

ORIGINAL ARTICLE

Gene and protein expression in human megakaryocytes derived from induced pluripotent stem cells

Kai Kammers¹ | Margaret A. Taub² | Rasika A. Mathias^{3,4} | Lisa R. Yanek³ | Kanika Kanchan⁴ | Vidya Venkatraman⁵ | Niveda Sundararaman⁵ | Joshua Martin³ | Senquan Liu⁶ | Dixie Hoyle⁶ | Koen Raedschelders⁵ | Ronald Holewinski⁵ | Sarah Parker⁵ | Victoria Dardov⁵ | Nauder Faraday³ | Diane M. Becker³ | Linzhao Cheng⁶ | Zack Z. Wang⁶ | Jeffrey T. Leek² | Jennifer E. Van Eyk⁵ | Lewis C. Becker³

¹Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

²Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

³The GeneSTAR Program, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

⁴Division of Allergy and Clinical Immunology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

⁵Advanced Clinical Biosystems Research Institute, Barbra Streisand Woman's Heart Center, The Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA

⁶Division of Hematology and Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

Correspondence

Lewis C. Becker, Johns Hopkins GeneSTAR Research Program, 1830 East Monument Street, Room 8028A, Baltimore, MD 21205, USA.
Email: lbecker@jhmi.edu

Abstract

Background: There is interest in deriving megakaryocytes (MKs) from pluripotent stem cells (iPSC) for biological studies. We previously found that genomic structural integrity and genotype concordance is maintained in iPSC-derived MKs.

Objective: To establish a comprehensive dataset of genes and proteins expressed in iPSC-derived MKs.

Methods: iPSCs were reprogrammed from peripheral blood mononuclear cells (MNCs) and MKs were derived from the iPSCs in 194 healthy European American and African American subjects. mRNA was isolated and gene expression measured by RNA sequencing. Protein expression was measured in 62 of the subjects using mass spectrometry.

Results and Conclusions: MKs expressed genes and proteins known to be important in MK and platelet function and demonstrated good agreement with previous studies in human MKs derived from CD34+ progenitor cells. The percent of cells expressing the MK markers CD41 and CD42a was consistent in biological replicates, but variable across subjects, suggesting that unidentified subject-specific factors determine differentiation of MKs from iPSCs. Gene and protein sets important in platelet function were associated with increasing expression of CD41/42a, while those related to more basic cellular functions were associated with lower CD41/42a expression. There was differential gene expression by the sex and race (but not age) of the subject. Numerous genes and proteins were highly expressed in MKs but not known to play a role in MK or platelet function; these represent excellent candidates for future study of hematopoiesis, platelet formation, and/or platelet function.

Manuscript handled by: Matthew T. Rondina

Final decision: Matthew T. Rondina, 19 February 2021

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *Journal of Thrombosis and Haemostasis* published by Wiley Periodicals LLC on behalf of International Society on Thrombosis and Haemostasis

Funding information

Barbra Streisand Women's Heart Center; National Heart, Lung, and Blood Institute, Grant/Award Number: R01 HL141944, U01 HL107446 and U01 HL72518; National Institutes of Health, Grant/Award Number: R01 GM105705; National Cancer Institute, Grant/Award Number: P30-CA006973; American Heart Association, Grant/Award Number: 15PSGC24470098

KEY WORDS

gene expression, induced pluripotent stem cells, mass spectrometry, megakaryocytes, platelets

Essentials

- There is interest in deriving megakaryocytes (MKs) from pluripotent stem cells (iPSC) for biological studies.
- We established a comprehensive dataset of genes and proteins expressed in iPSC-derived MKs in healthy African and European American men and women ($n = 194$).
- Genes and proteins important in platelet function were associated with greater expression of CD41 and CD42a in derived MKs.
- Numerous genes and proteins were highly expressed in MKs but not known to play a role in MK or platelet function; these represent excellent candidates for further study.

1 | INTRODUCTION

Blood platelets are generated from bone marrow megakaryocytes (MKs), which transfer messenger RNA and structural and functional proteins to developing platelets.¹ MKs are a critical “target tissue” for understanding the regulation of transcripts and proteins in anucleate platelets, which possess limited transcriptional and translational capacity.

MKs comprise less than 0.01% of nucleated cells in bone marrow and are difficult to obtain in adequate numbers from human subjects, even using invasive bone marrow sampling.² To overcome this barrier, MKs have been generated from CD34+ cells in umbilical cord blood,³⁻⁷ peripheral blood,⁴ or bone marrow taken from brain-dead organ donors,⁶ or from induced pluripotent stem cells (iPSCs) using chemically defined forward programming.² We have generated iPSCs from peripheral blood mononuclear cells using non-integrating episomal vectors and subsequently developed an efficient method for differentiating these iPSCs into MKs. The experimental approach utilizes feeder-free and xeno-free conditions to generate CD34+CD45+ hematopoietic progenitor cells, followed by generation and expansion of CD41+CD42a+ MKs.⁸ The MKs also express the MK markers CD42b and CD61 and display polyploidy. We have shown using RNA sequencing that megakaryocyte-related genes are highly expressed and that the transcriptome of these iPSC-derived MKs differs markedly from their parent iPSCs.⁹

In this work we extend our prior study⁹ by describing the complete transcriptome of MKs derived from iPSCs in 194 healthy European American (EA) and African American (AA) subjects. We relate gene expression to the expression of MK markers in each subject. In a subset of 62 subjects, we also compare the MK transcriptome with the expressed proteome measured by mass spectrometry. In future work, these results will be helpful to better understand how genetic variation influences platelet biology and function.

2 | EXPERIMENTAL PROCEDURES

2.1 | Study participants

Peripheral blood mononuclear cells (MNCs) were obtained from 219 subjects enrolled in GeneSTAR,^{10,11} a prospective study of EA and AA families with a history of early onset coronary artery disease (<60 years of age). Subjects were healthy family members of affected probands. The study was approved by the Johns Hopkins Medicine Institutional Review Board and all participants provided written informed consent.

2.2 | Generation of iPSC and derived MKs

The protocols used to generate iPSC and derived MKs are described in detail by Liu et al.⁸ Briefly, human iPSC lines were reprogrammed from MNCs using non-integrating episomal vectors and expanded in Essential 8 medium on either Matrigel (1:30; BD Biosciences) or vitronectin (5 $\mu\text{g}/\text{cm}^2$; Life Technologies). iPSCs were passaged until sufficient cells were present for MK differentiation (mean passage number 6.7 ± 2.8 (standard deviation [SD])) but not further, to limit somatic mutations. All iPSC lines were >85% viable by trypan blue. All were capable of forming embryoid bodies, although differentiation potential for the three germ layers was not tested. In the 152 subjects with paired donor/iPSC samples we confirmed perfect matching of the iPSC to donor DNA for all (identity-by-descent) minor genotype differences, and identified only two samples with minimal copy number variation (one with a 0.18 Mb deletion and one with a 0.82 Mb amplification), neither of which map to our top 100 gene results.

iPSCs were differentiated into CD34+CD45+ hematopoietic progenitor cells using the “spin-embryoid body” method in feeder- and serum-free conditions. For each subject, single iPSCs were suspended in serum-free medium (SFM), and on day 14,

cells were harvested and seeded for MK culture,⁸ generating a cell population enriched for CD41+CD42a+ (CD41 = ITGA2B, CD42a = GP9). MKs were cultured for 5 days and harvested by placing 500,000 to 1,000,000 cells into 1.5 ml microcentrifuge tubes in SFM, followed by centrifugation at 1000 rpm for 5 min at room temperature.

The percent of cells in each MK culture expressing both CD41 and CD42a was determined on days 14 and 19 by flow cytometry using anti-human CD41-APC (BD Biosciences), CD42-eFluor 450 (eBioscience), and CD42b-FITC (eBioscience). All samples were analyzed with FACSCalibur or LSRII flow cytometer (BD Biosciences). Ig isotype controls were used to define gating limits of the side scatter/forward scatter dot plots in each experiment. MK pellets were frozen at -80°C for further analysis at Johns Hopkins University (mRNA sequencing) or Cedars-Sinai Medical Center, Los Angeles (mass spectrometry).

2.3 | MK RNA isolation, library preparation, and sequencing

After thawing of MK pellets, Quick-RNA™ MicroPrep (Zymo Research, Cat# R1050) was used for total RNA isolation. An Agilent Bioanalyzer was used for quality control prior to library creation, with RIN (RNA Integrity Number) over 8.0.

TruSeq RNA Library Preparation Kit v2 (Illumina, Cat# RS-122-2001 and RS-122-2002) was used to generate libraries. Poly-A RNA was first purified from 10 to 200 ng RNA, fragmented to about 150 to 200 nucleotides in length, and then converted to cDNA. End repair was performed to remove 3' end overhangs and fill in 5' overhangs; next an "A" base was added to the 3' end for adaptor ligation and PCR amplification was performed. The resulting library was quantified and quality checked on an Agilent Bioanalyzer using DNA 1000 chips. Libraries were uniquely barcoded and pooled for sequencing.

DNA sequencing was performed on an Illumina® HiSeq 2500 instrument using standard protocols for paired end 100 bp sequencing. Average yield was ~15 Gb of raw sequencing data per lane, or ~300 million reads per lane.

