Published in final edited form as:

Circ Res. 2020 February 14; 126(4): 501–516. doi:10.1161/CIRCRESAHA.119.315215.

Longitudinal RNA-seq Analysis of the Repeatability of Gene Expression and Splicing in Human Platelets Identifies A Platelet *SELP* Splice QTL

Matthew T. Rondina[*,1,2,3], **Deepak Voora**[*,6], **Lukas M. Simon**[7], **Hansjörg Schwertz**[1,2,4], **Julie F. Harper**[1], **Olivia Lee**[1], **Seema C. Bhatlekar**[1], **Qing Li**[5], **Alicia S. Eustes**[1], **Emilie Montenont**[1], **Robert A. Campbell**[1,2], **Neal D. Tolley**[1], **Yasuhiro Kosaka**[1], **Andrew S. Weyrich**[1,2], **Paul F. Bray**[1,2], **Jesse W. Rowley**[1,2]

[1]Molecular Medicine Program

[2]Department of Internal Medicine

[3]George E. Wahlen VAMC Geriatric Research and Education Clinical Center, The University of Utah

[4]Rocky Mountain Center for Occupational and Environmental Health

[5]Huntsman Cancer Institute, Salt Lake City, Utah

[6]Duke Center for Applied Genomics & Precision Medicine, Durham, NC

[7]Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany.

Abstract

Rationale: Longitudinal studies are required to distinguish within versus between-individual variation, and repeatability of gene expression. They are uniquely positioned to decipher genetic signal from environmental noise, with potential application to gene variant and expression studies. However, longitudinal analyses of gene expression in healthy individuals—especially with regards to alternative splicing—are lacking for most primary cell types, including platelets.

Objective: To assess repeatability of gene expression and splicing in platelets and use repeatability to identify novel platelet eQTLs and sQTLs.

Methods and Results: We sequenced the transcriptome of platelets isolated repeatedly up to 4 years from healthy individuals. We examined within and between-individual variation and repeatability of platelet RNA-expression and exon skipping, a readily measured alternative splicing event. We find that platelet gene expression is generally stable between and within individuals over time—with the exception of a subset of genes enriched for the inflammation gene ontology. We show an enrichment among repeatable genes for associations with heritable traits,

Address correspondence to: Dr. Jesse W. Rowley, Eccles Institute of Human Genetics, University of Utah Health Sciences Center, 15 North 2030 East, Room 4220, Salt Lake City, UT 84112, jesse.rowley@u2m2.utah.edu.

*These authors contributed equally to this work

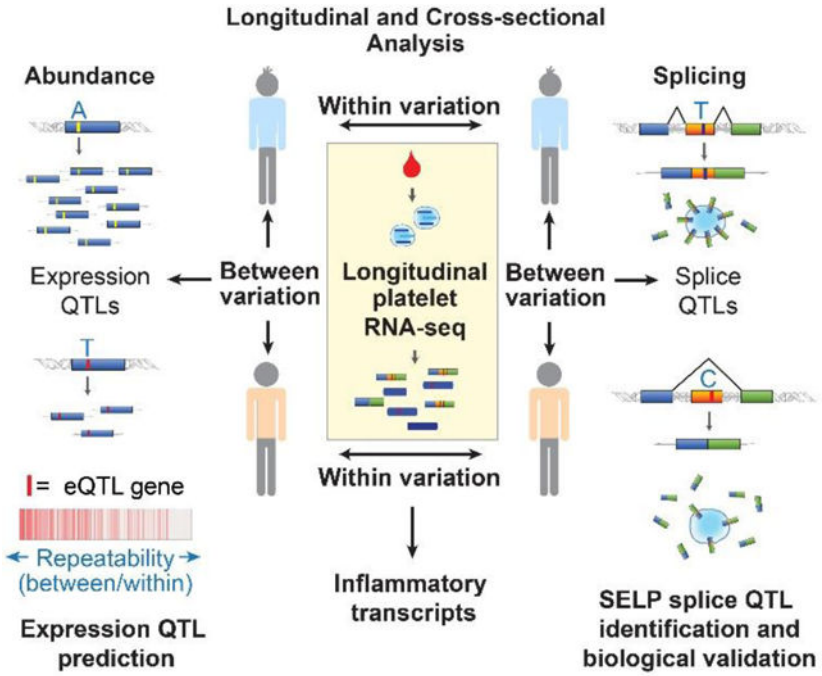
DISCLOSURES

The authors declare no conflicts or competing financial interests.

including known and novel platelet eQTLs. Several exon skipping events were also highly repeatable, suggesting heritable patterns of splicing in platelets. One of the most repeatable was exon 14 skipping of *SELP*. Accordingly, we identify rs6128 as a platelet sQTL and define an rs6128-dependent association between *SELP* exon 14 skipping and race. *In vitro* experiments demonstrate that this single nucleotide variant directly affects exon 14 skipping and changes the ratio of transmembrane versus soluble P-selectin protein production.

Conclusions: We conclude that the platelet transcriptome is generally stable over 4-years. We demonstrate the use of repeatability of gene expression and splicing to identify novel platelet eQTLs and sQTLs. rs6128 is a platelet sQTL that alters *SELP* exon 14 skipping and soluble versus transmembrane P-selectin protein production.

Graphical abstract



Keywords

Platelets; RNA-seq; transcriptome; longitudinal study genetics; longitudinal cohort study; alternative splicing repeatability; splicing; eQTL; sQTL

INTRODUCTION

Platelets are abundant, accessible blood cells that are increasingly used in gene expression studies. Like nucleated cells, platelets possess a diverse portfolio of RNAs, including coding mRNAs, small non-coding RNAs, lncRNAs, and others[1–3]. Yet their anucleate nature offers unique advantages over nucleated cells for studying gene expression. For one, ex vivo handling (cell isolation method, processing time, buffers, etc.) of nucleated cells can immediately affect the expression of thousands of transcripts[4–6]. On the other hand, platelets are transcriptionally unaffected by isolation[7,8], *allowing capture of the native in vivo gene*

expression signature. These attractive features render platelets an excellent choice for RNA diagnostics and gene expression studies.

Platelets have been used in GWAS, gene-phenotype[9,10], diagnostic[11], and differential expression studies with an emphasis on RNA abundance to elucidate mediators of platelet reactivity in health and disease[12]. Genetic modifiers of RNA abundance in platelets, called expression quantitative trait loci (eQTL), have also been described[13,14]. eQTLs are DNA sequence variants associated with gene expression that affect nearby (cis-) or remote (trans-) genes in a cell type specific manner. eQTLs are particularly important in genetic studies because they provide an intermediate and mechanistic link between a phenotype and gene association.

Beyond RNA abundance, it is now known that platelets and megakaryocytes harbor alternative structural features of RNA, including alternative start and stop sites, and alternative splicing[12], that diversify the transcriptome and proteome[15,16] and alter cellular function. In platelets, activation induces RNA splicing, and thereby modulates functional protein expression[15,17]. It is probable that genetic variants called splice QTLs (sQTLs) also influence basal and activation dependent RNA splicing levels in platelets. However, sQTLs for platelets have not yet been described.

Other major knowledge gaps exist regarding RNA abundance and structure in platelets. Nearly all platelet studies have been cross-sectional, examining gene expression at a single time point. Yet, gene expression can vary both between-individuals and within-individuals over time. Hormonal changes, circadian rhythm, inflammation, diet, and aging are examples of environmental cues that might alter gene expression within healthy individuals[18–20]. Such normal changes in gene expression can mask the ability to detect signal in differential gene expression, diagnostic, and genetic studies, and confound their analysis. Thus, understanding within individual versus between individual variation in gene expression is important for the design and interpretation of gene expression studies, and can be used to prioritize candidates in genetic studies.

With regards to genetic studies, several reports have suggested using repeatability to identify eQTL genes[21–23]. *In vivo* repeatability can only be calculated from multiple samplings from the same individual, and refers to the proportion of variation attributed to between-individual versus within-individual variation[24]. In the straightforward view, repeatability sets an upper-bound to broad sense heritability[25]: if between individual differences are not repeatable because of low between, and/or high within-individual variation, there will be insufficient power to detect heritable, genetic signal. For this reason, it has been recommended to measure the repeatability of a trait before performing GWAS[21]. Carlborg et. al.[22] found that censoring mouse eQTL data on repeatability was an effective method for prioritizing transcripts with a high a priori likelihood of successful eQTL identification. Hoffman et. al.[23] also demonstrated the potential use of within-individual *technical* variation to narrow candidates and facilitate eQTL prediction, although this study employed only single time point replicates. Together, these studies imply that longitudinal analysis of gene expression may facilitate prospective eQTL (and sQTL) gene discovery. Surprisingly, longitudinal analyses of gene expression are scarce in primary cells from healthy individuals, and absent

for platelets. To our knowledge, the repeatability of alternative splicing over time has not been established *for any primary cell type.*

In this study, we use longitudinal RNA-seq analysis to examine within (intra-) and between (inter-) individual variability of the human platelet transcriptome. We examine repeatability of gene abundance and exon skipping, a readily measured alternative splicing event. We retrospectively demonstrate the use of repeatability to decipher heritable signal from environmental noise, and identify eQTL genes. We prospectively use repeatability to prioritize eQTL and sQTL gene candidates for novel platelet eQTL and sQTL discovery.

METHODS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Human subjects and Platelet isolation.

Subjects were independently recruited from the University of Utah and Duke University (Table 1).

All subjects provided informed consent and this study was IRB approved by each institution. Subjects were healthy and without active medical conditions. No subjects had undergone surgery for the past 4 months. Any illness required resolution of symptoms for at least 7 days prior to sampling. Cohort 2 subjects were prospectively recruited. Cohort 1 subjects were part of a clinical study where they were previously exposed to aspirin, however no subjects were on aspirin for at least 4 weeks prior to each blood sampling. Blood was drawn by venipuncture into citrate tubes (cohort 1) or acid-citrate-dextrose (cohort 2) and sample processing initiated within 30 minutes of phlebotomy. Platelets were isolated via magnetic leukocyte depletion using CD45 microbeads (Miltenyi) as we have previously described[1,2,26].

RNA isolation and sequencing.

RNA was isolated using phenol-chloroform extraction (cohort 1) as previously described[1] or DirectZol kit (cohort 2, Zymogen). More details on RNA isolation are found in the Online Supplemental Methods. Sequencing libraries for cohorts 1 and 2 were barcoded and prepared using kits: KAPA Stranded mRNA-Seq Kit (Roche #KK8421) and TruSeq unstranded v2 with poly(A) selection (Illumina #RS-122) respectively. Libraries were sequenced 50 cycles, single end, on Illumina HiSeq 4000 (cohort 1) or Illumina HiSeq 2000 (cohort 2) to a depth of ~20–40 million mapped reads per sample. Fastq files have been deposited in the NIH Sequence Read Archive PRJNA531691.

RNA-seq analysis.

