Cell Genomics, Volume *3*

Supplemental information

Blood cell traits' GWAS loci colocalization with variation in PU.1 genomic occupancy prioritizes causal noncoding regulatory variants

Raehoon Jeong and Martha L. Bulyk

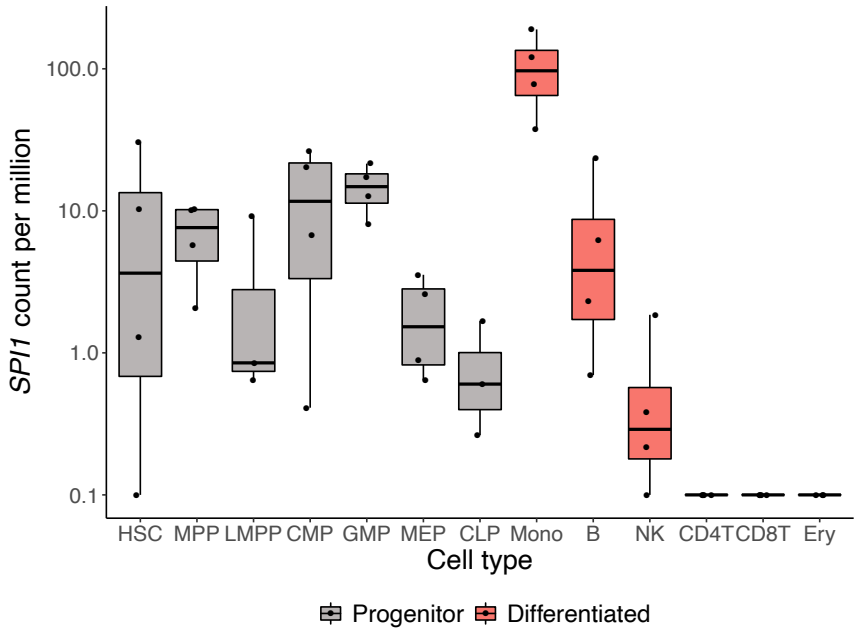

Figure S1. *SPI1* mRNA expression across blood cell types. Related to Figure 1.
Expression level is measured by bulk RNA-seq [S1]. The y-axis is log-scaled. Progenitor cell types (gray) and differentiated cell types (red) are colored accordingly. HSC: hematopoietic stem cell, MPP: multipotent progenitor, LMPP: lymphoid-primed multipotent progenitor, GMP: granulocyte-monocyte progenitor, CMP: common myeloid progenitor, MEP: megakaryocyte, CLP: common lymphoid progenitor, B: B cell, NK: natural killer cell, CD4T: $CD4^+$ T cell, CD8T: $CD8^+$ T cell, Ery: erythroid.

