# Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes

Thom CS[1,2,3,4], Voight BF[2,3,4]*

[1]Division of Neonatology, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[2]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Institute of Translational Medicine and Therapeutics, University of Pennsylvania, PA, USA

*Corresponding Author:
Benjamin F Voight
University of Pennsylvania Perelman School of Medicine
3400 Civic Center Blvd
10-126 Smilow Center for Translational Research
Philadelphia, PA 19104
bvoight@pennmedicine.upenn.edu
215-746-8083

## Abstract

*Background*
Genetic associations link hematopoietic traits and disease end-points, but most causal variants and genes underlying these relationships are unknown. Here, we used genetic colocalization to nominate loci and genes related to shared genetic signal for hematopoietic, cardiovascular, autoimmune, neuropsychiatric and cancer phenotypes.

*Results*
Our findings recapitulate developmental hematopoietic lineage relationships, identify loci associating traits with causal genetic relationships, and reveal novel associations. Out of 2706 loci with genome-wide significant signal for at least 1 blood trait, we identified 1779 unique sites (66%) with shared genetic signal for 2+ hematologic traits at a false discovery rate <5%. We could assign some sites to specific developmental cell types during hematopoiesis based on affected traits, including those likely to impact hematopoietic progenitor cells and/or megakaryocyte-erythroid progenitor cells. Through an expanded analysis of 70 human traits, we define 2+ colocalizing traits at 2007 loci from an analysis of 9852 sites (20%) containing genome-wide significant signal for at least 1 GWAS trait. In addition to variants and genes underlying shared genetic signal between blood traits and disease phenotypes that had been previously related through mendelian randomization studies, we define loci and related genes underlying shared signal between eosinophil count and eczema. We also identified colocalizing signals in a number of clinically relevant coding mutations, including in sites linking *PTPN22* with Crohns disease, *NIPA* with coronary artery disease and platelet trait variation, and the hemochromatosis gene *HFE* with altered lipid levels. Finally, we anticipate potential off-target effects on blood traits related novel therapeutic targets, including *TRAIL*.

*Conclusions*
Our findings provide a road map for gene validation experiments and novel therapeutics related to hematopoietic development, and offer a rationale for pleiotropic interactions between hematopoietic loci and disease end-points.

## Keywords

Hematopoiesis, Genetics, Colocalization

## Background

Identifying causal loci and genes from human genetic data is integral to elucidating novel disease insights and therapies approaches. Quantitative hematopoietic traits are relatively well studied, although relatively few causal variants and genes have been elucidated [1, 2]. Mendelian randomization studies have established causal relationships between hematopoietic traits and cardiovascular, autoimmune and neuropsychiatric disease [2]. Still, causal genes and loci remain elusive.

Genetic colocalization analysis permits identification of shared regulatory loci, with advances extending the scope of potential studies from two to >10 traits undergoing simultaneous analysis [3–5]. Recently, a colocalization algorithm was used to identify known and novel loci related to cardiovascular traits [5]. We reasoned that a similar analytical pipeline could help explain variants and genes underlying hematopoietic and other disease phenotypes. In this way, aggregated summary statistics might be used to specifically target loci with pleiotropic effects on multiple traits, enacted through one or a handful of genes.

Developmental cell types during hematopoiesis, the process that gives rise to all blood lineages, are relatively well mapped. We hypothesized that shared genetic signal impacting traits from multiple blood lineages might nominate genomic loci related to the stem and progenitor cells that spawned those types of blood cells. This approach is orthogonal to prior data that analyzed patterns in accessible chromatin to define genomic locations affecting multiple blood lineages [1]. For example, a shared single nucleotide polymorphism (SNP) related to quantitative variation in platelet, red blood cell (RBC), and white blood cell (WBC) counts might indicate a site or mechanism that is active in hematopoietic stem and progenitor cells (HSCs). SNPs related to platelet and RBC counts, but not WBC count, might reveal loci and related genes for megakaryocyte-erythroid progenitor (MEP) cells. We hypothesized that the directionality of such relationships might help elucidate lineage decisions during hematopoiesis, and help target loci and genes related to developmental hematopoiesis.

Blood traits are related to a number of human disease phenotypes [2]. Blood cells can cause disease (e.g. autoimmune traits) or be affected by therapies (e.g. anemia secondary to chemotherapy). For this reason, understanding pleiotropic associations between blood and other traits could reveal translationally relevant trait relationships or help predict off-target effects of gene-modifying therapies.

Here, we used genetic colocalization to define sites wherein 2 or more human traits shared genetic signal at genome-wide significant loci. We initially examined blood traits, and later expanded our analysis to include a total of 70 blood, autoimmune, cardiovascular, cancer, and neuropsychiatric traits. We then looked for quantitative trait loci impacting gene expression (eQTL) or exon splicing variation (sQTL) at or near sites of genetic colocalization. Our results identify sites that affect specific cell types during hematopoietic development, and reveal genetic variants underlying trait relationships between blood parameters and disease end-points.

## Results

*Genetic colocalization recapitulates hematopoietic lineage relationships*

Our first aim was to validate whether colocalization could effectively capture known trait relationships and genetic correlations between hematopoietic lineages [1]. We performed colocalization analysis [5] using summary statistics related to 34 quantitative hematopoietic traits for 2706 genome-wide significant loci [2], revealing a total of 1779 sites wherein 2 or more traits colocalized with a posterior probability > 0.7 (**Additional file 1: Table S1**). In simulations, these criteria identified the causal variant, or a variant in high LD with the causal variant, with a false discovery rate < 5% [5]. Colocalization sites specified 3.6±2.3 traits (mean±SD), with 22% of the loci (259 in total) representing highly pleiotropic sites where 6 or more traits colocalized (**Fig. 1a**). Hence, a substantial proportion of interrogated loci (66%) impacted multiple hematopoietic traits.

