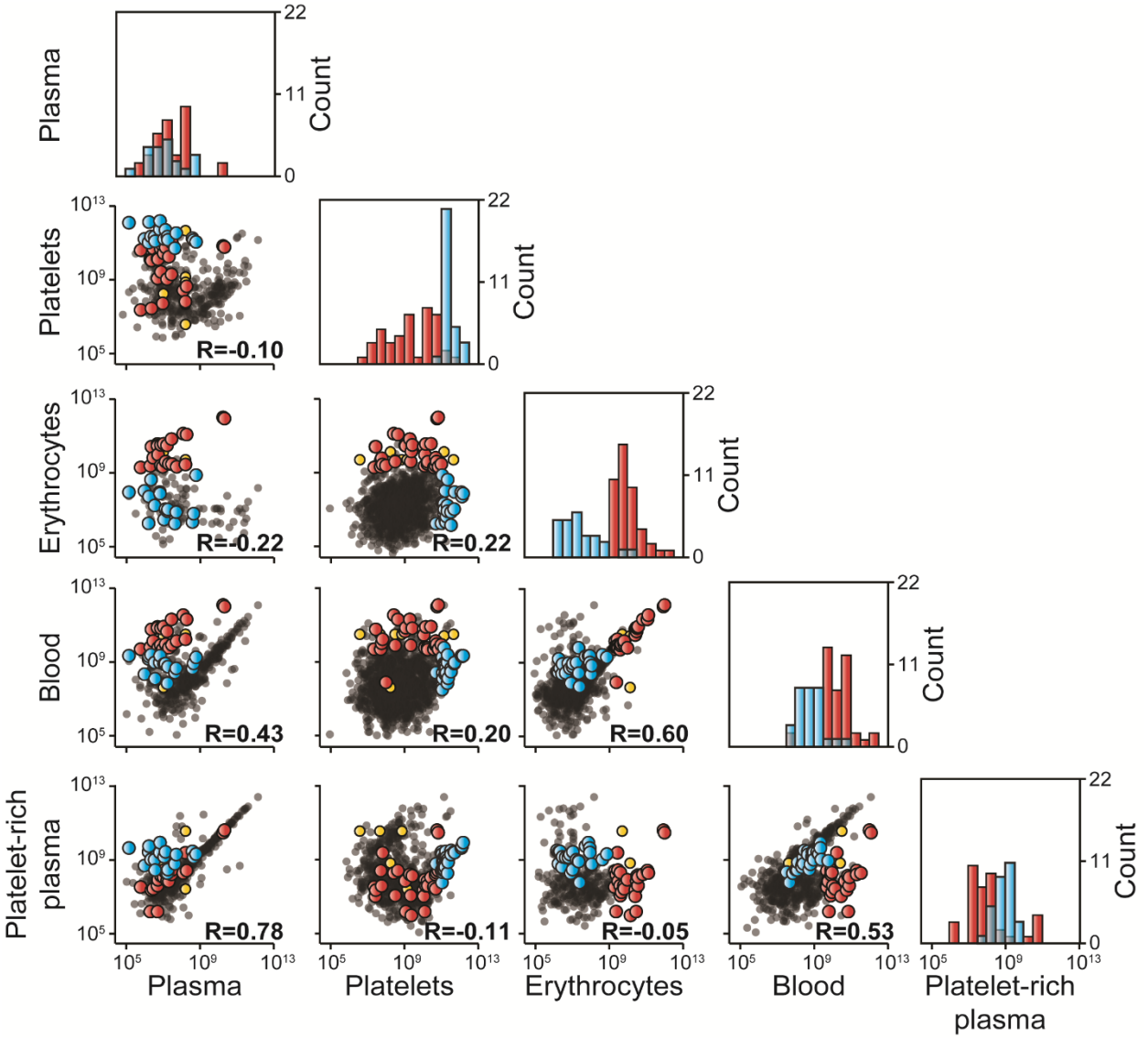


Fig. S2: Correlation of the main blood fractions. Correlation of the median proteomes of whole blood, erythrocytes, platelets, platelet-rich plasma and platelet-free plasma from 20 individuals. The 30 highest abundant proteins of erythrocytes and platelets are highlighted in red and blue, respectively. Proteins highlighted in yellow overlap between the top 30 proteins in erythrocytes and platelets. The histograms show the distribution of both marker panels over the abundance range of the plasma, platelet, erythrocyte, blood and platelet-rich plasma proteome, respectively.

