Online supplemental Methods, Figures, and Figure Legends

RNA isolation and sequencing

Following isolation, fresh platelet pellets were suspended in 1 mL of Trizol and frozen at -80c until RNA isolation. RNA from cohort 1 was isolated using phenol-chloroform extraction, isopropanol precipitation in the presence of glycogen, and 75% ethanol washes. Samples were DNAse treated (Invitrogen #AM1907), and RNA re-purified using ammonium acetate/isopropanol precipitation as described[1]. RNA from cohort 2 was isolated and DNAse treated using DirectZol kit and columns (Zymogen).

Novel platelet eQTL and sQTL analysis

RNA-seq fastq files for 234 previously published samples (Best et. al.[2,3], Netherlands cohort; hereafter referred to as NL cohort) were retrieved from NCBI short read archive PRJNA353588[2]. These, and fastq files for cohort 1, were aligned with STAR[4] to human reference genome (build HG38) in a splice-aware manner, and variants were called using the workflow built from the Genome Analysis Toolkit (GATK)[5] best practices for variant calling on RNA-seq. Variant call information and hard filters are reported in the "INFO" field of Online Datasets VII and XII. Variants tested for eQTLs were limited to within 2 kb of all genes not identified as eQTLs by PRAX1, 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. Variants were further excluded that were not identified in both cohort 1 and in the NL cohort. Combined filtering resulted in 641 variants across 181 genes. The RNA-seq allele frequencies of these variants was

comparable to DNA allele frequencies reported by the Genome of the Netherlands project (GoNL[6]) and 1000 genomes[7] (Online Figure XA), and clustered according to the expected populations by PCA analysis of allele frequencies (Online Figure XB). RNA-seq and expected allele frequencies for significant variants are reported in Online Datasets VII and XII. To assess population stratification, transcriptome wide variants were called. Multidimensional scaling (MDS) was implemented in Plink 1.9[8,9] on 1994 variants (filtered and pruned: FS > 30.0; QD < 2.0; clusterSize=3, clusterWindowSize=35; --geno 0.2; --hwe 10e-6; --maf .01; --indep-pairwise 50 5 0.2) co-identified from RNA-seq in cohorts 1, the NL cohort, and from 1000 genomes. Visual inspection of MDS plots indicated that individual RNA-seq samples clustered according to expected genetic ancestry, and that population structure was captured within the first 4 MDS components (Online Figure XC). Variance Stabilizing Transformation (VST) in the package DESeq2[10] was used to normalize gene expression. Gene-variant association was tested in the R package SNPassoc[11] using an additive model of variant-allele dosage (0,1,2), while controlling for the covariates sex, age, and population structure (first 4 MDS components). 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[12]. After significance testing, eQTLs were further limited to those reported in dbSNP to minimize the possibility of RNA-specific (i.e. RNA-editing) calls. At this threshold, 27 variants across 11 new platelet eQTL genes were identified (Online Dataset VII). These remained significant after further controlling for the first 5 latent variables estimated from surrogate variable analysis (SVA)[13] with explicit

adjustment for sex and age. The 27 significant gene-variant associations were then tested in Cohort 1 using the strategy described above for the NL cohort, with sex, age, and race included as covariates. However, because of the small sample size, a codominant model was allowed when 1 of the 3 genotypes required for additive model testing was missing. The results of cohort 1 eQTL analysis are provided as additional information in table VII, but given the comparatively small sample size, significance in cohort 1 was not expected and was not considered as criteria for novel platelet eQTL selection.

Caveats and limitations of RNA-seq for variant analysis

There are recognized strengths and limitations of using RNA-seq to infer genetic variants[14]. RNA variant calls are of different quality than genome calls. They are limited to expressed regions which has precluded fine mapping of causal eQTLs. Where there is extreme allelic imbalance, genetic variants can also be missed. As with any eQTL analysis, false positives are also possible, such as from confounding LD and genetic substructure not accounted for by large-scale population stratification. Because of the low density of variants in RNA-seq data, fine-scale structure analysis is not possible. With such limitations, additional observations are helpful in interpreting the results: the allele frequencies of RNA-seq calls for the significant variants aligned with expected allele frequencies (with the exception of rs879095052 in HBG1), 22/27 of the eQTLs (for 7/11 of the eGenes) have been reported in other tissues (The Genotype-Tissue Expression (GTEx) Project, see Online Dataset VII for specific tissues), and variants for

8/11 eGenes demonstrated allelic imbalance that was directionally consistent with the eQTL effect on expression.