To investigate trait relationships, we constructed a heat map to depict the percentage colocalization between trait pairs (**Fig. 1b**). Hierarchical clustering of colocalization results reflected blood lineage relationships, with platelet, erythroid, and white blood cell traits generally clustering as expected.

We then asked whether our colocalization findings mirrored genetic correlation between hematopoietic traits [6]. Indeed, more closely related traits colocalized more often (**Fig. 1c**, $r^2$=0.91 by quadratic regression with least squares fit). Directly correlated (e.g., 'neutrophil count' and 'neutrophil + eosinophil count'; 'granulocyte count' and 'myeloid white blood cell count'), and inversely correlated trait pairs (e.g., 'eosinophil percent of granulocytes' and 'neutrophil percent of granulocytes'; 'lymphocyte percent' and 'neutrophil percent'), essentially always colocalized. Several trait pairs fell outside the 95% prediction interval (e.g. 'mean platelet volume' and 'platelet count'; 'mean corpuscular hemoglobin concentration' and 'mean red cell volume') (**Fig. 1c**). Lineage-critical loci or genes might be expected to have more significant influence on these lineage-restricted trait pairs than would be captured by genetic correlation measure.

In sum, these results validated the notion that colocalization analysis results would mirror genetic correlation, and reflect known relationships among hematopoietic lineages and traits. Interestingly, trait pairs without genetic correlation frequently had some degree of colocalization (**Fig. 1c**, y-intercept = 0.077±0.123). This likely reflects horizontal pleiotropy, in which a given locus and related gene(s) impact traits that are not biologically related. In the context of hematopoietic development, our derived estimate of chance colocalization between unrelated traits is therefore ~8%.

*A genetic colocalization strategy to identify loci related to hematopoietic development*

We then leveraged our colocalization results to identify quantitative trait loci (QTLs) related to specific hematopoietic lineages and cell types. For example, loci where white blood cell (WBC), red blood cell (RBC), and platelet counts colocalize might indicate developmental perturbation in hematopoietic stem and progenitor cells (HSCs). We therefore looked for sites of colocalization between these quantitative blood traits, and identified overlapping genome-wide significant QTLs. Indeed, QTLs related to these loci pointed to known HSC regulatory genes *SH2B3* [7, 8], *ATM* [9], and *HBS1L-MYB* [10] (**Additional file 1: Table S2**).

We also parsed loci identified by colocalization to specifically affect platelet or red cell traits, with the hypothesis that these loci would relate to terminally differentiated blood cell biology. There were 439 sites nominated by colocalization analysis specifically for red cell traits (RBC, HCT, MCV, MCH, MCHC, RDW) but *not* platelet traits or WBC count. These sites, or highly linked loci, influenced expression of 614 genes (123 genes in whole blood, **Additional file 1: Table S3**). Gene ontology (GO) analysis [11] of these gene set revealed significant enrichment of genes related to cellular metabolic processes (**Additional file 1: Table S4**). A similar analysis of platelet trait-restricted sites (PLT, PCT, MPV, PDW), including highly linked loci, identified 270 sites impacting expression of 399 genes (77 genes in whole blood, **Additional file 1: Table S5**). Pathway analysis of these genes revealed enrichment of apoptotic cell clearance and metabolic processes (**Additional file 1: Table S6**). Complement-mediated apoptotic cell clearance mechanisms are indeed important for regulating platelet count [12].

To our surprise, pathways analyses of red cell and platelet lineage-restricted colocalization QTLs were not enriched processes ascribed to hematopoiesis, erythropoiesis, or megakaryopoiesis. This suggests that genes and processes linked to terminal red cell and platelet traits are largely impacted by cellular function and reactivity, rather than developmental perturbations. With notable exceptions whereby causal loci do impact hematopoietic development (e.g., [7, 13–15]), our findings suggest that this may not be true for the majority of identified genes and factors. In fact, our results indicate that cell-extrinsic properties (e.g. apoptotic cell clearance mechanisms) frequently impact quantitative hematopoietic traits. In sum, our findings reveal a multitude of known variants and genes, as well as novel QTL and related genes that warrant further study.

*Illuminating hematopoietic contributions and associations with disease phenotypes*

We then applied an extended colocalization analysis to summary statistics for 70 total hematopoietic, cardiovascular, autoimmune, cancer, and neuropsychiatric traits (**Additional file 1: Table S7**). Following allele harmonization, colocalization analysis using 9852 genome-wide significant loci from the NHGRI database [16] and blood traits [2] revealed a total of 2007 sites (20%) wherein 2 or more traits colocalized with a posterior probability > 0.7 (**Additional file 1: Table S8**). The average number of traits that colocalized at a given site was 3.3±2.4 (mean±SD), with 84 loci identified as a 'very pleiotropic' colocalization site for ≥9 traits (**Fig. 2a**). Known trait relationships were recapitulated among these colocalization sites (e.g. bipolar disorder and schizophrenia; **Fig. 2b-d**). These results again reflected genetic correlation between traits, estimating a small degree of pleiotropy (~4%) absent genetic correlation (**Fig. 2d**, $r^2$=0.86, y-intercept= 0.041±0.120).

Mendelian randomization analyses have established causal relationships between blood traits and some disease phenotypes [2]. Despite holding significant therapeutic potential, most causal loci and genes underlying these associations are unknown. Our results reveal putatively causal loci, related genes and molecular pathways related to these trait pairs (**Additional file 1: Table S9-S16**). For example, whole blood QTLs related to genes known to affect asthma pathogenesis or severity (e.g. *IL18R* [17–19]*, ZFP57* [20]*, BTN3A2* [21]*, NDFIP1* [22]*, SMAD3* [23]*, CLEC16A* [24]*,* and *TSLP* [25]) were associated with colocalization sites for asthma and neutrophil, eosinophil, monocyte and/or lymphocyte traits (**Additional file 1: Table S9-S12**). Similarly, QTLs for genes linked to coronary artery disease risk (e.g. *XRCC3* [26]*, SREBF1* [27]*, GIT1* [28, 29]*, SKIV2L* [29]*, MAP3K11/MLK3* [30]) were associated with colocalization sites linking coronary artery disease with lymphocyte and/or reticulocyte counts (**Additional file 1: Table S13-S15**). Other identified genes associated with colocalization loci represent novel

findings that could enhance understanding of the pathophysiology and/or treatment of these diseases, although functional validation remains necessary.

