scientific reports

Check for updates

OPEN

Assessment of a complete and classified platelet proteome from genome-wide transcripts of human platelets and megakaryocytes covering platelet functions

Jingnan Huang[1,2]✉, Frauke Swieringa[1,2,9], Fiorella A. Solari[2,9], Isabella Provenzale[1], Luigi Grassi[3], Ilaria De Simone[1], Constance C. F. M. J. Baaten[1,4], Rachel Cavill[5], Albert Sickmann[2,6,7,9], Mattia Frontini[3,8,9] & Johan W. M. Heemskerk[1,9]✉

Novel platelet and megakaryocyte transcriptome analysis allows prediction of the full or theoretical proteome of a representative human platelet. Here, we integrated the established platelet proteomes from six cohorts of healthy subjects, encompassing 5.2 k proteins, with two novel genome-wide transcriptomes (57.8 k mRNAs). For 14.8 k protein-coding transcripts, we assigned the proteins to 21 UniProt-based classes, based on their preferential intracellular localization and presumed function. This classified transcriptome-proteome profile of platelets revealed: (i) Absence of 37.2 k genome-wide transcripts. (ii) High quantitative similarity of platelet and megakaryocyte transcriptomes (R = 0.75) for 14.8 k protein-coding genes, but not for 3.8 k RNA genes or 1.9 k pseudogenes (R = 0.43–0.54), suggesting redistribution of mRNAs upon platelet shedding from megakaryocytes. (iii) Copy numbers of 3.5 k proteins that were restricted in size by the corresponding transcript levels (iv) Near complete coverage of identified proteins in the relevant transcriptome (log2fpkm > 0.20) except for plasma-derived secretory proteins, pointing to adhesion and uptake of such proteins. (v) Underrepresentation in the identified proteome of nuclear-related, membrane and signaling proteins, as well proteins with low-level transcripts. We then constructed a prediction model, based on protein function, transcript level and (peri)nuclear localization, and calculated the achievable proteome at ~ 10 k proteins. Model validation identified 1.0 k additional proteins in the predicted classes. Network and database analysis revealed the presence of 2.4 k proteins with a possible role in thrombosis and hemostasis, and 138 proteins linked to platelet-related disorders. This genome-wide platelet transcriptome and (non)identified proteome database thus provides a scaffold for discovering the roles of unknown platelet proteins in health and disease.

[1]Department of Biochemistry, CARIM, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. [2]Leibniz-Institut Für Analytische Wissenschaften-ISAS-E.V, Dortmund, Germany. [3]Department of Haematology, University of Cambridge, National Health Service Blood and Transplant (NHSBT), Cambridge Biomedical Campus, Cambridge, UK. [4]Institute for Molecular Cardiovascular Research (IMCAR), University Hospital RWTH, Aachen, Germany. [5]Department of Data Science and Knowledge Engineering, FSE, Maastricht University, Maastricht, The Netherlands. [6]Medizinische Fakultät, Medizinische Proteom-Center, Ruhr-Universität Bochum, Germany. [7]Department of Chemistry, College of Physical Sciences, University of Aberdeen, Aberdeen, UK. [8]Institute of Biomedical & Clinical Science, College of Medicine and Health, University of Exeter Medical School, Exeter, UK. [9]These authors contributed equally: Frauke Swieringa, Fiorella A. Solari, Albert Sickmann, Mattia Frontini and Johan W. M. Heemskerk. ✉email: j.huang@maastrichtuniversity.nl; jwmheem722@outlook.com

Platelets are generated in the bone marrow as cell fragments from hematopoietic stem cells that are differentiated into megakaryocytes. In the circulating, the mature platelets control many blood-related processes both in health and disease. These functions extend from blood vessel-lymph separation and maintenance of vascular integrity to allowing hemostasis, promoting arterial thrombosis, regulating inflammatory, immune and infection processes; and even facilitating tumor progression[1,2]. The ultrastructure and the protein/RNA composition of a platelet, determined during their ontogenesis, allows the execution of all these functions. However, comparative studies of the molecular composition and structure of platelets in relation to their functions and megakaryocytic origin are still missing.

Although platelets do not contain a nucleus, they are equipped with mitochondria, several types of storage granules and multiple intracellular membrane structures, including endoplasmic reticulum (smooth and rough), a likely rudimentary Golgi apparatus, lysosomes, peroxisomes and endosomes[3–5]. Characteristic large invaginations, designated as open canicular or dense tubular system, make up ~ 1% or the cell volume and are filled with blood plasma components. A well-developed actin-myosin and tubulin cytoskeleton is required for proplatelet formation, micro- organization of the membrane structures, and mediates activation-dependent structural changes[6–9]. Whether the full repertoire of metabolic enzymes is present in platelets is still unclear, while the glucose metabolism is well-developed[10,11]. Furthermore, the ribosomal mRNA translation machinery is retained as well as elements of protein processing and trafficking and a repertoire of proteolytic processes in the proteasome[12,13]. Overviews point to a battery of receptors and channels, multiple adaptor molecules and small molecule GTP-binding proteins (G-proteins), and large protein kinase and phosphatase networks[2,14].

Human genetic studies supported by mouse models show that hundreds and possibly thousands of platelet-expressed proteins contribute to thrombosis and hemostasis[15]. We reasoned that assembling the complete (quantitative) proteome and transcriptome of human platelets can provide a much better understanding of the molecules that determine platelet structure and functions in health and disease. As earlier platelet proteomes, reported in single articles, are limited in the numbers of identified proteins[16–18], there is a need to integrate multiple proteomic studies based on the same methodology. While the number of genes detected in available transcriptomes of platelets and megakaryocytes are a magnitude higher[19–21], these do not extend to the whole genome. Here, we combined multiple proteomes with the genome-wide RNA database of platelets and megakaryocytes generated by the Blueprint consortium[22,23], and integrated these into a platelet structure and function- based protein classification system, for defining the full platelet proteome. Detailed analysis of this database provided novel insights into the structure–function relations of platelets.

Results

Function-based classification of platelet proteins in merged proteome. Considering that the previously published (phospho)proteomics profiles of highly purified platelets from 22 healthy subjects in 6 cohorts were generated by the same analytical workflow[24–29], we decided to integrate these datasets (Suppl. Figure 1A). Primary sources of these datasets are listed in Table 1. The resulting, merged human platelet proteome—one of the largest described so far—contained a total of 5,211 identified proteins, of which 80% were present in at least 2 cohorts (Suppl. Datafile 2). For 3,629 of these proteins, also copy numbers per platelet were present. In order to obtain a useful knowledgebase, we then categorized these proteins into 21 classes, based on intracellular localization and function (Fig. 1A). For an objective classification, we used a dichotomous decision scheme together with human UniProt-KB assignments regarding the supposed primary location and/or function of that protein (Fig. 1B). Highest fractions of identified proteins were seen in the following classes (Suppl. Figure 1B): C_{20} (transcription & translation, $n = 488$ proteins), C_{12} (other metabolism, $n = 475$), C_{18} (signaling & adaptor proteins, $n = 471$), C_{11} (mitochondrial proteins, $n = 455$), and C_{10} (membrane receptors & channels, $n = 327$). Distribution profiles of the 3,629 proteins with copy numbers (Suppl. Figure 1C) showed highest abundance and gene expression levels of the classes: C_{01} (cytoskeleton actin- myosin), C_{07} (glucose metabolism) and C_{04} (cytoskeleton receptor-linked). This clustering analysis hence underscored the importance in platelets of signaling, mitochondrial and cytoskeletal proteins[2].

Relevant genome-wide transcriptomes of platelets and megakaryocytes. Based on well-purified human platelet and megakaryocyte preparations, the Blueprint consortium[30,31] has recently generated one of the largest databases with genome-wide, quantitative information on a total of 57.8 k transcripts in either cell type (Fig. 2, for source see Table 1). Examination of the distribution pattern of all gene-linked transcripts indicated that 37.2 k of these were essentially absent (log2fpkm 0.02–0.03 ± 0.03, mean ± SD) in platelets (Fig. 3A) and megakaryocytes (Fig. 3B). The residual presence of ~ 20 k expressed transcripts supports earlier analyses of the comparative transcriptomes of blood cells[19]. We then combined these Blueprint datasets with the combined proteome data to come to a draft full platelet proteome.

Based on a low threshold of log2fpkm ≥ 0.20 for relevant expression levels (see below), we obtained a defined set of 20.4 k transcripts, which was taken to assemble the relevant transcriptomes for platelets (17.6 k) and megakaryocytes (16.8 k). Comparison between cell types gave a same distribution pattern ($p > 0.10$, χ^2) for platelets and megakaryocytes (Fig. 3C,D). Filtering for transcripts of the 5.2 k identified platelet proteins, again resulted in similar distribution patterns (Fig. 3E,F). In either cell type, the lower level transcripts (log2fpkm < 1.00) were under-represented in comparison to the unfiltered genome-wide distribution ($p = 0.049$, χ^2).

Correlational analysis learned that the platelet and megakaryocyte transcriptomes were highly correlated; this was the case for both the 57.3 k genome-wide transcripts (log2fpkm ≥ 0.00, $R = 0.85$, $\beta > 0.99$) and the 20.4 k transcripts with relevant expression levels in either/both cell types (log2fpkm ≥ 0.20, $R = 0.75$, $\beta > 0.99$; Suppl. Figure 2A, B). This markedly revealed high similarity of the RNA species composition in human platelets and megakaryocytes. Concerning different RNA biotypes, this correlation remained high, when extracting only the

Cohort 1[24]
a. https://ashpublications.org/blood/article/120/15/e73/30645/The-first-comprehensive-and-quantitative-analysis
b. Supplemental Table S2 and S3: identified phosphopeptides and proteins
c. Pride repository (http://www.ebi.ac.uk/pride/), accessions 22201-22203, 22,206
Cohort 2[25]
a. https://ashpublications.org/blood/article/123/5/e1/32883/Time-resolved-characterization-of-cAMP-PKA
b. Supplemental Table S3 and S4: identified phosphopeptides and proteins
c. ProteomeXchange repositories PXD002883 and 10.6019/PXD002883
Cohort 3[26]
a. https://ashpublications.org/blood/article/129/2/e1/36101/Temporal-quantitative-phosphoproteomics-of-ADP
b. Supplemental Table 1 and 2: identified phosphopeptides and proteins
c. ProteomeXchange repository PXD001189
Cohort 4[27]
a. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5054341/
b. Supplemental Table 1 and 2: identified phosphopeptides and proteins
c. ProteomeXchange repositories PXD002883 and 10.6019/PXD002883
Cohort 5[28]
a. https://www.nature.com/articles/s41598-020-68379-3#Sec25
b. Datafile S1 and Datafile S2: identified phosphopeptides and proteins
c. ProteomeXchange repository PXD016534
Cohort 6[29]
a. https://ashpublications.org/blood/article/114/1/e10/26099/Platelet-membrane-proteomics-a-novel-repository
b. Table S4: list of proteins and peptides
c. Pride repository (http://www.ebi.ac.uk/pride/init.do), accessions 8127–8129
PLT and MGK transcriptomes
a. https://doi.org/10.3324/haematol.2019.238147
b. Transcript levels: https://blueprint.haem.cam.ac.uk/mRNA/
c. Deposited at BioRxiv https://doi.org/10.1101/764613

Table 1. Accessibility per proteome cohort of website link (a), used raw datasets (b) and deposited spectral data (c).

protein-coding genes (14.8 k, $R = 0.75$, $\beta > 0.99$), but it reduced for the 3.8 k RNA genes and 1.9 k pseudogenes ($R = 0.43–0.54$) (Suppl. Figure 2C-E).

For justification of the relevant transcript threshold for protein expression, we reduced this further from log2fpkm 0.20 to 0.15; this resulted in inclusion of no more than 8 extra proteins from the combined proteome, half of it being plasma-derived proteins and the other half with minimal copy numbers. This indicated that log2fpkm of 0.20, although arbitrary, provides a reasonable cutoff value for transcripts resulting in measurable proteins.

Using the combined knowledgebase of platelets and megakaryocytes, we assessed which of the 20.4 k expressed transcripts (log2fpkm ≥ 0.20) were also present in the 5.2 k platelet proteome (Fig. 2). It appeared that the majority of proteins had relevant transcription levels. In 19 of the 21 protein function classes only 1.6% of the protein transcripts were below the cut-off (77/4,907 with log2fpkm 0.04 ± 0.05, mean ± SD, $n = 19$) (Table 2). However, in the classes C_{02} (cytoskeleton intermediate) and C_{17} (secretory proteins), percentages of below cut-off were much higher, amounting to 58% and 24%, respectively.

Given the analysis above, we considered that the combined platelet and megakaryocyte transcriptome (either log2fpkm ≥ 0.20) may provide the most extensive list of mRNAs that can be translated into proteins. To evaluate this, we performed the same analysis as above for the platelet-only transcriptome. This resulted in a number of 'false' assignments of 181 (Table 2). For the megakaryocyte-only transcriptome data, this number increased to 329. Accordingly, the combined list of relevant platelet and megakaryocyte transcripts appeared to provide the best overlap with the proteomics dataset. By confining to proteins with relevant mRNA expression, the identified platelet proteome was therefore set at 5,050 proteins.

Comparison of (non-)identified parts of the platelet proteome. We then reasoned that starting from the genome-wide transcriptome of platelets and megakaryocytes (log2fpkm ≥ 0.20), it was possible to construct a 'full' theoretical platelet proteome and compare this with the identified platelet proteins. By thus comparing the identified proteins with the transcripts of protein-coding genes, we could calculate the remaining, non-identified part of the proteome at 9,721 proteins, *i.e.* 66% of all mRNA transcripts (Suppl. Figure 3A). Based on this analysis, the majority of the 14.8 k proteins in the theoretical proteome was still absent in the current platelet proteomes. A similar number of 14.3 k was obtained when only including the relevant transcripts of platelets (Suppl. Figure 3B,C).