Online Supplemental Figure Legends

Online Figure I. **Within and between individual stability of platelet non-coding RNA expression over 4 months (cohort 1) and 4 years (cohort 2)**. A and D: Unsupervised clustering and heatmaps of non-coding 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 non-coding 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 non-coding transcripts in B) cohort 1 or E) cohort 2. Each data point represents the regularized, log-transformed expression level (RLD) of a single non-coding 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. C and F: Boxplots summarizing the non-coding 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). In C, boxplots for cohort 1 are shown before and after adjusting for age, sex, and race, whereas only unadjusted are shown for cohort 2 (because of smaller sample size). P values from Wilcox test, adjusted.

Online Figure II. **Comparison of the within-individual 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-B: normalized expression (x-axis) plotted against within individual variation for each transcript (y-axis) for A) cohort 1 and B) cohort 2. C-D: normalized expression (x-axis) plotted against total variation for each transcript (y-axis) for C) cohort 1 and D) cohort 2. E-F: Total individual variation of each transcript (x-axis) plotted against the within-individual variation (y-axis) of each transcript for E) cohort 1 and F) cohort 2. Labeled points are representative transcripts with low-within and high total individual variation.

Online Figure III. **Variance partition analysis of platelet gene expression**. Violin plots showing the distribution of the percent of variance for each transcript (y-axis) attributable to the indicated covariates (x-axis). The width of the violin indicates the probability density of transcripts at each y-value. Boxplots indicate median and interquartile range, and outliers are plotted as individual points. For example, sex explains less than 50% of the variation for most transcripts, except for the Y chromosome genes EIF1AY, TMSB4Y, and UTY which vary almost exclusively according to sex. The plot on the far right indicates that for most transcripts (>50%), differences between individuals explain the majority of variation.

Online Figure IV. **Table of transcripts** with A) the highest expression, B) lowest within-individual variation, C) highest total variation, or D) highest repeatability (low within, high between individual variation) in cohort 1 RNA-seq data, and their reported association with race, sex, or eQTLs in PRAX1 microarray data. Associations with FDR < 1e-4 are highlighted in pink. NS = not significant.

Online Figure V. **For transcripts with a reported eQTL in PRAX1, the FDR is associated with repeatability.** On the x-axis is the lowest reported log FDR (i.e. -125 = 10^{-125}) for eQTLs associated with each transcript. On the y-axis is 1-repeatability (cohort 1) for each transcript.

Online Figure VI. Unsupervised clustering and heatmap based on the Exon PSI for all 245 identified exon skipping events in platelets within the 31 individuals in cohort 1 at T=0 and T=4 months. The histograms to the left 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 samples with the highest similarity in exon skipping levels. Nearest neighbor self-pairs are highlighted in yellow. Bars to the left are colored according to sequencing batch or lane.

Online Figure VII. PCR confirmation of *SELP* exon 14 skipping in platelets. Platelet RNA from 5 different individuals was reverse transcribed and cDNA amplified with

primers flanking exon 14 of *SELP*. Bands were extracted and sequenced by Sanger sequencing to confirm sequence.

Online Figure VIII. **Exon 14 skipping of *SELP* remains associated with rs6128 in disease.** Boxplots of *SELP* exon 14 mean PSI according to rs6128 genotype inferred from RNA-seq in all healthy and disease samples reported in the NL cohort when analyzed in A) aggregate or B) according to disease. In B, only diseases with multiple samples of at least 2 different genotypes are shown. *p adjusted for age, sex, smoking, hospital, and storage time.