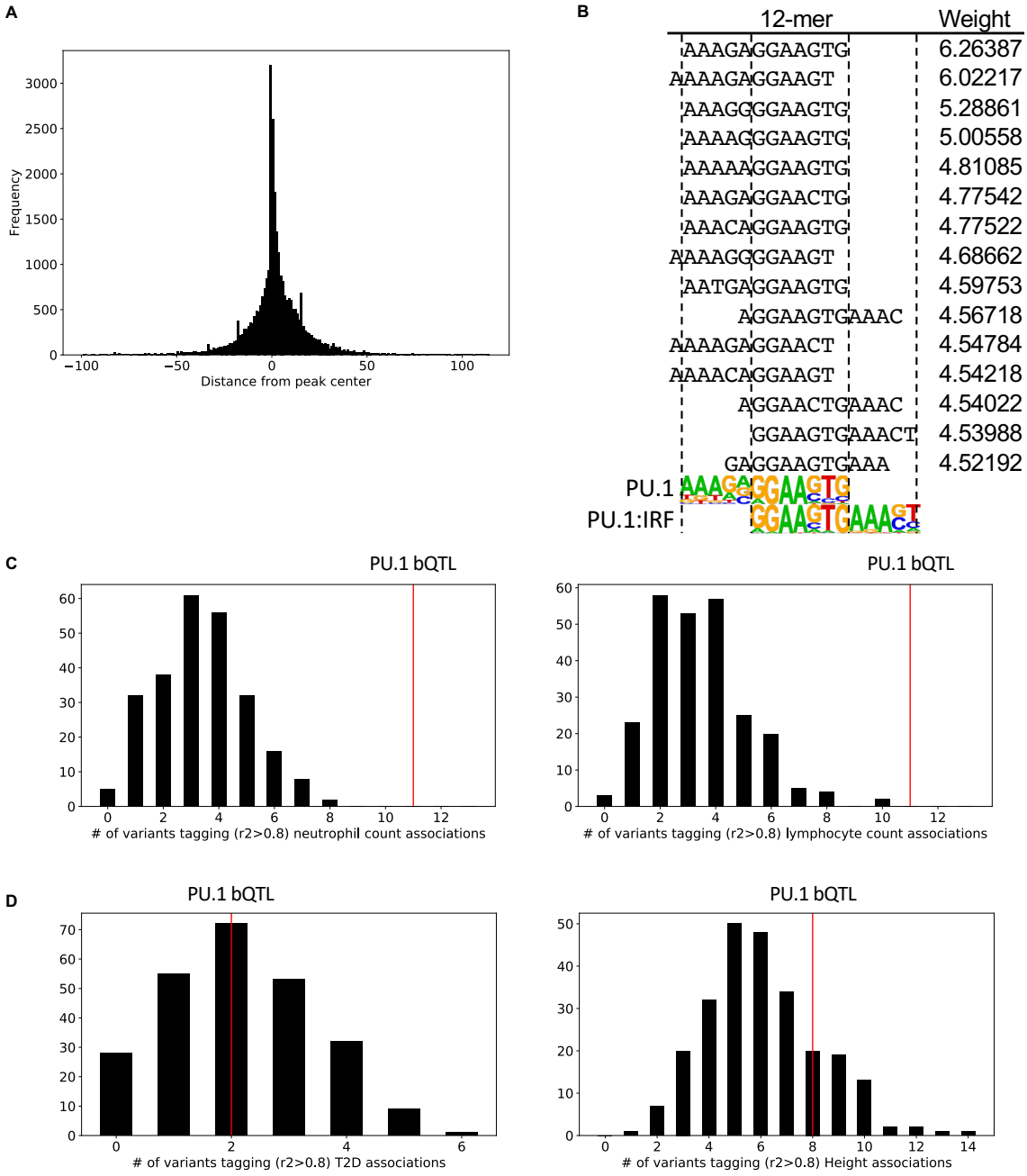

Figure S2. Properties of PU.1 binding sites and bQTLs. Related to Figures 1 and 2.
(A) Position of PU.1 motifs at PU.1 binding sites. The bp distance is measured from the center of a 200-bp PU.1 ChIP-seq peak.
(B) 12-mers with the highest (top 15) gkm-SVM weights aligned to PU.1 motif and PU.1:IRF composite motif.
(C and D) A subset of enrichment analysis results corresponding to Figure 2A. The histogram shows the number of variants tagging GWAS associations for each of 250 sets of null variants. The red lines indicate the number of PU.1 bQTL lead variants tagging GWAS associations. (C) Significant enrichment in PU.1 bQTL lead variants tagging (LD r^2 > 0.8) neutrophil and lymphocyte count associations. (D) Lack of enrichment in PU.1 bQTL lead variants tagging (LD r^2 > 0.8) type 2 diabetes (T2D) [S2] and height [S3] GWAS associations.