2.4 | mRNA-sequencing and data preprocessing

For alignment and assembly, we used the updated Tuxedo pipeline.¹² RNA-sequencing reads were aligned to the human genome (UCSC, hg19) using the spliced-read mapper HISAT2 in default mode (version 2.0.1¹³). Assembly of aligned RNA-seq reads into full-length transcripts representing multiple splice variants for each gene locus, including the *de novo* assembly option and the UCSC reference annotation genes.gtf (version archive-2014-06-02-13-47-56), was carried out by StringTie (version 1.3.3c¹⁴). Transcript abundances were quantified as FPKM (fragments per kilobase of transcript per million reads sequenced). For statistical analyses, we integrated

the results from StringTie into the software environment R (version 3.4.0¹⁵) and then used the software package Ballgown (version 2.8.0¹⁶). We ran the program gffcompare (<https://github.com/gpertea/gffcompare>) to obtain gene symbols for *de novo* transcripts that map to known genes stored in the reference annotation. To aggregate transcript abundances that belong to the same gene, we used the built-in function `gexpr` of the Ballgown package. We excluded low-abundance genes with median FPKM across all samples ≤ 1 . The filtered gene expression data were \log_2 (FPKM+1)-transformed for all downstream analyses.

2.5 | Differential mRNA expression analysis

Differential expression analyses were carried out at a gene level. For each gene we fitted a multivariable linear model to assess the effects of percent CD41+CD42a+ cells, as well as sex, age, and race, on gene expression. Statistical models were adjusted for the 10 RNA sequencing batches used and two surrogate variables were estimated directly from the filtered gene expression matrix capturing unknown, unmodeled, or latent sources of noise.¹⁷⁻¹⁹ For multiple comparison correction we calculated *q*-values from the observed moderated *P*-values.²⁰ Genes with calculated *q*-values <0.05 were considered statistically significant, controlling the expected false discovery rate (FDR) at 5%.

2.6 | Gene set enrichment analysis

Gene Ontology (GO)^{21,22} analysis was performed using the R package topGO (version 2.38.1²³) and results from the biological process ontology are reported. We used the algorithm "classic" in combination with the Fisher's exact test to assess gene group enrichment. A total of six analyses were performed, including: (1) genes positively associated with percent CD41+CD42a+ cells, (2) genes negatively associated with CD41+CD42a+ cells, (3) genes upregulated in females, (4) genes upregulated in males, (5) genes upregulated in AAs, and (6) genes upregulated in EAs. The total starting set of genes for all analyses included those expressed in the MKs and present in each of the six groups. For each of the six analyses we started with genes significant at an FDR of 0.05. Because we found very high numbers of genes significantly associated with percent CD41+CD42a+, for these two analyses we also limited the starting set of genes to those associated with at least a 2% change in CD41+CD42a+ to focus on those genes with larger effect sizes. The top 20 gene sets for each analysis where the fold enrichment for the gene set was >2 are reported.

2.7 | PINE network analysis

Using the network visualization tool PINE,²⁴ we also visualized a subnetwork of enriched GO categories/pathway terms to compare

the effect of percent CD41+CD42a+ across the 163 significantly expressed protein/gene pairs (see Figure 3). The resulting subnetwork consists of two central nodes for GO categories/pathway terms connected to its associated gene nodes. Differential expression data from the protein/gene pairs are represented as bar charts wherein the bar height indicates degree of fold change.

2.8 | Proteomics methods for mass spectrometry

Details regarding the proteomic methods are provided in supporting information. Briefly, after preparation of the MK pellets and liquid chromatography separation, data-independent acquisition mass spectrometry (DIA-MS) was performed using sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH®).²⁵ ProteoWizard v.3.0.6002,²⁶ DIA-Umpire,²⁷ X!Tandem Native v.2013.06.15.1,²⁸ X!Tandem Kscore v.2013.06.15.1,²⁹ and Comet v.2014.02 rev.2³⁰ were used to facilitate data conversion, generate pseudo spectra, and search databases. A final RT-normalized assay library was generated from all valid peptide spectrum matches filtered at peptide FDR of 1% with peptide probability cutoff ≥ 0.99 as previously described.³¹⁻³⁷ SWATH-targeted data analysis was carried out using OpenSWATH v.2.0.0.³⁸⁻⁴⁰ Data preprocessing and quantification was performed using mapDIA v2.4.1.⁴¹ Differential expression analyses were carried out as with gene expression, using multivariable linear models to examine the effects of percent CD41+CD42a+ positivity, sex, age, and race on expression of each protein. For multiple comparison correction we used *q*-values from the observed moderated

P-values. Proteins with calculated *q*-values < 0.05 were considered statistically significant, controlling the expected FDR at 5%.

3 | RESULTS

3.1 | Quality control and exclusion of samples

Twenty-four MK samples were excluded from the final dataset because of very low expression across all genes (\log_2 [FPKM +1] < 0.5). One sample was excluded due to uncertainty about identity. The remaining 194 samples were from 107 EAs (58 males, 49 females) and 87 AAs (35 males, 52 females). Mean age was 53 ± 13 years (range 29–86).

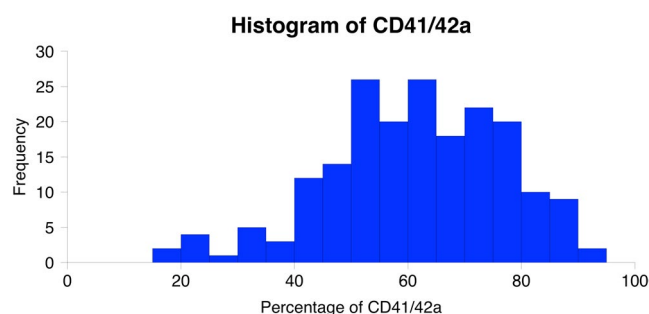


FIGURE 1 Distribution of percent CD41+CD42a+ in derived megakaryocyte (MK) samples. Percent CD41+CD42a+ positive cells in the 194 derived MK samples (one sample per subject). The mean percent CD41+CD42a+ positive cells was 61.3 ± 15.7 , range 16.3 to 93.0

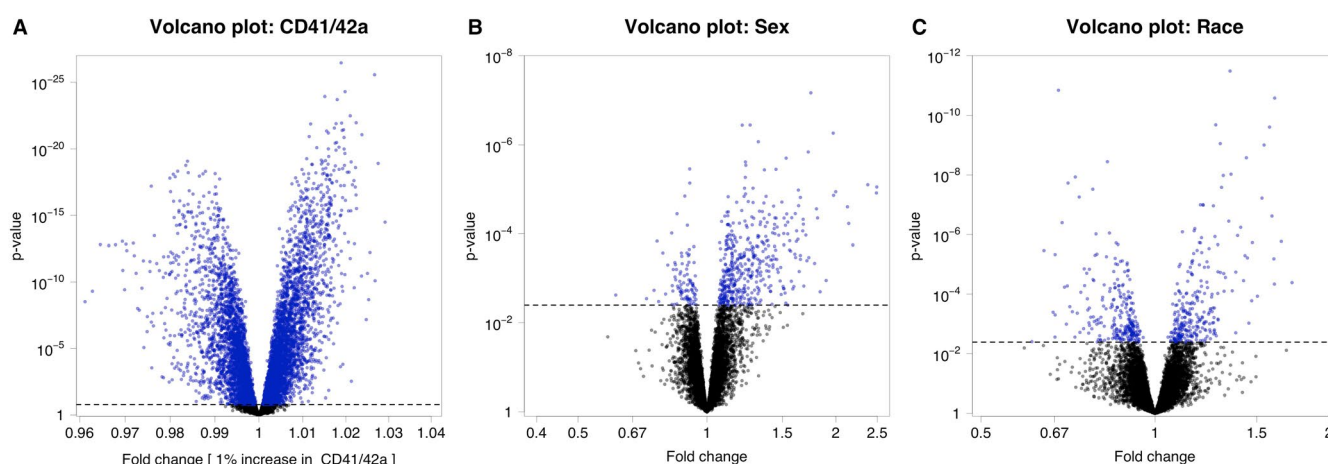


FIGURE 2 Volcano plots of differentially expressed genes in derived megakaryocytes (MKs) by (A) percent CD41+CD42a+, (B) sex, and (C) race. The false discovery rate--corrected significance level is indicated by the dashed line and statistically significant genes are colored in blue. Results are based on the multivariable linear models described in the Methods section. In (A), a value of, for example, 1.01 corresponds to an expected increase in gene expression of 1% per 1% increase in CD41+CD42a+. Among the 9596 genes expressed in MKs, 3868 genes were significantly positively related to percent CD41+CD42a+, while 3252 genes were significantly negatively related. In (B), a fold change value larger than 1 corresponds to an expected higher gene expression in women compared to men; 444 genes were expressed at a higher level in MKs derived from women than men, and 92 genes were expressed more in MKs from men than women. In (C), a fold change value larger than 1 corresponds to an expected higher gene expression in African Americans (AAs) compared to European Americans (EAs); 269 genes were expressed more in MKs derived from AAs than EAs, while 252 genes were expressed more in MKs from EAs than AAs

TABLE 1 Top genes associated with [A] increasing CD41+CD42a+ and [B] decreasing CD41+CD42a+