Online Figure IX. **rs6128 directly regulates exon 14 skipping in P-selectin protein**. Western blot analysis of P-selectin (antibody to c-terminal DYK tag) in HEK 293 cells following transfection of rs6128 C/C or T/T *SELP* (CMV promoter) constructs. Shown is a representative blot from 4 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=4 independent experiments. Note that anti-DYK antibody detected 2 distinct bands that differ by ~19 kDa whereas exon 14 encodes only 40 amino acids (< 5 kDa). The difference is from the heavy glycosylation of exon 14 as deglycosylation of lysates with PNGase rendered the bands indistinguishable in size (data not shown).

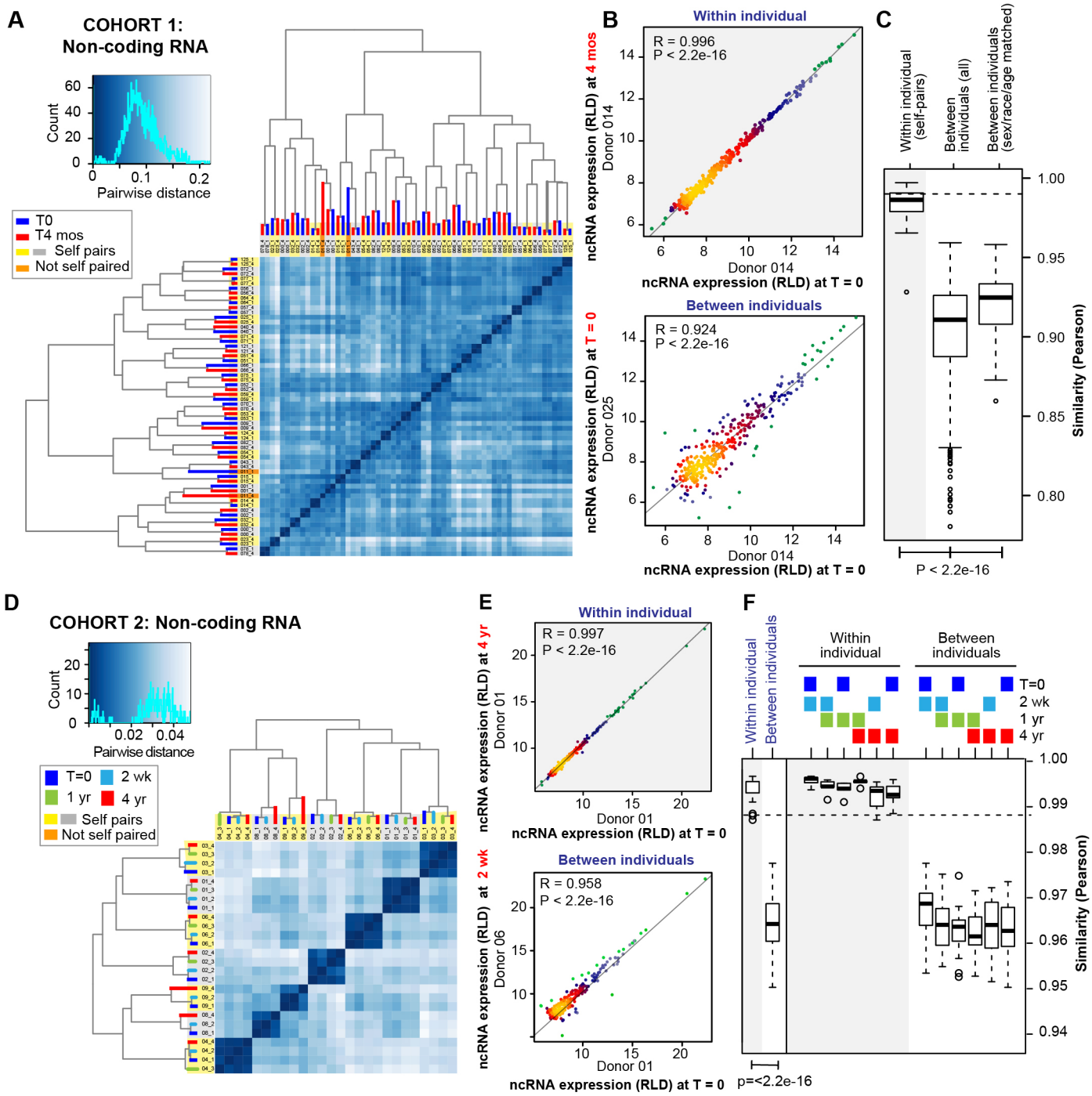
Online Figure X. **Comparison of RNA-seq with genomic variant calls and population stratification**. A) Allele frequencies of the 641 variants tested for eQTL