Our findings also revealed novel trait associations. Although eosinophil percentage and eczema did not reach statistical significance by Mendelian Randomization, these traits are clinically related [31] and colocalized at 10 sites (**Additional file 1: Table S16**). These results were localized near genes that regulate eosinophil biology (*ETS1* [32, 33] and *ID2* [33]) and autoimmune disease (*KIAA1109* [34–36] and *TAGAP* [37]), and indicate potential regulation of unexpected genes that warrant validation (*SNX32, ZNF652, KLC2*).

*Colocalization at coding variation sites identifies clinically relevant trait associations*

Next, we reasoned that colocalizing sites could help explain unexpected or pleiotropic effects of gene perturbations. Here, we focused on missense variation in coding regions to establish direct locus-gene relationships. This approach identified clinically relevant cross-trait associations.

Variation in rs2476601 causes a missense mutation in *PTPN22* (Cys1858Thr). This site has been linked to autoimmunity and Crohns disease phenotypes, but not ulcerative colitis [38]. Immune response dysregulation, including WBC biology, contributes to the Crohns phenotype [38]. We identified shared genetic signal for increased Crohns disease risk and decreased WBC count, but not ulcerative colitis, at this location (**Additional file 1: Table S17**). This finding supports a specific clinical association with Crohns for the *PTPN22* Cys1858Thr mutation.

Mean platelet volume (MPV) variation has previously been linked to altered risk of coronary artery disease, but understanding of genes underlying this association is lacking [2]. We identified colocalizing signals for increased coronary artery disease risk and increased MPV in a missense coding mutation for *ZC3HC1/NIPA* (**Additional file 1: Table S13**). This variant causes an Arg>His missense change in several *NIPA* isoforms. *NIPA* impacts heart disease risk and cell cycle regulation [39]. Further studies are needed to understand how this gene might coordinately impact platelet biology and coronary artery disease risk, as well as other traits linked to this locus.

Altered lipid and cholesterol levels have been clinically observed in patients with hereditary hemochromatosis due to mutations in *Human Factors Engineering* (*HFE*) [40]. Patients with hemochromatosis have lower cholesterol levels than normal, although an open question is whether this observation is due to manifestations of disease or *HFE* deficiency itself. Our data show that individuals heterozygous for the Cys282Tyr allele have lower reticulocyte count and higher total cholesterol and low density lipoprotein levels (**Additional file 1: Table S18**). This suggests that *HFE* haploinsufficiency increases cholesterol and lipid levels, and that decreased cholesterol in patients with hemochromatosis occurs secondary to myriad tissue manifestations of clinically significant hemochromatosis or iron overload [41].

Finally, we hypothesized that our analysis might also help predict off-target effects of novel therapeutic agents. For example, tumor necrosis factor (TNF)-related apoptosis inducing ligand (*TRAIL*) is a promising novel chemotherapeutic target [42]. A mutation in the *TRAIL* 3' UTR was associated with decreased total cholesterol and triglyceride levels, as well as decreased platelet counts (**Additional file 1: Table S19**). It will be interesting to see whether these traits are affected in upcoming clinical trials targeting *TRAIL*.

## Discussion

Genetic colocalization approaches have proven a powerful tool in revealing pleiotropic effects of certain loci on multiple traits [3, 4]. Here, we have adapted the colocalization methodology to reveal sites and genes related to specific cell stages in hematopoietic development, and identify relevant trait relationships between blood traits and human disease end-points. We present what we believe to be a minimal estimate of these associations, given our conservative threshold for colocalization (PP>0.7). This threshold revealed high-confidence targets, although future gene discovery studies might instead use a more relaxed threshold (e.g., PP>0.5) to enable a more encompassing set of loci.

GWAS have linked thousands of genomic sites with blood trait variation [2]. The biology related to each site could relate to developmental hematopoiesis, as has been shown for *CCND3* [13], *CCNA2* [14], *SH2B3* [7], and *RBM38* [15]. Alternatively, biology related to GWAS sites might impact terminally differentiated cell reactivity or turnover. For example, altered platelet reactivity can affect quantitative platelet traits [43, 44]. Cellular validation experiments might be streamlined if one could better parse relevant sites, genes and developmental stages based on GWAS information. Gene targets presented herein represent one approach to such a computational pipeline, and are orthogonal to previously published findings based on accessible chromatin patterns during hematopoietic development [1]. Future studies combining these computational modalities might be useful for those interested in evaluating specific genes or loci in blood progenitor biology.

Our expanded analysis of 70 human traits recapitulated known trait relationships between blood traits and human disease phenotypes, and identified sites with potential translational relevance. Understanding how missense coding mutations impact phenotypes offers the most direct relationship between genes and traits. An adaptation of our colocalization strategy might be employed to predict off-target effects of gene modulation, help understand cellular basis of disease, or investigate unexpected cellular developmental relationships (e.g., sites related to multiple mesoderm-derived tissues might triangulate to early mesodermal biology). We anticipate an expanded array of such targets could be revealed with larger, trans-ethnic GWAS.

## Conclusion

In an extensive genetic colocalization analysis, we have identified loci, genes and related pathways related to hematopoietic development. Further, our colocalization results identified loci relating 70 hematopoietic, cardiovascular, autoimmune, neuropsychiatric and cancer phenotypes. This repository of associations will be useful for mechanistic studies aimed at understanding biological links between phenotypes, for developing novel therapeutic strategies, and for predicting off-target effects of small molecule and gene therapies.