Detailed examination of the genes for which no protein products were detected revealed marked differences between function classes (Fig. 4A,B). Highest numbers and percentages of transcripts of the 'missing' proteins

A

No	Protein function class
01	Cytoskeleton actin-myosin
02	Cytoskeleton intermediate
03	Cytoskeleton microtubule
04	Cytoskeleton receptor-linked
05	Endosome proteins
06	ER & Golgi proteins
07	Glucose metabolism
08	Lysosome & peroxisome proteins
09	Membrane & protein trafficking
10	Membrane receptors & channels
11	Mitochondrial proteins
12	Other metabolism
13	Other nuclear proteins
14	Proteasomal proteins
15	Protein kinases & phosphatases
16	Protein processing
17	Secretory proteins
18	Signaling & adapter proteins
19	Small GTPases & regulators
20	Transcription & translation
21	Uncharacterized & other proteins
*	*pseudogene*
**	*RNA gene*

B

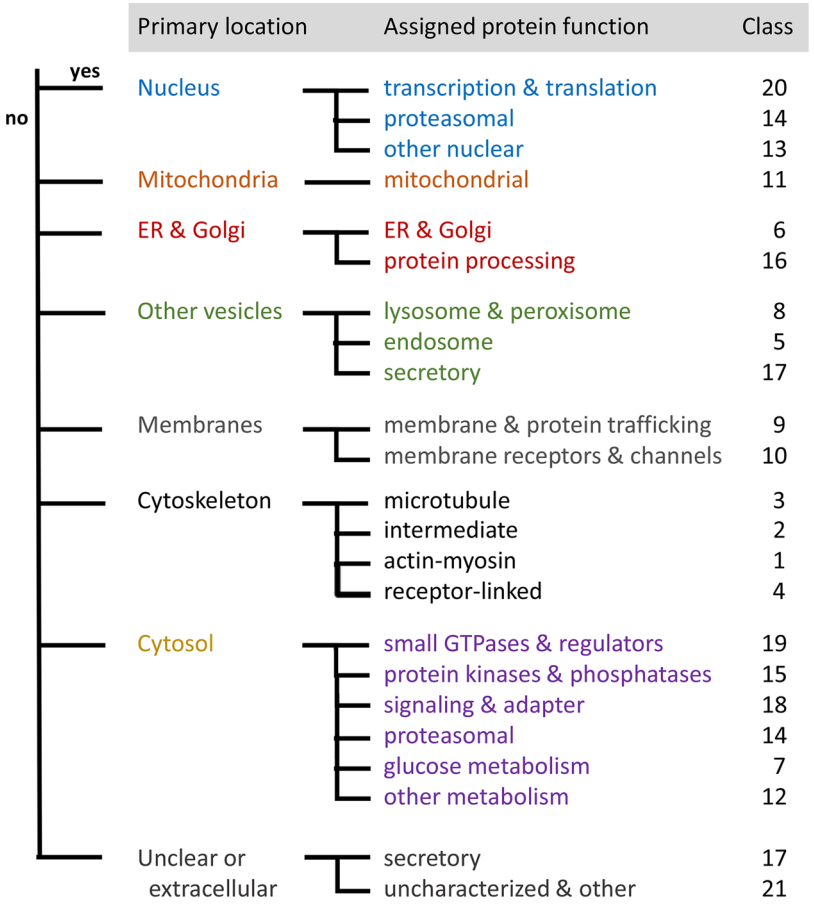


Figure 1. Classification scheme and decision tree for gene and protein assignment to 21 function classes. Assignment was based on primary subcellular localization of the protein and its assumed function according to UniProt-KB. (**A**) Class numbering in alphabetical order. (**B**) Hierarchical decision tree.

were obtained for: C_{20} (transcription & translation, $n = 1{,}795$), C_{21} (uncharacterized and other proteins, $n = 1{,}683$), C_{13} (other nuclear proteins, $n = 1{,}269$), C_{10} (membrane receptors & channels, $n = 1{,}112$), C_{17} (secretory proteins, $n = 583$), and C_{18} (signaling & adapter proteins, $n = 561$). This prompted us to investigate the reasons for these inter-class differences in coverage of the identified proteome.

Restraining factors for a complete platelet proteome. Acknowledging current mass-spectrometry limitations (see Suppl. Methods), we hypothesized that absence of mRNA products can be explained by three restraining factors: (i) low protein copy number, (ii) low mRNA level, and/or (iii) retaining of a protein in the megakaryocyte perinuclear region. The annotated platelet and megakaryocyte transcriptome knowledgebase allowed us to estimate these restraining factors.

The relation between platelet copy numbers and transcript levels is still unclear[32,33]. To reassess this issue, we compared the relevant Blueprint transcriptome (log2fpkm ≥ 0.20) with the 3.5 k proteins with known copy numbers. Correlative scatter plots showed a marked triangular pattern (Fig. 5A,B). This pattern indicated that the abundance of a protein was restricted by, but was not otherwise dependent of the transcript level. Given the high similarity of the platelet and megakaryocyte transcriptomes, this implied that the megakaryocytic mRNA levels in fact maximized the extent of protein expression in platelets.

To examine this further, we defined five regions in the proteome- transcriptome space, labeled as areas I-V (Fig. 5C). For each of 3.5 k quantified proteins, we performed a modeling analysis per function class in Matlab. This modelling revealed that—regardless of the use of platelet or megakaryocyte plots—several classes were significantly over-represented ($p = 10^{-2}$ to 10^{-10}) in some of these areas (Suppl. Table 1). As illustrated in Fig. 5D, for area I (high copy number and high mRNA), four classes were over-represented (i.e., cytoskeletal and glucose- metabolism proteins, $p < 10^{-2}$). For the areas II and III with low copy numbers ('low translation'), six and three classes were over-represented, respectively (*e.g.,* signaling-related, proteasomal, transcriptional and mitochondrial proteins). Thus, the classes accumulating in areas II-III appeared to be enriched in proteins with low copy numbers, irrespective of their corresponding transcript levels. Area V (low transcript levels) was enriched in keratin-like and secretory proteins (classes C_{02} and C_{17}); and area IV of medium mRNA levels contained most of the remaining classes.

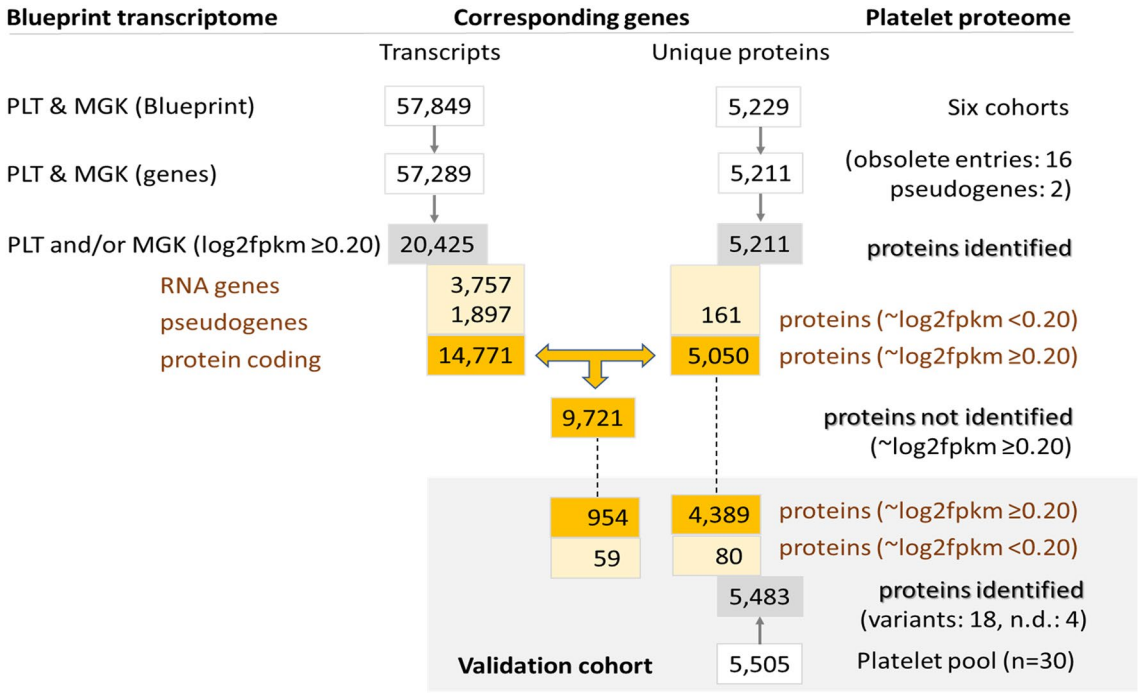


Figure 2. Dataflow of numbers of transcripts of proteome proteins. Relevant transcripts were defined as those of log2fpkm ≥ 0.20. Identified proteins refer to proteins present in the combined proteome from six cohorts. Non-identified proteins refer to proteins with relevant transcript levels in the combined PLT and MGK transcriptome. Data from validation cohort are also indicated.

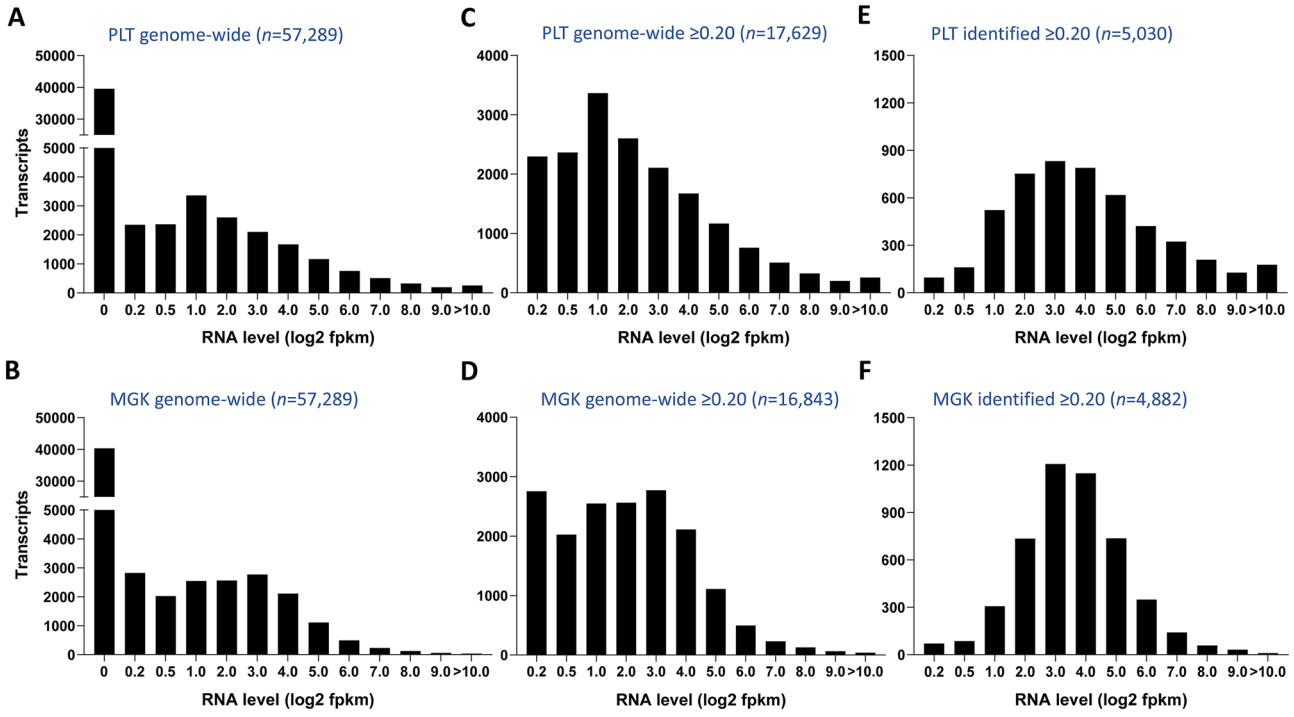


Figure 3. Histograms of RNA levels in transcriptome of platelets (PLT) or megakaryocytes (MGK). (**A**,**B**) Distribution of all 57,289 genome-wide transcripts. (**C**,**D**) Distribution of all relevant transcripts (log2fpkm ≥ 0.20) for PLT ($n = 17,629$) or MGK ($n = 16,843$). (**E**,**F**) Distribution of protein-coding transcripts, as identified in the proteome, for PLT ($n = 5,030$) or MGK ($n = 4,882$). Levels of RNA expression (log2fpkm) were binned as < 0.20, 0.20–0.50, 0.50–1.00, 1.00–2.00, etc. For flow of numbers of transcripts and proteins, see Fig. 2.

A	Proteins in transcriptome			B	Proteins in transcriptome		C	Proteins in transcriptome		D
PLT and/or MGK	yes ≥0.20	no <0.20	% False	PLT	yes ≥0.20	no <0.20	MGK	yes ≥0.20	no <0.20	Total
C01 Cytoskeleton actin-myosin	132	9	*6.4*	C01	132	9	C01	125	16	141
C02 Cytoskeleton intermediate	8	11	*57.9*	C02	8	11	C02	6	13	19
C03 Cytoskeleton microtubule	140	6	*4.1*	C03	139	7	C03	133	13	146
C04 Cytoskeleton receptor-linked	51	0	-	C04	51	0	C04	48	3	51
C05 Endosome proteins	52	0	-	C05	52	0	C05	52	0	52
C06 ER & Golgi proteins	190	1	*0.5*	C06	190	1	C06	190	1	191
C07 Glucose metabolism	46	1	*2.1*	C07	46	1	C07	45	2	47
C08 Lysosome & peroxisome proteins	74	1	*1.3*	C08	74	1	C08	74	1	75
C09 Membrane & protein trafficking	243	4	*1.6*	C09	242	5	C09	240	7	247
C10 Membrane receptors & channels	318	9	*2.8*	C10	315	12	C10	300	27	327
C11 Mitochondrial proteins	454	1	*0.2*	C11	454	1	C11	454	1	455
C12 Other metabolism	469	6	*1.3*	C12	467	8	C12	463	12	475
C13 Other nuclear proteins	200	3	*1.5*	C13	200	3	C13	196	7	203
C14 Proteasomal proteins	311	1	*0.3*	C14	310	2	C14	311	1	312
C15 Protein kinases & phosphatases	266	2	*0.7*	C15	266	2	C15	264	4	268
C16 Protein processing	199	1	*0.5*	C16	197	3	C16	198	2	200
C17 Secretory proteins	228	73	*24.3*	C17	223	78	C17	155	146	301
C18 Signaling & adapter proteins	463	8	*1.7*	C18	462	9	C18	446	25	471
C19 Small GTPases & regulators	284	3	*1.0*	C19	284	3	C19	281	6	287
C20 Transcription & translation	485	3	*0.6*	C20	484	4	C20	482	6	488
C21 Uncharacterized & other proteins	437	18	*4.0*	C21	434	21	C21	419	36	455
Total identified proteins	**5050**	**161**			5030	181		4882	329	
Pseudogenes	2	0			2	0		2	0	
Obsolete entries	0	16			0	16		0	16	
Together	5052	177	**5229**		5032	197		4884	345	

Table 2. Identified proteins in proteome in comparison to relevant transcriptome of platelets (PLT) and/or megakaryocytes (MGK). Indicated per function class are numbers of proteins with relevant (log2fpkm ≥ 0.20) or no relevant (log2fpkm < 0.20) mRNA expression. Analyzed were the combined PLT/MGK transcriptome (**A**), as well as the separate PLT (**B**) and MGK (**C**) transcriptomes. For the total of 5,232 identified proteins in the proteome, 2 appeared to be encoded by pseudogenes, and 16 were designated as obsolete entries in UniProt-KB. Also given are percentages of proteins without relevant expression level (% false). (**D**) Total numbers of assigned proteins per class independent of transcript level.