presence as called by RNA-seq in the NL cohort (x-axis) versus allele frequencies as reported in the Netherlands genome database (GoNL[6]); Pearson cor=0.93 ($p < 2.2e-16$). Allele frequencies for cohort 1 compared to those reported in 1000 genomes database[7] were also assessed but are not shown: cor=0.87 ($p < 2.2e-16$) for white individuals in cohort 1 RNA-seq calls vs European superpopulation genomes; cor = 0.89 ($p < 2.2e-16$) for black/African American individuals in cohort1 RNA-seq calls vs African superpopulation genomes B) PCA analysis comparing allele frequencies of the 641 variants as called by RNA-seq in Black/African American (AA) or white individuals in cohort 1, or the NL cohort, with allele frequencies reported in GoNL and 1000 genomes database superpopulations: East Asian (EAS), South Asian (SAS), Ad Mixed American (AMR), European (EUR), or African (AFR). C) MDS analysis[8,9] of population structure for each individual in cohort 1 and the NL cohort anchored to individuals from 1000 genomes. 1994 variants co-identified in the RNA-seq cohorts and 1000 genomes were used in the analysis. Results indicate that cohort 1 white individuals and the NL cohort RNA-seq individuals cluster almost exclusively with the EUR genome superpopulation, whereas the cohort 1 black/African American and other/unknown RNA-seq individuals cluster with the AFR superpopulation, and suggest admixture. The top three MDS components are plotted. Zoom-boxes are included for clarity where there is a high density of cohort1 white, NL cohort, and EUR individuals.