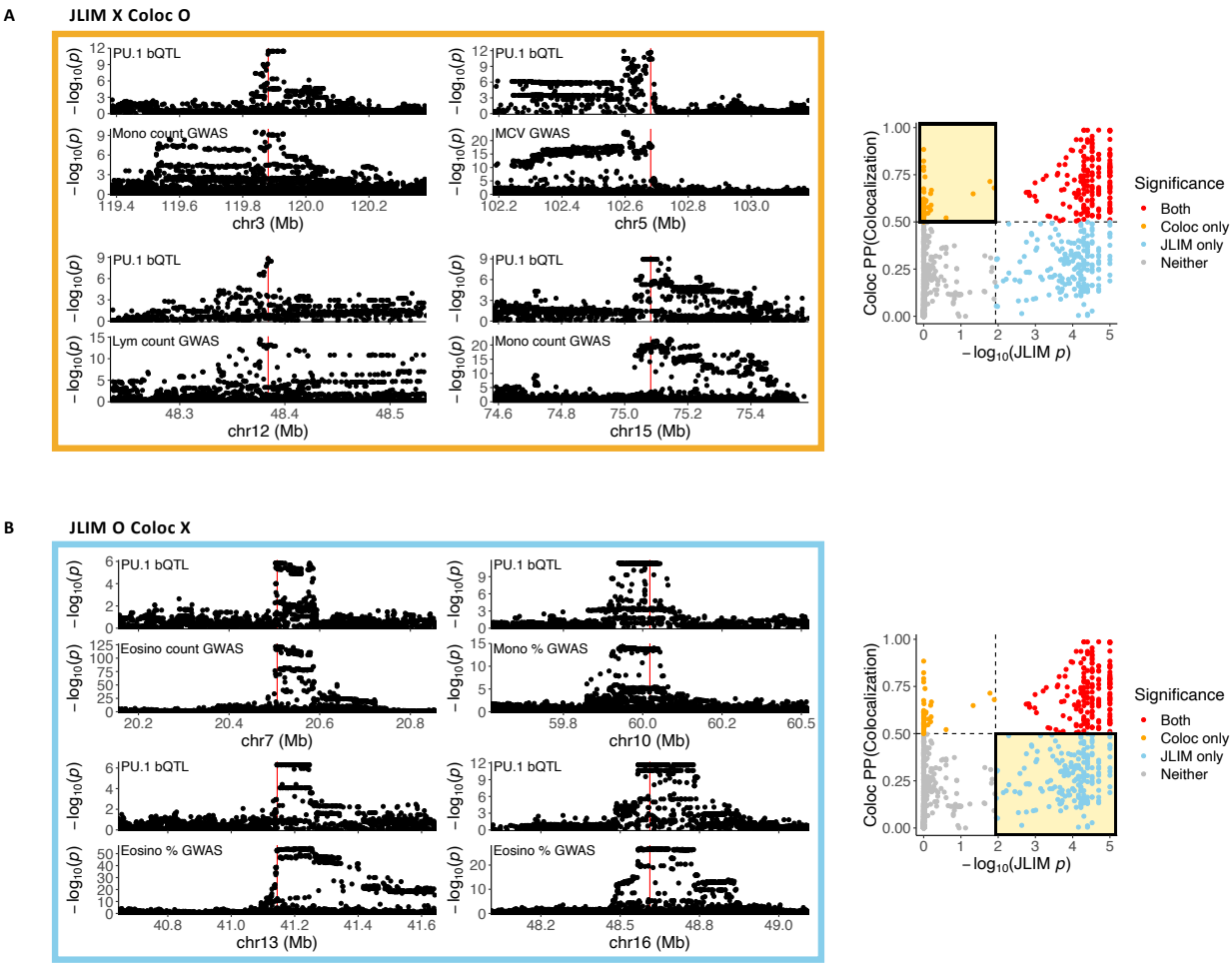

Figure S3. Examples of discordant colocalization results between JLIM and Coloc. Related to Figure 2.

(A and B) (Left) Example association plots for PU.1 bQTL and various blood cell traits. (Right) Colocalization results (same as Figure 2B) with yellow shading for the corresponding examples. (A) Loci with significant colocalization based on Coloc, but not JLIM. (B) Loci with significant colocalization based on JLIM, but not Coloc.