To categorize the low-level mRNAs, we examined the transcript level distributions per class, in which we separated the identified and non-identified parts of the theoretical proteome. Overall, the majority of the identified proteins showed relatively high corresponding transcript levels, regardless of their function class (Fig. 6A). On the other hand, the low-level mRNAs (log2fpkm 0.20–1.00) were enriched in the non-identified proteome (median $p = 0.0005$) (Fig. 6B). This held for 12 out of 21 classes, where transcripts of non-identified proteins appeared to be of a lower level.

To examine the low-level transcripts in these 12 classes, we searched for common elements ($n \geq 10$) in protein names. Examples are: for C_{01}: 'actin' or 'myosin'; for C_{03}: 'centromere', 'centrosomal' or 'dynein'; for C_{06}: 'AP1-3 complex subunit', 'Golgi' or 'trafficking protein particle' (Table 3). Close examination showed that, for all 12 classes with > 20% low-level mRNAs, the same > 20% also applied for elements of the non-identified proteome (Suppl. Table 2). As apparent from the listed most abundant transcripts of elements in almost all classes, the non-identified protein segments contained multiple isoforms or subunits of complexes that were also present in the identified segments, although the former had lower-level mRNAs (Table 3). Furthermore, sets of proteins seemed to be missing in almost all elements.

As a third restraining factor, we examined protein retainment in the megakaryocyte, by reasoning that in particular (peri)nuclear proteins will not move into a shedding proplatelet. This applied for the classes C_{20} (transcription & translation), C_{13} (other nuclear proteins) and C_{03} (cytoskeleton microtubule), containing multiple centromere/mitotic spindle proteins (Fig. 6A). Hence, these three classes were listed as providing additional explanation for low identification in the proteome (Suppl. Table 2).

Prediction model of the total platelet proteome. We then established an matrix for determining the three restraining factors per class (Fig. 7A). This matrix was then used to calculate weighted mean values of the fractions of identified proteins grouped per factor. The fractions of identified proteins for (i) low copy number, (ii) low mRNA > 20%, and (iii) retainment in megakaryocytes, amounted to 43%, 45% and 20%, respectively. For all other classes, the average fraction of identified proteins was 65% (Fig. 7A). By ratioing, this resulted in

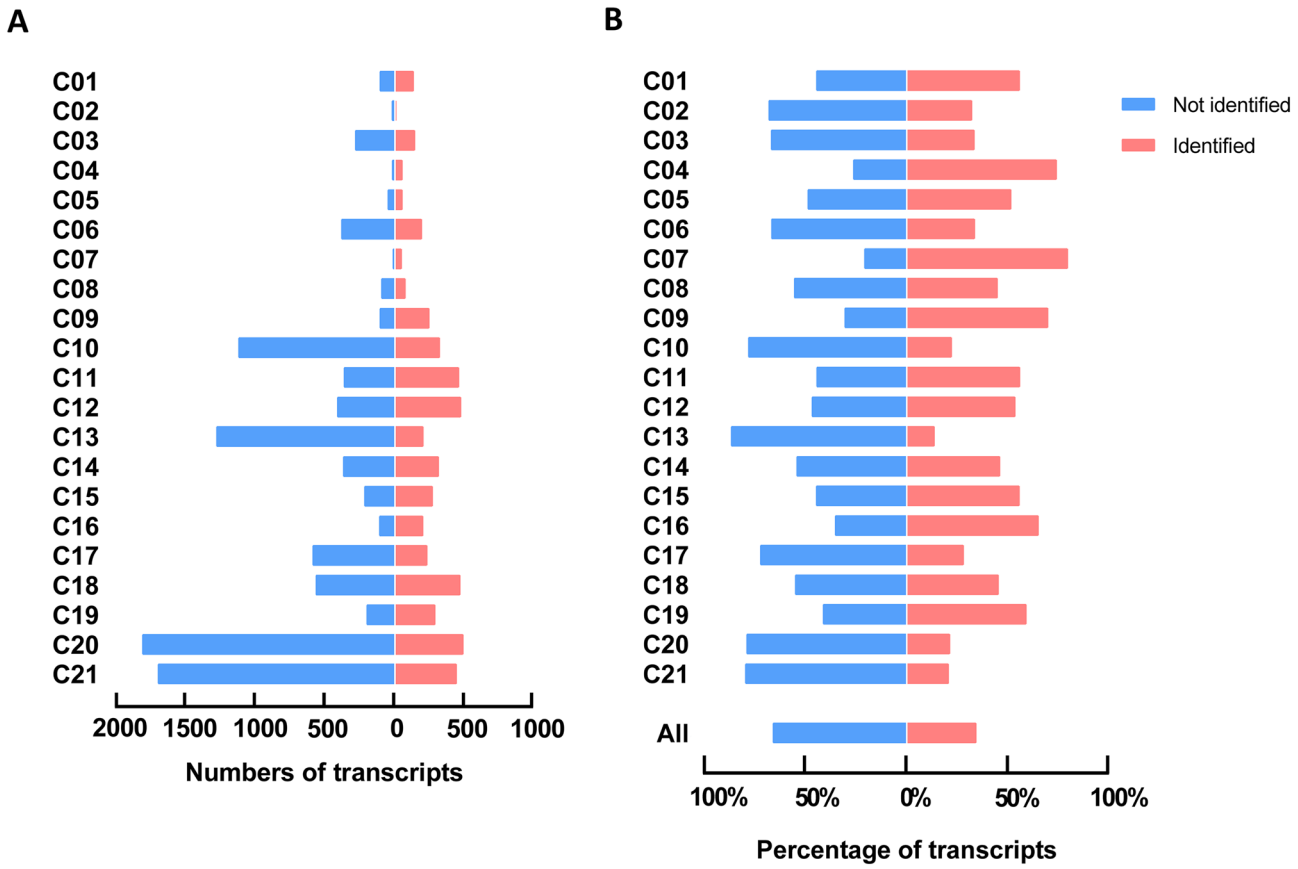


Figure 4. Transcript distribution of identified and not identified proteins in the platelet proteome per function class. Examined were all relevant protein-coding transcripts (log2fpkm ≥ 0.20) of the combined relevant PLT/MGK transcriptome, with separation of identified proteins ($n = 5{,}050$) and not identified proteins ($n = 9{,}721$). For full data, see Suppl. Figure 3. (**A**) Numbers of transcripts numbers per function class. (**B**) Percentage distribution of transcripts per function class.

correction factors (0.66, 0.69 and 0.31, respectively) for class predictions of the likeliness that additional proteins would appear in an enlarged proteome (Fig. 7B).

Summarizing, the prediction model indicated a greatly enlarged size of the platelet proteome up to 10 k proteins at a 1- or twofold higher detection efficacy. Markedly, apart from a consistent underrepresentation of classes of (peri)nuclear proteins (C_{03}, C_{13}, C_{20}), the model also predicted that a poor detection of proteins in the classes: C_{10} (membrane receptors & channels), C_{17} (secretory proteins), and C_{21} (uncharacterized & other proteins).

Proteome model validation. For validation of the model, we performed a new proteomic analysis with pooled platelets from 30 healthy subjects and the newest mass spectrometers. The obtained proteome included 4,389 of the previously identified proteins with relevant transcripts, as well as 954 previously not identified proteins (Fig. 2; details in Suppl. Datafile 3). Of additional 139 proteins without relevant transcript levels (log2fpkm < 0.20), the majority of 70% again appeared in C_{02} (intermediate cytoskeleton, $n = 15$, 11%) and C_{17} (secretory proteins, $n = 81$, 58%). This underscored the earlier observation that keratins and plasma proteins are present in the proteome of platelet samples.

Concerning the 954 novel obtained proteins, only small fraction of 3.8% showed low transcript levels with log2fpkm 0.20–1.00. Heatmap representation showed an similar distribution profile for all classes (Suppl. Figure 4). Markedly, inclusion of the novel proteins agreed with the prediction model for the majority of classes (Fig. 7C). Interestingly, higher than expected were the novel proteins for C_{20} (transcription & translation, additional 139 proteins) and C_{13} (other nuclear proteins $n = +121$); lower were those of C_{09} (membrane proteins, $n = +7$).

Coverage of genes associated with hemostasis and thrombosis. To further establish the clinical relevance of these datasets, we incorporated the identified proteome set into a Reactome-based protein–protein interaction network (267 core proteins and 2,679 new nodes) that was constructed to identify the roles of platelet and coagulation proteins in thrombosis and hemostasis[15]. As shown in Fig. 8, this network incorporated 1.3 k of the identified proteins (median protein copies 2,200, median transcript level log2fpkm 4.97), as well as a set of 1.1 k proteins/transcripts (median log2fpkm 1.97) not present in the combined proteome (Fig. 8A,B). Importantly, of the latter set, 172 proteins were obtained in the proteome of the validation cohort.

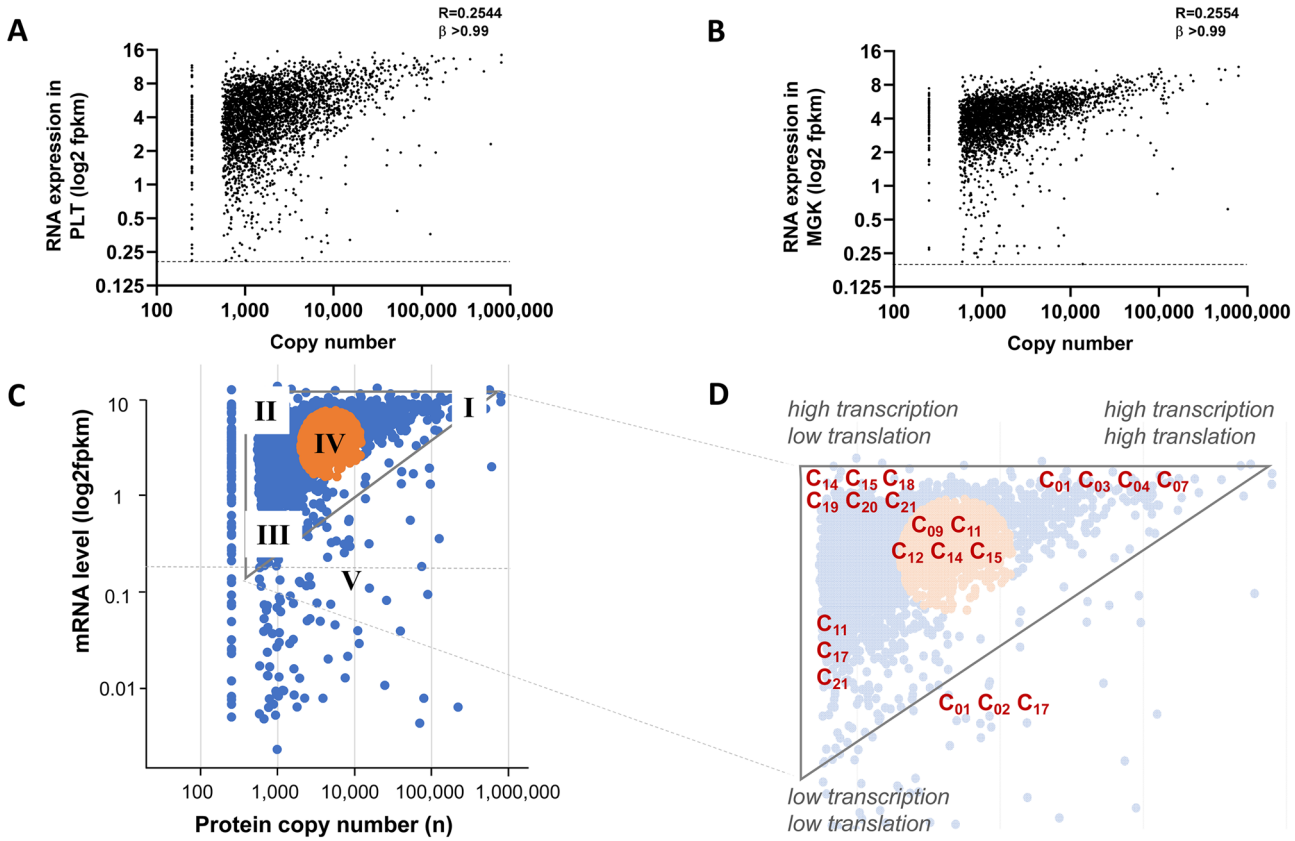


Figure 5. Comparison of protein copy numbers with mRNA levels and class-based analysis. (**A**,**B**) Protein copy numbers compared per gene to transcript levels (log2fpkm) for datasets of platelets (PLT, $n = 3{,}519$) (**A**) or megakaryocytes (MGK, $n = 3{,}442$). (**B**) Note triangular space, with low-abundance proteins (< 500 copies/platelet) were normalized to 150 copies. (**C**,**D**) Over-representation of protein function classes in quantitative proteome-transcriptome space per predefined area (I–V). Area I is considered to represent a condition of high translation (high mRNA level) and high transcription (high copy number); area II of high translation and low transcription; area III of low translation and transcription, and area IV an intermediate condition. Area V represents proteins without relevant transcript levels in PLT. Transcriptome-proteome triangle with analyzed areas (**C**). Enlarged space indicating function classes (C_{01}-C_{21}) with significant over-representation per area. Statistics in Suppl. Table 1.