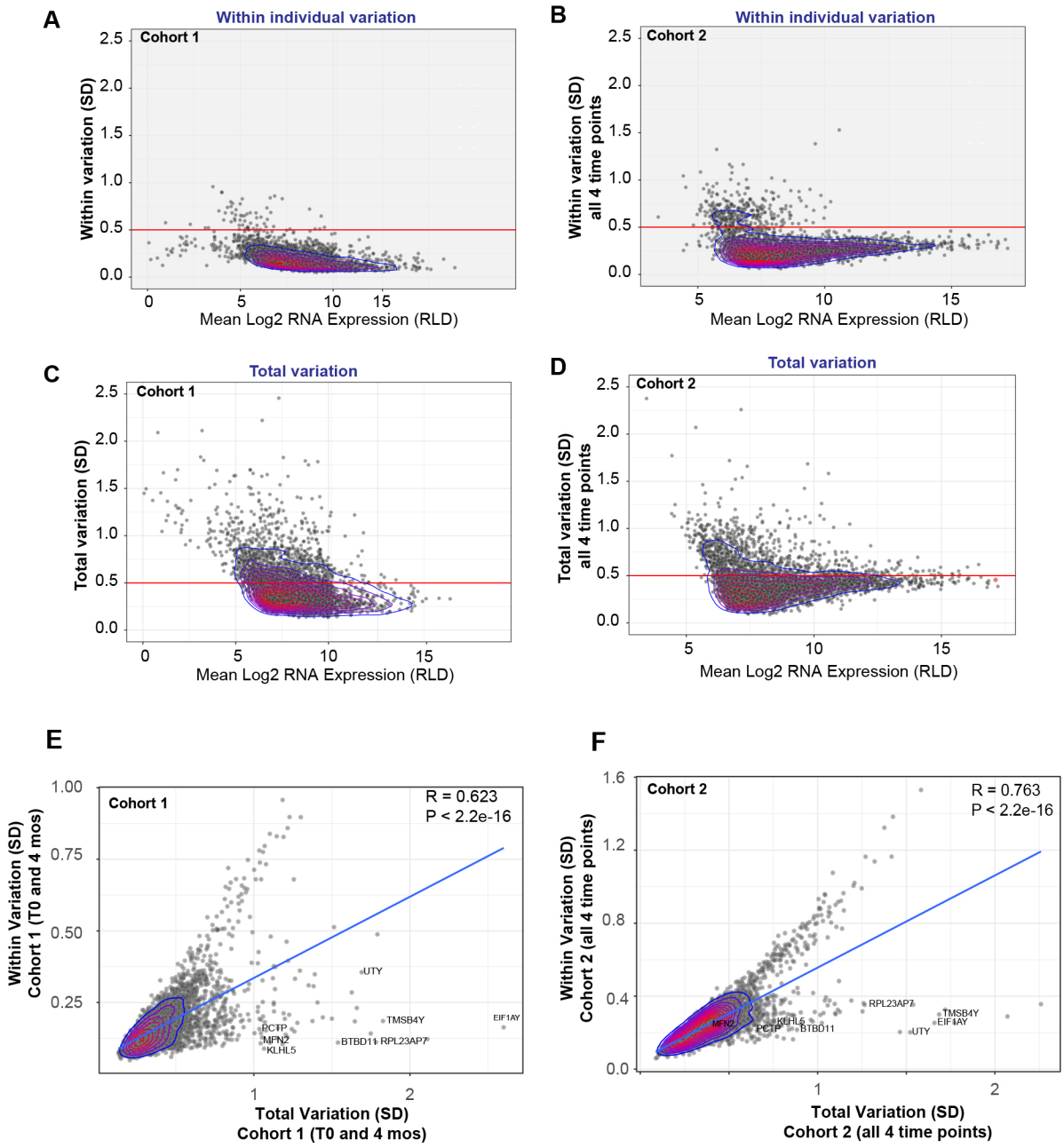
References

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.
2. Best MG, Sol N, In 't Veld SGJG, et. al. Swarm Intelligence-Enhanced Detection

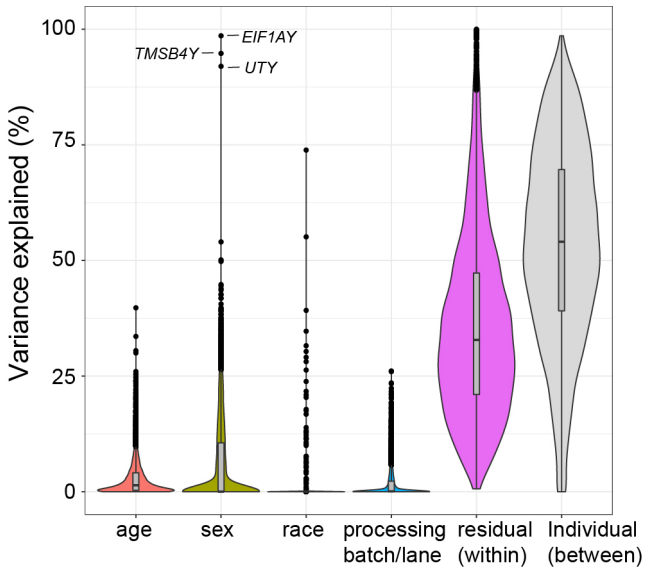
of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets. *Cancer Cell*. 2017;32:238-252.
3. Best MG, Sol N, Kooi I, Tannous J, et. al. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell*. 2015;28:666–676.
4. Dobin 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.
5. 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.
6. Genome of the Netherlands Consortium, Francioli LC, Menelaou A, et. al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014;46:818–825.
7. Gibbs RA, Boerwinkle E, Doddapaneni H, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
8. 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.
9. 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.
10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15:550.
11. 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.
12. 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.
13. Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet. 2007;3:e161.
14. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93:641–51.



Online Figure I



Online Figure II



Online Figure III

A

Top genes ranked by highest expression	Association With Trait in PRAX1		
	Sex FDR	Race FDR	PLT eQTL FDR
B2M	NS	NS	NS
NRGN	NS	NS	NS
PPBP	NS	NS	NS
HLA-E	NS	NS	NS
TMSB4X	NS	0.02	NS
FLNA	NS	NS	NS
SPARC	NS	NS	NS
TUBB1	NS	0.04	NS
FTH1	NS	4e-04	NS
TLN1	NS	NS	NS
F13A1	NS	NS	NS
ACTB	NS	NS	NS
GNAS	NS	0.03	NS
CCL5	NS	5e-03	NS
CLU	NS	NS	NS