	Gene symbol	Beta	P-value	q-value	Median FPKM	chr
[A]	PDLIM1	0.0270	3.42E-27	7.41E-24	211.3248	chr10
	THBS1	0.0380	2.70E-26	2.92E-23	625.8913	chr15
	LTBP1	0.0283	5.08E-25	3.67E-22	31.1222	chr2
	APBB1IP	0.0216	1.11E-24	6.03E-22	9.6560	chr10
	TOM1L1	0.0257	1.91E-24	8.26E-22	5.8751	chr17
	MYLK	0.0300	3.16E-23	1.14E-20	13.6728	chr3
	TUBB1	0.0319	1.05E-22	2.82E-20	109.8738	chr20
	CCND3	0.0272	1.13E-22	2.82E-20	41.1812	chr6
	LMNA	0.0253	1.24E-22	2.82E-20	38.7618	chr1
	PHKB	0.0170	1.30E-22	2.82E-20	16.1992	chr16
	P2RY1	0.0248	2.66E-22	5.24E-20	23.3552	chr3
	MFSD6	0.0278	3.58E-22	6.46E-20	5.3907	chr2
	ZMAT5	0.0273	4.02E-22	6.71E-20	32.8957	chr22
	FERMT3	0.0222	4.44E-22	6.87E-20	263.2393	chr11
	MTURN	0.0228	6.29E-22	9.09E-20	24.1897	chr7
	F2R	0.0338	8.41E-22	1.14E-19	47.9644	chr5
	KLHL20	0.0165	1.16E-21	1.48E-19	3.5662	chr1
	MAP1A	0.0280	1.36E-21	1.63E-19	14.6602	chr15
	RAP2A	0.0255	3.26E-21	3.72E-19	13.1104	chr13
	ITGA2B	0.0250	3.66E-21	3.96E-19	167.3767	chr17
	VCL	0.0263	7.40E-21	7.63E-19	41.2992	chr10
	LAT	0.0190	7.99E-21	7.87E-19	183.6901	chr16
	HERC2	0.0228	9.86E-21	9.15E-19	13.7259	chr15
	EHD3	0.0280	1.01E-20	9.15E-19	13.8968	chr2
	ATP2C1	0.0253	1.24E-20	1.07E-18	38.9764	chr3
[B]	B3GNT8	-0.0235	8.41E-20	5.88E-18	1.4258	chr19
	THEMIS2	-0.0240	1.72E-19	9.35E-18	7.2596	chr1
	PIK3CD	-0.0267	4.76E-19	2.10E-17	2.5723	chr1
	CAMKK2	-0.0176	4.87E-19	2.11E-17	2.7458	chr12
	TBC1D10C	-0.0232	6.13E-19	2.51E-17	3.4972	chr11
	ERP29	-0.0144	7.22E-19	2.78E-17	15.8252	chr12
	TRAF3IP3	-0.0184	7.51E-19	2.79E-17	1.2115	chr1
	CITED4	-0.0282	7.99E-19	2.84E-17	2.7809	chr1
	B3GNTL1	-0.0196	1.11E-18	3.65E-17	1.8419	chr17
	ATG16L2	-0.0275	1.24E-18	4.00E-17	1.7801	chr11
	SYTL1	-0.0293	1.57E-18	4.92E-17	1.9222	chr1
	IFT122	-0.0129	1.84E-18	5.61E-17	1.9093	chr3
	PRAM1	-0.0354	6.36E-18	1.55E-16	3.2520	chr19
	R3HCC1	-0.0143	7.24E-18	1.73E-16	8.1098	chr8
	RCSD1	-0.0229	1.42E-17	3.07E-16	1.7502	chr1
	GPR97	-0.0290	1.64E-17	3.48E-16	1.0923	chr16
	GRAMD4	-0.0164	1.89E-17	3.94E-16	2.1462	chr22
	RPL13	-0.0140	2.42E-17	4.91E-16	112.0632	chr16
	SNORD60	-0.0217	2.51E-17	5.04E-16	5.7640	chr16
	RPLP1	-0.0145	2.58E-17	5.13E-16	1607.6814	chr15
	AGTRAP	-0.0182	3.60E-17	6.89E-16	4.8317	chr1

(Continues)

TABLE 1 (Continued)

Gene symbol	Beta	P-value	q-value	Median FPKM	chr
MYB	-0.0260	8.18E-17	1.35E-15	3.3375	chr6
ADAM8	-0.0281	8.70E-17	1.42E-15	4.2200	chr10
MVB12A	-0.0154	9.83E-17	1.59E-15	1.7990	chr19
LRRC75A-AS1	-0.0196	1.06E-16	1.67E-15	64.4369	chr17

Abbreviation: FPKM, fragments per kilobase of transcript per million reads sequenced.

TABLE 2 Gene sets associated with [A] higher percent CD41+CD42a+, [B] lower percent CD41+CD42a+, and [C] gene sets upregulated in females

	Gene Ontology (GO) ID	Term	Fold enrichment	q-value
[A]	GO:0007596	Blood coagulation	5.96	2.45E-18
	GO:0050817	Coagulation	5.96	2.45E-18
	GO:0007599	Hemostasis	5.84	3.82E-18
	GO:0050878	Regulation of body fluid levels	5.23	9.54E-18
	GO:0009611	Response to wounding	4.34	4.77E-16
	GO:0042060	Wound healing	4.57	7.55E-16
	GO:0030168	Platelet activation	6.94	1.53E-14
	GO:0007155	Cell adhesion	3.20	2.52E-11
	GO:0022610	Biological adhesion	3.20	2.52E-11
	GO:0002576	Platelet degranulation	7.38	2.74E-11
	GO:0001775	Cell activation	2.60	9.32E-09
	GO:0016477	Cell migration	2.66	3.08E-08
	GO:0048870	Cell motility	2.50	2.64E-07
	GO:0051674	Localization of cell	2.50	2.64E-07
	GO:0034109	Homotypic cell--cell adhesion	7.01	7.00E-07
	GO:0040011	Locomotion	2.32	9.12E-07
	GO:1903034	Regulation of response to wounding	5.90	9.12E-07
	GO:0061041	Regulation of wound healing	6.25	1.00E-06
	GO:0070527	Platelet aggregation	7.33	1.00E-06
	GO:0098609	Cell--cell adhesion	3.28	1.13E-06
[B]	GO:0006955	Immune response	3.94	5.22E-28
	GO:0002376	Immune system process	3.24	5.22E-28
	GO:0002274	Myeloid leukocyte activation	5.76	5.22E-28
	GO:0006887	Exocytosis	5.26	5.22E-28
	GO:0042119	Neutrophil activation	6.13	5.22E-28
	GO:0036230	Granulocyte activation	6.10	5.22E-28
	GO:0002443	Leukocyte mediated immunity	5.30	5.22E-28
	GO:0043299	Leukocyte degranulation	5.95	5.22E-28
	GO:0001775	Cell activation	4.27	5.22E-28
	GO:0002275	Myeloid cell activation involved in immune response	5.92	5.22E-28
	GO:0002444	Myeloid leukocyte--mediated immunity	5.89	5.22E-28
	GO:0002283	Neutrophil activation involved in immune response	6.09	5.22E-28
	GO:0043312	Neutrophil degranulation	6.09	5.22E-28
	GO:0002366	Leukocyte activation involved in immune response	5.24	5.22E-28
	GO:0002446	Neutrophil-mediated immunity	6.00	5.22E-28
	GO:0002263	Cell activation involved in immune response	5.22	5.22E-28

(Continues)

TABLE 2 (Continued)

	Gene Ontology (GO) ID	Term	Fold enrichment	q-value
	GO:0045321	Leukocyte activation	4.30	5.22E-28
	GO:0046903	Secretion	4.27	5.22E-28
	GO:0002252	Immune effector process	4.37	5.22E-28
	GO:0045055	Regulated exocytosis	5.27	5.22E-28
[C]	GO:0030198	Extracellular matrix organization	3.98	7.29E-14
	GO:0043062	Extracellular structure organization	3.71	2.12E-13
	GO:0048870	Cell motility	2.21	1.08E-10
	GO:0051674	Localization of cell	2.21	1.08E-10
	GO:0040011	Locomotion	2.15	1.08E-10
	GO:0016477	Cell migration	2.24	2.18E-10
	GO:0007155	Cell adhesion	2.25	5.77E-09
	GO:0022610	Biological adhesion	2.24	7.35E-09
	GO:0072359	Circulatory system development	2.29	1.34E-08
	GO:0001944	Vasculature development	2.57	1.62E-08
	GO:0001568	Blood vessel development	2.58	1.94E-08
	GO:0072358	Cardiovascular system development	2.54	2.09E-08
	GO:0048646	Anatomical structure formation involved in morphogenesis	2.27	3.17E-08
	GO:0009887	Animal organ morphogenesis	2.32	3.58E-08
	GO:0048514	Blood vessel morphogenesis	2.70	4.40E-08
	GO:0035295	Tube development	2.26	9.05E-08
	GO:0030199	Collagen fibril organization	6.28	1.83E-07
	GO:0001525	Angiogenesis	2.75	3.74E-07
	GO:0035239	Tube morphogenesis	2.31	9.02E-07
	GO:0097435	Supramolecular fiber organization	2.43	2.24E-06

3.2 | Most highly expressed genes in MKs

Transcripts from a total of 9596 distinct genes were identified in the iPSC-derived MKs. The top 100 most highly expressed genes (median FPKM 167–3935) included a number of genes known to be involved in platelet function (PF4, PPBP, ITGA2B, CD9), cell adhesion (THBS1, FERMT3, LGALS1), and cell motility (ACTB, ACTG1, CFL1, PFN1, TPM4), and also included hemoglobins and ferritins, immune system proteins, antioxidant enzymes, metabolic enzymes, calcium binding proteins, G-protein binding molecules, transcription factors, and a large number of ribosomal proteins. There were also many genes not previously known to play a role in MK or platelet function. A complete list of genes expressed by iPSC-derived MKs is provided at the study website http://www.biostat.jhsph.edu/~kkammers/GeneSTAR_MK/index.html.