To further establish the coverage for platelet-related disorders, we extracted the databases Online Mendelian Inheritance in Man (OMIM)[34] and Bloodomics[23] in combination with a recent overview paper[35] for genes associated with bleeding, thrombocythemia or thrombophilia. This resulted in 138 genes, of which 9 were absent in the platelet transcriptome but present in the proteome (coagulation factor and other plasma proteins), and 5 were absent in both (Table 4). For the remaining set of 124 genes, transcript levels (log2fpkm 4.58 ± 3.70, mean ± SD) and copy numbers (22.8 ± 73.0 k) in platelets were relatively high. Markedly, the majority of these 124 genes encoded for proteins in the classes C_{10} (membrane receptors and channels, $n = 22$), C_{17} (secretory proteins, $n = 19$), C_{20} (transcription & translation, $n = 12$), C_{18} (signaling & adapter proteins, $n = 10$), with a lower presence in the other classes. In accordance with the network analysis, it is likely that many still unknown gene products link to a platelet quantitative or qualitative traits, and hence to bleeding or thrombosis. The near complete coverage of the theoretical platelet proteome for known hemostatic pathways was also checked in the Reactome database (not shown).

Discussion

In this paper, we integrated in a functional way the human platelet proteome, using data from six cohorts established in the same institute, with the recently composed genome-wide, > 57 k platelet and megakaryocyte transcriptomes from the Blueprint consortium[30]. By UniProt-aided categorization of all relevant transcripts (set at log2fpkm ≥ 0.20) into 21 protein function classes, we were able to generate a first full proteomic map of the sub-cellular, metabolic and signaling molecules in an average human platelet. Importantly, this analysis also provide a reference list of 37.2 k transcripts according to our lists are not or hardly expressed in platelets.

Overall, the manuscript covers six major novel aspects: (i) for the first time we established the full or theoretical platelet proteome based on a state-of-the-art genome-wide platelet and megakaryocyte transcriptome; (ii) using > 57 k transcripts we identified an unexpected high similarity of the quantitative platelet and megakaryocyte transcriptomes (including RNA gene transcripts), in spite of a weak correlation between the protein and transcript levels, providing insight into the distribution of RNA species upon platelet shedding; (iii) based on the

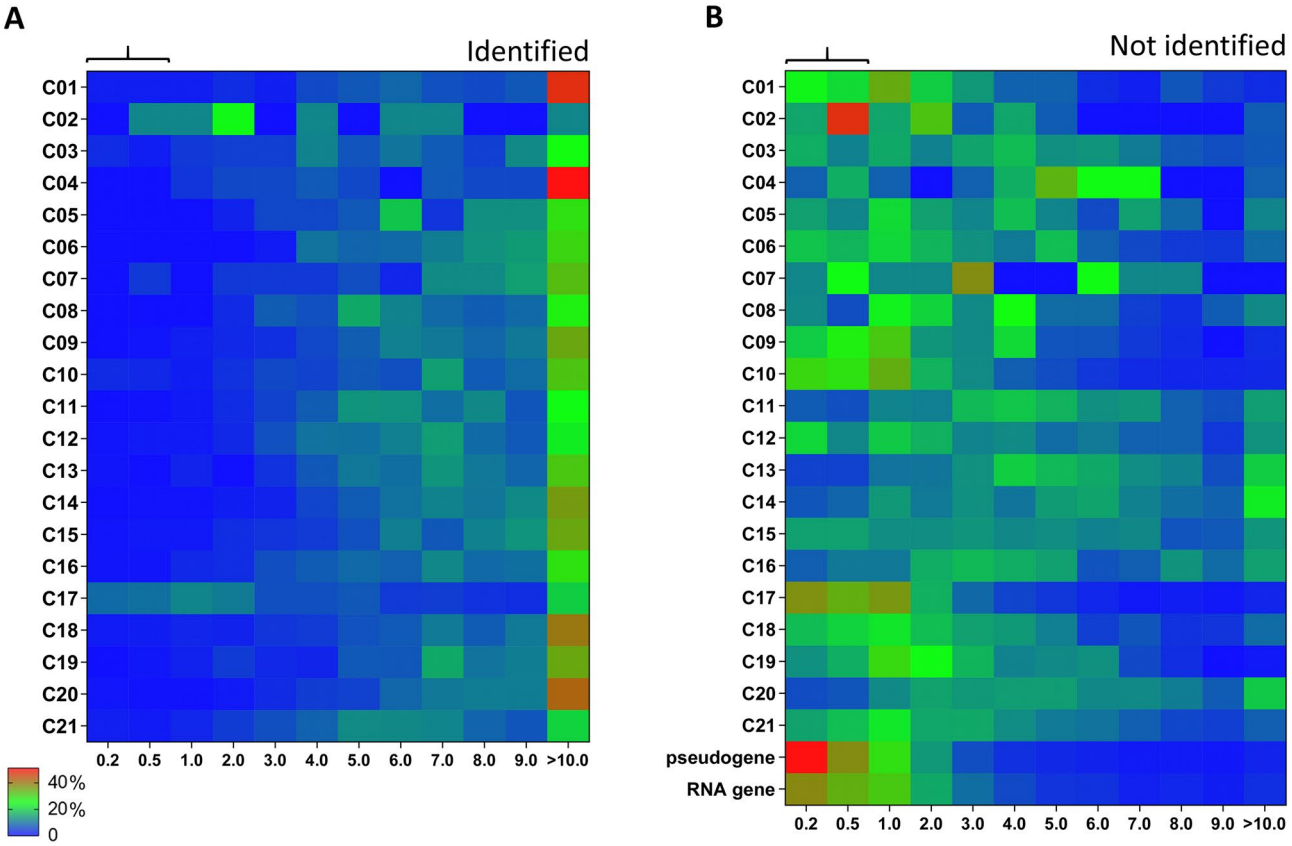


Figure 6. Distribution profile of relevant transcripts of per protein function class. For the relevant platelet transcriptome ($n = 17,629$), heatmaps were constructed of percentual distribution of transcript levels per function class (rainbow colors; blue = low, red = high). (**A**) Heatmap for transcripts of identified proteins ($n = 5,030$). (**B**) Heatmap for transcripts of non-identified proteins ($n = 9,267$); furthermore RNA genes ($n = 2,480$) and pseudogenes ($n = 852$). Expression levels (log2fpkm) were binned as 0.20–0.50, 0.50–1.00, 1.00–2.00, etc. For numbers of transcripts, see Suppl. Figure 3.

systematic protein classification, the collected data provide molecular understanding of the complexity of platelet structures and functions; (iv) based on the established theoretical proteome, we developed and also validated a prediction model for identifying missing proteins in the current proteome sample sets; (v) the combined datasets offer better understanding of protein adhesion and uptake of plasma proteins by platelets; (vi) the combination of quantitative transcriptomes and (partly) quantitative proteomes completes our knowledge of the roles of > 100 genes and proteins in diseases not limited to thrombosis and hemostasis.

Correlational analysis of the 20 k expressed transcripts in platelets and/or megakaryocytes indicated an overall high similarity between the transcriptomes of the two cell types. This particularly held for the 14.8 k transcripts of protein-coding genes (R = 0.75), while the correlation was lower for the 3.8 k RNA genes and 1.9 k pseudogenes (R = 0.43–0.54). Although inter-individual differences are expected, our findings indicate that the majority of mRNA species evenly spread from megakaryocytes to the formed proplatelets, with limited degradation during platelet ageing. The aberrant transcript profiles of pseudogenes and RNA genes, which in general were more abundant in megakaryocytes, may be due to retention or to enhanced degradation of such shorter RNA forms[36]. In agreement with our findings, also other authors presenting smaller-size and not genome-wide datasets (3.5 k proteins and 5.5 k mRNAs), have reported a low correlation between platelet protein and transcript levels[37,38]. This lack of correlation however does exclude a role of altered mRNA and protein levels in platelet-related diseases[21].

Based on the composition of the genome-wide transcriptomes of platelets and megakaryocytes, we calculated that the current proteome of 5,050 expressed proteins misses approximately 66% of the expected translation products. Highest percentages of missing proteins were seen in the classes C_{20} (transcription & translation 79%), C_{21} (uncharacterized proteins 79%), C_{13} (other nuclear proteins 86%), C_{10} (membrane receptors & channels 78%), C_{17} (secretory proteins 72%), and C_{18} (signaling & adapter proteins 55%). Especially low-level mRNAs (log2fpkm 0.20–1.00) appeared to be missing in the identified proteome, likely giving rise to only low copy numbers of proteins.

Proteomic technologies have been well developed, since the publication of the first draft human proteome, which revealed 17.3 k gene products and 4.1 k protein N-termini[39]. Accordingly, the present set of 5.0 k identified platelet proteins is higher than earlier published proteomes, *e.g.* of mouse platelets of 4.4 k proteins with copy numbers[40], or of the semi-quantitative 3.5–4.8 k proteins in human platelets[38,41]. Smaller size published platelet sub-proteomes are a 0.1 k secretome[42], and a 1.0 k sheddome[43]. Regarding platelet transcriptomes, which are more uniformly to construct, other authors have published a similar 20 k size with 16 k transcripts at > 0.3 fpkm[44].

Protein function class (n, log2fpkm≥0.20)	Identified in proteome (log2fpkm ≥0.20)		Not identified in proteome (high mRNA)		Not identified in proteome (low mRNA)		Sum all	Sum NI	Low mRNA
UniProt-KB: in protein name (n≥10)	(n)	Examples (top 4 genes)	(n)	Examples (top 4 genes)	(n)	Examples (top 4 genes)	(n)	(n)	%
C01 Cytoskeleton actin-myosin									
Actin	31	ACTB, ACTG1, ABLIM3, ABLIM1	3	FAM107A, AFAP1L1, IPP	2	ACTL10, ACTRT3	36	5	40%
Myosin	25	MYL6, MYL12A, MYL9, MYH9	9	MYLPF, MYO19, MYO1E, MYO10	11	MYO7B, MYO5C, MYO15A, MYBPC1	45	20	55%
C02 Cytoskeleton intermediate									
Keratin	5	KRT8, KRT1, KRT10, KRT73	3	KRT18, TCHP, KRT23	4	KRT7, KRTAP29-1, KRTCAP3, KRTAP10-6	12	7	57%
C03 Cytoskeleton microtubule									
Centromere	2	ZW10, CENPF	11	CENPI, CENPH, CENPJ, INCENP	3	CENPO, CENPM, CENPP	16	14	21%
Centrosomal	9	CEP162, CEP44, CEP128, CEP41	24	CEP70, CEP57L1, CEP350, CEP57	3	CEP126, CEP55, CEP295	36	27	11%
Dynein	16	DYNLRB1, DYNLL1, DNM3, DCTN2	15	DNAH14, EFCAB2, DYNLT1, DNAH12	10	CCDC65, DNAAF1, DNAH6, IQCA1	41	25	40%
Kinesin	11	KIF2A, KIFC3, KIF5B, KLC4	17	KIF28P, KIF3C, KIF22, KNL1	10	KIF21A, KIF25, KIF24, KLC3	38	27	37%
Mitotic spindle / HAUS	4	MZT2B, NUMA1, MZT1, HAUS6	7	SKA2, SKA3, HAUS2, HAUS7	2	MAD2L1, SKA1	13	9	22%
Tubulin	40	TUBB1, TUBA4A, TUBA8, TUBB4B	24	TUBA1B, MAP1LC3A, AKNA, TTLL7	10	MAP10, C16orf59, TPPP2, MAP9	74	34	29%
C04 Cytoskeleton receptor-linked									
LIM / Wiskott	7	LIMS1, PDLIM1, LASP1, WASF3	2	PDLIM2, WHAMMP3	1	FBLIM1	10	3	33%
C05 Endosome proteins									
Multivesicular body	9	CHMP3, CHMP4B, CHMP5, CHMP2A	2	MVB12A, CHMP4C	0	-	11	2	0%
WAS / WASH	6	WASHC3, WASHC2C, WASHC4, WASHC1	5	WASH6P, WASP3P, WASH2P, WASH4P	0	-	11	5	0%
C06 ER & Golgi proteins									
AP-1/3 complex subunit	15	AP1S1, AP2S1, AP2B1, AP3S1	3	AP1M2, AP3M2, AP1S3	1	AP3B2	19	4	25%
ER membrane / lumen protein	20	EMC3, KDELR2, EMC2, ERP29	9	KDELR1, EDEM3, HERPUD1, ERMARD	5	H6PD, SERP2, HRC, ERP27	34	14	36%
Golgi	31	TGOLN2, GOLGA7, GOLGA2, GOSR1	12	BLZF1, GOLGA8N, GOLGA8A, TVP23C	16	TVP23C-CDRT4, GOLGA8M, GOLGA7B, GOLGA8R	59	28	57%
Trafficking protein particle	12	TRAPPC1, TRAPPC6B, TRAPPC5, TRAPPC9	4	TRAPPC2, TRAPPC3L, TRAPPC6A,	0	-	16	4	0%
Transferase	4	MGST3, MGST2, GFPT1, ACAT2	50	CSGALNACT2, LFNG, ST8SIA4, CHST2	41	FUT1, UBIAD1, UGT2B11, GCNT2, PIGA	95	91	45%
C07 Glucose metabolism									
Glucose	12	GPI, G6PD, SLC37A1, PRKCSH	5	NUDT22, GANC, G6PC3, TGDS	1	G6PC2	18	6	17%
Fructose	7	ALDOA, GFPT1, PFKM, PFKP	3	PFKFB4, TIGAR, FBP1	1	PFKFB1	11	4	25%
C08 Lysosome & peroxisome proteins									
Lysosome / lysosomal	10	CTSA, LIPA, LAMP1, LAMP2	10	LAPTM5, LAPTM4B, LMBRD1, LAMP5	1	LAMP3	21	11	9%
Peroxisomal / peroxisomal	17	HSD17B4, ACOX1, ACAA1, PEX19	14	PXMP4, LONP2, PHYH, PEX26	6	TYSND1, PEX12, PEX11A, ACOX2	37	20	30%
V-type proton ATPase	12	ATP6V1E1, ATP6V1F, ATP6V1G1, ATP6V1D	3	ATP6V0C, ATP6V0E1, ATP6V0B	2	ATP6V1G2, ATP6V1B1	17	5	40%

Table 3. Subgroup analysis of non-identified proteins (n = 9,721) of the relevant PLT/MGK transcriptome. Per function class (C_{01}–C_{21}), the transcriptome database was searched for common elements in protein names ('actin', 'myosin'), and frequency was recorded as identified in the proteome, or not identified with a separation into high mRNA (log2fpkm > 1.00) or low mRNA (log2fpkm 0.20–1.00). Top-4 of most abundant transcripts were listed per element. Further indicated per element: numbers of all transcripts (Sum all), and numbers of transcripts not identified in proteome (Sum NI).