B

Top genes ranked by within variation	Association With Trait in PRAX1		
	Sex FDR	Race FDR	PLT eQTL FDR
SRSF8	NS	0.01	NS
GRB2	NS	NS	NS
ARPC3	NS	0.04	NS
ABLIM3	NS	NS	NS
HNRNPUL2	NS	NS	NS
UBE2F	NS	NS	NS
SUPT4H1	NS	NS	3e-8
HSBP1	NS	3e-03	3e-9
SLC35D2	NS	NS	NS
OSTF1	NS	NS	NS
PSMD2	NS	NS	NS
UIMC1	NS	NS	NS
RPL15	NS	0.02	NS
C6ORF106	NS	NS	NS
UBE2Q1	NS	NS	NS

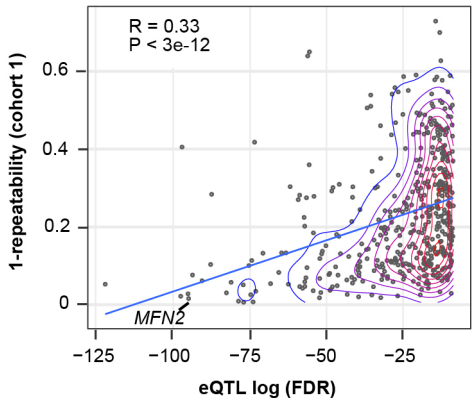
C

Top genes ranked by total variation	Association With Trait in PRAX1		
	Sex FDR	Race FDR	PLT eQTL FDR
EIF1AY	1e-78	NS	NS
TUBB2A	NS	2e-9	NS
TMSB4Y	5e-66	NS	NS
ALAS2	NS	NS	NS
RPL23AP7	NS	2e-04	2e-34
BTBD11	NS	1e-02	6e-33
MYO3B	NS	1e-13	NS
TMEM70	NS	NS	2e-8
SLC4A1	NS	NS	NS
C1orf87	NS	0.03	NS
PARD3B	NS	NS	NS
HBG1	NS	NS	NS
HBD	0.05	NS	NS
ZC3H6	NS	NS	NS
RSRC1	NS	NS	NS