## Methods

**SNP and study selection.** Human genome version hg19 was used for all analyses. GWAS summary statistics were obtained from publicly available repositories (**Additional file 1: Table S7**). We narrowed analysis to just those GWAS summary statistics for European populations with >1 x$10^6$ sites (i.e., those that were genome-wide). Analyzed SNPs were identified as genome-wide significant in the largest hematopoietic trait GWAS to date [2] or from a repository of genome-wide significant SNPs from a compilation of GWAS from the NHGRI [16] (downloaded January 2019). In addition, we analyzed quantitative trait locus data from GTEx V7 [45].

**Colocalization analyses.** We used HyPrColoc software to conduct colocalization experiments [5]. This software requires effect (e.g., beta or odds ratio) and standard error values for each analyzed SNP. We chose to analyze based on chromosome and position, given that multiple rsIDs might overlap at a given locus and be inconsistent between different GWAS. Although this removed duplicate rsIDs and may have caused some bias, we reasoned that this would be a minority of sites. This strategy optimized the number of individual positions that we were able to incorporate into our input dataset for colocalization analysis. We specifically looked at 500 kb regions (250 kb on either side of each site), in line with prior colocalization literature [5].

As the SNPs considered as input data varied between analyses, we presented separate results from analysis of 34 hematologic traits, and a composite 70 traits. GWAS summary statistics were harmonized prior to analyses (https://github.com/hakyimlab/summary-gwas-imputation/wiki). There were 29,239,447 genomic sites analyzed for colocalization among the hematologic traits. A total of 1,544,623 harmonized sites were analyzed from GWAS summary statistics for the 70 traits. The decreased number of sites included in this latter analysis resulted in decreased power to detect associations. This was reflected in the maximum number of traits colocalized in **Fig. 1a** was 25 (out of 34 traits) versus 23 traits in **Fig. 2a** (out of 70 traits).

After colocalization analysis, we narrowed our focus on only those loci with posterior probability for colocalization (PPFC) >0.7 based on empiric simulations results from the creators of this algorithm showing that this conservatively gave a false discovery rate <5% [5]. We note that a more relaxed PPFC (e.g. >0.5) yields substantially more loci. A less conservative threshold could in this way be used as a hypothesis-generating experiment for cellular follow up studies.

**Coding variant identification**. We used the Ensembl Variant Effect Predictor (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP) to identify coding variants and related gene consequences.

**Linkage disequilibrium and quantitative trait locus (QTL) analyses.** We wanted to assess comprehensively the potential gene expression or splicing changes related to colocalization sites. Thus, we analyzed each colocalization site together with all sites in high linkage disequilibrium (EUR $r^2$>0.90, PLINK version 1.9).

We used closestBed (https://bedtools.readthedocs.io) to identify the nearest gene to each SNP. Genes and positions were defined by BioMart (http://www.biomart.org/).

For each group of linked SNPs around a colocalization locus, we identified all eQTLs (GTEx V7 [45]), as well as all sQTLs as defined by two different algorithms (sQTLseekeR [46], Altrans [47]). In the manuscript and Additional file 1, the quantity of QTL SNPs and pathway analyses reflect a composite of all genes impacted by a given locus, or by highly linked SNPs. Note that a

given colocalization site might be linked with several SNPs, and that these SNPs might be in proximity and/or impact different genes. Affected genes shown are those with a unique Ensembl gene identifier (ENSG). In some cases, gene names may differ between Nearest Gene, eQTL and sQTL columns given that the underlying analyses were derived from different catalogues.

**Gene ontology analysis.** We submitted QTLs associated with specific traits for biological pathway assessment using the Gene Ontology (GO) resource (http://geneontology.org/). Statistical significance of GO Biological Process enrichments were assessed using binomial tests and Bonferroni correction for multiple testing. Presented data were those pathways with $p<0.05$.

**Empirical distribution for expected colocalization counts.** We used LDSC to estimate genetic correlation between traits (v1.0.1) [6]. Presented genetic correlation data reflect rg values obtained from LDSC analysis.

**Data presentation**
Data were created and presented using R, LocusZoom (http://locuszoom.org), Adobe Illustrator CS6 and GraphPad Prism 8.

**Statistics**
Statistical analyses were conducted using R and GraphPad Prism 8.

## Declarations

*Ethics approval and consent to participate*
Consent was obtained as part of original genetic studies. Hence, no further consent was necessary to use summary statistics as part of our study.

*Consent for publication*
Not applicable

*Availability of data and materials*
Code and scripts used to generate and analyze data are available on GitHub (https://github.com/thomchr/2019.Coloc).

*Competing interests*
The authors declare that they have no relevant conflicts of interest.

*Funding*
This work was supported through R01DK101478 (BFV), a Linda Pechenik Montague Investigator Award (BFV), T32HD043021 (CST), a Children's Hospital of Philadelphia Neonatal and Perinatal Medicine Fellow's Research Award (CST), an American Academy of Pediatrics Marshall Klaus Neonatal-Perinatal Research Award (CST) and a Children's Hospital of Philadelphia Foerderer Award (CST).

*Authors' contributions*
CST and BFV conceived of the project, generated, analyzed and interpreted data, and wrote the paper.