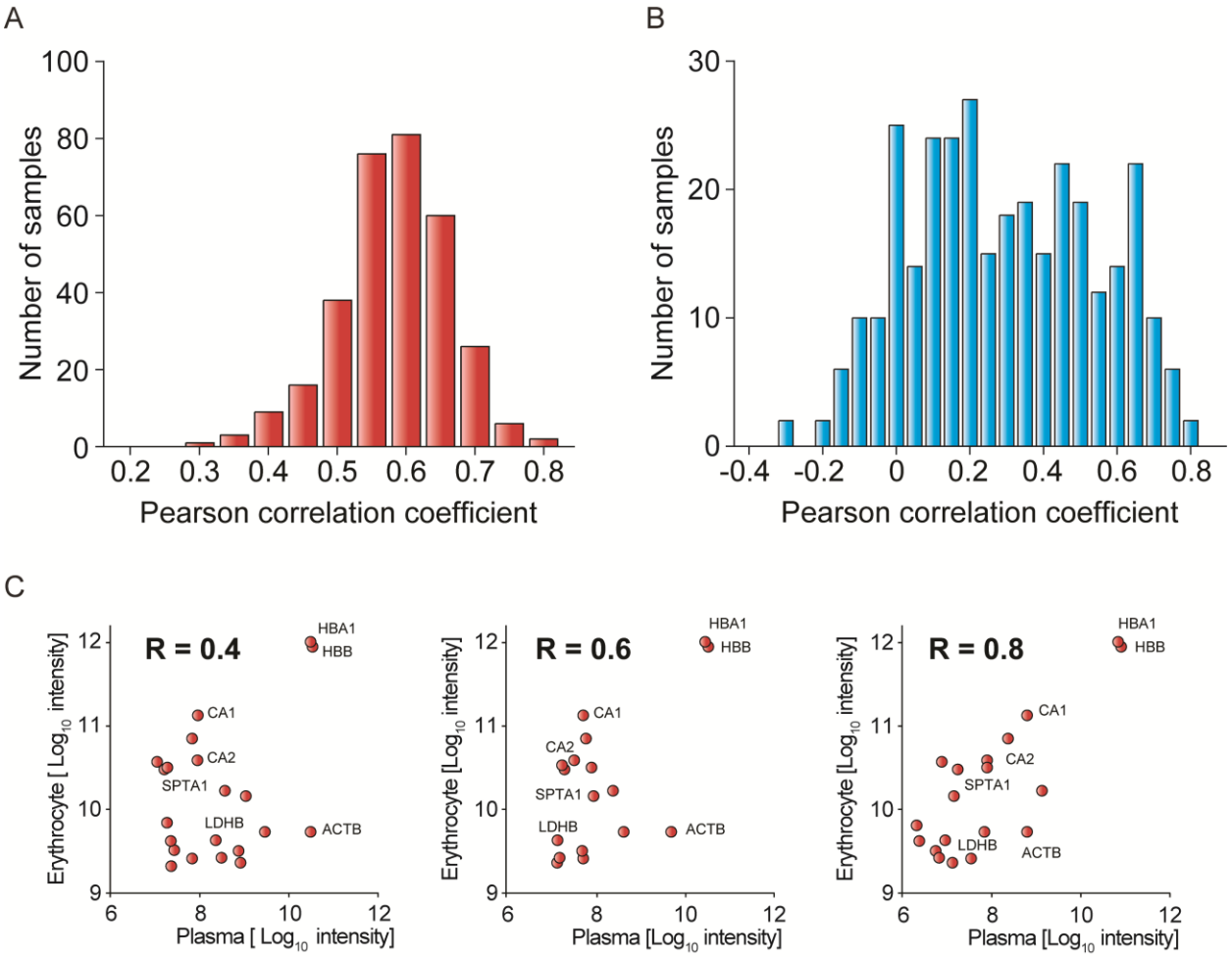


Fig. S3: Correlation of the contamination panels to plasma protein levels in the weight loss study. Histogram of Pearson correlation coefficients calculated between the reference cohort and the plasma samples in the weight loss study for the erythrocyte contamination panel (*23*). Distribution of the Pearson correlation coefficients for the platelet panel. (C) Exemplified correlations for the erythrocyte panel in three samples.