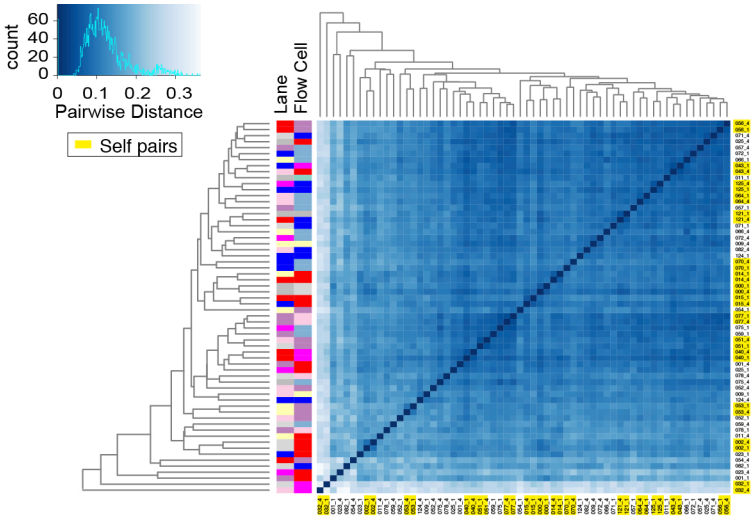
D

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

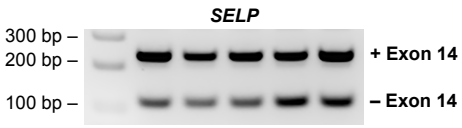
Online Figure IV



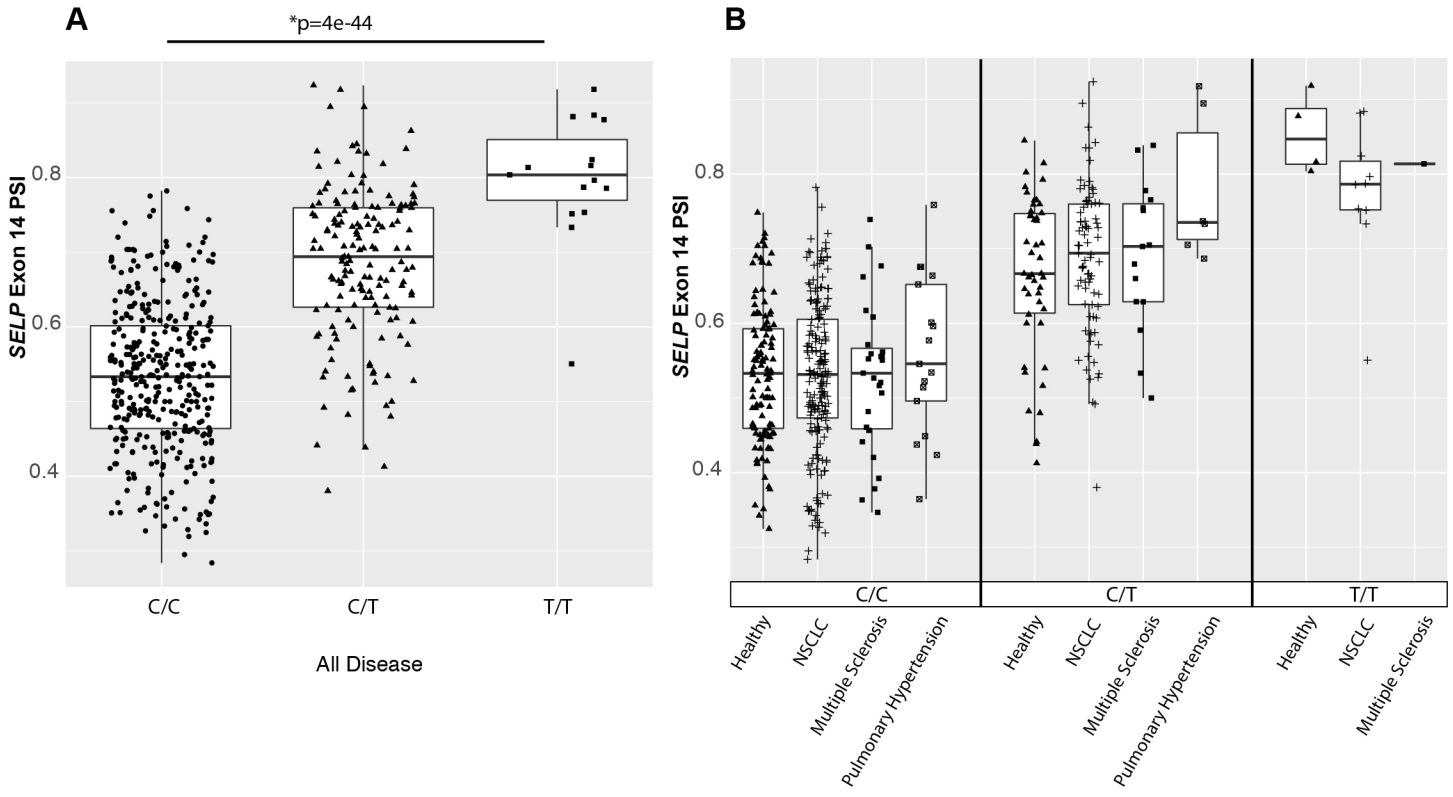
Online Figure V