*Acknowledgements*
Not applicable

## References

1. Ulirsch JC, Lareau CA, Bao EL, Ludwig LS, Guo MH, Benner C, et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat Genet. 2019;51:683–93. doi:10.1038/s41588-019-0362-6.
2. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167:1415-1429.e19. doi:10.1016/j.cell.2016.10.042.
3. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014;10:e1004383. doi:10.1371/journal.pgen.1004383.
4. Giambartolomei C, Zhenli Liu J, Zhang W, Hauberg M, Shi H, Boocock J, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics. 2018;34:2538–45. doi:10.1093/bioinformatics/bty147.
5. Foley CN, Staley JR, Breen PG, Sun BB, Kirk PD, Burgess S, et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. bioRxiv. 2019;:592238. doi:10.1101/592238.
6. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236–41. doi:10.1038/ng.3406.
7. Giani FC, Fiorini C, Wakabayashi A, Ludwig LS, Salem RM, Jobaliya CD, et al. Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. Cell Stem Cell. 2016;18:73–8. doi:10.1016/j.stem.2015.09.015.
8. Balcerek J, Jiang J, Li Y, Jiang Q, Holdreith N, Singh B, et al. Lnk/Sh2b3 deficiency restores hematopoietic stem cell function and genome integrity in Fancd2 deficient Fanconi anemia. Nat Commun. 2018;9:3915. doi:10.1038/s41467-018-06380-1.
9. Ito K, Hirao A, Arai F, Matsuoka S, Takubo K, Hamaguchi I, et al. Regulation of oxidative stress by ATM is required for self-renewal of haematopoietic stem cells. Nature. 2004;431:997–1002. doi:10.1038/nature02989.
10. Lieu YK, Reddy EP. Conditional c-myb knockout in adult hematopoietic stem cells leads to loss of self-renewal due to impaired proliferation and accelerated differentiation. Proc Natl Acad Sci U S A. 2009;106:21689–94. doi:10.1073/pnas.0907623106.
11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for The Unification of Biology. Nat Genet. 2000;25:25–9. doi:10.1038/75556.
12. Quach ME, Chen W, Li R. Mechanisms of platelet clearance and translation to improve platelet storage. Blood. 2018;131:1512–21. doi:10.1182/blood-2017-08-743229.
13. Sankaran VG, Ludwig LS, Sicinska E, Xu J, Bauer DE, Eng JC, et al. Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. Genes Dev. 2012;26:2075–87.
14. Ludwig LS, Cho H, Wakabayashi A, Eng JC, Ulirsch JC, Fleming MD, et al. Genome-wide association study follow-up identifies cyclin A2 as a regulator of the transition through cytokinesis during terminal erythropoiesis. Am J Hematol. 2015;90:386–91.
15. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell. 2016;165:1530–45. doi:10.1016/j.cell.2016.04.048.
16. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12. doi:10.1093/nar/gky1120.
17. Zhang Y, Moffatt MF, Cookson WOC. Genetic and genomic approaches to asthma: new insights for the origins. Curr Opin Pulm Med. 2012;18:6–13. doi:10.1097/MCP.0b013e32834dc532.

18. Reijmerink NE, Postma DS, Bruinenberg M, Nolte IM, Meyers DA, Bleecker ER, et al. Association of IL1RL1, IL18R1, and IL18RAP gene cluster polymorphisms with asthma and atopy. J Allergy Clin Immunol. 2008;122:651-654.e8. doi:10.1016/J.JACI.2008.06.030.
19. Zhang H, Wang J, Wang L, Xie H, Chen L, He S. Role of IL-18 in atopic asthma is determined by balance of IL-18/IL-18BP/IL-18R. J Cell Mol Med. 2018;22:354–73. doi:10.1111/jcmm.13323.
20. Ober C. Asthma genetics in the post-GWAS Era. Ann Am Thorac Soc. 2016;13:S85–90.
21. Guo Y, Wang AY. Novel Immune Check-Point Regulators in Tolerance Maintenance. Front Immunol. 2015;6:421. doi:10.3389/fimmu.2015.00421.
22. Yip KH, Kolesnikoff N, Hauschild N, Biggs L, Lopez AF, Galli SJ, et al. The Nedd4-2/Ndfip1 axis is a negative regulator of IgE-mediated mast cell activation. Nat Commun. 2016;7:13198. doi:10.1038/ncomms13198.
23. Lund RJ, Osmala M, Malonzo M, Lukkarinen M, Leino A, Salmi J, et al. Atopic asthma after rhinovirus-induced wheezing is associated with DNA methylation change in the *SMAD3* gene promoter. Allergy. 2018;73:1735–40. doi:10.1111/all.13473.
24. Ferreira MAR, Matheson MC, Tang CS, Granell R, Ang W, Hui J, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J Allergy Clin Immunol. 2014;133:1564–71. doi:10.1016/j.jaci.2013.10.030.
25. West EE, Kashyap M, Leonard WJ. TSLP: A Key Regulator of Asthma Pathogenesis. Drug Discov Today Dis Mech. 2012;9. doi:10.1016/j.ddmec.2012.09.003.
26. Gokkusu C, Cakmakoglu B, Dasdemir S, Tulubas F, Elitok A, Tamer S, et al. Association Between Genetic Variants of DNA Repair Genes and Coronary Artery Disease. Genet Test Mol Biomarkers. 2013;17:307–13. doi:10.1089/gtmb.2012.0383.
27. Bielicki P, Barnas M, Brzoska K, Jonczak L, Plywaczewski R, Kumor M, et al. Genetic determinants of cardiovascular disease in patients with obstructive sleep apnea (OSA). In: 4.2 Sleep and Control of Breathing. European Respiratory Society; 2015. p. OA1751. doi:10.1183/13993003.congress-2015.OA1751.
28. Pang J, Xu X, Getman MR, Shi X, Belmonte SL, Michaloski H, et al. G protein coupled receptor kinase 2 interacting protein 1 (GIT1) is a novel regulator of mitochondrial biogenesis in heart. J Mol Cell Cardiol. 2011;51:769–76. doi:10.1016/j.yjmcc.2011.06.020.
29. Yamada Y, Yasukochi Y, Kato K, Oguri M, Horibe H, Fujimaki T, et al. Identification of 26 novel loci that confer susceptibility to early-onset coronary artery disease in a Japanese population. Biomed Reports. 2018;9:383–404. doi:10.3892/br.2018.1152.
30. Gadang V, Konaniah E, Hui DY, Jaeschke A. Mixed-lineage kinase 3 deficiency promotes neointima formation through increased activation of the RhoA pathway in vascular smooth muscle cells. Arterioscler Thromb Vasc Biol. 2014;34:1429–36. doi:10.1161/ATVBAHA.114.303439.
31. Leiferman KM. Eosinophils in atopic dermatitis. J Allergy Clin Immunol. 1994;94 6 Pt 2:1310–7. doi:10.1016/0091-6749(94)90347-6.
32. Wang J, Shannon MF, Young IG. A role for Ets1, synergizing with AP-1 and GATA-3 in the regulation of IL-5 transcription in mouse Th2 lymphocytes. Int Immunol. 2006;18:313–23. doi:10.1093/intimm/dxh370.
33. Bochner BS. The eosinophil: For better or worse, in sickness and in health. Ann Allergy, Asthma Immunol. 2018;121:150–5. doi:10.1016/j.anai.2018.02.031.
34. Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJH, Franke B, et al. Novel Association in Chromosome 4q27 Region with Rheumatoid Arthritis and Confirmation of Type 1 Diabetes Point to a General Risk Locus for Autoimmune Diseases. Am J Hum Genet. 2007;81:1284–8. doi:10.1086/522037.
35. Hollis-Moffatt JE, Chen-Xu M, Topless R, Dalbeth N, Gow PJ, Harrison AA, et al. Only one independent genetic association with rheumatoid arthritis within the KIAA1109-TENR-IL2-IL21 locus in Caucasian sample sets: confirmation of association of rs6822844 with rheumatoid