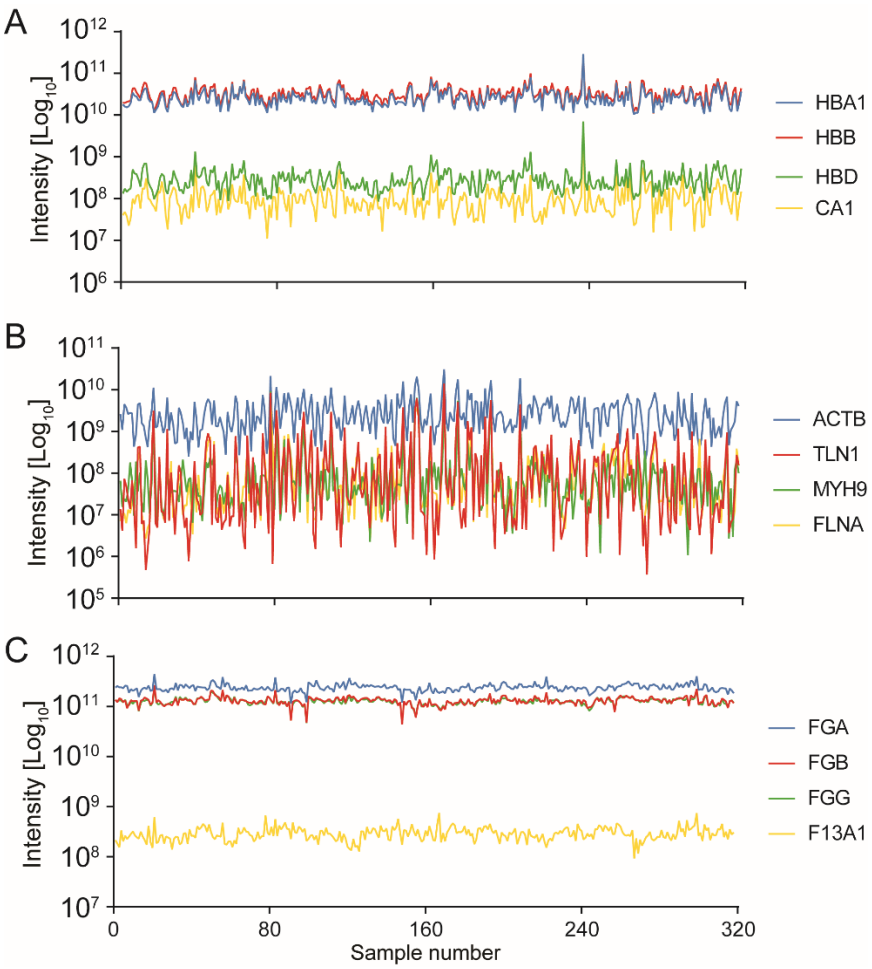


Fig. S4: Example trajectories for the top four proteins of each panel across all samples in the weight loss study. (A) Intensities of the four highest abundant erythrocyte specific proteins for all samples. (B) Intensities of the four highest abundant platelet proteins. (C) Intensities of the four most significantly regulated coagulation markers.