For analysis of expression variation, reads were aligned to GRCh38/hg38 using Novoalign (Novocraft) as we have previously described[1]. Reads were assigned to flattened Ensembl gene annotations using the USeq analysis package[27]. Read counts were normalized separately for each cohort using the DESeq2 analysis package[28]. Non-coding RNAs were

selected according to Ensembl transcript biotype. Heatmaps, clustering (complete linkage), density plots, boxplots, and scatterplots, were generated in R[29]. Read distribution plots were generated using integrative genomics viewer (IGV)[30]. Gene ontology enrichment was analyzed using DAVID[31].

Analysis of individual transcript variation.

RNA-seq data has a strong mean-variance relationship. The DESeq2 regularized log transformation (RLD)[28] was applied to counts which preferentially shrinks the overall variance among low abundant transcripts (while retaining outliers), thus allowing a more straightforward comparison of transcript variation across all expression levels. Lowest abundant transcripts were also arbitrarily removed. Within-individual variation was calculated as the standard deviation of all samples from the same individual. Total variation was calculated as the standard deviation of samples across all individuals in each cohort, thus within-individual variation is a sub-component of total individual variation. Sources of variance were quantified with a linear mixed model using the R package variancePartition[23]. Repeatability was calculated with the formula $\sigma_b^2/(\sigma_w^2 + \sigma_b^2)$ in the R package 'heritability'[32], including correction for sex in repeatability calculations for eQTL enrichment analysis and gene prioritization for eQTL and sQTL discovery.

Exon skipping analysis.

Exon skipping events were identified from triads of Exon/Exon and Exon/Intron junctions with > 5 reads per junction and the calculated percent spliced in (PSI; see Figure 5D) > .05 and < .95 in > 30% of samples. Junctions were excluded where the flanking Exon/Intron junction pairs varied by more than 10 fold in > 70% of samples. By this strategy we identified 245 exons (from 194 different transcripts) that were skipped in a significant fraction of the transcripts in some but not all samples.

SELP mini-gene.

The complete open reading frame for *SELP* transcript ENST00000263686.10 with a c-terminal DYK tag, and introns 13 and 14 that flank exon 14, were cloned into PCDNA-CMV and pCDH-MSCV-GFP vectors. A single nucleotide change C->T was made at rs6128. 293T cells (HEK 293T/17, ATCC CRL-11268) were maintained according to ATCC recommendations and used between passages 10–20. 293T cells were transfected with lipofectamine 2000. 24 hours after transfection, *SELP* RNA splicing was analyzed by PCR at cycles 25 and 30 using primers flanking exon 14 (5'-gtcaactaccgtgccaacct; 5'-taaggactcgggtcaaatgc). For flow cytometry experiments, cells were either co-transfected with a GFP plasmid (PCDNA-CMV) or GFP was contained within the same backbone as *SELP* (PCDH-MSCV). Surface expression of P-selectin was assessed by staining with Psel.KO2.3 APC antibody (ThermoFisher #17–0626-82) and analyzed by flow cytometry on a CytoFLEX analyzer (Beckman Coulter). MFI of P-selectin was normalized to GFP expression as assessed on live/transfected cells gated according to forward/side scatter (live) and FL-1 (GFP+) intensity. Soluble P-selectin was measured using a Quantikine ELISA kit (R&D Systems #DPSE00). For western blots, cells were lysed with RIPA, lysates denatured and reduced, proteins separated using an 10% SDS-PAGE, transferred onto a PVDF membrane, and blotted for tagged P-selectin with anti-DYK antibody (Cell Signaling Tech.

#2368S), followed by anti-rabbit HRP secondary antibody (Rockland #18–8816-33) and chemiluminescent detection (ThermoFisher #34580).

Novel platelet eQTL and sQTL analysis.

RNA-seq fastq files for 234 previously published samples (Best et. al.[8,33], Netherlands cohort; hereafter referred to as NL cohort) were retrieved from NCBI short read archive PRJNA353588[33]. These, and fastq files for cohort 1, were aligned with STAR[34] to human reference genome (build HG38) in a splice-aware manner, and variants were called and filtered using the workflow built from the Genome Analysis Toolkit (GATK)[35] best practices for variant calling on RNA-seq. Additional details on variant calls, filters used, and a note on the caveats and limitations of RNA-seq based variant analysis are included in the Online Supplemental Methods. Variants tested for eQTLs were limited to within 2 kb of all genes not identified as eQTLs by the previously published PRAX1[13] dataset, and with repeatability > 0.9 (238 genes) in the cohort 1 dataset. For comparison, the equivalent number of genes with lowest repeatability were also included. Combined filtering resulted in 641 variants across 181 genes that were tested for gene abundance–variant association. Gene-variant association was tested in the R package SNPassoc[36] using an additive model of variant-allele dosage (0,1,2), while controlling for the covariates sex, age, and population structure[37,38] (see Online Figure X and Online Supplemental Methods for details). Benjamini and Hochberg FDR correction for multiple testing (641 gene-variant tests) is reported. However, a conservative significance threshold of $p < 1e\text{-}6$ was used to filter novel eQTLs as if genome wide analysis had been performed[13]. Allelic imbalance of each significant eQTL was evaluated with the Wilcoxon rank sum test on the proportion of reference variant reads to total reads in each heterozygote individual.

A generalized linear model was used for logistic regression analysis in R to test the association of *SELP* exon 14 splicing and self-reported race or rs6128 variant-allele dosage. For this, *SELP* exon 14 inclusion counts versus total inclusion + exclusion counts were used as the binomial response variable. Where specified, models controlled for potential covariates including race or population structure, rs6128 genotype, sex, and age.

Statistics.

Significance for multimodality was calculated according to Hartigan's dip test statistic. Wilcoxon test with adjustment for multiple comparisons was used to test for similarity in distributions of between and within sample correlations. Kolmogorov–Smirnov test was used to test for enrichment in the rank of genes associated with genetic/heritable traits. Gene Set Enrichment Analysis (GSEA)[39] pre-ranked was used to evaluate enrichment among genes tested for the presence of significant eQTLs as identified in the PRAX1 cohort. For this, only the subset of genes tested by PRAX1 were included. The association between eQTL presence and different repeatability thresholds was estimated with odds ratio (odds at each threshold compared to no threshold) and significance evaluated with Chi-square test of independence. Two sided T-tests and correlation tests for significance (alpha = 0.05) were performed in R[29] using functions cor.test and t.test, which sets a lower p value limit of 2.2e-16.

RESULTS

To assess within and between individual variation of the platelet transcriptome, we performed RNA-seq analysis of leuko-depleted platelets isolated longitudinally from two independent cohorts of healthy individuals. For cohort 1, we analyzed platelets from 31 individuals at an initial visit (T0) and 4 months later. For cohort 2 we analyzed platelets from 7 individuals at T0 and then longitudinally over 4 years. The characteristics of the two cohorts are detailed in Table 1.

Platelets contain a stable within-individual gene expression signature.

Unsupervised clustering analysis of all pair-wise distances within and between individual transcriptomes in cohort 1 indicated a robust within-individual (self) RNA expression signature (Figure 1A), with most self-pairs clustering as nearest neighbor pairs. As depicted in Figures 1B-C, the mean within-individual correlation of platelet transcriptomes isolated 4 months apart was very high (r mean±sd=0.987±0.012). Between-individual correlations of the global platelet transcriptome were also high (0.947±0.024), but significantly lower ($p < 2.2e-16$) than within-individual correlations. Grouping samples by race, age, and sex only partly corrected the difference (Figure 1C). Within individual clustering of non-coding RNAs was also robust (Online Figure I A), with a mean *within*-individual correlation of 0.984±0.013. The mean *between* individual correlation for non-protein coding transcripts (0.905±0.031, Online Figure I B-C) was significantly weaker compared to protein coding transcripts ($p < 2.2e-16$), which is consistent with previous cross-sectional studies[40]. Raw and normalized counts for each transcript in cohort 1 are found in Online Datasets I and II.

Unsupervised clustering of total RNA transcriptomes in cohort 2 resulted in robust self-clustering and suggested minimal transcriptional drift over 4 years (Figure 1D). Within-individual correlations for samples isolated 4 years apart remained comparable to within-individual correlations for samples isolated only 2 weeks apart, and significantly higher than between-individual correlations regardless of time-point (Figure 1E-F). As shown in Online Figure I D, the within-self non-coding RNA signature was also robust, and uniquely identified individuals at all time points over the 4 years without exception—a reflection of the significantly higher within-self correlations in non-coding RNA expression compared to those between-individuals (Online Figure I E-F). Raw and normalized counts for each transcript in cohort 2 are found in Online Datasets III and IV.

Within and between individual transcript variation in platelets is reproducible across cohorts.

Together, the data in Figure 1 indicate similar patterns of gene expression variation between cohorts 1 and 2: the average within-individual correlations were similar (0.983±0.13 vs 0.987±0.12) for each cohort, as were the average between-individual correlations (0.958±0.021 vs 0.947±0.024). We further defined the specific transcripts in each cohort that displayed the least and most overall (total) variation compared to within individual variation (see Online Datasets V-VI). As depicted in Figure 2A and Online Figure II A-B, high within individual variation was limited to a small number of moderately expressed transcripts that were consistently variable in both cohorts 1 and 2. Transcripts that varied the most within

individuals in both cohorts were enriched for those within the inflammatory and defense response gene ontologies (GO; Figure 2B)[18,41]. As shown in Figure 2C and Online Figure II C-D, transcripts with high total variation spanned a broader range of expression levels, yet the extent of variation was still consistent between cohorts 1 and 2. The transcripts with the highest total variation predominantly overlapped those that varied the most within-individuals (Online Figure II E-F) leading to an enrichment in the inflammatory and defense response GO (Figure 2D). In addition to inflammatory transcripts, we noted several genes with high total variation that were previously associated with sex, race, or platelet eQTLs[10,13,42]. However, unlike the inflammatory transcripts, most of these did not demonstrate high within individual variation (see bottom right quadrants of Online Figure II E-F).

Variance partition analysis[23] was used to further partition and quantify for each gene the amount of variation attributable to sex, race, and other covariates, and to decouple between from within and total variation. As shown in Online Figure III, for more than half of all genes, between individual variation was responsible for the majority (>50%) of gene expression variation, followed by residual within individual variation. On the other hand, sex and race affected only a small number of genes. Other known covariates including age and sample processing contributed to only a minor portion of variation for each gene.

Repeatability defines heritable platelet gene expression and predicts eQTL genes.

As discussed in the introduction, we hypothesized that repeatability, which captures the within and between individual variation in a single index (see methods), could be used to prioritize genes for eQTL analysis. Therefore, we calculated repeatability for each gene (Online Datasets V-VI), and retrospectively tested whether repeatability is associated with genetic and heritable regulation of gene expression in platelets. To do this we utilized the published PRAX1[13] dataset as an independent (no overlap with the current cohorts) and cross-platform (microarray) validation dataset. PRAX1 previously associated platelet transcripts with sex and race, and identified 612 platelet eQTL genes. We used cohort 1 for the analysis because it was larger in size than cohort 2 and better matched the diversity and demographics of PRAX1. We ranked all genes in the PRAX1 microarray according to the repeatability calculated by RNA-seq in cohort 1. Ranking by repeatability resulted in a significant enrichment for genes associated with sex, race, or eQTLs ($p < 2e\text{-}16$), that was significantly greater than ranking genes by abundance, within-individual, or total variation ($p < 2e\text{-}6$; also see Online Figure IV). As shown in Figure 3A, 100% of the top 15 repeatability ranked genes are significantly associated in expression with sex, race, or a cis-eQTL. As an example, *MFN2* is an established platelet eQTL gene[13] that was among the most repeatable genes (repeatability = 0.98). This is because *MFN2* expression varies according to eQTL genotype[13] by more than 8 fold between individuals, while remaining relatively constant within individuals over time (Figure 3B).