arthritis at a genome-wide level of significance. Arthritis Res Ther. 2010;12:R116. doi:10.1186/ar3053.
36. Dorra B, Hajer F, Ali A, Isabel M, Abida O, Nabil T, et al. Autoimmune diseases association study with the KIAA1109-IL2-IL21region in a Tunisian population. Front Immunol. 2013;4. doi:10.3389/conf.fimmu.2013.02.00529.
37. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. 2013;45:1353–62. doi:10.1038/ng.2770.
38. Hedjoudje A, Cheurfa C, Briquez C, Zhang A, Koch S, Vuitton L. Rs2476601 polymorphism in PTPN22 is associated with crohn's disease but not with ulcerative colitis: A meta-analysis of 16,838 cases and 13,356 controls. Ann Gastroenterol. 2017;30:197–208.
39. Jones PD, Kaiser MA, Najafabadi MG, McVey DG, Beveridge AJ, Schofield CL, et al. The coronary artery disease-associated coding variant in zinc finger C3HC-type containing 1 (ZC3HC1) affects cell cycle regulation. J Biol Chem. 2016;291:16318–27.
40. Adams PC, Pankow JS, Barton JC, Acton RT, Leiendecker-Foster C, McLaren GD, et al. HFE C282Y Homozygosity Is Associated With Lower Total and Low-Density Lipoprotein Cholesterol. Circ Cardiovasc Genet. 2009;2:34–7. doi:10.1161/CIRCGENETICS.108.813089.
41. Pilling LC, Tamosauskaite J, Jones G, Wood AR, Jones L, Kuo CL, et al. Common conditions associated with hereditary haemochromatosis genetic variants: Cohort study in UK Biobank. BMJ. 2019;364.
42. Stuckey DW, Shah K. TRAIL on Trial: Preclinical advances for cancer therapy. Trends Mol Med. 2013;19. doi:10.1016/J.MOLMED.2013.08.007.
43. Park Y, Schoene N, Harris W. Mean platelet volume as an indicator of platelet activation: Methodological issues. Platelets. 2002;13:301–6. doi:10.1080/0953710022201483332.
44. Giustino G, Kirtane AJ, Généreux P, Baber U, Witzenbichler B, Neumann FJ, et al. Relation between Platelet Count and Platelet Reactivity to Thrombotic and Bleeding Risk: From the Assessment of Dual Antiplatelet Therapy with Drug-Eluting Stents Study. Am J Cardiol. 2016;117:1703–13.
45. Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80- ). 2015;348:648–60. doi:10.1126/science.1262110.
46. Monlong J, Calvo M, Ferreira PG, Guigó R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. Nat Commun. 2014;5:4698. doi:10.1038/ncomms5698.
47. Ongen H, Dermitzakis ET. Alternative Splicing QTLs in European and African Populations. Am J Hum Genet. 2015;97:567–75. doi:10.1016/J.AJHG.2015.09.004.