C09 Membrane & protein trafficking									
Exocyst complex	9	EXOC5, EXO6B, EXOC4, EXOC6	0	-	1	EXOC3L1	10	1	100%
Protein transport	10	SEC31A, SEC61G, SEC61B, SEC23A	1	SEC31B	0	-	11	1	0%
Sorting nexin	17	SNX3, SNX6, SNX1, SNX5	6	SNX13, SNX19, SNX20, SNX10	4	SNX25, SNX33, SNX7, SNX32	27	10	40%
Synapto	5	SNAP23, ESYT2, SYLT4, SNAP29	6	SYLT1, SYLT3, SYLT11, SNAP47	8	SYT17, SYT15, SYT1, SYNGR2	19	14	57%
Syntaxin	16	STXBP2, STX7, STXBP3, STX11	4	STX3, STXBP4, STXBP6, STX1B	0	-	20	4	0%
Vacuolar protein sorting-associated	26	VPS28, VPS37A, VPS41, VPS37B	0	-	0	-	26	0	0%
C10 Membrane receptors & channels									
Calcium / Cal	9	ORAI2, ORAI1, SLC8A3, ITPR2	15	SLC24A3, ORAI3, CACNB1, CALHM6	18	CALCRL, KCNMA1, CACNA1D, CACNA1A	42	33	55%
Chemokine / Interleukin receptor	3	IL6ST, CXCR4, CCR4, IL17RA	6	IL7R, IL10RA, IL2RG, CXCR2	9	CCR7, IL1R1, IL18R1, CXCR3	18	15	60%
C-type lectin domain family	1	CLEC1B	8	CLEC7A, CLEC2D, CLEC12A, CLEC10A	10	CLEC2A, CLEC12B, CLEC5A, CLEC17A	19	18	56%
Glycoprotein	8	GP1BB, GP9, GP1BA, CD36	15	CD28, CD3E, CD8A, CD3G	9	CD1E, CD1C, KEL, MCAM	32	24	38%
G-protein coupled	1	ADGRG1	14	ADGRE1, ADGRE2, GRP97, GRP183	30	GPRC5B, GPR135, GPR162, ADGRL4	45	44	68%
Integrin	13	ITGB1, ITGB3, ITG2B, ITGB5	7	ITGAM, ITGAX, ITGAE, ITGAL	10	ITGA7, ITGB4, ITGA3, ITGBL1	30	17	59%
Olfactory receptor	0	-	4	OR2W3, OR2L13, OR2T8, OR2B6	24	OR2L2, OR14L1P, OR2M4, OR52N4	28	28	86%
Purinoceptor	3	P2RY12, P2RX1, P2RY1,	7	P2RY10, P2RX5, P2RX4, P2RX7	4	P2RY11, P2RY14, P2RX6, P2RY6	14	11	36%
Solute carrier family / SLC	52	SLC40A1, SLC2A3, SLC44A2, SLC35E1	51	SLC38A2, SLC35F5, SLC11A1, SLC2A11	43	SLC12A7, SLC6A9, SLC16A6, SLC24A4	146	94	46%
Voltage-dependent/gated	6	VDAC3, KCNA3, VDAC2, VDAC1	12	KCNE3, KCND3, HVCN1, KCNQ1	20	KCNH3, KCNC3, KCNQ5, KCNA2	38	32	63%
C11 Mitochondrial proteins									
ATP synthase	17	MT-ATP6, MT-ATP8, ATP5MPL, ATP5L	5	ATP5MC3, ATP5MC1, ATPSCKMT, ATPAF2	1	ATP5MGL	23	6	17%
Cytochrome b/c	31	MT-CO2, UQCRH, COX6B1, UQCR11	14	MT-CO3, MT-CO1, MT-CYB, COX17	4	UQCC3, COX4I2, COA7, UQCRHL	49	18	22%
Import	21	TOMM6, TOMM5, TOMM20, TIMM9	6	TIMM17B, TIMM10B, TIMM23, TIMM17A	0	-	27	6	0%
NADH dehydrogenase	32	NDUFA6, NDAFA5, NDUFB4, NDUFS5	5	NDUFAF3, NDUFA1, NDUFC1, NDUFAF6	2	NDUFAF4, NDUFA4L2	39	7	29%
Ribosomal protein	47	MRPL41, MRPL35, MRPS28, MRPS23	37	MRPL28, MRPL47, MRPL30, MRPL34	0	-	84	37	0%
tRNA	19	CARS2, FARS2, RARS2, MRPL58	11	GATB, TRMT61B, MTFMT, GATC	1	VARS2	31	12	8%
C12 Other metabolism									
(Metabolite) kinase	44	CMPK1, GUK1, GK, DGKD	15	PANK3, ETNK1, AK9, ITPK1	10	ETNK2, PANK1, TK1, RBKS	69	25	40%
(Metabolite) phosphatase	30	INPP5K, PFKFB3, ACYP1, FIG4	15	CTDNEP1, PLPP5, IMPA2, PHOSPHO1	8	NUDT15, INPP5E, PAH, PHOSPHO2	53	23	35%
(Metabolite) reductase	31	CYB5R3, GMBR, CYB5R1, MRSB3	10	PYCR2, SHRS3, WWOX, FAR2	8	PYCRL, AKR1C2, TXNRD3, AKR1C4	49	18	44%
(Metabolite) synthase	29	GLUL, SERS2, SMS, PAPSS1	20	MTR, FDFT1, MTRR, DHPS	8	CARNS1, ADSSL1, GYS2,	57	28	29%
(Metabolite) transferase	76	AGPAT1, MBOAT2, GSTO1, UGP2	49	SAT1, PCMTD1, NAT8B, SGMS1	37	HS6ST1, GSTA4, PRMT10, TGM1	162	86	43%
(Metabolite) epimerase	6	GALM, RPE, GALE, GNE	1	DSE	1	DSEL.	8	2	50%
C13 Other nuclear proteins									
Chromatin	4	SMARCA5, CHRAC1, ACIN1	12	HP1BP3, BAP18, MEAF6, PDS5A	1	CHAF1B	17	13	8%
Histone	16	H1-2, H2AC11, H1-4, KMT2C	121	H2AC6, H3-3A, H2BC4, H3C10	13	-	150	134	10%
Nuclear pore complex	6	NUP54, NOP58, NUP88, NUP214	19	NUP50, NPIPB5, NUP35, NUP98	9	NPIPB2, NPIPA7, NUP210L, NPIPA3	34	28	32%
Polymerase	4	PARP14, PARP9, PARP10, PARP12	24	POLD4, POLR2J3, POLR1D, POLE3	0	-	28	24	0%
Repair protein	4	MRE11, RAD50, MMS19, MSH6	22	MLH3, SPIDR, SWI5, RAD21	3	XRCC3, RAD51, ERCC6L	29	25	12%
C14 Proteasomal proteins									
COP9 signalosome	9	COPS9, COPS3, COPS4, COPS8	1	COPS9	0	-	10	1	0%
E3 ubiquitin-protein ligase	55	RBX1, MARCH2, TRIM58, RNF13	104	MARCH6, MKRN1, RBBP6, SIAH2	21	RNF43, PELI3, TRIM32, SNURF	180	125	17%
Kelch-like	0	-	14	KLHL18, KLHL8, KLHL2, KLHL7	6	KLHL15, KLHL3, KBTBD8, KLHL17	20	20	30%
NEDD	12	UBE2F, NEDD8, NAE1, UBA3	2	NDFIP1, WWP1	1	NDFIP2	15	3	33%
Proteasome / proteasomal	37	PSMB9, PSMB8, PSME1, PSMF1	7	SEM1, POMP, PSME3, PSGM2	1	C1orf105	45	8	13%
Ubiquitin-conjugating	15	UBE2D3, UBE2E3, UBE2H, UBE2K	15	UBE2Q2, UBE2D2, UBE2C, UBE2D	2	UBE2T, UBE2Q2L	32	17	12%

Table 3. (continued)

C15 Protein kinases & phosphatases										
Mitogen-activated protein kinase	23	MAP2K3, MAPKAPK2, MAPK1, MAP4K5	12	MAPK6, MAP2K5, MAPK8, MAP3K3	6	MAPK11, MAPK12, MAP3K12, MAP3K21	41	18		33%
Serine/threonine-protein kinase	83	RIOK3, TLK1, STK24, AKT3	40	STK40, PRKD3, TLK2, MKNK2, UHMK1	21	STK31, PRKY, NEK8, PLK1	144	61		34%
Serine/threonine-protein phosphatase	25	ANKRD28, PPP1CB, PPP6C, PPP6R1	11	PPP4C, PPP1R10, PPP4R2, PPP4R3A	4	PPP4R4, PPP2R3A, PPEF1, PPEF2	40	15		27%
Tyrosine (protein) phosphatase	16	PTPN18, PTPN12, PTP4A2, PTPRA	6	PTPRU, PTPN2, PTPN4, PTPN22	7	PTPRB, PTPN13, PTPN20B, PTPN20A	29	13		54%
Tyrosine-protein kinase	19	LYN, FYN, BTK, CSK	7	TXK, HCK, LCK, ZAP70	3	STYK1, BMX, TNK1	29	10		30%
C16 Protein processing										
Dol / dolichol	14	DAD1, STT3B, DPM3, STT3A	2	DPM2, ALG6	0	-	16	2		0%
Galactosyltransferase (C12)	5	C1GALT1, B4GALT1, B4GALT4, COLGALT1	4	B4GALT6, B3GALT4, B4GALT3, B4GALT5	2	B4GALT2, C1GALT1C1	11	5		40%
Glucosyl / glycosyltransferase	14	ALG6, RPN2, UGGT1, RPN1	3	ALG8, ALG10B, OST4	0	-	17	3		0%
Methyltransferase	5	PRMT2, PRMT1, PRMT3, PRMT5	8	EEF1AKMT1, DPH7, PRMT7, NTMT1	0	-	13	8		0%
Palmitoyl / sialyltransferase	1	ST3GAL6	14	ST3GAL3, SPTSSA, ST3GAL4, ST3GAL1	5	PORCN, ZDHHC23, ZDHH9, SPTSSB	20	19		26%
Pept / peptidyl	37	PPIA, FKBP1A, FKBP8, XPNPEP1	9	PPIG, XPNPPE3, PGPEP1, FKBP9	4	CPA3, FBKP10, CPA1, FOLH1	50	13		31%
C17 Secretory proteins										
Collagen	9	COL4A3BP, COL6A3, COL4A2, COL6A1	7	COL24A1, COL10A1, COL18A1, COL25A1	17	COL4A1, COL9A3, COL15A1, COL17A1	33	24		71%
Growth factor	14	TGFB1, EGF, HDGF, PDGFA	8	VEGFB, FGF2, IGFBP4, VEGFA	14	LTBP2, TGFA, FGF12, TGFB3	36	22		64%
Interleukin / chemokine	7	CCL5, CXCL5, CXCL3, IL16	16	IL32, CXCL8, CKLF, CXCL16	18	IL33, IL6, CLCF1, CCL7	41	34		83%
Metalloproteinase	5	TIMP1, ADAM8, TIMP2, TIMP3	9	ADAMTS6, MMP25, ADAM15, MMP28	14	TIMP4, ADAMTS9, MMP11, ADAMTS10	28	23		81%
Protease	4	SPINT2, SERPING1, MASP1, SERPINA5	7	PRSS27, PRSS53, CTRL, PRADC1	10	KAZALD1, HTRA3, HPN, PRSS57	21	17		89%
C18 Signaling & adapter proteins										
14-3-3 / S100 protein	11	YWHAZ, YWHAE, YWHAH, S100A8	2	S100A10, S100P	2	S100A2, S100A1	15	4		50%
Adapter / adaptor protein	13	GRAP2, SLA2, FAM89B, SH2B3	13	STRADB, PRAM1, SH2B1, GRAP	9	TICAM2, GULP1, STAP2, SHF	35	22		41%
Bcl / Bax	9	BNIP2, BNIP3L, BCL2L1, BAD, BAK1	5	TMBIM6, BCL7B, BMF, BCL9, BCL2	4	BNIPL, BIK, BCL2L14	18	9		44%
Calcium / calpain	15	CAPNS1, CAPN1, ATP2A3, CAPN2	11	CALM3, CALM2, CABP5, CAPN11	12	CADPS2, SLC24A1, CABP4, PDE1B	38	23		52%
cAMP / A-kinase	8	C2orf88, AKAP13, AKAP2, ADCY3	14	PKIG, PDE4D, AKAP7, AKIP1	12	ADCY4, PDE4C, ARPP21, AKAP3	34	26		46%
Caspase	11	CARD19, CARD8, CASP4, CASP2	7	CARD16, CASP1, CARD11, CASP10	1	CARD14	19	8		13%
cGMP / G kinase	7	PDE5A, GUCY1B3, GUCY1A3, GMPS	7	GKAP1, PDE6D, PDE6H, PDE6B	3	PDE9A, GUCY2D, GUCY1B2	17	10		30%
Guanine nucleotide / G-protein	25	RGS18, RGS10, GNAS, GNG11	6	RGS2, GNG2, RGS9, GNGT2	12	RGS5, RGS22, RGS17, GNG12	43	18		67%
Phosphatidyl /phosphoinositide	41	PIK4K2A, DAPP1, TMEM55A, PIK3C3	20	PIK3R5, PLCD4, PIK3CD, PIK3R3	5	PLCZ1, PIP5KL1, NYAP2, PREX1	66	25		20%
C19 Small GTPases & regulators										
Arf-GAP	7	ACAP2, GIT2, ACAP1, GIT1	8	AGAP4, ADAP1, ARAP3, ADAP2	3	AGAP8, AGAP6, AGAP7	18	11		27%
Rab	63	RABGAP1L, RAB11A, RAB27B, RAB31	11	RABL2A, RAB11FIP1, RGP1, RIN2	4	RAB44, RAB39B, RAB26, RAB36	78	15		27%
Ral	8	RALB, RALBP1, RALGAPA1, RALGAPB	2	RALGDS, RALGPS1	2	RGL1, RGL3	12	4		50%
Rap	9	RAP1B, RAP1A, RAP1GDS1, RAP1GAP2	7	RAPGEF1, RAP1GAP, RAPGEF3, RAPGEF5	1	GARNL3	17	8		13%
Ras *(non RAb, Ral, Rap)*	6	RSU1, RASGRP2, RASA3, RASA1	7	RASSF5, RASGRF2, RASA4, RASGRP4	7	RASEF, IQGAP3, RASGEF1A, RASAL1	20	14		50%
Rho	38	ARHGDIB, RHOA, ARHGAP18, ARHGAP21	28	ARHGAP30, RHOQ, ARHGAP26, TAGAP	22	ARHGAP22, ARHGAP25, ARHGEF5, SYDE1	88	50		44%
C20 Transcription & translation										
Nuclear	42	NCOR1, TIA1, NCOA7, NFAT5	82	NCOA4, HNRNPUL1, NCOA2, NFYC	3	NPIPB15, NFATC4, NARF	127	85		%
Ribosomal protein	34	RPS20, RPLP1, RPS11, RPS18	60	RPL41, RPS27, RPS29, RPL13	2	RPL17, RPS6KL1	96	62		%
Transcription	40	SLTM, BTF3, FLI1, NFE2	48	ATF4, CAMTA1, SUPT4H1, RUNX1	215	YAP1, TEAD4, E2F5, POU6F1	303	263		%
Translation	44	EIF1, TMA7, EIF4G2, EIF4G3	11	EIF2AK1, EIF1B, EIF1AX, EIF4E3	1	EIF3CL	56	12		%
tRNA	31	RARS, WARS, NARS, DARS	31	ADAT1, TRIT1, DUS1L, CDKAL1	9	METTL2A, TRMT10A, TRMT61A, CTU1	71	40		%
Zinc finger	9	ZC3H7A, HELZ, ZC3HAV1, ZBTB7A	409	GFI1B, ZEB2, ZNF664, ZNF271	161	ZNF483, ZNF589, IKZF4, ZNF563	579	570		%
C21 Uncharacterized & other proteins										
Coiled-coil	15	CCDC92, CCDC9B, CCDC25, CCDC9	55	CCDC7, CCDC126, CCDC175, CCDC180	21	CCDC157, CCDC136, CCDC121, CCDC81	91	76		28%
FAM	58	FAM110A, FAM122B, FAM177A1, FAM114A2	118	FAM13B, FAM81B, FAM228B, FAM193B	79	FAM13C, FAM209A, FAM71F2, FAM161B	255	197		40%
Leucine-rich repeat	6	LRCH3, LRCH4, LRRC47, C10orf11	12	LRRC28, LRRC37A3, LRRC23, LRRC37B	18	LRRC61, LRFN4, LRIG2, LRRC1	36	30		60%
Putative	10	INAFM2, UQCRFS1P1, SNX29P2, HSPA7	49	PLEKHA8P1, NBPF8, C15orf54, NSHG12	31	ALG1L2, CCDC144CP, ZNF815, C22orf34	90	80		39%
Transmembrane	40	TMEM50A, CMTM5, CMTM6, TM6SF1	56	TMEM91, TMAM158, TMEM185A, TMEM248	37	TMEM268, TMEM267, TMEM185B, TMEM273	133	93		40%
Uncharacterized	19	C1orf198, KIAA0513, C1orf43, KIAA2013	119	C4orf3, C12orf76, C6orf62, LOC100287036	111	ZSWIM9, ENS00000173366, C16orf74, C3orf80	249	230		48%