When assessing specifically the enrichment for eQTL genes, GSEA indicated a significant enrichment ($p = 0$) of known eQTLs as repeatability increased, with those with repeatability >0.68 (leading edge) accounting for the enrichment (Figure 3C). Significant enrichment was also observed when ranking by mean expression abundance or total variation, although the

enrichment score for these measures was lower than for repeatability. According to binned analysis of odds ratios, the odds of identifying an eQTL for a gene with a repeatability <0.5 is 3.2 fold lower (p = 5e-15) than testing a gene at random (Figure 3D). The odds of identifying an eQTL for a gene with a repeatability > 0.9 is 6.2 fold higher (p = 6e-45) than random and 1.8 fold higher (p=0.006, adjusted) than ranking according to total variation. Furthermore, for known platelet cis-eQTLs, there was a significant correlation between the eQTL FDR and the repeatability of its associated transcript (Online Figure V). Thus, repeatability indicates an enrichment for and strength of cis-eQTL signal in platelets and may be a useful filtering and prioritization strategy to identify genes with an eQTL signal.

Microarrays differ from RNA-seq in accuracy, sensitivity, and comprehensiveness, and some eQTL genes identifiable by RNA-seq may have been missed by PRAX1. Therefore, we used RNA-seq data to re-interrogate 238 genes with high repeatability (>0.9), yet with no cis-eQTLs previously found. We tested these for cis-eQTLs using a publicly available dataset[13,42] collected in the Netherlands, of platelet RNA-seq from 234 healthy individuals (NL cohort). Genetic variants were called from RNA-seq reads across each gene, and tested for association with RNA-seq abundance. Despite the known limitations of RNA-seq in calling genetic variants (e.g. most variants are located in promoters and introns), 11 new probable platelet eQTL genes (Online Dataset VII) were identified. In contrast, when analyzing the same number of genes with the lowest repeatability, we did not identify any additional eQTLs—a difference which was statistically significant (11/238 vs 0/238, p=0.0009, Fisher Exact). Allele Specific Expression (ASE) analysis of allelic imbalance, which measures the ratio of read counts coming from each allele within heterozygotes, confirmed a significant and directionally consistent within-sample eQTL effect on allelic imbalance for 8 of the 11 genes (Online Dataset VII). An example of one of the novel platelet eQTL genes is long non-coding RNA *LINC01089*—one of the most repeatable (0.95) and abundant (top 10% by RNA-seq) transcripts in platelets. As shown in Figure 3E, there is a strong additive allele dosage effect of rs1168663 on *LINC01089* expression among cohort 1 individuals at both time points, and among individuals in the NL cohort. As shown in Figure 3F, *LINC01089* expression demonstrates significant allelic imbalance in cohort 1 and in the NL cohort. Together these data define several novel platelet cis-eQTLs, demonstrating the utility of repeatability from longitudinal expression data to predict heritable gene expression variation and prioritize targets for prospective identification of cis-eQTL genes.

Alternative exon skipping in platelets is maintained within-individuals over time.

To assess within-versus between-individual stability of alternative splicing, we focused on an alternative splicing event that is readily measured in RNA-seq data: exon skipping. We stringently identified exon skipping events in cohort 1 (see methods), and calculated Percent of exon Spliced In (PSI; Figure 4A) for each (Online Datasets VIII-XI). As shown in Figure 4B, there was a broad range of exon skipping levels among the different exon skipping events that was mostly consistent between individuals and within-individuals over time. Unsupervised clustering analysis using PSI resulted in a preference for within-individual clustering compared to between-individual clustering (Online figure VI), although this was not as robust as clustering based on expression. Nonetheless, the within-individual

correlation of PSI was significantly higher than between-individual correlation, independent of age, race, or sex (Figure 4B-C), suggesting a heritable component of exon skipping levels in platelets.

Identification of a platelet splice QTL associated with race that affects exon 14 skipping in SELP.

Unlike eQTLs, platelet sQTLs have not previously been identified. We therefore used repeatability to prioritize exon skipping events, with the goal of identifying novel, robust, and physiologically relevant platelet cis-sQTLs. To this end, we assessed the within/between individual variation of PSI for each exon skipping event, and ranked each by repeatability (Online Dataset VIII for full table). As shown in Figure 5A, Exon 14 of *SELP*, which codes the leukocyte adhesion and platelet activation marker P-selectin, ranked second among repeatable exon skipping events. The difference in exon 14 exon skipping between donors, but stability within individuals is illustrated by the alignment plots in Figure 5B and the correlation plot in Figure 5C.

Exon 14 skipping predicts an in-frame deletion of the transmembrane domain of P-selectin. An exon 14 deficient isoform of P-selectin was previously detected in endothelial cells and platelets[43–45] and at significant levels in the human circulation[44–46]. PCR (Online Figure VII), cloning, and Sanger sequencing (data not shown) verified that the RNA isoform predicted by RNA-seq in our own cohorts matches the previously described soluble protein isoform in plasma. Together, this implicates exon 14 skipping as a heritable source of variability in P-selectin protein cell surface and soluble plasma levels between individuals.

Previous studies have associated soluble P-selectin in the plasma with a variety of clinical and genetic factors including race and single nucleotide polymorphisms (SNPs)[47–50]. As shown in Figure 5D, we observed a significant increase of *SELP* exon 14 inclusion among blacks/African Americans compared to whites at both time points. A search for the most likely responsible genetic variants identified a SNP, rs6128, within exon 14 of *SELP* with a homozygous MAF (T/T) that is much higher among Africans (0.29) compared to Europeans (0.04)[51]. Intriguingly, rs6128 has been associated with plasma P-selectin[52] levels and diabetic retinopathy, especially among African Americans[50]. However, the direct relationship of rs6128 to soluble P-selectin is unclear since the C to T transition does not change the protein sequence or modify canonical splice sites. Bioinformatic analysis[53] of the sequence surrounding rs6128 predicted a net loss of an exonic splicing silencer (ESS) motif and a net gain of 2 exonic splicing enhancer motifs (ESE) (Figure 6A). To determine whether rs6128 is associated with *SELP* exon 14 splicing in platelets, we inferred rs6128 genotypes from RNA-seq reads in cohort 1 and the NL cohort[33] (variant details are in Online Dataset XII). As shown in Figure 6B-C, the rs6128 SNP is significantly associated in both cohorts with *SELP* exon 14 skipping in platelets. The association of rs6128 with *SELP* exon 14 skipping was independent of age, sex, race, or population structure.

We tested whether the difference in rs6128 MAF between Africans and Europeans might account for the association of exon 14 skipping with race indicated in Figure 5D. Consistent with this, there was no difference in exon 14 skipping levels between blacks/African

Americans and whites after correcting for rs6128 (p = .4 and .3 for T=0 and T=4 months respectively).

Given the importance of P-selectin in disease, we extended the *SELP* exon 14 splicing analysis to additional diseases also available through Best et. al.[8,33]. As shown in Online Figure VIII, none of the diseases examined (non-small cell lung cancer, multiple sclerosis, or pulmonary hypertension) was significantly associated with *SELP* exon 14 splicing, or the effect of rs6128 on splicing. This indicates that the levels of exon 14 skipping are very stable within the individual, even in the context of environmental stressors which have been shown to trigger changes in platelet transcript abundance[11,33].

Rs6128 directly affects SELP exon 14 skipping and the proportion of soluble to surface P-selectin in vitro.

Non-causal markers are commonly falsely identified in genetic association studies because of linkage with other unobserved variables[54]. To specifically test a causative effect of rs6128 on *SELP* exon 14 splicing, we generated mini-gene constructs of the *SELP* ORF with rs6128 C/C or T/T, and included introns flanking exon 14 (Figure 7A). Since it is known that promoter differences can influence splicing[55], two different promoters (CMV or MSCV) were tested for each minigene construct. Constructs were expressed in HEK 293 cells, which lack endogenous P-selectin. Following transfection, RT-PCR analysis confirmed that the single nucleotide change from C/C to T/T resulted in a significant shift in the ratio of *SELP* RNA isoforms (Figure 7B), in the direction consistent with RNA-seq results. This occurred for both promoters, but with a more pronounced shift observed for the CMV promoter. Western blot analysis indicated that the T/T variant resulted in a shift toward exon 14 inclusion in P-selectin protein (Online Figure IX). As shown in Figure 7C, and consistent with inclusion of the transmembrane domain, the single nucleotide change from C/C to T/T significantly increased (2 fold) the amount of surface P-selectin on HEK 293 cells. In contrast, the T/T variant significantly decreased the amount of soluble P-selectin in the supernatant as measured by ELISA (Figure 7D). Together this data demonstrates a causal relationship between rs6128 genotype, the amount of exon 14 inclusion in *SELP* RNA, and the proportion of soluble to surface P-selectin expression.

DISCUSSION

Analysis of within-individual versus between individual variation in two independent cohorts indicated that the human platelet transcriptome is highly stable in healthy individuals for up to 4 years. There are very few longitudinal studies available on primary nucleated cells for comparison. One study with similar design by Radich[56] et. al. observed a 30% within-individual misclassification rate for leukocyte transcriptomes *even after selecting for a gene signature that maximized variation between individuals*. Although differences exist between this published study and our current work, we identified that platelets had a lower within-individual misclassification rate (10–12%) *without signature selection*. We speculate that their anucleate nature and 7–10 day lifespan moderate *in vitro* and *in vivo* RNA changes in platelets, promoting a stable and defined *in vivo* healthy gene expression signature.

Platelet gene expression profiling via RNA-sequencing is emerging as a relevant tool for platelet function studies, for defining the consequences and causes of disease, and for disease diagnostics[8–10,12,14,33,57]. However, almost all published studies to date, have relied on single-time point comparisons of the platelet transcriptome between a disease cohort and healthy subjects. As healthy subjects are often used as the “baseline” or “control” condition in these studies, understanding whether the platelet transcriptome is durable – or not – in health is critical to understanding the robustness of these comparisons. Our finding that the platelet transcriptome is generally stable over 4 years in healthy individuals lends validity to these comparisons. In depth analysis of the individual transcripts that do vary within and between individuals also suggest some limitations and caveats.

Although most transcripts were stable, a few transcripts varied substantially within individuals. Most within-variable transcripts were related to inflammation. These may inform studies evaluating the effects of inflammation on platelet gene expression, and may be of relevance to clinical findings that inflammatory stress is associated with platelet counts and function[58]. The range of variation for inflammatory transcripts in health compared to overt inflammatory disease may be worth investigating when assessing the impact of inflammatory gene changes on disease.