3.3 | Effect of CD41+CD42a+ on gene expression

The percent of cells with surface expression of both CD41 and CD42a, an index of MK differentiation,⁴² was determined by flow cytometry for each sample. The mean \pm SD percent CD41+CD42a+ positive cells was 61.3 ± 15.7 for the 194 samples,

and there was no difference by sex or race (mean \pm SD 60.9 ± 16.6 for EA men, 59.1 ± 14.5 for EA women, 62.6 ± 15.2 for AA men, and 62.8 ± 16.1 for AA women, $P > 0.05$ in all cases) or by age (Spearman correlation 0.025, $P = 0.68$). Figure 1 shows the number of samples with specified CD41+CD42a+ values. In 22 pairs of biological replicates mean difference in CD41+CD42a+ was only 0.82 ± 10.4 (absolute difference 7.09 ± 7.5 ; Figure S1 in supporting information).

To determine whether gene expression in MKs was related to the percent of CD41+CD42a+ positive cells, we performed linear regression analysis for each gene across all samples. We included sex, age, and race, as well as known and unknown batch variables as covariates; 3868 genes were significantly positively related to percent CD41+CD42a+ (q -value < 0.05), while 3252 genes were significantly negatively related. Figure 2A is a volcano plot of expression of each gene (shown as fold change for each unit change of CD41+CD42a+; a value of 1.01 means that the expected increase in expression of genes is 1% per 1% increase in CD41+CD42a+) against $-\log_{10}$ of the P -value. All points above the dotted horizontal line are statistically significant at an FDR of 5%. Table 1A/B shows the 25 most significantly expressed genes, positively and negatively related to increasing percent CD41+CD42a+. The top 25 genes positively related to CD41+CD42a+ include some familiar genes important in platelet function (THBS1, P2RY1, F2R, and

TABLE 3 Top genes expressed [A] more in MKs from women than men and [B] more in MKs from men than women

	Gene symbol	Beta	P-value	q-value	Median FPKM	chr
[A]	MT2A	0.8065	6.83E-08	0.000455	11.8523	chr16
	MAPRE3	0.3351	3.58E-07	0.000799	1.1107	chr2
	TSPYL5	0.2735	3.60E-07	0.000799	1.2419	chr8
	CTHRC1	0.9797	5.44E-07	0.000907	3.8431	chr8
	RPP25	0.3985	8.50E-07	0.001133	1.6830	chr15
	TMEM176B	0.7860	1.45E-06	0.001606	2.1834	chr7
	TPM2	0.6143	1.99E-06	0.001892	20.3649	chr9
	RDH10	0.2986	2.43E-06	0.002021	1.9361	chr8
	PNPO	0.3016	2.88E-06	0.002030	6.2156	chr17
	BMP1	0.5356	3.56E-06	0.002030	1.9601	chr8
	SLC39A4	0.4454	3.66E-06	0.002030	4.2206	chr8
	PLCD3	0.2977	5.40E-06	0.002769	1.6482	chr17
	COL1A2	1.2483	7.91E-06	0.003511	15.6864	chr7
	COL3A1	1.3180	8.88E-06	0.003511	27.7951	chr2
	NPDC1	0.3571	9.46E-06	0.003511	1.0606	chr9
	DCBLD1	0.2917	9.48E-06	0.003511	2.1580	chr6
	GFPT2	0.7095	1.04E-05	0.003634	1.1680	chr5
	COL6A2	0.9991	1.14E-05	0.003634	10.3902	chr21
	COL1A1	1.3149	1.21E-05	0.003634	42.2856	chr17
	GPSM1	0.2993	1.27E-05	0.003634	1.2589	chr9
	COL6A1	0.9828	1.37E-05	0.003634	9.7299	chr21
	CERCAM	0.4585	1.45E-05	0.003634	3.2731	chr9
	SLC9A3R2	0.2876	1.45E-05	0.003634	1.4234	chr16
	FUCA2	0.1645	1.47E-05	0.003634	6.2667	chr6
	S100A12	0.7252	1.53E-05	0.003634	3.1477	chr1
[B]	COIL	-0.1342	3.49E-06	0.00203	10.1785	chr17
	INTS2	-0.1335	7.25E-06	0.00345	2.1349	chr17
	POT1	-0.1721	1.42E-05	0.00363	2.1872	chr7
	TCERG1L	-0.2324	3.55E-05	0.00531	1.1619	chr10
	ICA1	-0.1521	6.43E-05	0.00576	3.2449	chr7
	SKAP2	-0.2740	9.64E-05	0.00755	10.9814	chr7
	ARHGAP21	-0.3873	0.000147	0.00864	18.9704	chr10
	CNOT10	-0.1015	0.000194	0.00964	5.5947	chr3
	CEP104	-0.1371	0.000225	0.01032	2.6777	chr1
	PHF10	-0.1558	0.000268	0.01137	2.8991	chr6
	CYTL1	-0.3424	0.000269	0.01137	7.0313	chr4
	ANP32B	-0.1983	0.00027	0.01137	61.9684	chr9
	CKAP2	-0.1390	0.000289	0.01177	7.9531	chr13
	GTF3C5	-0.1449	0.00031	0.01206	16.6675	chr9
	TOP1	-0.1952	0.00032	0.01215	31.9450	chr20
	RSF1	-0.1402	0.000434	0.01402	3.4499	chr11
	MFSD11	-0.1003	0.000441	0.01409	3.4336	chr17
	SYK	-0.2496	0.000442	0.01409	14.6902	chr9
	FHOD1	-0.1924	0.000468	0.01453	22.7506	chr16
	TMIGD2	-0.2448	0.000493	0.01485	1.7628	chr19
	CERS2	-0.3042	0.000503	0.01490	46.9461	chr1

(Continues)

TABLE 3 (Continued)

Gene symbol	Beta	P-value	q-value	Median FPKM	chr
INPP5A	-0.1513	0.000505	0.01490	4.7767	chr10
HDHD2	-0.1579	0.000619	0.017055	10.9950	chr18
PYCR2	-0.2370	0.000631	0.017055	39.9532	chr1
TRIQQ	-0.2616	0.000632	0.017055	2.1452	chr8

Abbreviations: FPKM, fragments per kilobase of transcript per million reads sequenced; MK, megakaryocyte.

ITGA2B), genes involved in cell-cell adhesion (FERMT3, VCL), and genes related to actin and myosin (PDLIM1, MYLK).

The top 25 genes negatively related to CD41+CD42a+ include a variety of cellular regulatory genes (numerous ribosomal and endoplasmic reticulum genes, transcriptional activators, as well as calmodulin dependent protein kinase, and a subunit of PI3 K). Of the top 500 significant genes associated with CD41+CD42a+, all were still significant after deletion of samples with low percent CD41+CD42a+ cells (<50%). A complete list of genes associated with CD41+CD42a+ can be obtained from the study website.

Gene set enrichment analysis was done to determine functional groups of genes whose expression was related positively and negatively to percent CD41+CD42a+, using as the start set of genes those with an effect size of >2% and an FDR of 5%. Using the "biological process" ontology, we detected 233 gene sets at an FDR of 5% showing enrichment for the genes that were significantly upregulated in MKs with a higher percentage of CD41+CD42a+. These sets included many processes biologically relevant to MK and platelet function, and the top 20 most significant (all with fold enrichment >2) are shown in Table 2A. There were 407 gene sets that were significantly enriched with the 202 genes upregulated with decreasing percentage of CD41+CD42a+ (i.e., MKs less differentiated). These sets included a number of processes related to immune and inflammatory responses and leukocyte function, and the top 20 most significant gene sets (all with fold enrichment >2) are in Table 2B.

3.4 | Effect of age, sex, and race on gene expression

With the same linear regression model as described before we investigated whether age, sex, or race of subjects were associated with derived MK gene expression after adjustment for CD41+CD42a+ positivity. At an FDR threshold of 5%, no genes had differential expression in MKs by age. Sex, on the other hand, did have a significant effect on differential gene expression: 444 genes were expressed at a higher level in MKs derived from women than men, and 92 genes were expressed more in MKs from men than women. This is shown graphically in the volcano plot in Figure 2B, and the top 25 genes expressed more in MKs from women than men and vice versa are shown in Table 3A/B. The top 25 genes expressed more in MKs from women include five types of collagen, BMP1 (bone morphogenic protein), and TPM2

(β -tropomyosin, which stabilizes actin filaments). In contrast, the top 25 genes differentially expressed in MKs from men include a telomere protection gene (POT1), a secretory gene expressed in CD34⁺ hematopoietic cells (CYTL1), a tyrosine kinase involved in cell adhesion (SYK), and a cell adhesion receptor (TMIGD2). All of the genes differentially expressed by sex can be obtained from the searchable tables on the study website. Gene set enrichment analysis showed 265 gene sets significantly upregulated in females compared to males. The top 20 gene sets included cell motility and adhesion, angiogenesis, and collagen fibril organization (Table 2C). No gene sets were found for those genes significantly upregulated in males compared to females.