## Figures and Figure Legends

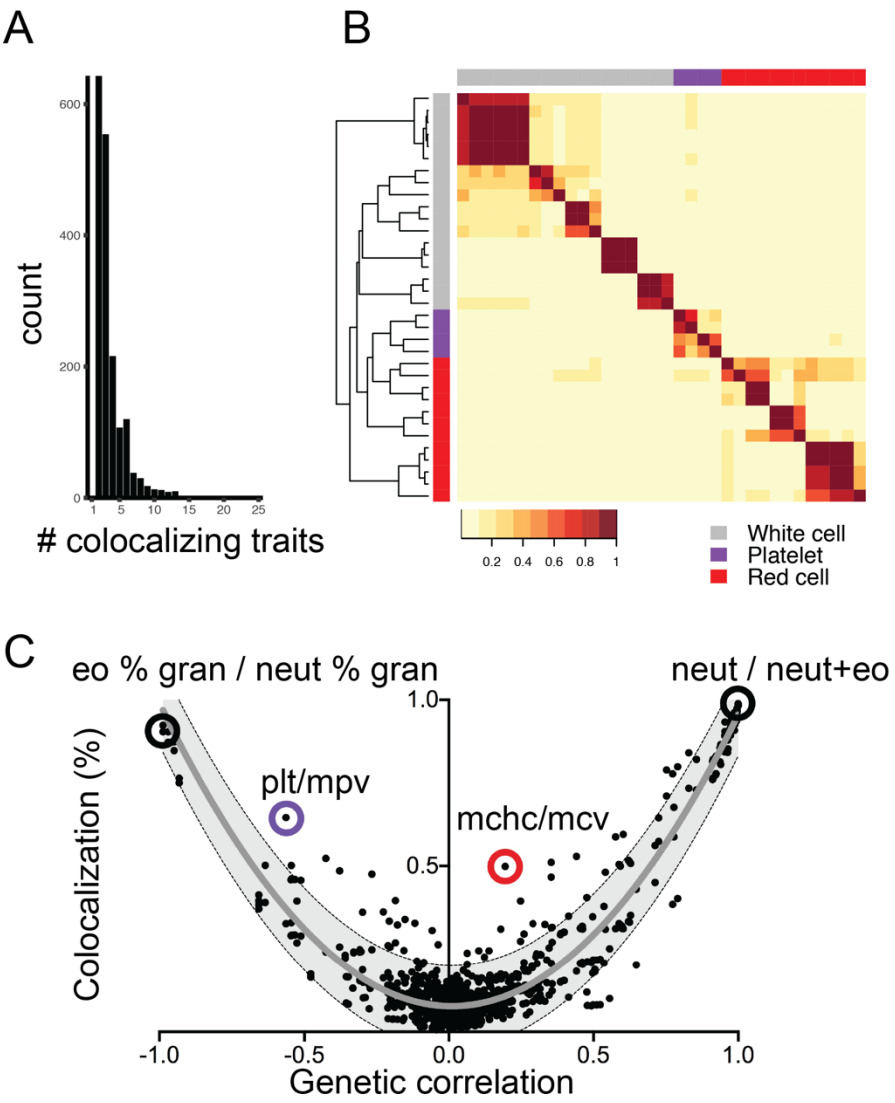


**Figure 1. Genetic colocalization between blood traits reflects hematopoietic lineage relationships.**
**a** Number of traits identified at each colocalization site (max = 25). **b** Heat map depicting percent overlap at colocalization sites between each hematopoietic trait pair. In each box, the number of sites where the row-specified trait and column-specified trait colocalized was normalized to the total number of colocalization sites for the 'row trait'. For this reason, the heat map is asymmetric. Color scale represents the proportion of loci where each pair of traits colocalized. To the left of the heat map, hierarchical clustering accurately segregated red cell, platelet, and white cell traits in general agreement with blood lineage relationships. **c** Degree of colocalization (% overlap) generally reflects genetic correlation between trait pairs. Shaded area depicts 95% prediction, with gray line at mean. Exemplary trait pairs are highlighted. eo % gran, percentage of granulocytes that are eosinophils. neut % gran, percentage of granulocytes that are neutrophils. plt, platelet count. mpv, mean platelet volume. mchc, mean corpuscular hemoglobin content. mcv, mean red cell corpuscular volume. neut, neutrophil count. neut+eo, total neutrophil plus eosinophil count.