We observed major platelet expression differences between individuals. Sex and race accounted for some major differences. Other sources of individual variation contributed more. Our data suggest a prominent role for cis-eQTLs. Regardless of the source of variation, the propensity of a gene to vary between (or within) healthy individuals might be taken into consideration when interpreting differential disease-gene studies and designing experiments for validation. Genes with high inherent variation require larger sample sizes to reach statistical confidence. When sample sizes are small, false positives are more likely for genes with high inherent variation. Correction for known covariates might be helpful in this regard. While corrections for sex, race, and age are often considered in differential gene expression analyses, eQTLs are normally unavailable or ignored.

Stability information may also guide (along with the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE[59]) guidelines) the selection of reference genes for normalization controls. Within/between stable genes such as SYK, AKT1/2, GP1BA, ACTB might be good choices. On the other hand, highly within-variable genes, such as the gene TUBB1, should generally be avoided as reference genes.

As an application of our longitudinal dataset, we used repeatability as a filtering strategy to identify new platelet eQTLs. Because of the limitation of multiple testing[60], even large scale gene expression association studies employ a filtering and prioritization strategy to circumvent testing thousands of genes (and even more alternative splice events) against millions of genetic loci. Common filtering strategies include hard filtering on abundance[61] or total variance. However, filtering for total variance alone will also enrich for transcripts overly-influenced by technical or environmental noise. To avoid this, several studies[21–23] (see introduction) have suggested using repeatability instead. Here we have used platelet longitudinal data to experimentally test this idea. We observed a significant enrichment for

cis-eQTLs among the most repeatable genes that significantly improved the ability to identify eQTL genes compared to using abundance or total variation.

Although significantly enriched for eQTLs, the association with repeatability was not perfect. Cohort 1 was assayed on a different platform (RNA-seq vs microarray), was smaller (31 vs 154), and had similar, but not identical demographics to PRAX1 (race (Blacks/African Americans): 45% vs 42%, ns; sex (M): 49% vs 32%, ns; age: 42+/−11 vs 29+/−7, $p < .05$). Repeatability measured in the same samples used for eQTL analysis would presumably result in the best predictions. However, large longitudinal studies are often impractical because of the costs and challenges of repeated sampling. Moreover, repeatability can add a layer of confidence to eQTL results if measured in an independent cohort. For this, larger sample sizes that reflect the demographics and environment of the test cohorts will presumably fare better. Cohorts that are too small to capture genetic variation (i.e. low MAF eQTLs), or are subject to systematic environmental perturbations, will suffer from overall lower repeatability, and lack sensitivity to predict eQTLs. Additional studies are needed to determine how repeatability might be applied more generally in the analysis of additional datasets, cell types, and to handle gene-environment interactions.

We used an additional RNA-seq cohort (NL cohort) to test for unreported platelet eQTL genes among the most repeatable genes in cohort 1. Of the identified eQTLs, 22/27 (for 7/11 eQTL genes) have been reported in other tissues (The Genotype-Tissue Expression (GTEx) Project, see Online Dataset VII). Noteworthy among eQTL candidate genes is *TECPR2*, which was previously associated with platelet counts by GWAS[62]. Long non-coding RNA eQTL genes were also identified: *LINC01089*, *MAGI2-AS3*, *KANSL1-AS1*. Like these 3 genes, we found that non-coding RNAs are generally stable within individuals, yet more variable between individuals compared to protein-coding RNAs, suggesting more genetic diversity among non-coding RNAs. Long non-coding RNAs have gained attention for their multi-faceted ability to regulate gene expression, but are understudied in platelets. How diversity in expression of long non-coding RNAs relates to functional gene expression differences between individuals is a subject of interest for future investigation.

Repeatability was further applied to prioritize exon skipping events to find those with the greatest likelihood of identifying a biologically tractable sQTL signal. A robust association between rs6128 and *SELP* exon 14 skipping was identified. During the course of this work, a significant association between rs6128 splicing and exon 14 skipping was also identified in whole blood samples[63], further strengthening the findings.

To establish causality, we performed transfection experiments in HEK 293 cells, which advantageously do not express endogenous *SELP*. The results strongly implicate rs6128 as causal for differential SELP exon 14 splicing in platelets, but do not rule out potential contributing effects of the endogenous promoter, or of additional associated or linked variants. The confirmation of a platelet observation in an unrelated cell line suggests rs6128 may affect *SELP* splicing in multiple tissues such as endothelial cells—another major producer of P-selectin.

Surface P-selectin mediates leukocyte interactions and inflammation, is involved in atherogenesis, and plays a role in tumor metastasis. Soluble P-selectin is a functionally and clinically relevant platelet protein in the circulation associated with a variety of diseases[47,64]. While a major source of soluble P-selectin is related to activation induced shedding, a significant amount is heritable. Associations between rs6128 and soluble P-selectin protein levels in plasma have been reported[50,52]. The observed effects of rs6128 on exon 14 *SELP* splicing and soluble versus surface P-selectin localization establishes a link between these observations. They may explain previous clinical studies that have associated rs6128 with plasma P-selectin and diabetic retinopathy in African Americans[50]. Finally, the finding that *SELP* is differentially spliced according to race may be therapeutically relevant in light of promising clinical trials that effectively used P-selectin blockade to treat pain crisis in Sickle Cell Anemia, a disease that predominantly affects individuals of African descent[47].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank Brad Demarest for insights into statistics. We also thank Diana Lim for preparation of the figures, critical comments, and consultation regarding effective display of the images.

SOURCES OF FUNDING

This work was supported by HL066277 and HL112311 (ASW), AG040631, HL126547, and HL092161 (MTR), HL118049 (DV), HL144957, and GM103806 (JWR). HS was supported by a Post-Doctoral fellowship (0625098Y), a Beginning-Grant-in-Aid (09BG1A2250381) from the American Heart Association Western States Affiliate, and a Lichtenberg-Professorship from the Volkswagen Foundation. This investigation was supported by the University of Utah Population Health Research (PHR) Foundation, with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 5UL1TR001067–05 (formerly 8UL1TR000105 and UL1RR025764). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This material is the result of work supported with resources and the use of facilities at the George E. Wahlen VA Medical Center, Salt Lake City, Utah. The sponsor had no role in the design or preparation of paper. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal (V8) on 10/01/19.

Nonstandard Abbreviations and Acronyms:

QTL	quantitative trait loci
eQTL	expression QTL
sQTL	splice QTL
GWAS	genome wide association study
ASE	allele specific expression
SNP	single nucleotide polymorphism

REFERENCES

1. Rowley JW, Oler AJ, Tolley ND, Hunter BN, Low EN, Nix DA, Yost CC, Zimmerman GA, Weyrich AS. Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. Blood. 2011;118:e101–e111. [PubMed: 21596849]
2. Bray PF, McKenzie SE, Edelstein LC, Nagalla S, Delgrosso K, Ertel A, Kupper J, Jing Y, Londin E, Loher P, Chen HW, Fortina P, Rigoutsos I. The complex transcriptional landscape of the anucleate human platelet. BMC Genomics. 2013;14:1. [PubMed: 23323973]
3. Gnatenko D V, Dunn JJ, McCorkle SR, Weissmann D, Perrotta PL, Bahou WF. Transcript profiling of human platelets using microarray and serial analysis of gene expression. Blood. 2003;101:2285–93. [PubMed: 12433680]
4. Beliakova-Bethell N, Massanella M, White C, Lada S, Du P, Vaida F, Blanco J, Spina CA, Woelk CH. The effect of cell subset isolation method on gene expression in leukocytes. Cytometry A. 2014;85:94–104. [PubMed: 24115734]
5. Bhattacharjee J, Das B, Mishra A, Sahay P, Upadhyay P. Monocytes isolated by positive and negative magnetic sorting techniques show different molecular characteristics and immunophenotypic behaviour. F1000Research. 2017;6:2045. [PubMed: 29636897]
6. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Moser K, Ortmann WA, Espe KJ, Balasubramanian S, Hughes KM, Chan JP, Begovich A, Chang SY, Gregersen PK, Behrens TW. Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. Genes Immun. 2004;5:347–53. [PubMed: 15175644]
7. Angénieux C, Maître B, Eckly A, Lanza F, Gachet C, de la Salle H. Time-Dependent Decay of mRNA and Ribosomal RNA during Platelet Aging and Its Correlation with Translation Activity. PLoS One. 2016;11:e0148064.
8. Best MG, Sol N, Kooi I, Tannous J, et al. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. Cancer Cell. 2015;28:666–676. [PubMed: 26525104]
9. Kondkar AA, Bray MS, Leal SM, Nagalla S, Liu DJ, Jin Y, Dong JF, Ren Q, Whiteheart SW, Shaw C, Bray PF. VAMP8/endobrevin is overexpressed in hyperreactive human platelets: suggested role for platelet microRNA. J Thromb Haemost. 2010;8:369–78. [PubMed: 19943878]
10. Edelstein LC, Simon LM, Montoya RT, Holinstat M, Chen ES, Bergeron A, Kong X, Nagalla S, Mohandas N, Cohen DE, Dong J, Shaw C, Bray PF. Racial differences in human platelet PAR4 reactivity reflect expression of PCTP and miR-376c. Nat Med. 2013;19:1609–16. [PubMed: 24216752]
11. Best MG, Wesseling P, Wurdinger T. Tumor-Educated Platelets as a Noninvasive Biomarker Source for Cancer Detection and Progression Monitoring. Cancer Res. 2018;78:3407–3412. [PubMed: 29921699]
12. Schubert S, Weyrich AS, Rowley JW. A tour through the transcriptional landscape of platelets. Blood. 2014;124:493–502. [PubMed: 24904119]
13. Simon LM, Chen ES, Edelstein LC, Kong X, Bhatlekar S, Rigoutsos I, Bray PF, Shaw CA. Integrative Multi-omic Analysis of Human Platelet eQTLs Reveals Alternative Start Site in Mitofusin 2. Am J Hum Genet. 2016;98:883–97. [PubMed: 27132591]
14. Kong X, Simon L, Holinstat M, Shaw C, Bray P, Edelstein L. Identification of a functional genetic variant driving racially dimorphic platelet gene expression of the thrombin receptor regulator, PCTP. Thromb Haemost. 2017;117:962–970. [PubMed: 28251237]
15. Nassa G, Giurato G, Cimmino G, Rizzo F, Ravo M, Salvati A, Nyman TA, Zhu Y, Vesterlund M, Lehtiö J, Golino P, Weisz A, Tarallo R. Splicing of platelet resident pre-mRNAs upon activation by physiological stimuli results in functionally relevant proteome modifications. Sci Rep. 2018;8:498. [PubMed: 29323256]
16. Schwertz H, Tolley ND, Foulks JM, Denis MM, Risenmay BW, Buerke M, Tilley RE, Rondina MT, Harris EM, Kraiss LW, Mackman N, Zimmerman GA, Weyrich AS. Signal-dependent splicing of tissue factor pre-mRNA modulates the thrombogenicity of human platelets. J Exp Med. 2006;203:2433–40. [PubMed: 17060476]