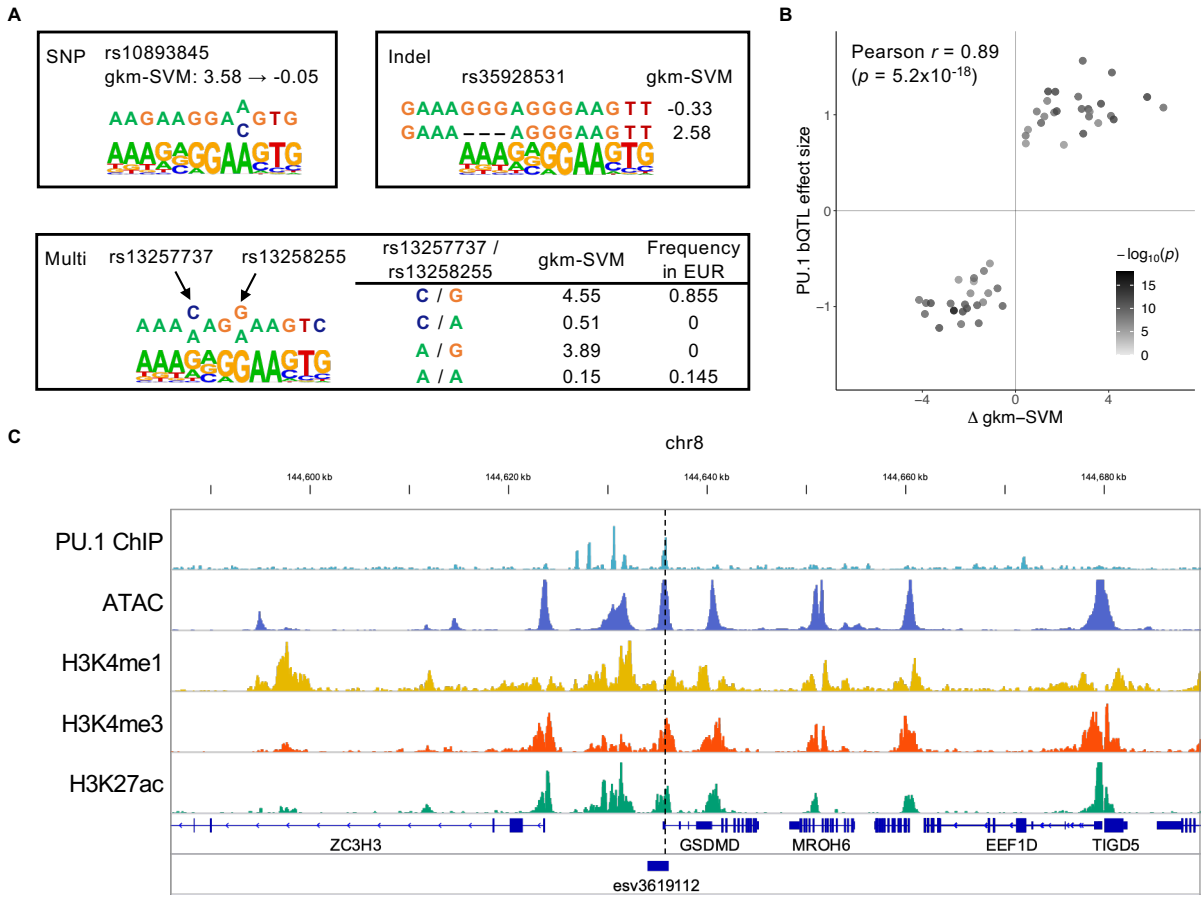

Figure S4. Examples of variants affecting PU.1 binding. Related to Figure 2.
(A) Examples of PU.1 motif-altering variants. Categorization of the variants correspond to Figure 2B. At the variant position, the top and bottom bases are reference and variant alleles, respectively. EUR: European ancestry population in the 1000 Genomes Project.
(B) Comparison of changes in motif score (Δ gkm-SVM) and estimated bQTL effect sizes of PU.1 motif-altering variants (SNPs and indels) at 49 colocalized loci.
(C) An example of a copy number variation (esv3619112) affecting a PU.1 binding site. The vertical dotted line indicates the location of the affected PU.1 binding site.

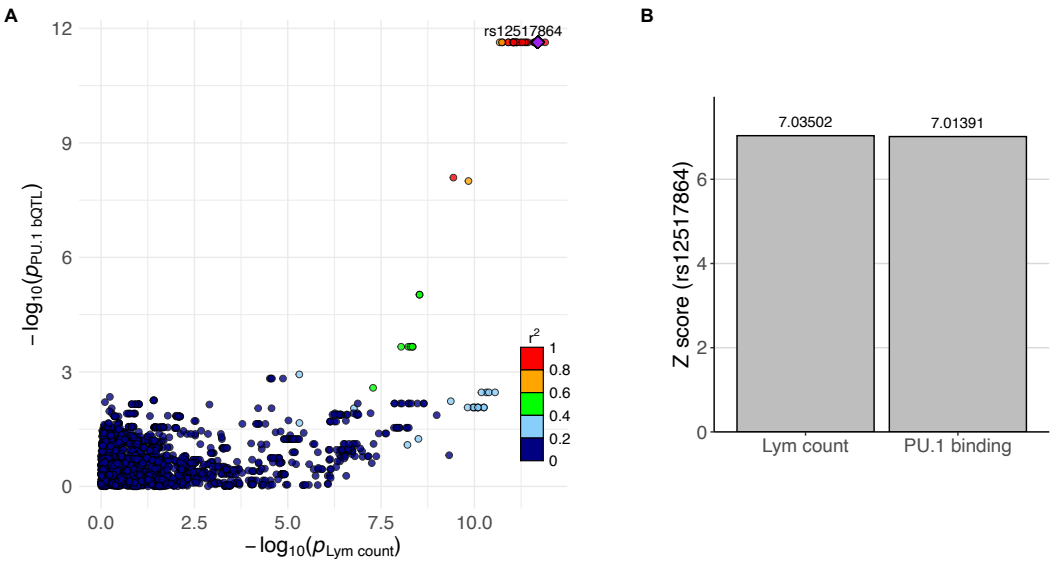

Figure S5. Colocalization of PU.1 bQTL and lymphocyte count association signals at *ZNF608* locus. Related to Figure 5.

(A) Merged association plot for PU.1 bQTL and lymphocyte count association signals. Points are colored by LD r^2 with respect to rs12517864, which is labeled with a purple diamond.

(B) Z scores of rs12517864 for lymphocyte count and PU.1 bQTL association.