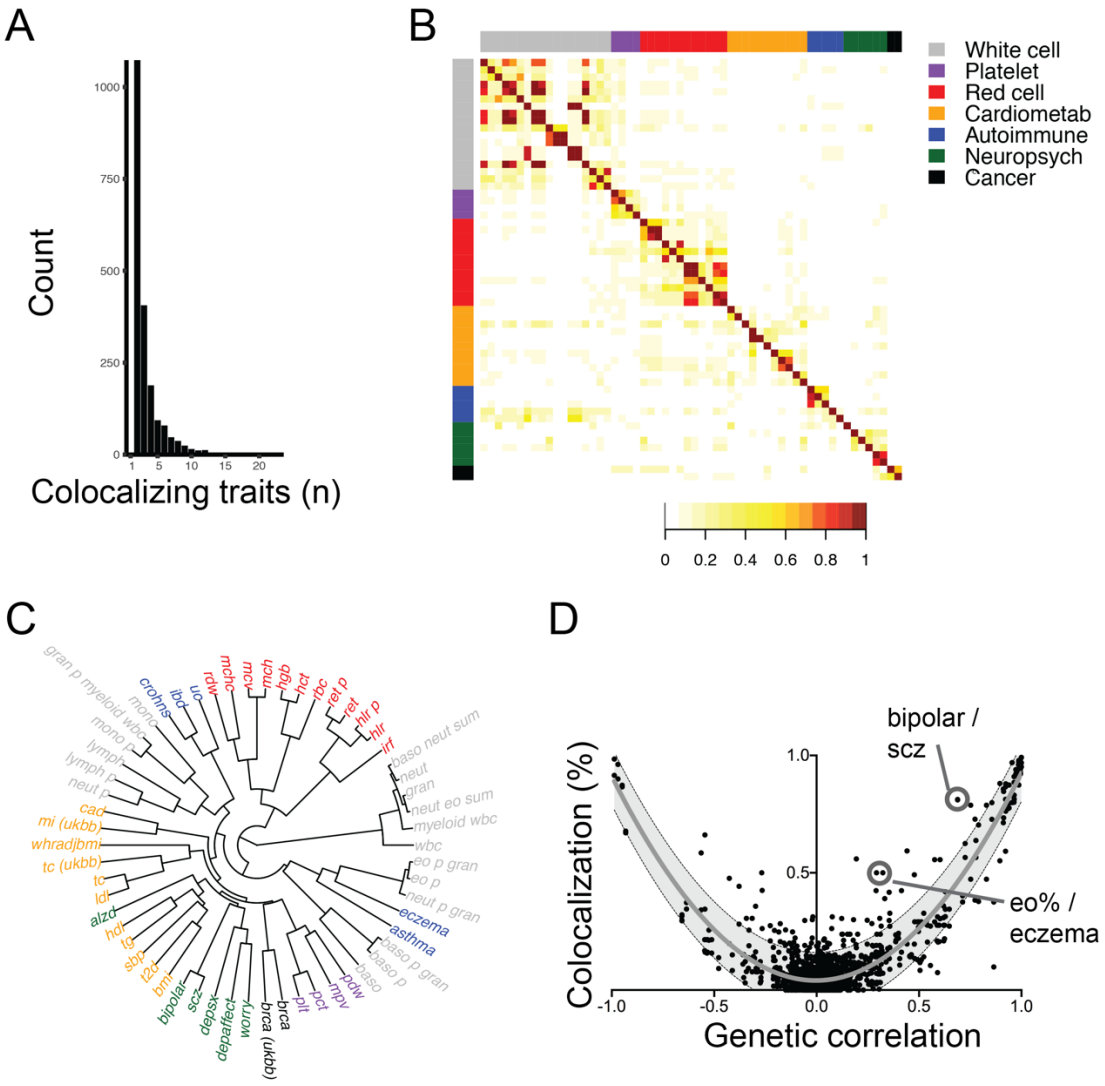


**Figure 2. Genetic colocalization reveals shared regulatory loci and implicates causal genes underlying genetic associations between hematopoietic traits and disease end-points.**

**a** Number of traits identified at each colocalization site (max = 23). **b** Heat map depicting percent overlap at colocalization sites between each trait pair. In each box, the number of sites where the row-specified trait and column-specified trait colocalized was normalized to the total number of colocalization sites for the 'row trait'. For this reason, the heat map is asymmetric. **c** Hierarchical clustering based on colocalization results accurately associates related traits, which are color coded according to the key in part **b**. **d** Degree of colocalization (% overlap) reflects genetic correlation between trait pairs. Shaded area depicts 95% prediction, with gray line at mean. Exemplary trait pairs are circled. scz, schizophrenia. eo%, percentage of white blood cells that are eosinophils.

## Additional file 1 Table Legends

**Table S1.** Traits and SNPs identified by colocalization analysis [5] of hematopoietic traits. All identified sites are shown in this table. Candidate SNPs are indicated as chr:pos. The posterior probability of colocalization, regional (genomic) probability of colocalization, and posterior probability explained at each locus are indicated.

**Table S2.** 'Hematopoietic stem cell' sites at which white blood cell (wbc), red blood cell (rbc), and platelet (plt) counts colocalize. Sites were specified by chromosome and position. The rsID(s) associated with each site are shown. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S3.** 'RBC trait only' sites at which only the indicated red blood cell traits colocalized, excluding platelet or white blood cell traits. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9). rbc, red blood cell count. hct, hematocrit. mcv, mean red cell corpuscular volume. rdw, red cell distribution width.

**Table S4.** Gene ontology pathway analysis of genes regulated by eQTLs linked to 'RBC trait only' sites. Shown are pathways with p<0.05 by Binomial test using Bonferroni correction for multiple testing.

**Table S5.** 'Platelet trait only' sites at which only the indicated platelet traits colocalized, excluding red blood cell or white blood cell traits. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9). plt, platelet count. pct, platelet-crit. mpv, mean platelet volume. pdw, platelet distribution width.

**Table S6.** Gene ontology pathway analysis of genes regulated by eQTLs linked to 'platelet trait only' sites. Shown are pathways with p<0.05 by Binomial test using Bonferroni correction for multiple testing.

**Table S7.** Genome wide association study summary statistics used in our analysis. The trait(s) queried are shown, along with study Pubmed identification number (PMID). UK BioBank studies can be found using the link provided.

**Table S8.** Traits and SNPs identified by colocalization analysis [5] of 70 human traits. All identified sites are shown in this table. Candidate SNPs are indicated as chr:pos. The posterior probability of colocalization, regional (genomic) probability of colocalization, and posterior probability explained at each locus are indicated.

**Table S9.** Colocalization sites for Lymphocyte count (lymph) and Asthma. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S10.** Colocalization sites for neutrophil count (neut) and Asthma. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S11.** Colocalization sites for eosinophil percentage (eo%) and Asthma. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S12.** Colocalization sites for monocyte count (mono) and Asthma. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S13.** Colocalization sites for mean platelet volume (mpv) and coronary artery disease (cad). Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S14.** Colocalization sites for reticulocyte count (ret) and coronary artery disease (cad). Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S15.** Colocalization sites for lymphocyte count (lymph) and coronary artery disease (cad). Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S16.** Colocalization sites for eosinophil percentage (eo%) and Eczema. Gene symbols for the nearest gene, all eQTLs, all eQTLs in whole blood, and all sQTLs (sQTLseekeR and Altrans methods) are shown in the indicated columns related to the indicated SNP (rsID) and any SNPs in high linkage disequilibrium ($r^2$>0.9).

**Table S17.** Analysis of a coding variant (rs2476601) that causes a missense mutation in *PTPN22* shows significant colocalization white white blood cell (wbc) count and Crohns disease. The effect sizes and direction (+/-) are shown.

**Table S18.** Colocalization analysis for a coding variant (rs1800562) in Human Factors Engineering (*HFE*), mutations in which cause hereditary hemochromatosis. Effects on total cholesterol (TC), low density lipoprotein (LDL), and red blood cell traits (high light scatter reticulocyte count, hlr; high light scatter reticulocyte percentage, hlr_p; mean corpuscular hemoglobin concentration, mchc; red cell distribution width, rdw ; reticulocyte count, ret ; reticulocyte percentage, ret_p), with significant colocalization signal at this locus, are shown.

**Table S19.** Colocalization analysis for a coding variant (rs17600346) in Tumor necrosis factor (TNF)-related apoptosis inducing ligand (*TRAIL*, also known as *TNF10*). Effects on colocalized traits total cholesterol (TC), triglyceride (TG), white blood cell (granulocyte percentage of myeloid white blood cells, gran_p_myeloid_wbc; monocyte percentage, mono_p) and platelet traits (platelet-crit, pct; platelet count, plt) are shown.