Table 3. (continued)

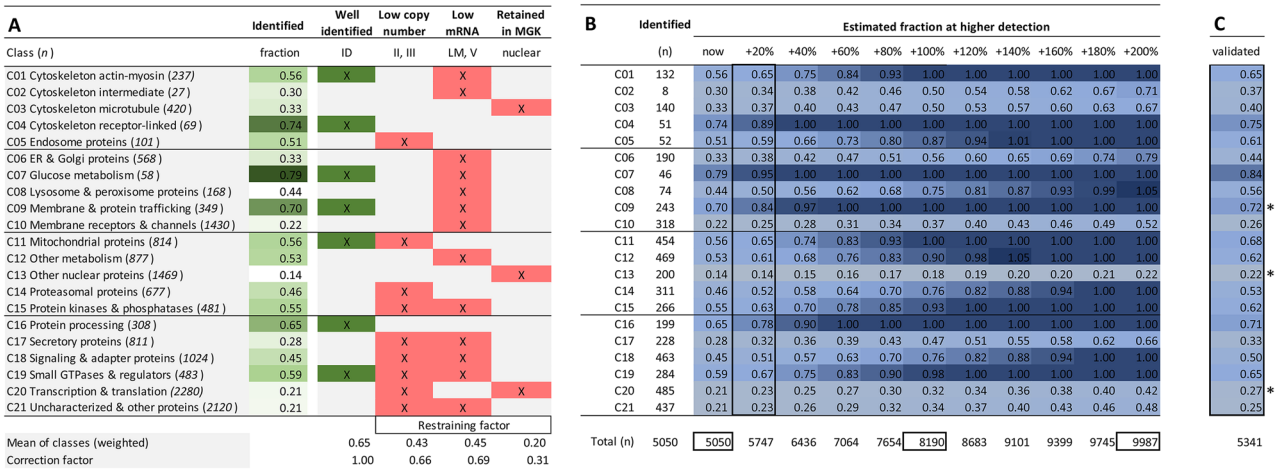
As a check of the present concept—starting from genome-wide platelet and megakaryocyte transcriptomes to determine the theoretical proteome—we evaluated the proteomes reported in three papers, using the current GeneCards gene designations. The proteomes of platelets from Dengue patients[45] or from platelet concentrates[46] were found to contain 93.1% (1,769/1,901) and 98.4% (2,466/2,505) proteins that were present in our protein database. Proteins without relevant transcripts were quite low, 2.1% and 0.1%, respectively. A paper analyzing the proteomes from cord blood and adult peripheral blood platelets[47] showed lower overlap of 79.9% (3,950/4,941) with the current proteome, supplemented with 16.4% proteins with relevant transcripts and 3.7% (183/4,941) without relevant transcripts in dataset. For the last fraction, it is unclear if residual presence of neonatal transcripts contributes to this higher percentage.

In platelet proteomics, the detection of proteins from blood plasma or other blood cells is a continuous point of attention. Our analysis based on highly purified, washed platelet preparations indicated the invariable present presence of plasma proteins. This can be explained by the fact that platelets exhibit an extensive open canicular system (estimated at 1 vol%) in open contact with the plasma, and furthermore also endocytose plasma proteins. The list includes 73 proteins classified as C_{17} (secretory proteins) without corresponding mRNAs, of which at least fibrinogen and β2-glycoprotein 1 are known to be taken up by platelets[48]. Of note, fibrinogen levels are greatly reduced in the proteome of patients with Glanzmann's thrombasthenia, lacking integrin αIIbβ3. At the other hand, we find that multiple 'plasma proteins' can also be expressed by platelets themselves. Hence, even with the development of quality checks of 'plasma contamination', it may be difficult to rate many secretory proteins as platelet or non-platelet.

Apart from the inevitable presence of plasma proteins in platelet preparations, also other conditions may influence the obtained platelet protein composition. One relevant condition is that of macro-thrombocytopenia (e.g., Bernard-Soulier syndrome), often resulting in more fragile platelets, where obtaining of the high quality platelet preparation is a challenge. Another factor is emperipolesis, such as engulfment of hematopoietic cells by megakaryocytes in malign disorders, also affecting the platelet proteome.

To explain the missing of proteins in the identified proteome, we considered three restraining factors: (i) low protein copy number, (ii) low mRNA level, and (iii) protein retainment in the megakaryocyte perinuclear region. By estimating these restraining factors per protein function class, we calculated the technically achievable proteome of ~ 10 k proteins. The assumption is that improved technical developments will generate larger size proteomes (Suppl. Methods).

For validation of the function class-based prediction model of the remaining part of the proteome, we generated an additional proteomic set, which revealed 1.0 k new proteins in the predicted classes, of which 97%



A	Identified	Well identified	Low copy number	Low mRNA	Retained in MGK
Class (*n*)	fraction	ID	II, III	LM, V	nuclear
C01 Cytoskeleton actin-myosin (*237*)	0.56	X		X	
C02 Cytoskeleton intermediate (*27*)	0.30			X	
C03 Cytoskeleton microtubule (*420*)	0.33				X
C04 Cytoskeleton receptor-linked (*69*)	0.74	X			
C05 Endosome proteins (*101*)	0.51		X		
C06 ER & Golgi proteins (*568*)	0.33			X	
C07 Glucose metabolism (*58*)	0.79	X		X	
C08 Lysosome & peroxisome proteins (*168*)	0.44			X	
C09 Membrane & protein trafficking (*349*)	0.70	X		X	
C10 Membrane receptors & channels (*1430*)	0.22			X	
C11 Mitochondrial proteins (*814*)	0.56	X	X		
C12 Other metabolism (*877*)	0.53			X	
C13 Other nuclear proteins (*1469*)	0.14				X
C14 Proteasomal proteins (*677*)	0.46		X		
C15 Protein kinases & phosphatases (*481*)	0.55		X	X	
C16 Protein processing (*308*)	0.65	X			
C17 Secretory proteins (*811*)	0.28		X	X	
C18 Signaling & adapter proteins (*1024*)	0.45		X	X	
C19 Small GTPases & regulators (*483*)	0.59	X	X	X	
C20 Transcription & translation (*2280*)	0.21		X		X
C21 Uncharacterized & other proteins (*2120*)	0.21		X	X	
			Restraining factor		
Mean of classes (weighted)		0.65	0.43	0.45	0.20
Correction factor		1.00	0.66	0.69	0.31

B	Identified	Estimated fraction at higher detection										
	(n)	now	+20%	+40%	+60%	+80%	+100%	+120%	+140%	+160%	+180%	+200%
C01	132	0.56	0.65	0.75	0.84	0.93	1.00	1.00	1.00	1.00	1.00	1.00
C02	8	0.30	0.34	0.38	0.42	0.46	0.50	0.54	0.58	0.62	0.67	0.71
C03	140	0.33	0.37	0.40	0.43	0.47	0.50	0.53	0.57	0.60	0.63	0.67
C04	51	0.74	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C05	52	0.51	0.59	0.66	0.73	0.80	0.87	0.94	1.01	1.00	1.00	1.00
C06	190	0.33	0.38	0.42	0.47	0.51	0.56	0.60	0.65	0.69	0.74	0.79
C07	46	0.79	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C08	74	0.44	0.50	0.56	0.62	0.68	0.75	0.81	0.87	0.93	0.99	1.05
C09	243	0.70	0.84	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C10	318	0.22	0.25	0.28	0.31	0.34	0.37	0.40	0.43	0.46	0.49	0.52
C11	454	0.56	0.65	0.74	0.83	0.93	1.00	1.00	1.00	1.00	1.00	1.00
C12	469	0.53	0.61	0.68	0.76	0.83	0.90	0.98	1.05	1.00	1.00	1.00
C13	200	0.14	0.14	0.15	0.16	0.17	0.18	0.19	0.20	0.20	0.21	0.22
C14	311	0.46	0.52	0.58	0.64	0.70	0.76	0.82	0.88	0.94	1.00	1.00
C15	266	0.55	0.63	0.70	0.78	0.85	0.93	1.00	1.00	1.00	1.00	1.00
C16	199	0.65	0.78	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
C17	228	0.28	0.32	0.36	0.39	0.43	0.47	0.51	0.55	0.58	0.62	0.66
C18	463	0.45	0.51	0.57	0.63	0.70	0.76	0.82	0.88	0.94	1.00	1.00
C19	284	0.59	0.67	0.75	0.83	0.90	0.98	1.00	1.00	1.00	1.00	1.00
C20	485	0.21	0.23	0.25	0.27	0.30	0.32	0.34	0.36	0.38	0.40	0.42
C21	437	0.21	0.23	0.26	0.29	0.32	0.34	0.37	0.40	0.43	0.46	0.48
Total (n)	5050	5050	5747	6436	7064	7654	8190	8683	9101	9399	9745	9987

C
validated
0.65
0.37
0.40
0.75
0.61
0.44
0.84
0.56
0.72 *
0.26
0.68
0.62
0.22 *
0.53
0.62
0.71
0.33
0.50
0.65
0.27 *
0.25
5341

Figure 7. Restraining factors per function class and prediction model of full platelet proteome. Analysis of non-identified proteins ($n = 9{,}721$) from the relevant, combined PLT/MGK transcriptome per function class. Full dataset is provided in Suppl. Table 2. (**A**) Fraction of identified proteins in green. Well-identified classes with fractions > 0.55 labeled as ID. Indicated in red are each of three restraining factors per class: (i) over- represented low copy number (areas II-III in Fig. 5D), (ii) low mRNA level (area V, LM = low mRNA > 45%); (iii) retainment in megakaryocyte (peri)nucleus upon platelet shedding. Bottom: means of identified fractions (weighted for the presence of multiple factors); and correction factor in comparison to 'well-identified'. (**B**) Based on identified proteins ($n = 5{,}050$), modelled prediction of increased identification of missing proteins per class at higher proteomic detection. Shown per class are fractions of total relevant transcripts (heatmapped), and total expected proteins (bottom line). (**C**) Validation of prediction model based on novel proteome with 5,341 identified proteins.

with relevant transcript levels. Interestingly, nuclear-related proteins were more frequently present than was predicted, thus pointing to a more prominent incorporation of (peri)nuclear proteins in megakaryocyte-shed platelets than was anticipated.