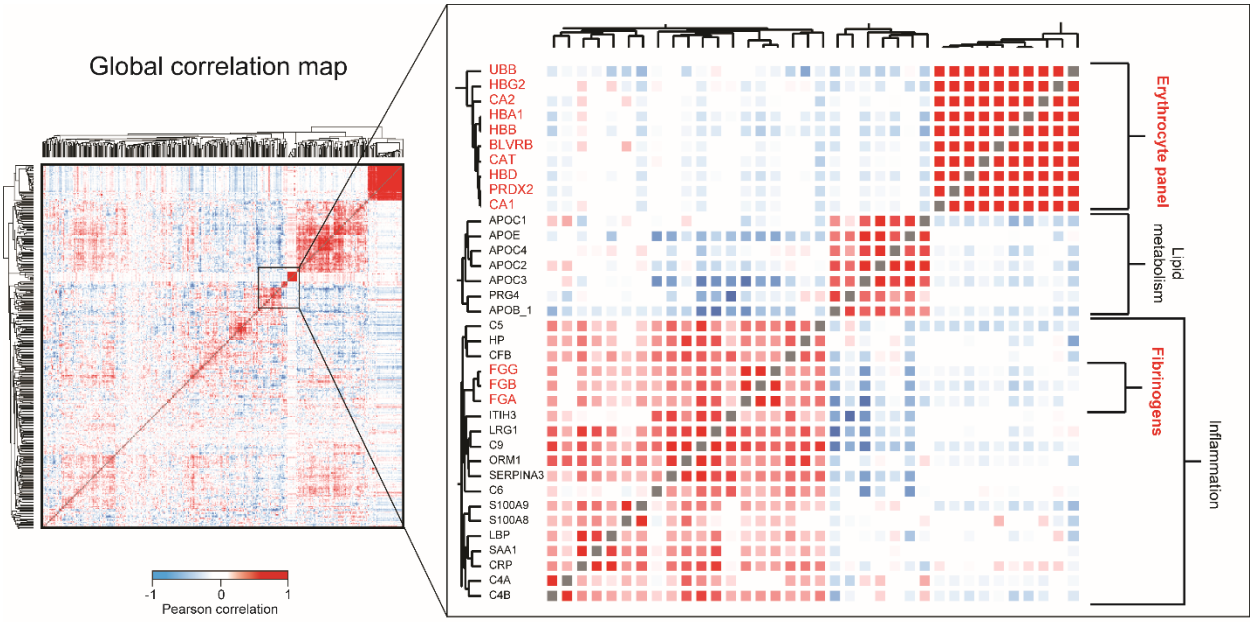


Fig. S5: Erythrocyte and coagulation marker in the global correlation map. The global correlation map is shown on the left and the magnified inset shows three clusters of correlating proteins. The erythrocyte panel and the fibrinogens are highlighted in red. The color-code for the Pearson correlation coefficient is indicated.