Race also had a significant effect on differential gene expression: 269 genes were expressed more in MKs derived from AAs than EAs, while 252 genes were expressed more in MKs from EAs than AAs. This is shown in the volcano plot in Figure 2C, and the top 25 genes expressed more in MKs from EAs than AAs and vice versa are shown in Table 4A/B. The top genes expressed significantly more in AAs include RHCE (an Rh blood group gene), TNNT1 (slow skeletal muscle troponin T), PF4V1 (a cytokine similar to platelet factor 4), GSTM4 (glutathione S transferase), and NMRK1 (nicotinamide riboside kinase 1, involved in the synthesis of NAD⁺). In contrast, the top genes expressed significantly more in MKs from EAs include CCS (copper chaperone for superoxide dismutase), SDHA (succinate dehydrogenase), MYH10 (myosin heavy chain), and EPHB4 (ephrin type B receptor, a subgroup of TK receptors). All genes differentially expressed by race can be obtained from the study website. Gene set enrichment analysis showed no gene sets significantly different by race.

3.5 | Most highly expressed proteins

After exclusions for quality control, a protein was included for quantification if it could be detected in >50% of the 62 MK samples. Other less consistently expressed proteins could reflect inter-individual variability, but were not included in the dataset. We identified 1229 distinct proteins meeting this criterion, and a complete list of these proteins is provided at the study website. Linear regression analyses showed that CD41+CD42a+ expression was significantly associated with protein quantity for 197 proteins; 178 proteins were associated with increasing CD41+CD42a+ at an FDR of 5%. The top 50 proteins positively associated with CD41+CD42a+ are shown in Table 5A and include several known to be important

TABLE 4 Top genes expressed [A] more in MKs from African Americans than European Americans and [B] more in MKs from European Americans than African Americans

	Gene symbol	Beta	P-value	q-value	Median FPKM	Chr
[A]	RHCE	0.4314	3.27E-12	2.07E-08	1.5236	chr1
	DNAAF3	0.6878	2.59E-11	5.46E-08	3.6528	chr19
	PSPH	0.3492	2.06E-10	3.10E-07	2.4287	chr7
	TNNT1	0.6585	2.45E-10	3.10E-07	3.0800	chr19
	RASGRP3	0.3752	8.83E-10	8.88E-07	5.6681	chr2
	PROK2	0.6271	9.83E-10	8.88E-07	1.1947	chr3
	RSU1	0.5253	2.64E-09	2.09E-06	48.6804	chr10
	ATP6V0E2	0.4369	9.41E-09	5.95E-06	2.8470	chr7
	NAA38	0.3925	1.04E-08	5.97E-06	26.0722	chr17
	IRS2	0.3828	2.61E-08	1.18E-05	9.7000	chr13
	NOTCH2NL	0.6150	6.01E-08	2.23E-05	4.8227	chr1
	C4orf33	0.2753	1.00E-07	3.28E-05	3.4482	chr4
	MRPL35	0.2609	1.01E-07	3.28E-05	4.4603	chr2
	PCTP	0.2773	1.04E-07	3.28E-05	12.4859	chr17
	DOCK10	0.3355	1.09E-07	3.29E-05	1.8223	chr2
	PF4 V1	0.6723	2.41E-07	6.92E-05	68.0776	chr4
	ATP8A1	0.4241	3.58E-07	9.85E-05	5.4787	chr4
	PLA2G4C	0.4928	5.75E-07	0.000145	5.7795	chr19
	ADI1	0.2289	6.38E-07	0.000155	8.7925	chr2
	NHLRC2	0.2251	7.54E-07	0.000177	5.1179	chr10
	LIPT2	0.1578	8.17E-07	0.000185	1.1703	chr11
	IFITM3	0.4747	1.08E-06	0.000221	41.3854	chr11
	GSTM4	0.3445	1.17E-06	0.000232	10.3624	chr1
	NMRK1	0.2672	1.28E-06	0.000246	5.3224	chr9
	SV2C	0.7260	1.72E-06	0.000319	7.5199	chr5
[B]	CCS	-0.5552	1.45E-11	4.58E-08	9.4941	chr11
	SDHA	-0.2738	3.58E-09	2.52E-06	11.2003	chr5
	PPIL3	-0.4585	1.17E-08	6.17E-06	5.0189	chr2
	LAMA5	-0.5002	1.87E-08	9.11E-06	2.1665	chr20
	TRIM52-AS1	-0.3578	3.02E-08	1.27E-05	3.5586	chr5
	PPM1H	-0.4359	5.52E-08	2.18E-05	1.4975	chr12
	SLC39A4	-0.5342	4.01E-07	0.000106	4.2206	chr8
	SORD	-0.2618	8.95E-07	0.000195	2.0281	chr15
	WNT5B	-0.3391	9.59E-07	0.000202	2.0332	chr12
	MRPS7	-0.2234	1.92E-06	0.00032	18.4382	chr17
	MCCC1	-0.2129	2.17E-06	0.000346	2.9210	chr3
	SPATC1L	-0.6386	3.44E-06	0.000506	3.3551	chr21
	NEFH	-0.3340	4.54E-06	0.000624	1.3608	chr22
	MYH10	-0.5743	4.70E-06	0.000631	3.9504	chr17
	MZT2A	-0.3448	4.98E-06	0.000656	13.3315	chr2
	DHRS4	-0.2794	5.55E-06	0.000702	6.9083	chr14
	LCMT2	-0.2292	6.30E-06	0.000765	2.2515	chr15
	ABCB8	-0.1880	7.62E-06	0.000875	2.0641	chr7
	C7orf13	-0.1987	7.97E-06	0.000885	1.6348	chr7
	EIF6	-0.1595	8.95E-06	0.000958	16.7769	chr20

(Continues)

TABLE 4 (Continued)

Gene symbol	Beta	P-value	q-value	Median FPKM	Chr
RRP12	-0.2264	9.58E-06	0.000993	3.3077	chr10
HEBP2	-0.2155	1.01E-05	0.001006	1.4548	chr6
RNF121	-0.1610	1.03E-05	0.001006	1.7547	chr11
EPHB4	-0.4071	1.42E-05	0.001278	2.3440	chr7
LINC00116	-0.3774	1.46E-05	0.001298	5.8081	chr2

Abbreviations: FPKM, fragments per kilobase of transcript per million reads sequenced; MK, megakaryocyte.

in platelet function (THBS1, ITGA2B, PPBP, PF4, PTGS1 [COX1], ITGB3, and three components of the von Willebrand factor (VWF) receptor: GP1BB, GP1BA, and GP9); proteins involved in cell adhesion (LGALS3BP, LGALS1, ESAM, FERMT3); actin-binding proteins (COTL1, VASP, PSTPIP2, PFN1); and a number of proteins involved in G-protein signaling (GNB1, RAC2, RAB27B, RGS10), ion channels, and transport of various molecules. The 19 proteins significantly negatively related to CD41+CD42a+ included annexins; endoplasmic reticulum proteins; actin and cytoskeleton regulatory proteins; and proteins involved in transcriptional pausing, DNA repair, cell migration, cilia function, and proteasome function (Table 5B). Age, sex, and race were not associated with differential protein expression.

The 1229 expressed proteins were mapped to gene symbols via uniprot.org, and the results for protein and gene expression were compared: 1148 genes corresponding to these proteins were present in the gene expression results. The expressed proteins with essentially no gene expression included albumin, which was most likely endocytosed by MKs from the culture medium and not actually transcribed/translated from the MK genome. Endocytosis of plasma proteins may explain why certain alpha granule proteins, but not their genes, are present *in vivo* in platelets.⁴³ This includes factor XI (GDF11), HGF, IGF1, PLG (from which angiostatin is cleaved), COL18A1 (from which endostatin is cleaved), BMP2, and BMP4.⁴³ We found that some of these genes are expressed in MKs (GDF11, COL18A1, BMP2), even though they were not present in platelets in the previous study,⁴³ consistent with the transcripts either not being transferred from the MKs to platelets or being degraded once transferred.

We compared the direction and magnitude of CD41+CD42a+ expression (i.e., MK differentiation) on protein and gene expression using the 163 overlapping proteins/genes for which both were significantly associated with CD41+CD42a+. In the great majority of cases (92.6%), changes in protein and transcript expression were in the same direction (Figure 3A). Spearman correlation of median expression levels of these shared proteins/genes was 0.52 ($P < 2.2 \times 10^{-16}$). The lack of a higher correlation suggests that there are other regulatory effectors of cellular protein quantity than mRNA levels, which could include variations in translation, proteolysis, and transcript degradation.

We used PINE analysis to visualize the subnetwork of gene transcript/protein pairs associated with CD41+CD42a+ expression. An example of network analysis for the two GO categories of “platelet

activation, signaling, and aggregation” and “integrin-mediated cell adhesion” is shown in Figure 3B. The direction of effect sizes (shown as bar charts for each gene/protein pair) was the same for all expressed gene/protein pairs, and in most cases, effect sizes were comparable for gene and protein. Expression of genes and proteins was increased for many molecules well recognized to be important to platelet function, including molecules comprising the fibrinogen (itga2b, itgb3) and VWF (gb1ba, gb9) receptors, g-proteins (gna12, gna13, gnb1), the cytoskeleton (actn1, tuba4a, tln1, plek, pfn1), guanosine triphosphate (GTP)-related signaling molecules (rap1a, rac2, rasgrp2), protein kinase activators (pkca, lyn, rab27b), and proteins secreted from platelets (pf4, thbs1, ppbp, f13a1, mmrn1). Perhaps less anticipated were increases in expression of transcript and protein for: mesencephalic astrocyte derived neurotrophic factor (manf, a protein localized to endoplasmic reticulum and golgi and important to dopaminergic neuron survival), transgelin (tagln2, a protein found in smooth muscle cells whose function is unclear), and galectin 3 binding protein (lgals3 bp). While galectins 1 and 8 are strong platelet agonists known to be present and released from activated platelets and play a role in the uptake of factor V by MKs, the role of lgals3 bp in platelet and MK function is unclear.⁴⁴

3.6 | Comparison to other hematopoietic progenitors

Ideally, MKs would be available from bone marrow for comparison to iPSC-derived MKs, but their scarcity in marrow has limited direct study of MK gene and protein expression. Prior studies have therefore relied on generation of MKs from CD34+ hematopoietic cells³⁻⁷ or from iPSCs using chemically defined forward programming.² The Bloodomics Consortium described gene expression using microarrays in CD34+ derived MKs from four cord blood samples as part of the HaemAtlas project of human differentiated blood cells.³ The numbers of genes expressed in derived MKs was similar between our study ($n = 9596$) and the Bloodomics study ($n=10,444$ probe sets mapping to 9089 unique genes, their Table S3), but of these, there were only 6867 genes with matching annotated gene symbols; 5958 of the 6867 genes (87%) were expressed in both our study and the Bloodomics study.