The function class-based analysis of (non)identified platelet proteome, based on relevant transcript levels (log2fpkm ≥ 0.20) as well as the listing of 37.2 k genome-wide not expressed transcripts provides novel and detailed information on the presence of protein isoforms, subunits of complexes and metabolic, protein processing and signaling pathways (see Table 3). For instance, regarding the apoptosis-related Bcl/Bax proteins (C_{18}) involved in platelet clearance[49], the isoforms BNIP2, BCL2L1 (BCL-XL or BIM), BAD and BAK1 are present in the current proteome, while also the transcripts of BLC7B, BCL9 and BCL2 are highly expressed. As another example, regarding the glycosyl transferases (C_{16}) and epimerases (C_{12}) implicated in the surface glycosylation pattern and thereby in platelet survival time[50], prominently present in the proteome (transcriptome) are GALM, GALE, GNE, C1GALT1 and B4GALT1/3/4/5/6, while C1GALT1C1 (COSMC) is only lowly transcribed.

In this Covid-19 era, our list also provides information on ACE2, BSG and TMPRSS2. In platelets and megakaryocytes, ACE2 expression levels appear to be very low (log2fpkm 0.00–0.03), similar to the levels in other blood cells (https://blueprint.haem.cam.ac.uk/bloodatlas). On the other hand, BSG (basigin) with high transcript levels is present in the platelet proteome, but not the marginally expressed TMPRSS2.

Both network analysis and OMIM-based evaluation of the genes/proteins known to contribute to platelet count, hemostasis and thrombosis showed high coverage by the current platelet proteome and transcriptome dataset. Since still little is known of many of the proteins, the list of 20 k transcripts reveals a wealth of novel information on proteins that will influence platelet structure and function. Knowledge for understanding disease processes is still limited, as prior work from our and other labs describe only small-size alteration in platelet (phospho)proteomes of patients with Scott (*ANO6*)[27] or Glanzmann (*ITGA2B*)[48] disorders or with pseudo-hypoparathyroidism (*GNAS*)[28]. Altogether, this underscores that our approach to define a complete platelet proteome provides a valuable scaffold for further exploring and understanding platelet traits in and beyond thrombosis and hemostasis.

The current approach to define a classified full or theoretical platelet proteome from transcriptomes of platelets and megakaryocytes offers new insights into platelet composition and function, but also has limitations. As discussed above, platelets and megakaryocytes can bind and incorporate proteins from plasma, extracellular matrix or other cells, where the corresponding transcripts can be missing. In case of low transcript levels, copy numbers of proteins in platelets can be too low to be detected by mass spectrometric techniques (for detailed discussion on technical limitations, see supplementary methods). Furthermore, the source (individual healthy, diseased subject) and purification method of platelets and megakaryocytes can influence the specific composition of proteome and transcriptome, especially regarding the more rare molecules. It is noted here, that a subset of proteins expressed at very low copy numbers may be relevant for platelet ontogenesis, but have limited impact on platelet functions.

Earlier analyses indicated that the platelet proteome from healthy subjects is quite stable with < 15% of changes[51]. Similarly, the global platelet proteomes from the few patients, extensively studied so far—such as

Figure 8. Network-based potential roles of (non)identified proteins in platelet proteome in arterial thrombosis and hemostasis. Using a published meta-analysis of mouse genes in thrombosis and bleeding, the network was built in Cytoscape, containing 267 core genes (bait nodes), 2679 new nodes, connected by 19.7 k interactions[15]. (**A**) Redrawn network visualization with color-coded proteins identified (green) or not identified (red) in the platelet proteome, with relevant transcript levels (node size, log2fpkm). Names are listed of 40 proteins with highest mRNA expression levels. (**B**) Distribution profile of (non)identified proteins with transcript levels (median copy numbers, median log2pkm). No mRNA = below relevant threshold. Attribute lists are given in Suppl. Datafile 4.

Albright hereditary osteodystrophy, Glanzmann or Scott syndrome patients—showed only minor changes compared to that of control subjects[27,28,48]. the technical abilities to study this in the future is made in the revised discussion (page 16). In the near future, with the use of roboting techniques allowing higher throughput analysis of large sample sets and with the application of stable isotope markers[17], we expect to know more on the variable part of the platelet proteome in health and disease.

Methods

Subject cohorts and platelet samples. Washed, purified blood platelets were obtained in the same laboratories from six cohorts of healthy control donors, anonymized for medical-ethical reasons after informed consent. For each cohort, platelet samples were freshly isolated from anticoagulated blood by first collecting platelet-rich plasma, and removing plasma by a double wash step. Contamination was < 0.02% for red blood cells and leukocytes, presence of plasma about 1 vol%. Raw proteomic data per cohort are provided in the following papers. Cohort 1 ($n = 3$) in Burkhart et al.[24], cohort 2 ($n = 3$) in Beck et al.[25], cohort 3 ($n = 3$) in Beck et al.[26], cohort 4 ($n = 2$) in Solari et al.[27], cohort 5 ($n = 8$) in Swieringa et al.[28], and cohort 6 ($n = 3$) in Lewandrowski et al.[29]. Platelets were always derived from anonymous healthy donors, due to ethics restrictions also not revealing age

Class	Genes/proteins
C01 Cytoskeleton actin-myosin (*7/237*)	ANKRD26, **ACBT**, **ACTN1**, **ARPC1B**, **FLNA**, **MYH9**, **TPM4**, **WIPF1**
C03 Cytoskeleton microtubule (*1/420*)	**TUBB1**
C04 Cytoskeleton receptor-linked (*1/69*)	**FERMT3**
C05 Endosome proteins (*1/101*)	**LYST**
C06 ER & Golgi proteins (*9/568*)	**AP3B1**, **AP3D1**, **CISD2**, **COG6**, **HPS1**, **HPS4**, **LIMAN1**, **MCFD2**, **SLC35A1**
C08 Lysosome & peroxisome proteins (*3/168*)	ABCD4, **GBA**, HPS6
C09 Membrane & protein trafficking (*7/349*)	DTBBP1, HPS3, HPS5, **BLOC1S3**, **BLOC1S5**, **BLOC1S6**, **NBEAL2**
C10 Membrane receptors & channels (*23/1430*)	ACVRL1, **ANO6**, **CD36**, **CD40LG**, **CD55**, **CD81**, **CLCN7**, **ENG**, **FCGR2A**, **FCGR2C**, **GP1BA**, **GP1BB**, **GP6**, **GP9**, **ITG2A**, **ITG2B**, **ITGA2B**, **ITGB3**, **MPL**, **ORAI1**, **P2RY12**, **TBXA2R**, THBD, **TRPM7**
C11 Mitochondrial proteins (*5/814*)	**AGK**, **CYCS**, DGUOK, DHFR, **GPX4**
C12 Other metabolism (*5/877*)	**ADA**, ADA2, **DGKE**, **GALE**, **KDSR**
C13 Other nuclear proteins (*3/1469*)	ACD, CTC1, CDRE1C
C15 Protein kinases & phosphatases (*6/481*)	MASTL, **MPIG6B**, NBEA, **PTPN11**, **PTPRJ**, **SRC**
C16 Protein processing (*7/308*)	ACP5, ALG8, CHST14, GGCX, **GNE**, **HLCS**, **VKORC1**
C17 Secretory proteins (*19/811*)	**A2M**, ADAMTS13, **C2**, **C3**, **CFH**, ELANE, **F13A1**, **F5**, F8, FCGR2B, **KLKB1**, PLAT, PLAU, **PLG**, **PROC**, **PROS1**, **SERPINE1**, **SERPINF2**, **VWF**
C18 Signaling & adapter proteins (*11/1024*)	**CALR**, **DIAPH1**, **FYB**, GDF2, **GNAS**, **PLA2G4A**, **PTGS1**, **SH2B3**, **SMAD4**, **STIM1**, **TBXAS1**
C19 Small GTPases & regulators (*3/483*)	**CDC42**, **RASGRP2**, **WAS**
C20 Transcription & translation (*13/2280*)	**DKC1**, **DNAJC21**,ETV6, **FLI1**, FOXP3, GATA1, GATA2, GFI1B, IKZF5, MECOM, **RBM8A**, RUNX1, **SLFN14**
Absent in transcriptome (C17)	**F13B**, **F2**, **FGA**, **FGB**, **FGG**, **HABP2**, **HRG**, **SERPINC1**, **SERPIND1**
Absent in transcriptome and proteome	ABCG5, ABCG8, EPHB2, HOXA11, PRKACG

Table 4. Platelet-expressed proteins in whole-genome transcriptome implicated in hemostasis and thrombosis. Listed are per platelet function class genes expressed in the (non)identified platelet proteome, which according to recent OMIM, Bloodomics and overviews[23,34,35] in man contribute to bleeding, thrombocytopenia or thrombophilia. Coding as follows. **Bold**: identified in platelet proteome; green: bleeding or thrombocytopenia; red: thrombophilia; black: either reported.

or sex. New experimental work was approved by the Ethics Committee of Maastricht University and Maastricht University Medical Centre[28].

The genome-wide Blueprint gene expression data were generated from platelets obtained from venous blood ($n \geq 3$ per transcript, NHS Blood and Transplant healthy blood donors), and depleted from leukocytes[23,31]. Primary data are public accessible via https://blueprint.haem.cam.ac.uk/mRNA/ or htt ps://blueprint/haem.cam.ac.uk/bloodatlas/. [31]. Purity of platelets was checked by Sysmex, hemocytometer and from transcriptional signatures. Culturing of megakaryocytes ($n \geq 3$ per transcript) from cord blood, and check by flow cytometry (CD41 and CD42 double-positive) were as described[19]. Blood samples from healthy volunteers were obtained after full informed consent according to the Declaration of Helsinki.

Proteomes. In all reported studies, platelet lysates were analyzed according to a common bottom-up mass-spectrometry proteomics approach in the same laboratory. Experiments details are in the original papers[24–29]. Briefly, purified lysed platelets were subjected to a filter-aided sample preparation or ice-cold ethanol precipitation procedure. Isolated proteins were then trypsin-digested in guanidinium HCl or urea and (triethyl) ammonium bicarbonate (incubated over night at 37 °C). For global proteome analysis, complex peptide mixtures were fractionated by high-pH reversed phase chromatography (pH 6 or 8). For detection and quantification of platelet phospho-peptides, an enrichment procedure was included using TiO_2 beads, followed by hydrophilic interaction liquid chromatography (HILIC) fractionation. Fractions of peptides or phosphopeptides were analyzed by nano-liquid chromatography (LC)-MS/MS using QExactive (QStar Elite) and Orbitrap Velos mass spectrometers. Raw data were processed with Proteome Discoverer, SearchGui and Peptide Shaker implemented with Mascot and Sequest and X!Tandem search algorithms. Spectra were searched against a human UniProt-KB database. For database versions, see the original papers[24–29]. In all cases, a false discovery rate (FDR) of 1% was set.

Primary data deposits and links. Primary datasets were downloaded per proteome cohort via the website links of Table 1, also providing information on the deposited spectral datasets. In cohort one ($n = 3$ subjects), relative protein abundance levels[52] were determined in combination with a protein abundance estimate to give protein copy numbers per platelet[51]. In brief, protein copy numbers were assessed based on a normalized spectral abundance factor (NSAF) method. First, absolute quantification information was obtained from a set of 24 reference proteins (providing reference copy numbers), which then was used to correct NSAF indexes and was extrapolated to copy numbers of remaining proteins with known NSAF values.

In cohorts 2–5 ($n = 3, 3, 2, 8$ subjects, respectively), additional proteins were obtained without copy numbers, obtained from either global proteome analysis and/or phosphoproteome analysis[25–28]. In cohort 6 ($n = 3$ subjects), platelet membrane proteins were identified[29]. Presence of individual proteins per cohort is indicated in Suppl. Datafile 2.

Proteome tabling construction. The summative identified proteins with or without copy numbers, derived from global proteome or sub-proteome/enrichment (phospho-proteins or membrane proteins) analysis, were all checked in UniProt-KD (consulted January 2019—January 2020) and listed per corresponding gene (GeneCards). If no match between UniProt-KD assignment and gene name was found, additional gene databases were consulted (Biomart, Ensembl).

Transcriptomes. Genome-wide quantitative data of 57,849 transcripts assessed in human platelets and human megakaryocytes were established via a guided procedure by the Blueprint consortium[23,31]. For link to sources, see Table 1. For establishing relevant transcription levels, we used an arbitrary, low expression cut-off of log2fpkm ≥ 0.20, which included lowly abundant transcripts, to include all theoretical proteins presumably with very low levels (Suppl. Datafile 1).

Functional classification of protein-coding and other transcripts. The knowledge bases GeneCards (consulted January 2019—January 2020) was used to primarily separate protein-coding genes, RNA genes and pseudogenes. GeneCards provides comprehensive information on the annotated and predicted human genes, integrating gene-centered data from ~ 150 web sources[53]. Gene annotation was performed for all 20,425 gene transcripts (out of 57,849) with log2fpkm ≥ 0.20 in platelets and/or megakaryocytes.