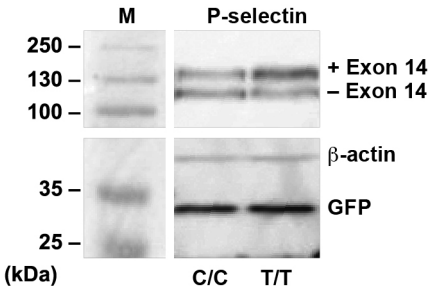
Online Figure VI

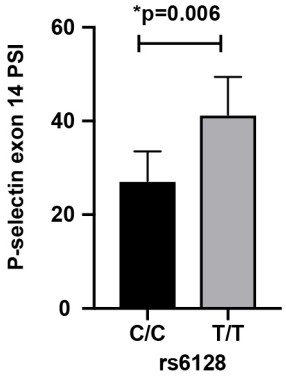


Online Figure VII

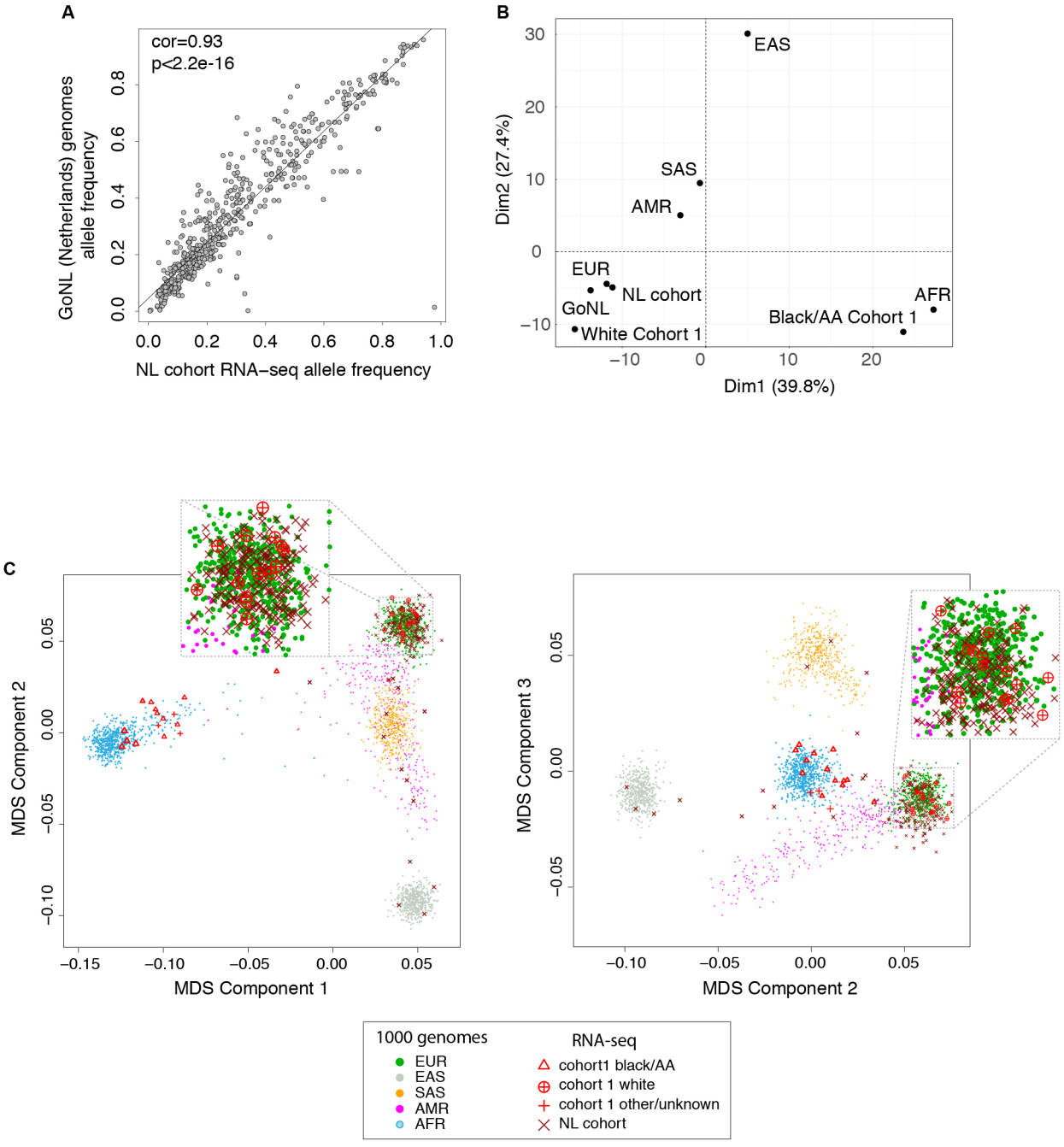


Online Figure VIII





Online Figure IX



Online Figure X