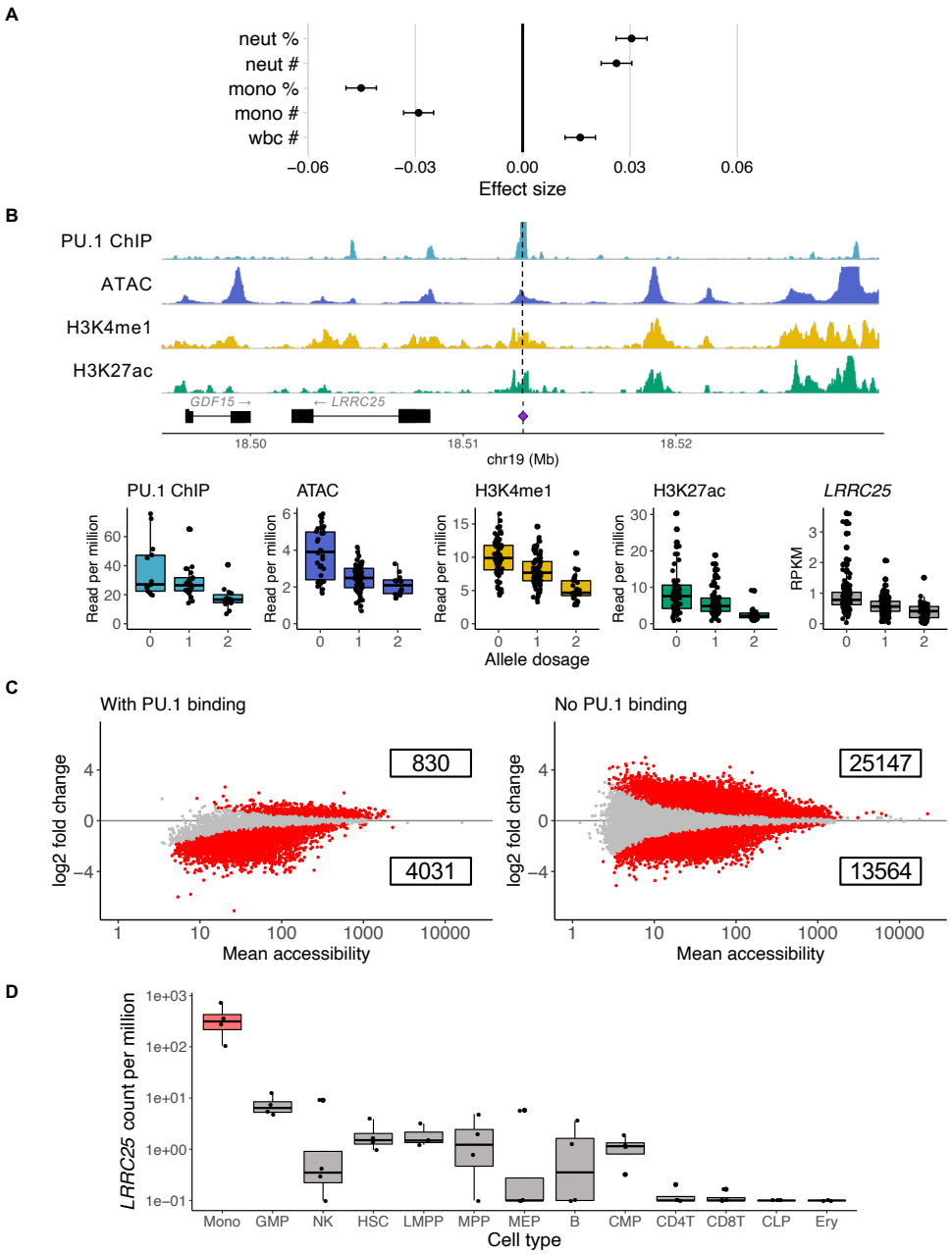

Figure S6. Effects of PU.1 motif-altering deletion rs5827412. Related to Figure 6.
(A) GWAS effect size estimates for rs5827412 on 5 blood cell traits. The error bars indicate 95% confidence interval. wbc #: white blood cell count, neut % & #: neutrophil percentage & count, mono % & #: monocyte percentage & count. Abbreviations of blood cell traits are further described in Table S3.
(B) Regulatory QTL effects of rs5827412. (top) Genome tracks show PU.1 ChIP-seq, ATAC-seq, and H3K4me1 and H3K27ac ChIP-seq data from LCLs, respectively. The dotted vertical line and the purple diamond mark the location of rs5827412. (bottom) 4 phenotype values in read per million for each genome track and reads per kilobase million for *LRRC25* expression levels. Allele dosage corresponds to that of the deletion allele. On top of the box plots, all the data points are shown.
(C) PU.1-dependent loss of chromatin accessibility. Log_2 fold change in chromatin accessibility in *SPI1*, the gene encoding PU.1, knock-out RS4;11 cell line for regions with PU.1 occupancy measured by ChIP-seq (left) and those without (right). Red points are accessible regions with significant gain or loss (p_{adj} < 0.05) of accessibility in knock-out mutants. Numbers in boxes represent the number of differentially accessible regions that show increase or decrease, respectively, in accessibility in *SPI1* knock-outs.
(D) *LRRC25* mRNA expression level across 13 blood cell types. Monocyte is colored red, and the rest are colored in gray. The y-axis is log-scaled. Cell types are abbreviated as in Figure S1.