For all relevant transcripts of protein-coding genes (log2fpkm ≥ 0.20), a supervised classification procedure was developed to combine the corresponding proteins into function classes. The classification was hierarchical, according to a yes/no decision tree (Fig. 1), instructed by the EMBL UniProt-KB knowledgebase (visited January 2019–January 2020)[54]. UniProt-based decisions were based on the general description in Uniprot-KB of the (putative) protein's intracellular location and cellular function. Priority order of decision assignment was according to classical cell biology, *i.e.* from central' to 'peripheric: nucleus → mitochondria → endoplasmic reticulum and Golgi apparatus → cell → other cellular vesicles (lysosomes, peroxisomes, endosomes, secretory vesicles) → (plasma) membrane interactions → cytoskeleton structures → cytosolic protein types. When no relevant information was available, proteins were classified as 'Uncharacterized and other proteins'. Note that (assumed) extracellular proteins were classified as secretory proteins, as these are considered to be released into the blood plasma by gland cells.

Area analysis of proteome-transcriptome space. For the matrix of 3,626 proteins with information on copy numbers and transcript levels in platelets (log2fpkm × 1000), a rectangular triangle was obtained, in which five areas (I-V) were pre-defined as follows. Top right corner, I ($x = 100{,}000$, $y = 8$, x-radius = 0.4, $n = 58$ PLT); top left corner, II ($x = 1000$, $y = 8$, x-radius = 0.3, n = 776 PLT), bottom left corner, III ($x = 1000$, $y = 0.75$, x-radius = 0.3, $n = 137$ PLT); middle of triangle, IV ($x = 5000$, $y = 4$, x-radius = 0.4, $n = 928$ PLT), and all below the triangle, V ($x = 600$–$200{,}000$, $y = 0.6$–10.2, $n = 185$ PLT). For each dot (protein) in the matrix, using Matlab the distance (in log space) was determined to each of the predefined areas; and recordings were made as in/out. Subsequently, for the proteins per function class, p-values of over-representation in pre-defined areas were calculated, employing a native Matlab function.

Proteome prediction modelling. For prediction of the 'missing' (non-identified) part of the platelet proteome, we generated a model that was based on the definition, per protein class of three restraining factors: (i) low protein copy number, (ii) low mRNA level, and (iii) protein retainment in megakaryocytes upon proplatelet formation. Therefore, per function class, the fraction of non-identified proteins was calculated from all transcripts with log2fpkm ≥ 0.20 in platelets and/or megakaryocytes, with an arbitrary setting of well-identified classes having < 45% 'missing proteins'. Classes with low copy numbers were obtained from the proteome-transcriptome matrix (over-representation in areas II and III); or when no other explanation for low identification was present. Classes with low mRNA levels were also taken from the proteome-transcriptome space (over-representation in area V); or when the transcript fraction with log2fpkm 0.20–1.00 was > 22.5% (arbitrary set at half of 45%). Classes with supposed protein retainment in megakaryocytes came from handbook knowledge, i.e. the 'nuclear classes' C_{13} and C_{20}; and furthermore C_3-cytoskeleton microtubule, given the retainment of mitotic spindle and centromere structures. Mean restraining factors were calculated from the averages of non-identified proteins in the corresponding classes. See further Suppl. Methods. Coverage of hemostatic pathways was checked in the Reactome database[55].

Model validation using extended novel proteome. To validate our model, platelet samples were collected as above from 30 healthy subjects, digested with trypsin, and analyzed by liquid chromatography-mass spectrometry. See further Suppl. Methods. Mass spectrometry proteomics data were deposited to the ProteomeXchange Consortium via the PRIDE partner repository[56] with the dataset identifier PXD022011 (username: reviewer_pxd022011@ebi.ac.uk; password: 7BeFQOxP).

Bioinformatics and statistics. Statistical comparison was by probability analysis in Excel (Mann–Whitney U-test or Student t-test for continuous variables). Distribution profiles were compared by a χ^2 test. Values of $p<0.05$ were considered significant.

Received: 5 March 2021; Accepted: 26 May 2021
Published online: 11 June 2021

References

1. Versteeg, H. H., Heemskerk, J. W., Levi, M. & Reitsma, P. S. New fundamentals in hemostasis. *Physiol. Rev.* **93**, 327–358 (2013).
2. Van der Meijden, P. E. & Heemskerk, J. W. Platelet biology and functions: New concepts and clinical perspectives. *Nat. Rev. Cardiol.* **16**, 166–179 (2018).
3. Werner, G. & Morgenstern, E. Three-dimensional reconstruction of human blood platelets using serial sections. *Eur. J. Cell. Biol.* **20**, 276–282 (1980).
4. Van Nispen tot pannerden, H. *et al.* The platelet interior revisited: Electron tomography reveals tubular alpha-granule subtypes. *Blood* **116**, 1147–1156 (2010).
5. Thon, J. N. & Italiano, J. E. Platelets: production, morphology and ultrastructure. *Handb. Exp. Pharmacol.* **210**, 3–22 (2012).
6. Pertuy, F. *et al.* Myosin IIA is critical for organelle distribution and F-actin organization in megakaryocytes and platelets. *Blood* **123**, 1261–1269 (2014).
7. Poulter, N. S. & Thomas, S. G. Cytoskeletal regulation of platelet formation: Coordination of F-actin and microtubules. *Int. J. Biochem. Cell. Biol.* **66**, 69–74 (2015).
8. Bender, M. *et al.* Dynamin 2-dependent endocytosis is required for normal megakaryocyte development in mice. *Blood* **125**, 1014–1024 (2015).
9. Becker, I. C. *et al.* Actin/microtubule crosstalk during platelet biogenesis in mice is critically regulated by Twinfilin1 and Cofilin1. *Blood Adv.* **26**, 2124–2134 (2020).
10. Akkerman, J. W. Regulation of carbohydrate metabolism in platelets: A review. *Thromb. Haemost.* **39**, 712–722 (1978).
11. Kramer, P. A., Ravi, S., Chacko, B., Johnson, M. S. & Darley-Usmar, V. M. A review of the mitochondrial and glycolytic metabolism in human platelets and leukocytes: implications for their use as bioenergetic biomarkers. *Redox Biol.* **2**, 206–210 (2014).
12. Nayak, M. K., Kulkarni, P. P. & Dash, D. Regulatory role of proteasome in determination of platelet life span. *J. Biol. Chem.* **288**, 6826–6834 (2013).
13. Colberg, L., Cammann, C., Greinacher, A. & Seifert, U. Structure and function of the ubiquitin-proteasome system in platelets. *J. Thromb. Haemost.* **18**, 771–778 (2020).
14. Boyanova, D., Nilla, S., Birschmann, I., Dandekar, T. & Dittrich, M. PlateletWeb: A systems biologic analysis of signaling networks in human platelets. *Blood* **119**, e22-34 (2012).
15. Baaten, C. C. *et al.* A synthesis approach of mouse studies to identify genes and proteins in arterial thrombosis and bleeding. *Blood* **132**, e35–e46 (2018).
16. Burkhart, J. M. *et al.* What can proteomics tell us about platelets?. *Circ. Res.* **114**, 1204–1219 (2014).
17. Loosse, C., Swieringa, F., Heemskerk, J. W., Sickmann, A. & Lorenz, C. Platelet proteomics: From discovery to diagnosis. *Exp. Rev. Proteomics* **15**, 467–476 (2018).
18. Van der Meijden, P. E. & Heemskerk, J. W. Platelet protein shake as playmaker. *Blood* **120**, 2931–2932 (2012).
19. Chen, L. *et al.* Transcriptional diversity during lineage commitment of human blood progenitors. *Science* **345**, 6204 (2014).
20. Wright, J. R., Amisten, S., Goodall, A. H. & Mahaut-Smith, M. P. Transcriptomic analysis of the ion channelome of human platelets and megakaryocytic cell lines. *Thromb. Haemost.* **116**, 272–284 (2016).
21. Davizon-Castillo, P., Rowley, J. W. & Rondina, M. T. Megakaryocyte and platelet transcriptomics for discoveries in human health and disease. *Arterioscler. Thromb. Vasc. Biol.* **40**, 1432–1440 (2020).
22. Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
23. Petersen, R. *et al.* Platelet function is modified by common sequence variation in megakaryocyte super enhancer. *Nat. Commun.* **8**, 16058 (2017).
24. Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. & Zahedi, R. P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* **75**, 1454–1462 (2012).
25. Beck, F. *et al.* Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. *Blood* **123**, e1–e10 (2014).
26. Beck, F. *et al.* Temporal quantitative phosphoproteomics of ADP stimulation reveals novel central nodes in platelet activation and inhibition. *Blood* **129**, e1–e12 (2017).
27. Solari, F. A. *et al.* Combined quantification of the global proteome, phosphoproteome, and proteolytic cleavage to characterize altered platelet functions in the human Scott syndrome. *Mol. Cell. Proteomics* **15**, 3154–3169 (2016).
28. Swieringa, F. *et al.* Diagnostic potential of phosphoproteome of prostaglandin-treated platelets from patients with confirmed or suspected pseudohypoparathyroidism type 1a linked to platelet functions. *Sci. Rep.* **10**, 11389 (2020).
29. Lewandrowski, U. *et al.* Platelet membrane proteomics: a novel repository for functional research. *Blood* **114**, e10–e19 (2009).
30. Stunnenberg, H. G. The International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
31. Grassi, L. *et al.* Cell type specific novel lncRNAs and circRNAs in the blueprint haematopoietic transcriptomes atlas. *Haematologica* https://doi.org/10.3324/haematol.2019.238147 (2021).
32. Geiger, J. *et al.* Response: platelet transcriptome and proteome: Relation rather than correlation. *Blood* **121**, 5257–5258 (2013).
33. Rowley, J. W. & Weyrich, A. S. Coordinate expression of transcripts and proteins in platelets. *Blood* **121**, 5255–5256 (2013).
34. Online Mendelian Inheritance in Man (OMIM): an online catalog of human genes and genetic disorders. https://omim.org (2020).
35. Palma-Barqueros, V. *et al.* Inherited platelet disorders: An updated overview. *Int. J. Mol. Sci.* **22**, 4521 (2021).
36. Schubert, S., Weyrich, A. S. & Rowley, J. W. A tour through the transcriptional landscape of platelets. *Blood* **124**, 493–502 (2014).
37. Frobel, J. *et al.* Platelet proteome analysis reveals integrin-dependent aggregation defects in patients with myelodysplastic syndromes. *Mol. Cell. Proteomics* **12**, 1272–1280 (2013).
38. Londin, E. R. *et al.* The human platelet: Strong transcriptome correlations among individuals associate weakly with the platelet proteome. *Biol. Direct* **9**, 3 (2014).
39. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
40. Zeiler, M., Moser, M. & Mann, M. Copy number analysis of the murine platelet proteome spanning the complete abundance range. *Mol. Cell. Proteomics* **13**, 3435–3445 (2014).
41. Sabrkhany, S. *et al.* Exploration of the platelet proteome in patients with early-stage cancer. *J. Proteomics* **177**, 65–74 (2018).

42. Van Holten, T. C. *et al.* Quantitative proteomics analysis reveals similar release profiles following specific PAR-1 or PAR-4 stimulation of platelets. *Cardiovasc. Res.* **103**, 140–146 (2014).
43. Fong, K. P. *et al.* Deciphering the human platelet sheddome. *Blood* **117**, e15–e26 (2011).
44. Middleton, E. *et al.* Sepsis alters the transcriptional and translational landscape of human and murine platelets. *Blood* **134**, 911–923 (2019).
45. Trugilho, M. R. *et al.* Platelet proteome reveals novel pathways of platelet activation and platelet-mediated immunoregulation in dengue. *Plos Pathog.* **13**, e1006385 (2017).
46. Salunkhe, V. *et al.* A comprehensive proteomics study on platelet concentrates: Platelet proteome, storage time and Mirasol pathogen reduction technology. *Platelets* **30**, 368–379 (2019).
47. Stokhuijzen, E. *et al.* Differences between platelets derived from neonatal cord blood and adult peripheral blood assessed by mass spectrometry. *J. Proteome Res.* **16**, 3567–3575 (2017).
48. Loroch, S. *et al.* Alterations of the platelet proteome in type I Glanzmann thrombasthenia caused by different homozygous delG frameshift mutations in ITGA2B. *Thromb. Haemost.* **117**, 556–569 (2017).
49. Quach, M. E., Chen, W. & Li, R. Mechanisms of platelet clearance and translation to improve platelet storage. *Blood* **131**, 1512–1521 (2018).
50. Lee-Sundlov, M. M., Stowell, S. R. & Hoffmeister, K. M. Multifaceted role of glycosylation in transfusion medicine, platelets, and red blood cells. *J. Thromb. Haemost.* **18**, 1535–1547 (2020).
51. Burkhart, J. M. *et al.* The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. *Blood* **120**, e73-82 (2012).
52. Colaert, N., Gevaert, K. & Martens, L. RIBAR and xRIBAR: methods for reproducible relative MS/MS-based label-free protein quantification. *J. Proteome Res.* **10**, 3183–3189 (2011).
53. Stelzer, G. *et al.* The GeneCards suite: from gene data mining to disease genome sequence analyses www.genecards.org. *Curr. Protoc. Bioinformatics* **54**, 1.30.31–33 (2016).
54. Dogan, T. *et al.* UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. *Bioinformatics* **32**, 2264–2271 (2016).
55. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
56. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: IMPROVING support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

Acknowledgements

JH, IP and IDS are supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement TAPAS No. 766118. JH is enrolled in a joint PhD program of the Universities of Maastricht and Santiago de Compostela (Spain); IP and IDS are enrolled in a joint PhD program of the Universities of Maastricht and Reading (UK). MF is supported by the British Heart Foundation (FS/18/53/22863). Research support by the Ministerium für Innovation, Wissenschaft und Forschung from Nordrhein-Westfalen, the Cardiovascular Centre (HVC) of Maastricht University Medical Centre+, the Centre for Molecular Translational Medicine (INCOAG, MICRO-BAT), the German Federal Ministry of Education and Research (BMBF 01EO1503) and the Deutsche Forschungsgemeinschaft (ZA 639/4-1 and JU 2735/2-1).

Author contributions

F.S., F.A.S., I.P. and I.D.S. analyzed and interpreted data and revised the manuscript; F.A.S., L.G., R.C., C.B., A.S., M.F. provided essential tools and revised the paper; J.H., M.F. and J.W.H. designed research, analyzed and interpreted data and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-91661-x.

Correspondence and requests for materials should be addressed to J.H. or J.W.M.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.