17. Denis MM, Tolley ND, Bunting M, Schwertz H, Jiang H, Lindemann S, Yost CC, Rubner FJ, Albertine KH, Swoboda KJ, Fratto CM, Tolley E, Kraiss LW, McIntyre TM, Zimmerman GA, Weyrich AS. Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets. Cell. 2005;122:379–91. [PubMed: 16096058]
18. Bryois J, Buil A, Ferreira PG, Panousis NI, Brown AA, Viñuela A, Planchon A, Bielser D, Small K, Spector T, Dermitzakis ET. Time-dependent genetic effects on gene expression implicate aging processes. Genome Res. 2017;27:545–552. [PubMed: 28302734]
19. Waaseth M, Olsen KS, Rylander C, Lund E, Dumeaux V. Sex hormones and gene expression signatures in peripheral blood from postmenopausal women-the NOWAC postgenome study. BMC Med Genomics. 2011;4:29. [PubMed: 21453500]
20. Arnardottir ES, Nikonova EV, Shockley KR, Podtelezhnikov AA, Anafi RC, Tanis KQ, Maislin G, Stone DJ, Renger JJ, Winrow CJ, Pack AI. Blood-gene expression reveals reduced circadian rhythmicity in individuals resistant to sleep deprivation. Sleep. 2014;37:1589–600. [PubMed: 25197809]
21. Barendse W. The effect of measurement error of phenotypes on genome wide association studies. BMC Genomics. 2011;12:232. [PubMed: 21569388]
22. Carlborg O, De Koning DJ, Manly KF, Chesler E, Williams RW, Haley CS. Methodological aspects of the genetic dissection of gene expression. Bioinformatics. 2004;21:2383–93.
23. Hoffman GE, Schadt EE. variancePartition: interpreting drivers of variation in complex gene expression studies. BMC Bioinformatics. 2016;17:483. [PubMed: 27884101]
24. Lessells CM, Boag PT. Unrepeatable Repeatabilities: A Common Mistake. Auk. 1987;104:116–121.
25. Dohm MR. Repeatability estimates do not always set an upper limit to heritability. Funct Ecol. 2002;16:273–280.
26. Voora D, Cyr D, Lucas J, Chi JT, Dungan J, McCaffrey TA, Katz R, Newby LK, Kraus WE, Becker RC, Ortel TL, Ginsburg GS. Aspirin exposure reveals novel genes associated with platelet function and cardiovascular events. J Am Coll Cardiol. 2013;62:1267–76. [PubMed: 23831034]
27. Nix DA, Di Sera TL, Dalley BK, Milash BA, Cundick RM, Quinn KS, Courdy SJ. Next generation tools for genomic data generation, distribution, and visualization. BMC Bioinformatics. 2010;11:455. [PubMed: 20828407]
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550. [PubMed: 25516281]
29. R Development Core Team. R: A Language and Environment for Statistical Computing. 2018;Available from: http://www.r-project.org
30. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92. [PubMed: 22517427]
31. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57. [PubMed: 19131956]
32. Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, Keurentjes JJB, van Eeuwijk FA. Marker-based estimation of heritability in immortal populations. Genetics. 2015;199:379–98. [PubMed: 25527288]
33. Best MG, Sol N, In ‘t Veld SGJG, et al. Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets. Cancer Cell. 2017;32:238–252. [PubMed: 28810146]
34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. [PubMed: 23104886]
35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–1303. [PubMed: 20644199]

36. Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V. SNPassoc: an R package to perform whole genome association studies. Bioinformatics. 2007;23:654–655. [PubMed: 17237056]
37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75. [PubMed: 17701901]
38. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. [PubMed: 25722852]
39. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
40. Popadin K, Gutierrez-Arcelus M, Dermitzakis ET, Antonarakis SE. Genetic and epigenetic regulation of human lincRNA gene expression. Am J Hum Genet. 2013;93:1015–26. [PubMed: 24268656]
41. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A. 2003;100:1896–901. [PubMed: 12578971]
42. Simon LM, Edelstein LC, Nagalla S, Woodley AB, Chen ES, Kong X, Ma L, Fortina P, Kunapuli S, Holinstat M, McKenzie SE, Dong JF, Shaw CA, Bray PF. Human platelet microRNA-mRNA networks associated with age and gender revealed by integrated plateletomics. Blood. 2014;123(16):e37–45. [PubMed: 24523238]
43. Johnston GI, Bliss GA, Newman PJ, McEver RP. Structure of the human gene encoding granule membrane protein-140, a member of the selectin family of adhesion receptors for leukocytes. J Biol Chem. 1990;265:21381–5.
44. McEver RP. Properties of GMP-140, an inducible granule membrane protein of platelets and endothelium. Blood Cells. 1990;16:73–80. [PubMed: 1693535]
45. Ishiwata N, Takio K, Katayama M, Watanabe K, Titani K, Ikeda Y, Handa M. Alternatively spliced isoform of P-selectin is present in vivo as a soluble molecule. J Biol Chem. 1994;269:23708–15.
46. Semenov A V, Romanov YA, Loktionova SA, Tikhomirov OY, Khachikian M V, Vasil'ev SA, Mazurov A V. Production of soluble P-selectin by platelets and endothelial cells. Biochem Biokhimiîâ. 1999;64:1326–35.
47. Ataga KI, Kutlar A, Kanter J, et al. Crizanlizumab for the Prevention of Pain Crises in Sickle Cell Disease. N Engl J Med. 2017;376:429–439. [PubMed: 27959701]
48. Lee DS, Larson MG, Lunetta KI, Dupuis J, Rong J, Keaney JF, Lipinska I, Baldwin CT, Vasan RS, Benjamin EJ. Clinical and genetic correlates of soluble P-selectin in the community. J Thromb Haemost. 2007;6:20–31. [PubMed: 17944986]
49. Burger PC, Wagner DD. Platelet P-selectin facilitates atherosclerotic lesion development. Blood. 2003;101:2661–2666. [PubMed: 12480714]
50. Penman A, Hoadley S, Wilson JG, Taylor HA, Chen CJ, Sobrin L. P-selectin Plasma Levels and Genetic Variant Associated With Diabetic Retinopathy in African Americans. Am J Ophthalmol. 2015;159:1152–1160.e2. [PubMed: 25794792]
51. Gibbs RA, Boerwinkle E, Doddapaneni H, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. [PubMed: 26432245]
52. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558:73–79. [PubMed: 29875488]
53. Raponi M, Kralovicova J, Copson E, Divina P, Eccles D, Johnson P, Baralle D, Vorechovsky I. Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. Hum Mutat. 2011;32:436–444. [PubMed: 21309043]
54. Platt A, Vilhjálmsson BJ, Nordborg M. Conditions under which genome-wide association studies will be positively misleading. Genetics. 2010;186:1045–52. [PubMed: 20813880]
55. Cramer P, Pesce CG, Baralle FE, Kornblihtt AR. Functional association between promoter structure and transcript alternative splicing. Proc Natl Acad Sci U S A. 1997;94:11456–60.

56. Radich JP, Mao M, Stepaniants S, Biery M, Castle J, Ward T, Schimmack G, Kobayashi S, Carleton M, Lampe J, Linsley PS. Individual-specific variation of gene expression in peripheral blood leukocytes. Genomics. 2004;83:980–988. [PubMed: 15177552]
57. Campbell RA, Schwertz H, Hottz ED, et al. Human megakaryocytes possess intrinsic anti-viral immunity through regulated induction of IFITM3. Blood. 2019;blood-2018–09-873984.
58. Koupenova M, Clancy L, Corkrey HA, Freedman JE. Circulating Platelets as Mediators of Immunity, Inflammation, and Thrombosis. Circ Res. 2018;122:337–351. [PubMed: 29348254]
59. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clin Chem. 2009;55:611–622. [PubMed: 19246619]
60. Altshuler D, Daly MJ, Lander ES. Genetic Mapping in Human Disease. Science. 2008;322:881–888. [PubMed: 18988837]
61. Kumar V, Westra HJ, Karjalainen J, et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet. 2013;9:e1003201.
62. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167:1415–1429.e19. [PubMed: 27863252]
63. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49:139–145. [PubMed: 27918533]
64. Ludwig RJ, Schön MP, Boehncke WH. P-selectin. Expert Opin Ther Targets. 2007;11:1103–1117. [PubMed: 17665981]
65. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57. [PubMed: 19131956]

NOVELTY AND SIGNIFICANCE

What Is Known?

- Human platelets have a rich repertoire of RNAs.
- Platelet RNA expression differs between individuals in health and disease, and is associated with platelet function.
- Single time-point studies of the platelet transcriptome are increasingly utilized for biological discovery in human health and disease.

What New Information Does This Article Contribute?

- Platelet RNA expression is stable and repeatable for up to 4 years when assessed over time in healthy human donors.
- The integrated use of longitudinal repeatability metrics significantly enhances the discovery of genetic variants that affect gene expression and splicing.
- A genetic variant in the SELP gene directs the removal of the P-selectin transmembrane domain.

Although anucleate, platelets possess a rich and dynamic transcriptome. Platelet transcriptomics are increasingly used, with applications ranging from cancer diagnostics to novel gene discovery. Our study adds to the field by establishing, for the first time, the stability – or reproducibility – of platelet gene expression and splicing in healthy donors assessed repeatedly for up to four years. This type of longitudinal assessment has been lacking for any primary human cell, let alone platelets. We found that the platelet transcriptome is exquisitely stable in health, which may aid comparisons in disease settings, and enhance diagnostics and prognostics that use platelet RNA. Moreover, we show that integrating measures of repeatability (e.g. between versus within-individual variation of platelet RNA expression and splicing) results in improved detection of genes affected by nearby genetic variants. We apply this technique in the discovery of a platelet SELP sQTL. Functionally, this splice QTL directs the removal of the transmembrane domain of P-selectin. We show that this transmembrane domain deletion is reduced in blacks compared to whites. In vitro, this increases surface, but decreases soluble, P-selectin. We suggest that this may have implications in diseases more common in blacks where P-selectin is a therapeutic target (e.g. sickle cell disease).