Bloodomics found 289 probe sets mapping to 263 unique genes expressed specifically by MKs in comparison to other blood cell types (their Table S5). Of these 263 genes, there were 208 with

TABLE 5 Top proteins associated with [A] increasing CD41+CD42a+ and [B] decreasing CD41+CD42a+

	Gene symbol	Beta	P-value	q-value	Median expression
[A]	B2 M	0.035863	5.65E-06	0.003352	346.7
	LGALS3BP	0.038867	3.08E-05	0.004993	714.1
	STOM	0.02458	4.03E-05	0.004993	39154.7
	GNB1	0.016162	5.53E-05	0.004993	2347.6
	ATP2C1	0.031223	5.99E-05	0.004993	2062.8
	THBS1	0.035832	6.68E-05	0.004993	504056.5
	HPSE	0.031029	7.85E-05	0.004993	3575.8
	COTL1	0.021648	8.11E-05	0.004993	7465.8
	TTYH3	0.025442	8.88E-05	0.004993	513.9
	TIMP1	0.021845	9.68E-05	0.004993	6369.6
	VASP	0.018307	0.000101	0.004993	16465.9
	STXBP2	0.026981	0.000115	0.004993	3000.3
	SLC44A1	0.031071	0.000123	0.004993	968
	ESAM	0.026414	0.000145	0.00537	610.6
	INF2	0.03094	0.000175	0.005771	472.8
	ITGA2B	0.028288	0.000184	0.005771	103782.5
	PPBP	0.036772	0.000185	0.005771	105275.5
	LTBP1	0.021508	0.000202	0.005771	5410
	EPS15	0.014599	0.000204	0.005771	146
	BTK	0.026331	0.000227	0.006135	4802.7
	ITGB3	0.02838	0.000241	0.006229	38273.1
	RAC2	0.027483	0.000266	0.006395	9458.8
	RAB27B	0.031002	0.00028	0.006395	16682.8
	MMRN1	0.029524	0.000293	0.006395	8515.8
	GP1BB	0.034905	0.000298	0.006395	25123.1
	RG510	0.021232	0.000312	0.006395	6490.6
	EMD	0.011212	0.000313	0.006395	2069.5
	GP1BA	0.035435	0.00034	0.00672	7413.6
	OXCT1	0.01956	0.0004	0.007252	212.3
	CLIC1	0.017771	0.000403	0.007252	48946.1
	PTGS1	0.027477	0.000422	0.00736	8812.9
	LAP3	0.01264	0.000444	0.007532	5323.6
	PF4	0.03772	0.000461	0.007606	21666
	NRGN	0.025664	0.000494	0.007763	3525.1
	ASAH1	0.018808	0.000508	0.007763	811
	PPIF	0.015367	0.000531	0.007763	8305.8
	LGALSL	0.03695	0.000532	0.007763	3273.6
	CORO1C	0.017928	0.000536	0.007763	18280.6
	PTPRJ	0.028775	0.000552	0.007795	2866.3
	LY6G6F	0.020041	0.000592	0.007934	823.3
	DIAPH1	0.020883	0.000612	0.007934	20427.9
	FERMT3	0.026229	0.000612	0.007934	58259.5
	GP9	0.026693	0.000624	0.007934	10325.2
	PSTPIP2	0.023908	0.000628	0.007934	2603.5
	PFN1	0.014849	0.000683	0.008439	149442
	WDR1	0.018022	0.000758	0.009185	43066.3

(Continues)

TABLE 5 (Continued)

	Gene symbol	Beta	P-value	q-value	Median expression
	FAH	0.013117	0.000828	0.009614	6423.8
	EMILIN1	0.020691	0.000831	0.009614	4196.9
	FLNA	0.014951	0.000887	0.00975	309848
[B]	ANXA1	-0.02817	0.000123	0.004993	8231.19
	MARCKSL1	-0.02945	0.000126	0.004993	129.47
	NELFB	-0.02077	0.000392	0.007252	166.19
	RPN2	-0.01752	0.000843	0.009614	2398.2
	ANXA6	-0.02123	0.000861	0.009642	4249.09
	PHGDH	-0.02287	0.001215	0.012016	3445.365
	MSH6	-0.01726	0.001371	0.012917	166.28
	AHNAK	-0.02902	0.002594	0.019736	809.92
	DNAH5	-0.05604	0.003543	0.023457	513.83
	BASP1	-0.01955	0.003597	0.023457	615.165
	UGP2	-0.0074	0.003886	0.02427	614.08
	JUP	-0.02516	0.007875	0.035726	83.195
	BPI	-0.03993	0.008235	0.035726	987.88
	PFN2	-0.03442	0.008345	0.035883	196.455
	CAPG	-0.0175	0.011177	0.040941	3201.575
	CPSF2	-0.01117	0.01164	0.041164	217.64
	PRMT1	-0.01092	0.01191	0.041725	484.98
	CKAP4	-0.02484	0.012089	0.041725	2097.675
	PSMB5	-0.01378	0.013324	0.043681	845.125

annotated gene symbols in common with ours, and of these we identified 175 (84%) in iPSC-derived MKs with a median FPKM threshold >1. We also identified 44/50 lineage-specific MK genes reported by Macaulay et al.⁵ We identified all 39 autosomal plasma membrane receptor genes reported by Sun et al.⁶ to be expressed by MKs (derived from CD34+ bone marrow cells) during megakaryocyte development. The most common of these genes included ADAM10, CD55, CD63, F2R, ICAM2, IL6ST, ITGA2B, ITGB1, ITGB3, PTG1R, and SELP, but genes less well known to have a role in MK and/or platelet biology (CCRL2, SEMA4D, ADIPOR2, IL21R) were also found and confirmed in our study.

4 | DISCUSSION

This is the first study to comprehensively characterize the MK transcriptome and proteome in a large number of healthy subjects comprising both sexes, African and European Americans, and a wide range of ages. MKs were derived from iPSCs reprogrammed from MNCs in each subject, allowing us to compare gene and protein expression profiles by sex, race, and age of the MNC donor. In addition, we found that the percent of cells expressing the MK surface markers CD41 and CD42a varied among subjects despite a consistent derivation protocol, and that the expression profile of genes and proteins was strongly dependent on MK

CD41+CD42a+ expression. We previously showed in 14 subjects that megakaryocyte- and platelet-related genes are highly expressed in iPSC-derived MKs and that their transcriptome differs markedly from their parent iPSCs.⁹ Our prior study verified that there is high genomic structural integrity and genotype concordance between MKs and their parent iPSCs. Our current study expands on these prior findings to more fully understand gene and protein expression in iPSC-derived MKs, which are currently being studied as a donor-free source of platelet production for transfusion medicine.⁴⁵⁻⁴⁷

We found nearly 4000 genes and 178 proteins that were significantly positively associated with percent CD41+CD42a+. The most significant of these genes and proteins included many well known to be important in platelet function (Tables 1A and 5A). In contrast, about 3000 genes and 19 proteins were significantly negatively associated with percent CD41+CD42a+, the most significant of which included genes/proteins involved in basic cellular functions, immune and inflammatory responses, and leukocyte function (Tables 1B and 5B). These genes and proteins are all consistent with less mature MKs. Although we infer that as MKs mature, genes associated with increasing percent CD41+CD42a+ are expressed more while genes associated with decreasing CD41+CD42a+ are expressed less, our study is limited by the lack of a temporal analysis of gene expression as the MKs pass through different stages of maturity.