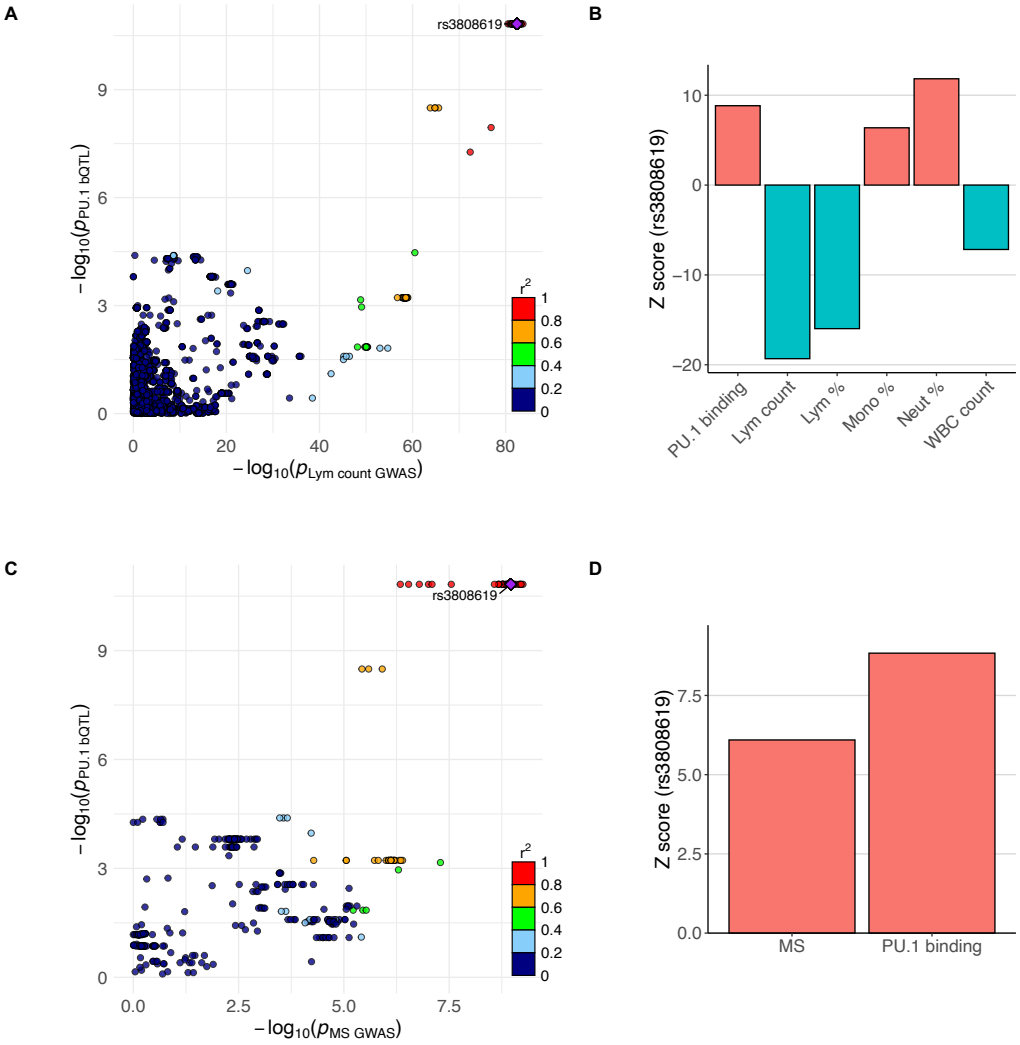

Figure S7. Colocalization of PU.1 bQTL and multiple sclerosis association signals at *ZC2HC1A* locus. Related to Figure 7.

(A) Merged association plot for PU.1 bQTL and lymphocyte count association signals. Points are colored by LD r^2 in the 1000 Genomes Project European population, with respect to rs3808619, which is labeled with a purple diamond.

(B) Z scores of rs3808619 for PU.1 bQTL and 5 blood cell traits association.

(C) Merged association plot for PU.1 bQTL and multiple sclerosis (MS) association signals [S4]. Points are labeled and colored as in (A).

(D) Z scores of rs3808619 for MS and PU.1 bQTL association.

Table S1. Summary of PU.1 ChIP-seq data. Related to Figure 1.
CEU: Utah residents (CEPH) with Northern and Western European ancestry

Sample	Ancestry	Sex	Number of mapped and filtered reads	Reference	EMBL-EBI ArrayExpress Accession
NA06985	CEU	Female	33654054	Waszak et al. [S5] (*n* = 45)	E-MTAB-3657
NA06986	CEU	Male	24382793		
NA06994	CEU	Male	25705372		
NA07037	CEU	Female	28128352		
NA07048	CEU	Male	22419180		
NA07051	CEU	Male	14588459		
NA07056	CEU	Female	27771303		
NA07357	CEU	Male	20551802		
NA10847	CEU	Female	25676020		
NA10851	CEU	Male	28145921		
NA11829	CEU	Male	25063350		
NA11830	CEU	Female	5188895		
NA11831	CEU	Male	21836117		
NA11832	CEU	Female	31885381		
NA11840	CEU	Female	24872788		
NA11881	CEU	Male	13183993		
NA11894	CEU	Female	32952045		
NA11918	CEU	Female	28771879		
NA11920	CEU	Female	56161783		
NA11931	CEU	Female	30173305		
NA11992	CEU	Male	22578216		
NA11994	CEU	Male	22222710		
NA12005	CEU	Male	29794021		
NA12043	CEU	Male	26253001		
NA12154	CEU	Male	23808118		
NA12156	CEU	Female	24555856		
NA12234	CEU	Female	27479585		
NA12249	CEU	Female	7879613		
NA12275	CEU	Female	8835732		
NA12282	CEU	Male	35200108		
NA12286	CEU	Male	39649385		
NA12287	CEU	Female	28933471		
NA12383	CEU	Female	41835685		
NA12489	CEU	Female	23405252		
NA12750	CEU	Male	31094939		
NA12760	CEU	Male	26128031		
NA12761	CEU	Female	13982870		
NA12762	CEU	Male	4512149		
NA12763	CEU	Female	14502734		
NA12776	CEU	Female	33261968		
NA12812	CEU	Male	21479602		
NA12813	CEU	Female	12761678		
NA12814	CEU	Male	12867291		
NA12815	CEU	Female	24083021		
NA12873	CEU	Female	26526926		
NA07346	CEU	Female	12677089	Kilpinen et al. [S6] (*n* = 4)	E-MTAB-1884
NA11993	CEU	Female	36953245		
NA12891	CEU	Male	25834224		
NA12892	CEU	Female	26772335		