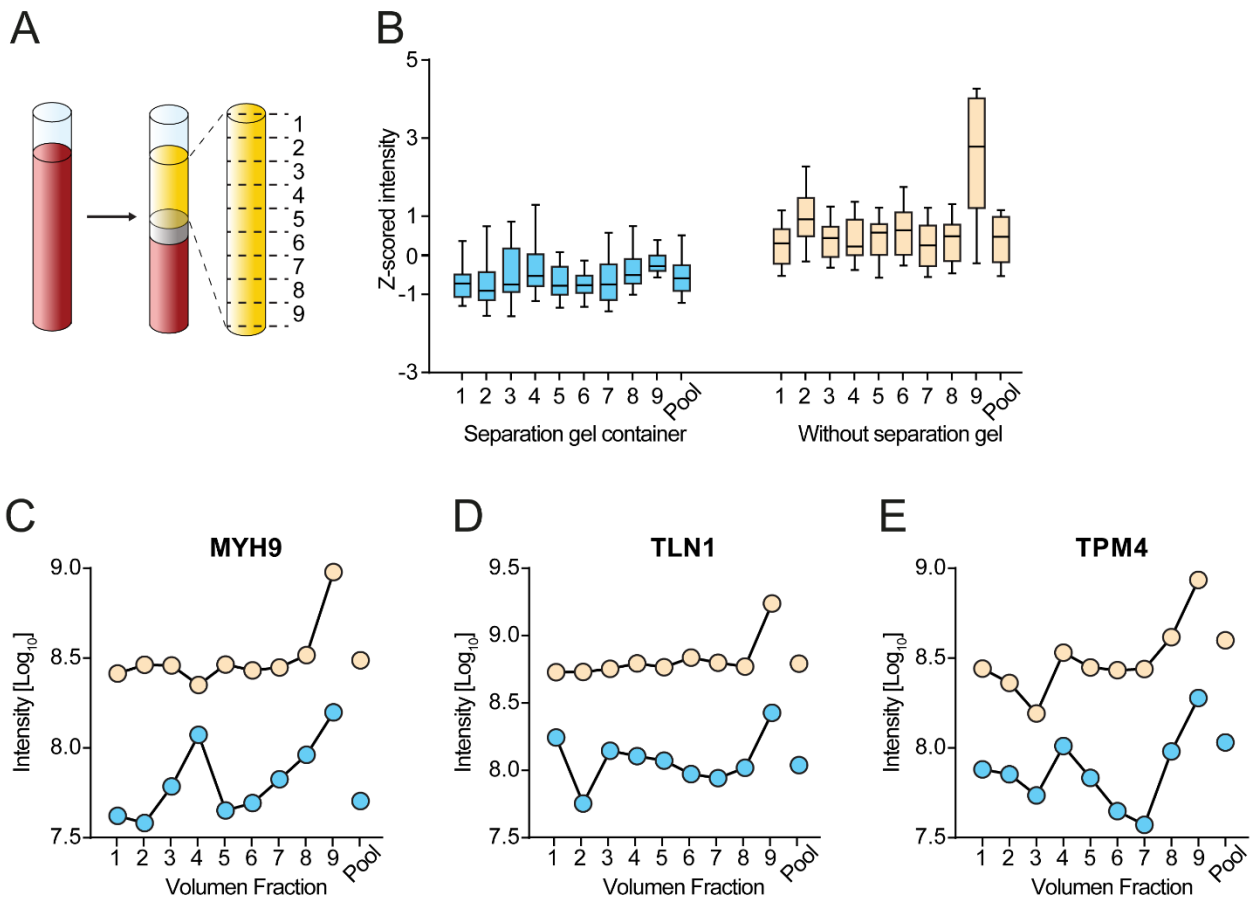


Fig. S6: Distribution of platelet proteins in different plasma volume fractions after centrifugation. (A) Plasma from nine different layers were harvested starting from the top after centrifugation to the top above the buffy coat in 500 µl steps. (B) The boxplots indicate the Z-scores of 27 of the top 30 platelet proteins that were quantified with at least 50% valid values in this experiment for the volume fractions 1-9 and the pool of all layers. The whiskers indicate the 10-90% quartile and the horizontal line within the boxplots is the median. (C-E) Examples for three platelet proteins for the three plasma collection protocols with intensity values, illustrating the changes within protocol and the volume fractions. Data points of the centrifugation container with and without a gel plug are color coded in blue and beige, respectively.