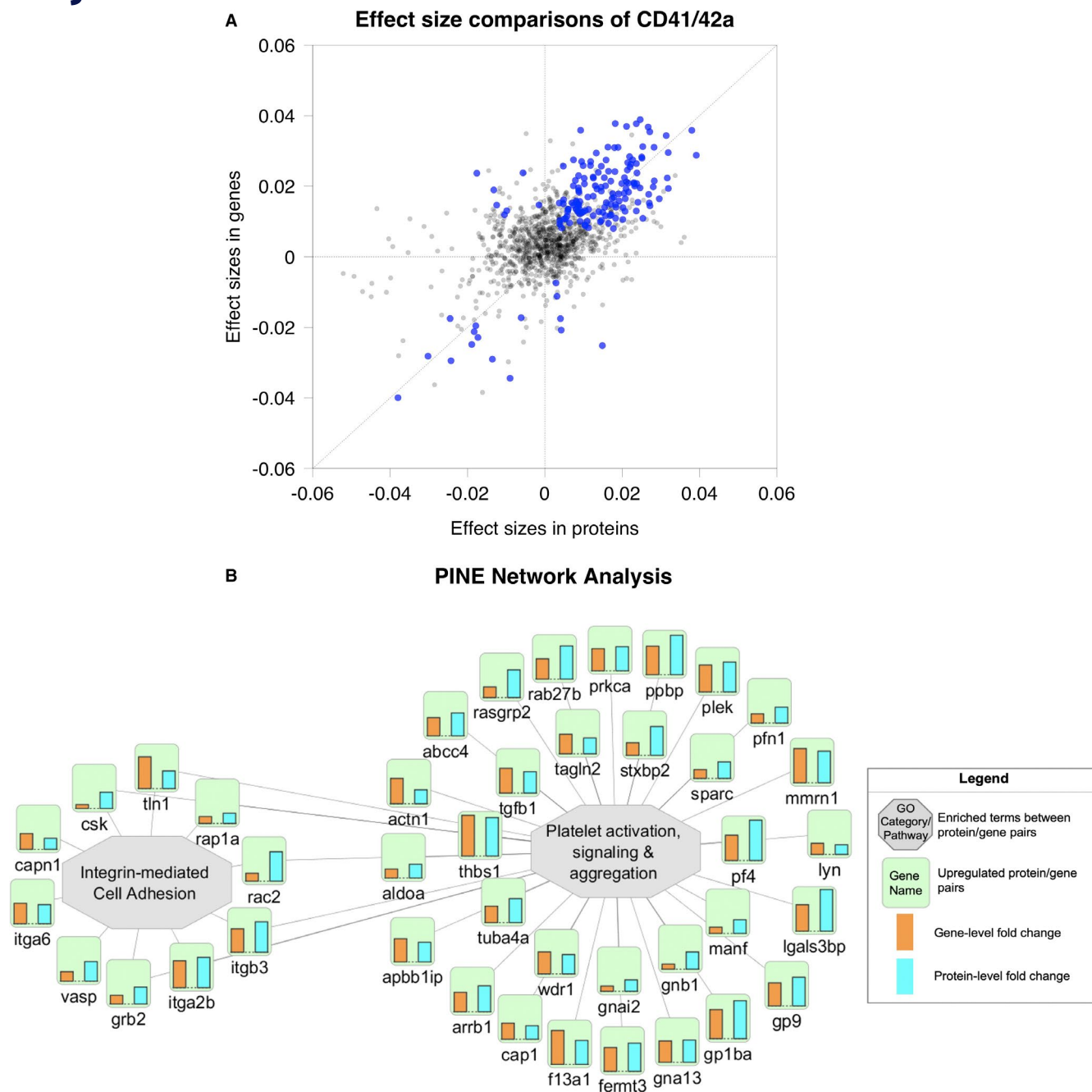


FIGURE 3 Comparison of gene and corresponding protein expression in derived megakaryocytes (MKs). A, Scatterplot of the effect sizes (beta coefficient) of percent CD41+CD42a+ on gene expression and corresponding protein expression using the 1148 overlapping proteins/genes; 163 protein/genes were both significantly associated with CD41/42a, either positively or negatively (highlighted in blue). The direction of effect sizes was the same for 92.6% of these expressed proteins/gene pairs. B, PINE-generated subnetwork of effect sizes of CD41+CD42a+ for the 163 gene/protein pairs associated with CD41+CD42a+. The direction of effect sizes (shown as bar charts with orange indicating gene-level fold change and blue indicating protein-level fold change) was the same for all expressed proteins/gene pairs (highlighted in green) involved in integrin-mediated cell adhesion and platelet activation, signaling, and aggregation (shown in gray).

We found that after adjusting for percent CD41+CD42a+ positivity, there was significant differential expression of genes in iPSC-derived MKs based on sex of the donor, even though MNCs from both sexes were reprogrammed to iPSCs with identical protocols. MKs from female donors significantly expressed 444 more genes than MKs from male donors ($q < 0.05$); the top genes included five

types of collagen, BMP1, and TPM2. Significant gene sets included cell motility and adhesion and collagen fibril organization. If similar sex-related differences in gene expression occur in natural bone marrow MKs, they may provide a clue as to why platelet aggregability is greater in women than men.⁴⁸⁻⁵² We have previously reported that platelets from premenopausal women demonstrate more GPIIb-IIIa

activation than those from men following stimulation with ADP or TRAP⁴⁸ and in more than 1200 subjects from GeneSTAR, women's platelets were significantly more reactive to arachidonic acid, ADP, collagen, and epinephrine.⁴⁹ MKs from male donors significantly expressed 92 more genes than MKs from females, but it is not clear how these specific genes would differentially affect platelet or MK function.

Race of the donor also had an effect on differential gene expression in iPSC-derived MKs. MKs derived from AAs significantly expressed 269 more genes than MKs derived from EAs, while MKs from EAs differentially expressed 252 more genes than MKs from AAs ($q < 0.05$) (Table 3A/B, Figure 2C). No genes were differentially expressed by age, even though the donor age range was large (29–86 years). From the proteomic analysis, no proteins were differentially expressed by sex, race, or age. It is likely that the relatively small sample size available for proteomics was responsible for our inability to find statistically significant differences in protein expression.

Differential expression of autosomal genes in MKs by sex and race is consistent with incomplete reprogramming of the parent cell, which may continue to carry an epigenetic signature of the donor MNCs. MKs derived from these iPSCs may retain epigenetic characteristics, and as a result express a different gene profile based on sex and race. Lister et al.⁵³ reported that insufficient reprogramming resulting in memory of progenitor somatic cell methylation state is common, and there appear to be hotspots for failed epigenetic reprogramming in iPSCs. Kim et al.⁵⁴ found that iPSC derived from factor-based reprogramming harbor residual DNA methylation signatures characteristic of their somatic tissue of origin. Ronen and Benvenisty⁵⁵ detected more than 200 differentially expressed autosomal genes in reprogrammed iPSCs from male and female subjects. In the case of sex differences in MK gene expression, sex chromosomes could contribute to the differences through both cis and trans mechanisms. It is also possible that deriving MKs from iPSCs with more passages would have reduced epigenetic memory in the iPSCs, and that our finding of differentially expressed genes by sex or race would have been different. Another limitation is that we did not validate our findings of differential gene expression by sex and race in primary bone marrow MKs or in circulating platelets.

We quantified a much greater number of genes than proteins expressed by MKs derived from the same subjects. This was most likely due to a difference in sensitivity provided by the two approaches (RNAseq of extracted mRNA vs. SWATH-DIA mass spectrometry of trypsin digested peptides). Both approaches balanced sensitivity and specificity, while minimizing false positives. It is unlikely biologically that large numbers of mRNA transcripts would be expressed without corresponding protein translation. Gene expression by RNAseq has been validated in many prior studies against qPCR.^{56–59} Previous studies have examined the relation between steady state gene expression and protein expression in a wide variety of organisms, including mammalian cells, with most studies finding a correlation of >0.50 ,^{60–63} which is similar to what we found. Differences between gene expression and protein abundance have been attributed to

variations in the rates of translation, protein degradation, and/or transcript degradation.

We found very good agreement between gene expression in iPSC-derived MKs and previous studies in which MKs were derived from CD34⁺ progenitor cells in cord blood or bone marrow, suggesting that either can be used to generate MKs *in vitro*. Originating from peripheral blood specimens, iPSC-derived MKs are more accessible than MKs derived from cord or marrow CD34⁺ cells, and they have potential to generate platelets *in vitro*. Moreau et al.² reported the large-scale production of MKs from iPSCs using chemically defined forward programming. Gene expression showed enrichment of categories consistent with MK/platelet function (platelet activation, platelet degranulation, response to wounding, vesicle mediated transport) and downregulation of pluripotency features.² However, differential expression analysis comparing cord blood and iPSC-derived MKs showed distinct differences. MKs derived from more primitive progenitors (cord blood, fetal liver, or embryonic stem cells) have been shown to reflect more primitive hematopoiesis⁶⁴ than MKs derived from adult CD34⁺ cells.⁴ MKs derived from iPSCs provide the potential for chronic autologous platelet production because iPSCs can be reprogrammed from each individual needing repeated platelet transfusions.

In summary, our study showed that iPSC-derived MKs expressed genes and proteins known to be important in MK and platelet function and demonstrated very good agreement with previous studies of gene expression in human MKs derived from CD34⁺ progenitor cells. We provide a unique comprehensive dataset of genes and proteins expressed in iPSC-derived MKs, which can be sorted and downloaded by investigators. We found many genes expressed highly in MKs but not known to play a role in MK or platelet function, and these might be excellent candidates for further study to determine their effect on hematopoiesis, platelet formation, or platelet function. This may provide a fruitful approach for identifying new biological pathways important for MK formation or function.

ACKNOWLEDGMENTS

This work was supported by NHLBI projects U01 HL72518 and U01 HL107446 and 1R01HL141944-01. KK and MAT were supported from R01 GM105705. KK was also supported by a Center Core Grant from the National Cancer Institute (P30-CA006973). JVE was supported by the Erika J. Glazer Endowed Chair in Women's Heart Health from the Barbra Streisand Women's Heart Center, and by the American Heart Association award number 15GPGSC24470098. Illumina sequencing was conducted at the Genetic Resources Core Facility, Johns Hopkins Institute of Genetic Medicine, Baltimore, MD. Proteomics was carried out by the Cedars-Sinai Proteomics and Metabolomics Core, Los Angeles, CA. Genotyping was performed through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under U.S. Federal Government contract number HHSN268201100037C from the NHLBI.