Table S3. Description of blood cell traits. Related to Figure 2.
This table is adapted from Table S1 of Vuckovic et al. [S7].

Cell Type	Abbreviation	Blood cell trait	Description
Granulocyte	Baso count	Basophil count	Count of basophils per unit volume of blood
	Baso %	Basophil percentage of white cells	Percentage of white cells that are basophils
	Eosino count	Eosinophil count	Count of eosinophils per unit volume of blood
	Eosino %	Eosinophil percentage of white cells	Percentage of white cells that are eosinophils
	Neut count	Neutrophil count	Count of neutrophils per unit volume of blood
	Neut %	Neutrophil percentage of white cells	Percentage of white cells that are neutrophils
	WBC count	White blood cell count	Aggregate count of white cells per unit volume of blood
Monocyte	Mono count	Monocyte count	Count of monocytes per unit volume of blood
	Mono %	Monocyte percentage of white cells	Percentage of white cells that are monocytes
Lymphocyte	Lym count	Lymphocyte count	Aggregate count of lymphoid cells per unit volume of blood
	Lym %	Lymphocyte percentage of white cells	Percentage of white cells that are lymphocytes
Mature red cell	Hb conc	Hemoglobin concentration	Concentration of hemoglobin with respect to unit of volume of blood
	Ht %	Hematocrit	Volume fraction of blood occupied by red cells
	MCH	Mean corpuscular hemoglobin	Average mass of hemoglobin per red cell
	MCV	Mean corpuscular volume	Mean volume of red blood cells
	MSCV	Mean sphered corpuscular volume	Mean volume of sphered red cells
	RBC count	Red blood cell count	Count of red blood cells per unit volume of blood
	RBC dist width	Red cell distribution width	Coefficient of variation of red cell volume distribution
Immature red cell	HLSR count	High light scatter reticulocyte count	Count of high RNA content (immature) reticulocytes per unit volume of blood
	HLSR %	High light scatter reticulocyte percentage of red cells	Immature reticulocyte count as a percentage of red blood cell count
	Imm ret frac	Immature fraction of reticulocytes	Fraction of reticulocytes with high RNA content, as measured by light scatter
	MRV	Mean reticulocyte volume	Mean volume of reticulocyte cells
	Ret count	Reticulocyte count	Count of reticulocytes per unit volume of blood
	Ret %	Reticulocyte fraction of red cells	Percentage of red blood cells that are reticulocytes
Platelet	MPV	Mean platelet volume	Mean volume of platelets
	Plt count	Platelet count	Count of platelets per unit volume of blood
	Plt crit	Plateletcrit	Volume fraction of blood occupied by platelets
	Plt dist width	Platelet distribution width	The spread of the platelet volume distribution. Note that Sysmex and Coulter use different statistics to measure spread.

Note S1. Note about discordant results from JLIM and Coloc. Related to Figure 2.

Although we didn't aim to rigorously investigate the differences between JLIM [S9] and Coloc [S10], we looked through the examples where the two methods showed discordant results (Figures 2B and S3). First, we visually inspected the association plots for some of the loci, where only Coloc showed significant colocalization. Here, we could not clearly determine whether they are false positives by Coloc or false negatives by JLIM (Figure S3A). It is possible that the LD structure is different enough between the GWAS cohort and the PU.1 bQTL samples to cause JLIM to fail to reject the null hypothesis. On the other hand, loci that only JLIM showed colocalization often had a large set of variants in LD (Figure S3B). This trend is likely due to JLIM's model specification, where the JLIM statistics is higher if the lead variants for the two traits show high LD [S9], even if the LD block includes more variants. In sum, some of the loci with discordant results can be false negatives, but we decided to focus on loci with significant colocalization from both methods.

Note S2. Note about the two blood cell traits GWAS data. Related to Figure 6.