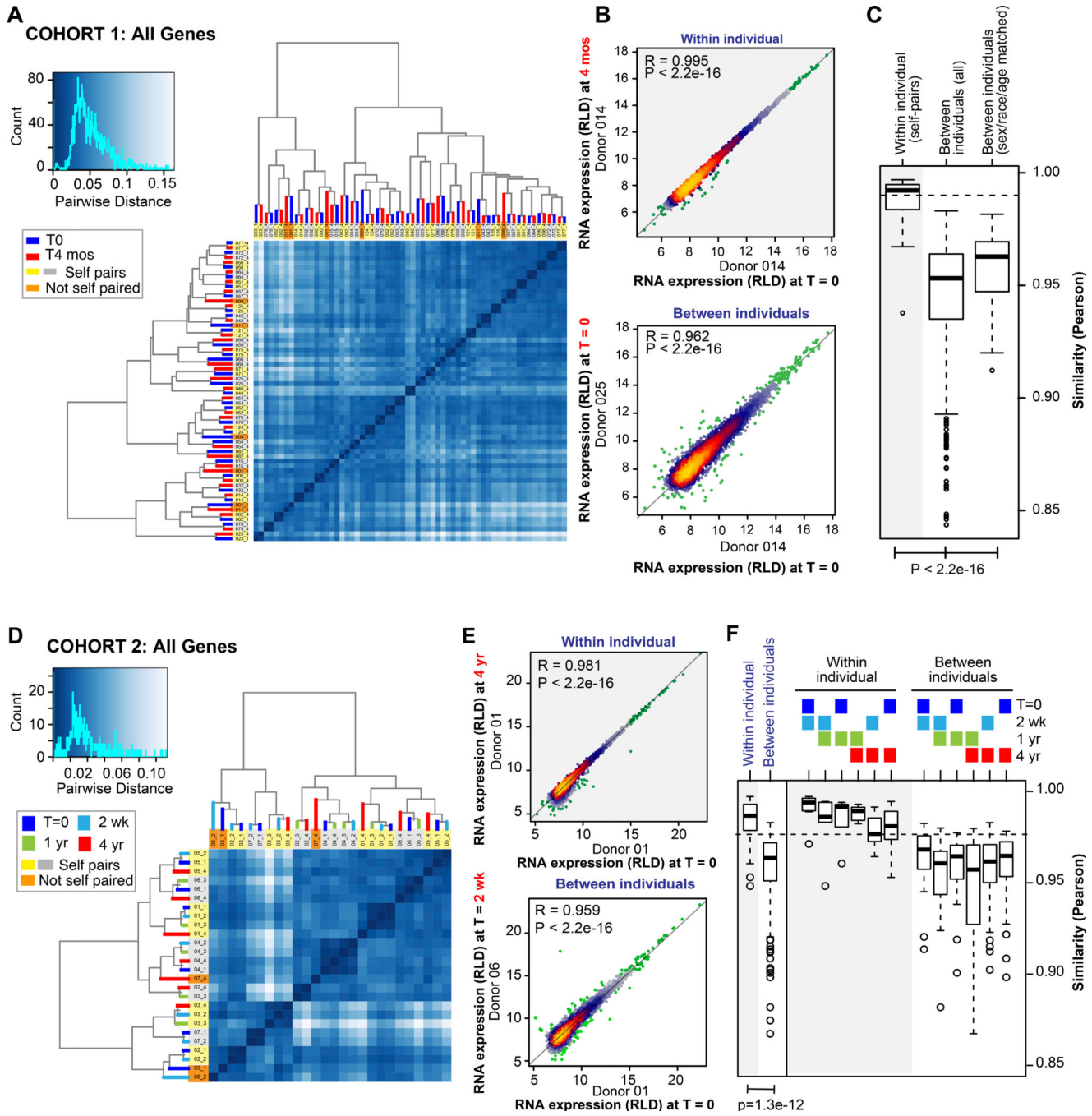


Figure 1. Within and between individual stability of platelet RNA expression over 4 months (cohort 1) and 4 years (cohort 2).

A and D: Unsupervised clustering and heatmaps of total RNA expression in platelets from all samples in A) cohort 1 and D) cohort 2. The histograms to the left of each heatmap show the distribution of distances between all pairs of samples, and the darkness of blue indicates the degree of similarity between pairs of samples. Samples that cluster as neighbors in the heatmap dendrograms reflect transcriptomes with the highest similarity. Nearest neighbor self-pairs are highlighted in yellow and gray, whereas nearest neighbor non-self pairs are highlighted in orange. B and E: Example individual correlation plots of all transcripts in B)

cohort 1 or E) cohort 2. Each data point represents the regularized, log-transformed expression level (RLD) of a single transcript from the specified donor at time 0 (x axis) versus 0, 2 wk, 4 months, or 4 years (y axis) within the same individual (top panels) or a different individual (bottom panels). Points are heat-colored according to density. P values are from Pearson correlation. C and F: Boxplots summarizing the RNA expression Pearson correlation between all within versus between-individual pairs at C) time 0 and 4 months or F) in aggregate at all time points (left) or at the individually specified time points (right). With regards to specified time points in F, note that the average within-individual correlation did not significantly decrease as samples taken farther apart were compared. For example, there was not a significant difference when comparing the average within-individual correlation of T0 versus 2 weeks with the average within-individual correlation of T0 versus 4 years. Boxplots for cohort 1 (C) are shown before and after adjusting for age, sex, and race, whereas they are not adjusted for cohort 2 (F), because of the smaller sample size. P values are from Wilcox test, adjusted.

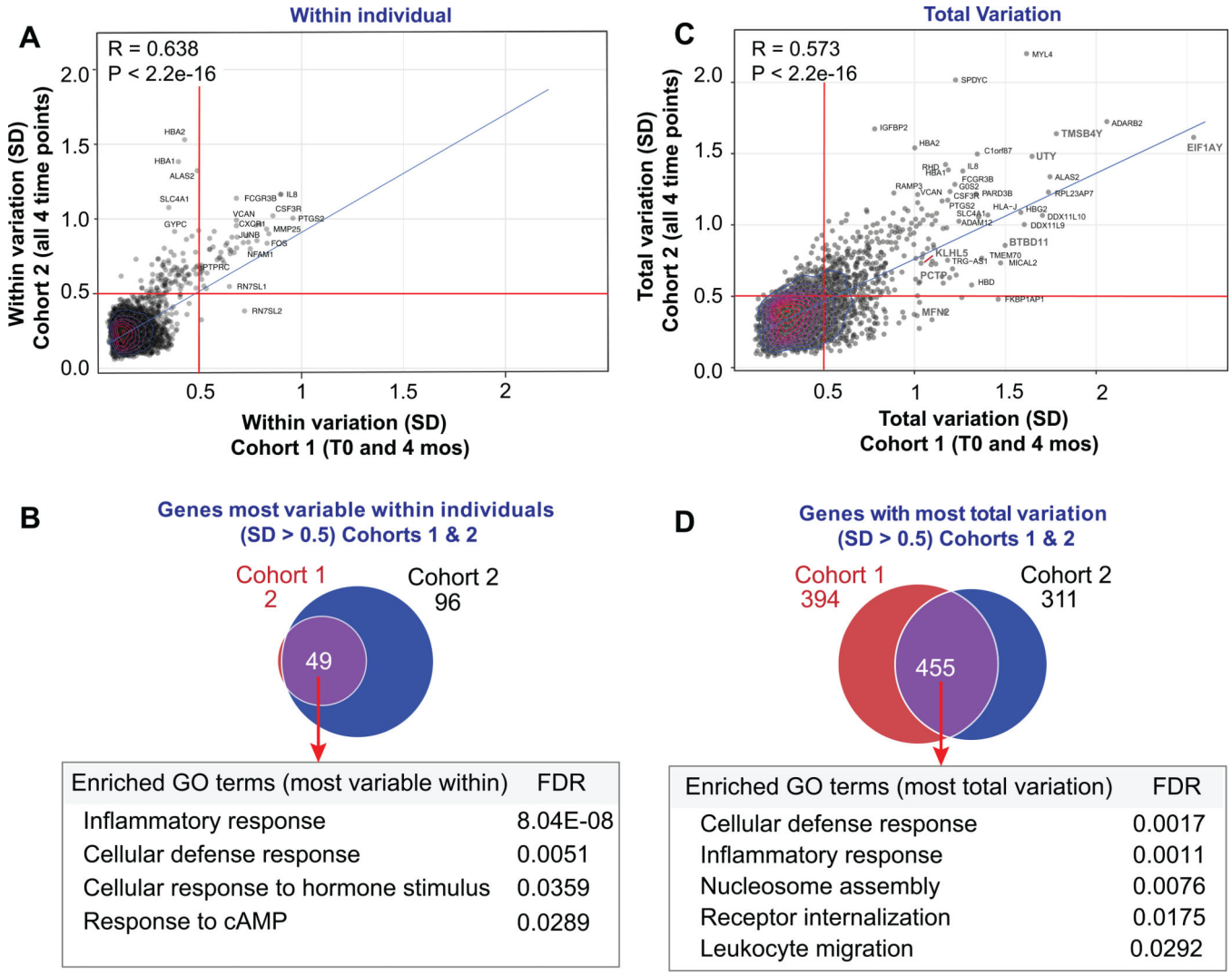


Figure 2. Comparison between cohorts of the within and total variation of each transcript in platelets.

The mean within and total individual variation (standard deviation, SD) was calculated from the regularized log transformed expression (RLD) for each transcript. A and C: the A) within or C) total individual variation of each transcript in cohort 1 (x-axis) plotted against the respective variation of each transcript in cohort 2 (y-axis). The horizontal and vertical lines at 0.5 marks an arbitrary threshold of variation used for the Venn diagrams in B and D. B and D: Venn diagrams of the overlap in the transcripts with highest B) within and D) total variation. Listed below each Venn are the significantly enriched GO terms for the transcripts overlapping both cohorts. FDR = Benjamini False Discovery Rate calculated by David[65].

A

Top Genes Ranked By Repeatability	Association With Trait in PRAX1 Sex FDR	Race FDR	PLT eQTL FDR
RPL23AP7	NS	2e-04	2e-34
EIF1AY	1e-78	NS	NS
BTBD11	NS	8e-04	6e-33
KLHL5	NS	NS	5e-35
TUBB2A	NS	2e-9	NS
TESPA1	NS	NS	1e-23
TMSB4Y	5e-66	NS	NS
MFN2	NS	NS	4e-42
MYO3B	NS	1e-13	NS
TUBB6	0.03	NS	7e-15
RHD	NS	0.04	3e-21
FHL3	NS	NS	3e-43
SPDYC	NS	NS	2e-12
MEST	NS	NS	4e-16
HTATIP2	NS	NS	3e-33