CONFLICTS OF INTEREST

None of the authors report any conflicts of interest.

AUTHOR CONTRIBUTIONS

L Becker, R Mathias, N Faraday, D Becker, L Cheng, Z Wang, and J Van Eyk designed and conceived of the research questions and study design. K Kammers, M Taub, K Kanchan, J Leek, J Martin, K Raedschelders, V Venkartraman, N Sundararaman, S Parker, and J Van Eyk designed the analysis protocols and performed the data analysis, interpretation, and summarization. L Cheng, Z Wang, S Liu, D Hoyle, S Parker, V Dardov, R Holewinski, and K Raedschelders designed and performed the laboratory experiments. All authors were involved in the writing of the manuscript.

REFERENCES

- Cecchetti L, Tolley ND, Michetti N, Bury L, Weyrich AS, Greslele P. Megakaryocytes differentially sort mRNAs for matrix metalloproteinases and their inhibitors into platelets: a mechanism for regulating synthetic events. *Blood*. 2011;118:1903-1911.
- Moreau T, Evans AL, Vasquez L, et al. Large-scale production of megakaryocytes from human pluripotent stem cells by chemically defined forward programming. *Nat Commun*. 2016;7:11208.
- Watkins NA, Gusnanto A, de Bono B, et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*. 2009;113:e1-9.
- Bluteau O, Langlois T, Rivera-Munoz P, et al. Developmental changes in human megakaryopoiesis. *J Thromb Haemost*. 2013;11:1730-1741.
- Macaulay IC, Tijssen MR, Thijssen-Timmer DC, et al. Comparative gene expression profiling of in vitro differentiated megakaryocytes and erythroblasts identifies novel activatory and inhibitory platelet membrane proteins. *Blood*. 2007;109:3260-3269.
- Sun S, Wang W, Latchman Y, Gao D, Aronow B, Reems JA. Expression of plasma membrane receptor genes during megakaryocyte development. *Physiol Genomics*. 2013;45:217-227.
- Kim JA, Jung YJ, Seoh JY, Woo SY, Seo JS, Kim HL. Gene expression profile of megakaryocytes from human cord blood CD34(+) cells ex vivo expanded by thrombopoietin. *Stem Cells*. 2002;20:402-416.
- Liu Y, Wang Y, Gao Y, et al. Efficient generation of megakaryocytes from human induced pluripotent stem cells using food and drug administration-approved pharmacological reagents. *Stem Cells Transl Med*. 2015;4:309-319.
- Kammers K, Taub MA, Ruczinski I, et al. Integrity of induced pluripotent stem Cell (iPSC) derived megakaryocytes as assessed by genetic and transcriptomic analysis. *PLoS One*. 2017;12:e0167794.
- Faraday N, Yanek LR, Mathias R, et al. Heritability of platelet responsiveness to aspirin in activation pathways directly and indirectly related to cyclooxygenase-1. *Circulation*. 2007;115:2490-2496.
- Johnson AD, Yanek LR, Chen M-H, et al. Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. *Nat Genet*. 2010;42:608-613.
- Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT. *StringTie and Ballgown*. *Nat Protoc*. 2016;11:1650-1667.
- Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12:357-360.
- Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33:290-295.
- R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017. <https://www.R-project.org/>
- Frazee AC, Pertea G, Jaffe AE, et al. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol*. 2015;33:243-246.
- Leek JT, Storey JD. A general framework for multiple testing dependence. *Proc Natl Acad Sci USA*. 2008;105:18718-18723.
- Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. 2007;3:1724-1735.
- Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28:882-883.
- Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. 2003;100:9440-9445.
- Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25:25-29. <https://doi.org/10.1038/75556>.
- Gene Ontology Consortium. Going forward. *Nucleic Acids Res*. 2015;43:D1049-1056.
- Adrian A, Rahnenführer J. topGO: Enrichment analysis for Gene Ontology. R package version 2.28.0, 2016.
- Sundaraman N, Go J, Robinson AE, Mato JM, Lu SC, Van Eyk JE, Venkatraman V. PINE: an automation tool to extract & visualize protein-centric functional networks. *J Am Soc Mass Spectrom*. 2020;31:1410-1421.
- Parker SJ, Chen L, Spivia W, et al. Identification of putative early atherosclerosis biomarkers by unsupervised deconvolution of heterogeneous vascular proteomes. *J Proteome Res* 2020;19. <https://doi.org/10.1021/acs.jproteome.0c00118>
- Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008;24:2534-2536.
- Tsou C-C, Avtonomov D, Larsen B, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods*. 2015;12:258-264.
- Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007;4:207-214.
- Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 2004;20:1466-1467.
- Eng JK, Jahan TA, Hoopmann MR, Comet: an open-source MS/MS sequence database search tool. *Proteomics*. 2013;13:22-24.
- Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol*. 2005;1:2005 0017.
- Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002;74:5383-5392.
- Shteynberg D, Deutsch EW, Lam H, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics*. 2011;10(12):M111.007690.
- Collins BC, Gillet LC, Rosenberger G, et al. Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat Methods*. 2013;10:1246-1253.
- Lam H, Deutsch EW, Eddes JS, et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*. 2007;7:655-667.
- Escher C, Reiter L, MacLean B, et al. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics*. 2012;12:1111-1121.
- Parker SJ, Rost H, Rosenberger G, et al. Identification of a set of conserved eukaryotic internal retention time standards for data-independent acquisition mass spectrometry. *Mol Cell Proteomics*. 2015;14:2800-2813.
- Röst HL, Rosenberger G, Navarro P, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol*. 2014;32:219-223.

39. Weisser H, Nahnsen S, Grossmann J, et al. An automated pipeline for high-throughput label-free quantitative proteomics. *J Proteome Res.* 2013;12:1628-1644.
40. Mallick P, Schirle M, Chen SS, et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol.* 2007;25:125-131.
41. Teo G, Kim S, Tsou C-C, et al. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J Proteomics.* 2015;129:108-120.
42. Sato N, Kiyokawa N, Takada K, et al. Characterization of monoclonal antibodies against mouse and rat platelet glycoprotein V (CD42d). *Hybridoma.* 2000;19:455-461.
43. Rowley JW, Oler AJ, Tolley ND, et al. Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. *Blood.* 2011;118:e101-111.
44. Schattner M, Rabinovich GA. Galectins: new agonists of platelet activation. *Biol Chem.* 2013;394:857-863.
45. Borst S, Sim X, Poncz M, French DL, Gadue P. Induced pluripotent stem cell-derived megakaryocytes and platelets for disease modeling and future clinical applications. *Arterioscler Thromb Vasc Biol.* 2017;37:2007-2013.
46. Sugimoto N, Eto K. Platelet production from induced pluripotent stem cells. *J Thromb Haemost.* 2017;15:1717-1727.
47. Takayama N, Eto K. Pluripotent stem cells reveal the developmental biology of human megakaryocytes and provide a source of platelets for clinical application. *Cell Mol Life Sci.* 2012;69:3419-3428.
48. Faraday N, Goldschmidt-Clermont PJ, Bray PF. Gender differences in platelet GPIIb-IIIa activation. *Thromb Haemost.* 1997;77:748-754.
49. Becker DM, Segal J, Vaidya D, et al. Sex differences in platelet reactivity and response to low-dose aspirin therapy. *JAMA.* 2006;295:1420-1427.
50. Otahbachi M, Simoni J, Simoni G, et al. Gender differences in platelet aggregation in healthy individuals. *J Thromb Thrombolysis.* 2010;30:184-191.
51. Zwierzina WD, Kunz F, Kogelnig R, Herold M. Sex-related differences in platelet aggregation in native whole blood. *Thromb Res.* 1987;48:161-171.
52. Khetawat G, Faraday N, Nealen ML, et al. Human megakaryocytes and platelets contain the estrogen receptor beta and androgen receptor (AR): testosterone regulates AR expression. *Blood.* 2000;95:2289-2296.
53. Lister R, Pelizzola M, Kida YS, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature.* 2011;471:68-73.
54. Kim K, Doi A, Wen B, et al. Epigenetic memory in induced pluripotent stem cells. *Nature.* 2010;467:285-290.
55. Ronen D, Benvenisty N. Sex-dependent gene expression in human pluripotent stem cells. *Cell Rep.* 2014;8:923-932.
56. Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 2008;320:1344-1349.
57. Griffith M, Griffith OL, Mwenifumbo J, et al. Alternative expression analysis by RNA sequencing. *Nat Methods.* 2010;7:843-847.
58. Brooks MJ, Rajasimha HK, Roger JE, Swaroop A. Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl(-/-) retinal transcriptomes. *Mol Vis.* 2011;17:3034-3054.
59. Wu AR, Neff NF, Kalisky T, et al. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods.* 2014;11:41-46.
60. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. *Cell.* 2016;165:535-550.
61. Vogel C, de Sousa Abreu R, Ko D, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol.* 2010;6:400.
62. Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ.* 2014;2:e270.
63. Maier T, Guell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 2009;583:3966-3973.
64. Galloway JL, Zon LI. Ontogeny of hematopoiesis: examining the emergence of hematopoietic cells in the vertebrate embryo. *Curr Top Dev Biol.* 2003;53:139-158.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kammers K, Taub MA, Mathias RA, et al. Gene and protein expression in human megakaryocytes derived from induced pluripotent stem cells. *J Thromb Haemost.* 2021;19:1783-1799. <https://doi.org/10.1111/jth.15334>