We utilized two blood cell traits GWAS data for this work. They are both statistics for the UK Biobank data with notable differences. Canela-Xandri and colleagues analyzed data for 452,264 White British individuals [S8], whereas Vuckovic and colleagues analyzed those from 408,112 individuals of British ancestry [S7]. They both applied linear mixed models. We incorporated Canela-Xandri et al. data for colocalization analyses because we expected greater statistical power due to larger sample sizes. However, Canela-Xandri and colleagues imputed the genotypes using the Haplotype Reference Consortium panel, which only includes SNPs and not indels, leading to SNP-only data. On the other hand, Vuckovic and colleagues imputed the genotypes using 1000 Genomes Project Phase 3 [S11] and UK10K [S12] panel, which includes SNPs and short indels. Therefore, we used Vuckovic et al. data for plotting Figure 6, where a short deletion alters the PU.1 motif, and for determining credible set sizes based on their fine-mapping results.

Note S3. Note about lymphocyte count association at *ZC2HC1A* locus. Related to Figure 7.

We pinpointed the PU.1 motif-altering SNP rs3808619 as the likely regulatory variant for colocalized PU.1 bQTL and lymphocyte count association at *ZC2HC1A* locus. Since the variant affects a PU.1 motif at its binding site at *ZC2HC1A* promoter, and the variant is significantly associated with increased *ZC2HC1A* expression, we hypothesized that the direct consequence of the variant is *ZC2HC1A* upregulation. As ZC2HC1A has no known function yet, we investigated this locus further. *IL7* gene is located downstream of *ZC2HC1A*, and a multi-ancestry blood cell trait GWAS study [S13] demonstrated that a South Asian ancestry-specific missense mutation (rs2014122253) in *IL7* that increased IL-7 protein secretion in a heterologous cellular system was associated with increased lymphocyte count. rs2014122253 is extremely rare in the European population, so it is not in LD with rs3808619. Interestingly, in eQTLGen data, rs3808619 was significantly, but relatively weakly, associated (p=9.45x10^{-14}) with lower *IL7* expression [S14] (this is compared to p=3.27x10^{-310} for *ZC2HC1A*). Although our analysis with GEUVADIS European LCL samples [S15] didn't show significant association (p > 0.1), eQTL Catalogue data[S16] showed that rs3808619 is significantly associated with lower *IL7* expression in multi-ancestry GEUVADIS LCL eQTL analysis [S15] (p = 2.85x10^{-9}) and TwinsUK LCL eQTL analysis [S17] (p=2.32x10^{-10}); only the latter analysis showed rs3808619 within the credible set of 41 variants. As Chen and colleagues showed that increased IL-7 secretion is associated with increased lymphocyte count [S13], rs3808619's association with lower *IL7* expression and lower lymphocyte count is plausible. How rs3808619 increases regulatory activity by increasing affinity to PU.1 binding leading to increased *ZC2HC1A* expression potentially lowers *IL7* expression is yet unresolved.

Supplemental Information Reference

[S1] Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet *48*, 1193–1203. https://doi.org/10.1038/ng.3646

[S2] Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet *50*, 1505–1513. https://doi.org/10.1038/s41588-018-0241-6

[S3] Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet *46*, 1173–1186. https://doi.org/10.1038/ng.3097

[S4] International Multiple Sclerosis Genetics Consortium (IMSGC), Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., et al. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet *45*, 1353–1360. https://doi.org/10.1038/ng.2770

[S5] Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell *162*, 1039–1050. https://doi.org/10.1016/j.cell.2015.08.001

[S6] Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., et al. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science *342*, 744–747. https://doi.org/10.1126/science.1242463

[S7] Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.-H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell *182*, 1214-1231.e11. https://doi.org/10.1016/j.cell.2020.08.008

[S8] Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. Nat Genet *50*, 1593–1599. https://doi.org/10.1038/s41588-018-0248-z

[S9] Chun, S., Casparino, A., Patsopoulos, N.A., Croteau-Chonka, D.C., Raby, B.A., De Jager, P.L., Sunyaev, S.R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet *49*, 600–605. https://doi.org/10.1038/ng.3795

[S10] Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383

[S11] Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393

[S12] Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–89. https://doi.org/10.1038/nature14962

[S13] Chen, M., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell

Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198-1213.e14. https://doi.org/10.1016/j.cell.2020.06.045

[S14] Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet *53*. https://doi.org/10.1038/s41588-021-00913-z

[S15] Lappalainen, T., Sammeth, M., Friedländer, M.R., 'T Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511. https://doi.org/10.1038/nature12531

[S16] Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M.P., Kuzmin, I., Trevanion, S.J., et al. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet *53*, 1290–1299. https://doi.org/10.1038/s41588-021-00924-w

[S17] Buil, A., Brown, A.A., Lappalainen, T., Viñuela, A., Davies, M.N., Zheng, H.-F., Richards, J.B., Glass, D., Small, K.S., Durbin, R., et al. (2015). Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat Genet *47*, 88–91. https://doi.org/10.1038/ng.3162