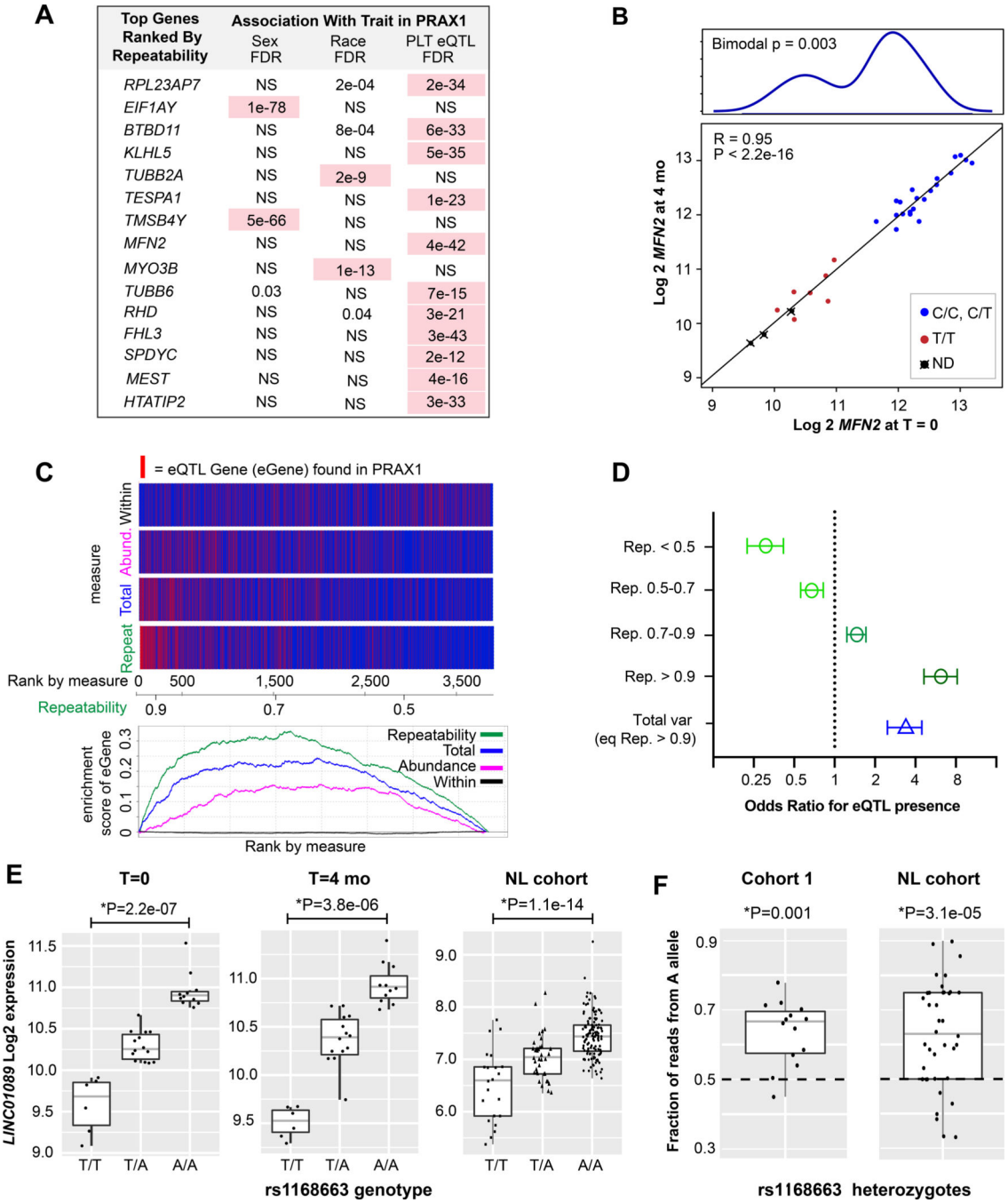


Figure 3. Transcripts ranked by repeatability are enriched in heritable traits and eQTLs.

A) Table of transcripts with the highest repeatability in cohort 1 RNA-seq data, and their reported association with race, sex, or eQTLs in PRAX1[13,42] microarray data. Associations with FDR < 1e-4 are highlighted in pink. NS = not significant. B) Correlation plot of RNA expression (log normalized) of *MFN2* at time 0 (x-axis) and 4 months (y-axis). Points are colored according to rs1474868 genotype (ND = not determined). Above is a density histogram showing a bimodal distribution according to genotype. Bimodal P value from Hartigan's diptest for multi-modality. C) Top: enrichment plots for the presence of eQTLs

ranked according to different measures: within variation, mean expression abundance, total variation, or repeatability. The axis below the plot indicates the gene rank according to each measure, and indicates the value of the repeatability measure (the values of the other measures are not noted on the axis). Genes with a known eQTL are in red, those without are in blue. Thus, genes with the highest repeatability are nearly 100% eQTL genes, whereas those with the lowest repeatability are nearly 0%. Bottom: plot of cumulative enrichment scores for each metric. D) Odds ratios for the likelihood of identifying an eQTL for genes at the indicated repeatability thresholds compared to the same number of genes ranked by total variation. E) Boxplots of *LINC01089* expression according to rs1168863 genotype in cohort 1 at time 0 and 4 months and in the NL cohort[33]. *P values adjusted for age, sex, and cohort 1) race or cohort 2) population structure (inferred genetic ancestry[37,38]). F) Boxplots demonstrating allelic imbalance of rs1168863 within heterozygotes in cohort 1 and the NL cohort. The proportion of RNA-seq reads with A nucleotide versus T nucleotide was calculated and plotted for each heterozygote individual.

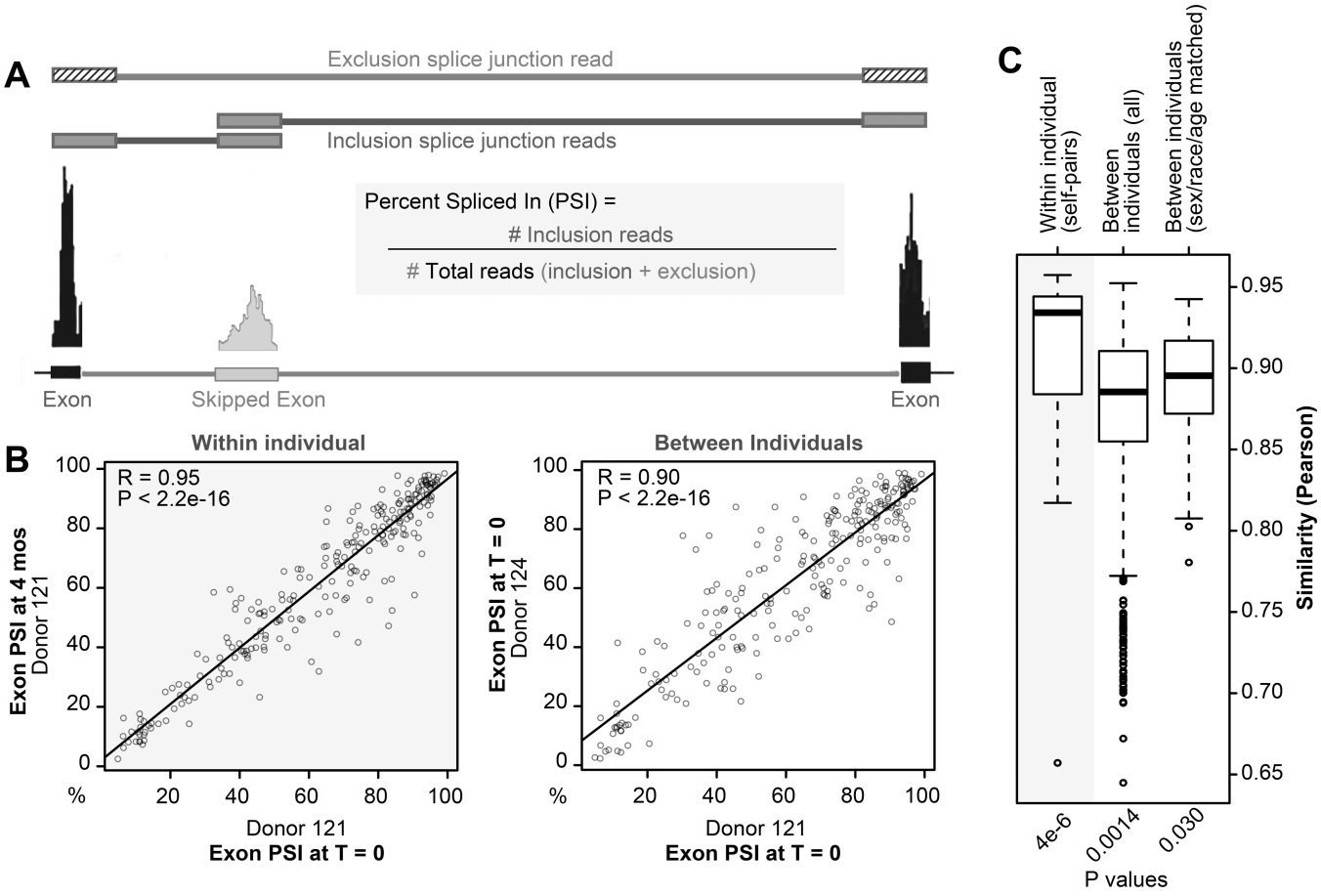
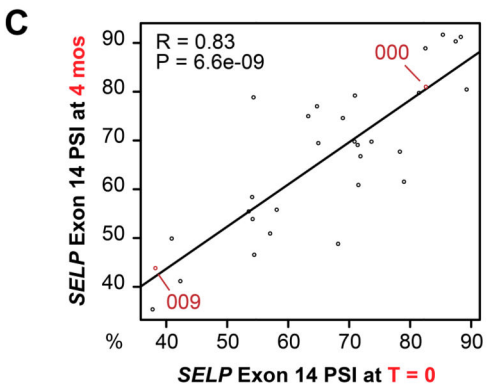


Figure 4. Within and between individual stability of exon skipping in platelets.

A) Schematic of how exon skipping events are defined. Percent exon Spliced In (PSI) is calculated using splice junction reads and is the ratio of exon inclusion junction reads over total junction reads. B) Correlation plots the PSI for all exon skipping events within (left panel) and between (right panel) individuals. Each point represents a single exon skipping event from the specified donor at time 0 (x axis) versus 0 or 4 months (y axis). C) Boxplots summarizing Pearson correlations of within versus between-individual pairs when analyzing PSI of all exon skipping events at time 0 and 4 months, and after adjusting for age, sex, and race. *Wilcox test, adjusted.

A

Top exons by repeatability	Median PSI	PSI repeatability
TSPOAP1-AS1 exon 2	0.10	0.86
SELP exon 14	0.69	0.84
TMEM40 (2-5) exon 4	0.53	0.83
LINC00152 exon 3	0.08	0.83
FYB exon 12	0.89	0.83
MAX exon 5	0.42	0.82
ENKUR exon 2	0.16	0.79
RBX1 exon 4	0.88	0.78
MTURN exon 2	0.31	0.77
METTL22 exon 10	0.49	0.77



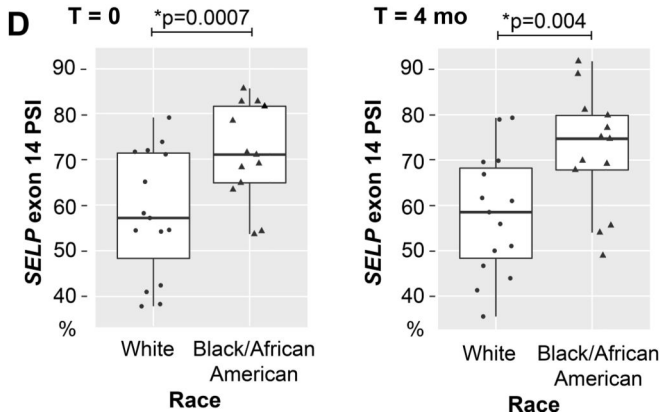


Figure 5. Repeatability of Exon 14 skipping in *SELP* and association with race.