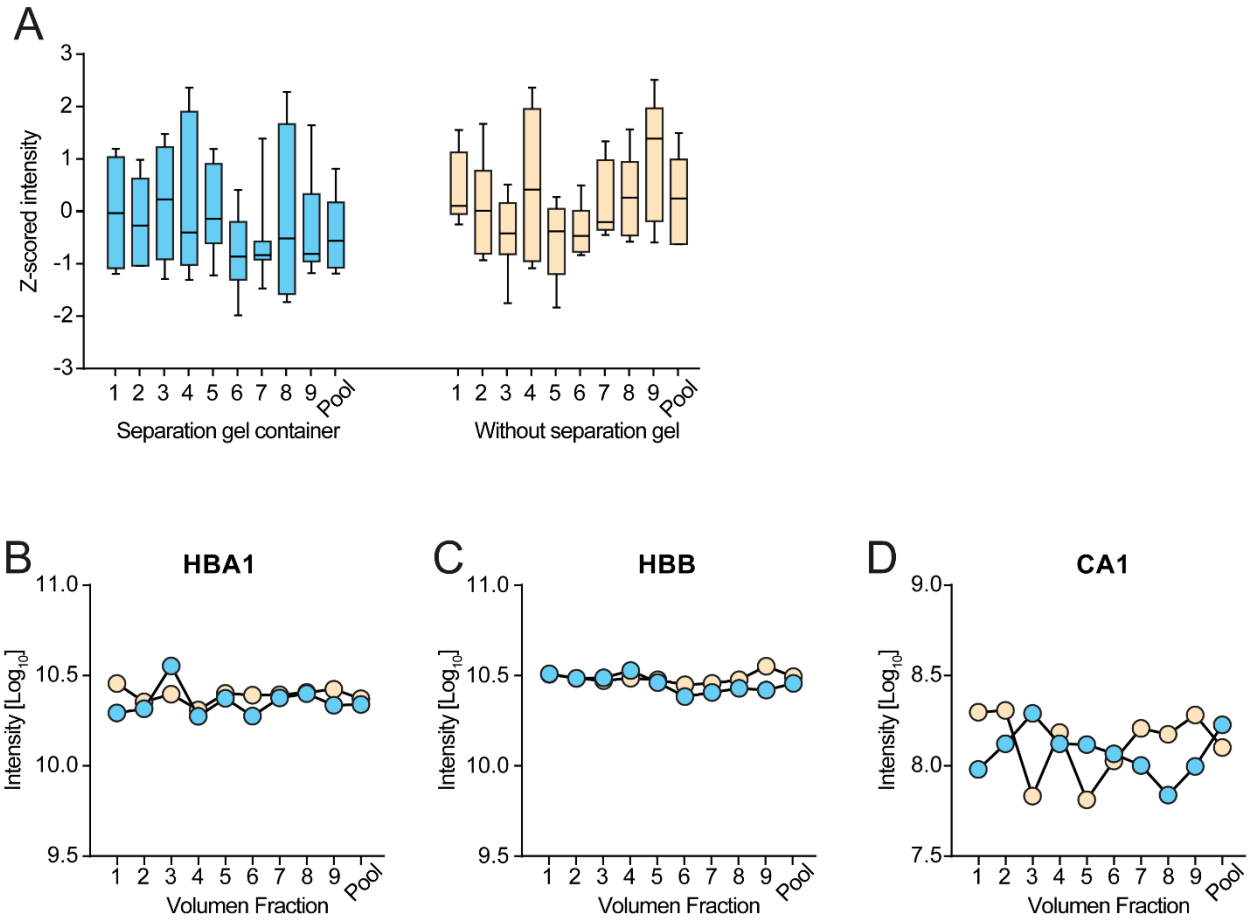


Fig. S7: Distribution of erythrocyte proteins in different plasma volume fractions after centrifugation. (A) Volume fractions 1-9 and the pool of all layers for three different plasma collection protocols. The boxplots indicate the Z-scores of 7 of the top 30 erythrocyte proteins that were quantified with at least 50% valid values in this experiment. The whiskers indicate the 10-90% quartile and the horizontal line within the boxplots is the median. (C-E) Examples for three erythrocyte proteins for the three plasma collection protocols with intensity values, illustrating the changes within protocol and the volume fractions. The color code is according to panel B indicating the three protocols.