A: Table of the most repeatable exon skipping events in platelets. B: Representative IGV plots of sequencing reads from two different individuals at time 0 and 4 months, showing the differential distribution of reads between individuals that align to or skip exon 14 of *SELP*. The histograms indicate the cumulative abundance of reads that aligned to each exon. A subset of individual reads is shown below each histogram that indicate split splice junction reads by thin lines (absent in read) that connect to thick lines (mapped portion of read). Red and blue reads are splice junction reads that align to or skip exon 14 respectively. C: Correlation plot of *SELP* exon 14 PSI. Each point represents the PSI for an individual donor at time 0 (x-axis) and 4 months (y-axis). Donors represented in the IGV plots in B are labeled in red text. D) Boxplot of *SELP* exon 14 mean PSI according to race at T=0 and T=4 months.

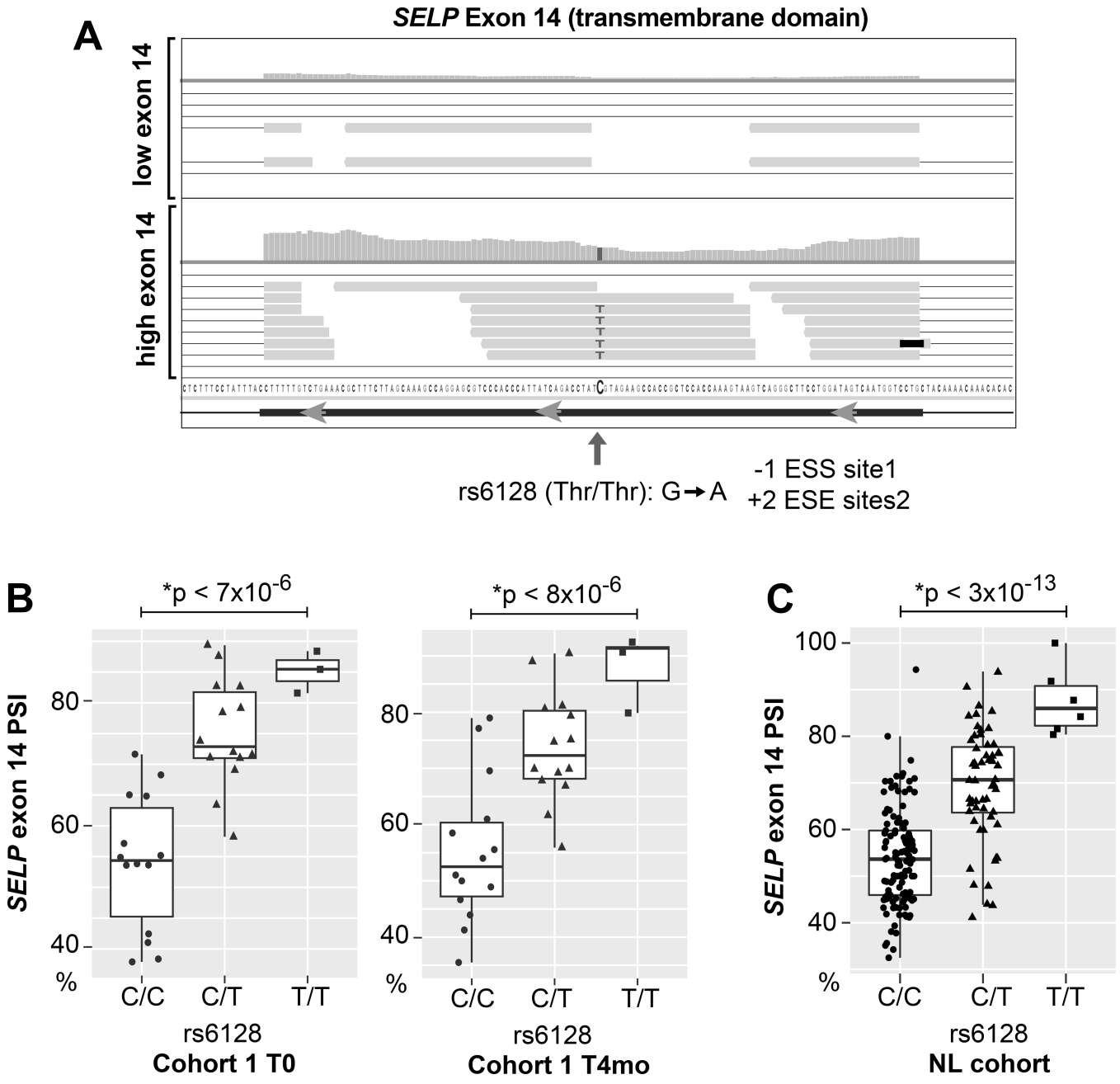


Figure 6. rs6128 is a platelet *SELP* exon 14 splice QTL.

A) Close up IGV plot showing read distribution across *SELP* exon 14 for Top) an individual with rs6128 A/A and relatively high levels of exon skipping reads or Bottom) an individual with rs6128 (T/T) variant. The C->T change does not change amino acid sequence, but alters exonic splicing silencer and enhancer sites as predicted by Ex-Skip[53]. B) Boxplot of *SELP* exon 14 mean PSI according to rs6128 genotype inferred from RNA-seq in cohort 1 at time 0 and 4 months. C) Boxplot of *SELP* exon 14 mean PSI according to rs6128 genotype inferred from RNA-seq data published in the NL cohort. *P values adjusted for age, sex, and B) race or C) population structure (inferred genetic ancestry[37,38]).

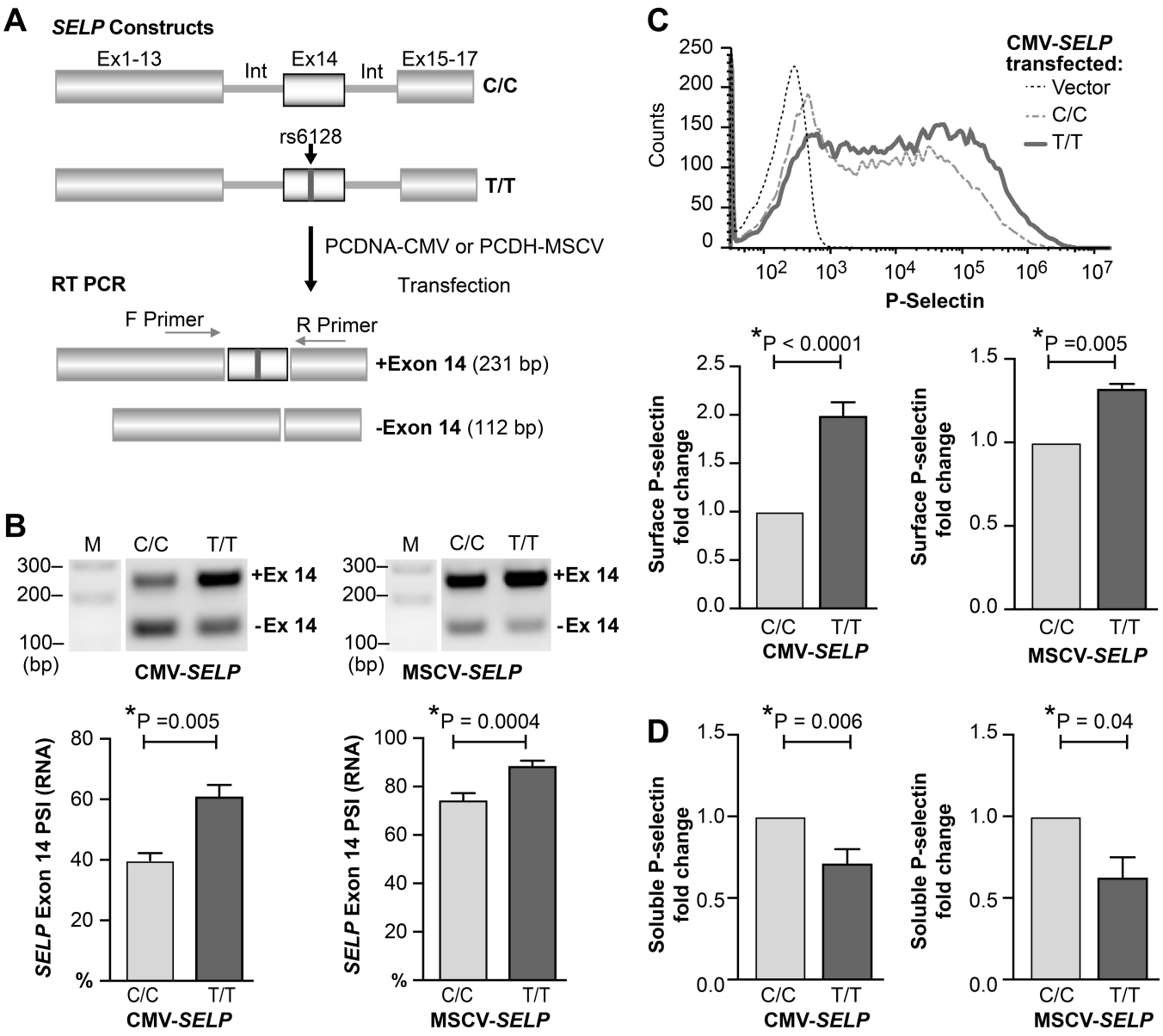


Figure 7. rs6128 directly regulates exon 14 skipping in *SELP* and alters the ratio of surface to soluble P-selectin protein expression.

A) Schematic of mini-gene constructs of *SELP* that include the ORF of *SELP*, and the introns flanking exon 14. The C/C and T/T constructs vary by a single nucleotide at rs6128. Constructs were cloned into vectors with 2 different promoters (CMV or MSCV). After transfection into HEK 293 cells, the introns are spliced out and exon 14 is variably spliced out (skipped). The extent of exon 14 skipping is measured by PCR via exon 14 flanking primers that generate two PCR products of different sizes. B) RT-PCR analysis of *SELP* exon 14 skipping following transfection of HEK 293 cells with rs6128 C/C or T/T vectors. Shown is a representative result from 5 independent experiments. Below are bar graphs and standard error summary of PSI calculated according to densitometry analysis of the exon 14 inclusion band (upper band) divided by the sum of the upper and lower bands (total).*paired t-test, n=5 independent experiments. C) Flow cytometry analysis of P-selectin surface

expression following transfection of HEK 293 cells with rs6128 C/C or T/T vectors. Top is a representative histogram overlay of P-selectin surface expression 24 hours after transfection with CMV promoter empty vector, rs6128 C/C, or T/T. Below are bar graph and standard error summaries of the fold change (normalized to transfection) of surface P-selectin MFI following transfection. *paired T test, n=5–6 pairs per group. D) ELISA analysis of soluble P-selectin in supernatants of HEK 293 cells following transfection with rs6128 C/C or T/T vectors. *paired T test, n=12–14 pairs per group.

Table 1.

Characteristics of cohorts 1 and 2 and timeline of platelet collection.

	Cohort 1 (=31)	**Cohort 2 (n=7)**
Center	Duke University	University of Utah
Age (yrs)	42 ± 11	47 ± 9.1
Male Gender, n (%)	10 (32%)	5 (71%)
Race/Ethnicity, n (%)		
Black or African American	13 (42%)	0%
Caucasian	15 (48%)	7 (100%)
Unknown or Not Reported	3 (10%)	0%
Platelet RNA-sequencing