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
Genetic associations link hematopoietic traits and disease end-points, but most causal variants and genes underlying these relationships are unknown. We used genetic colocalization to pinpoint loci regulating hematopoietic, cardiovascular, autoimmune, neuropsychiatric and cancer phenotypes. Our findings recapitulate developmental relationships between hematopoietic traits, identify quantitative trait loci (QTLs) relating traits with known causal genetic relationships, and reveal novel genetic associations between disease phenotypes, providing a road map for gene validation experiments.

Main

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 more > 10 traits undergoing simultaneous analysis.[3–5] Recently, this colocalization algorithm was used to identify known and novel loci related to cardiovascular traits.[5] We reasoned that similar analysis could help explain variants and genes underlying hematopoietic and other disease phenotypes. Specifically, we looked for quantitative trait loci impacting gene expression (eQTL) or exon splicing variation (sQTL) at or near sites of genetic colocalization. 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.

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 (**Supplementary Table 1**). 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), although 22% of the loci (259 in total) represented sites where 6 or more traits colocalized (**Fig. 1a**). Hence, a substantial proportion of interrogated loci impact quantitative perturbation in multiple hematopoietic cell types.

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 expected lineage and trait relationships (**Fig. 1c**).

We then asked whether colocalization mirrored genetic correlation between hematopoietic traits[6]. Indeed, more closely related traits colocalized more often (**Fig. 1d**, r^2=0.90 by quadratic regression with least squares fit). Directly correlated (e.g., 'neutrophil count' and 'neutrophil + eosophil 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. 1d**). 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 cell types and traits. However, trait pairs without genetic correlation frequently had some degree of colocalization (**Fig. 1d**, y-intercept = 0.082). 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 ~1%.

We then leveraged our colocalization results to identify quantitative trait 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). Indeed, QTLs related to these loci pointed to known HSC regulatory genes *SH2B3,*[7,8] *ATM*,[9] and *HBS1L-MYB*[10] (**Supplementary Table 2**).

We also parsed loci identified by colocalization to specifically affect platelet or red cell traits, with the hypothesis that these loci would relate to differentiated 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, **Supplementary Table 3**). Gene ontology (GO) analysis[11] of these gene set revealed significant enrichment of genes related to cellular metabolic processes (**Supplementary Table 4**). 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, **Supplementary Table 5**). Pathway analysis of these genes revealed enrichment of apoptotic cell clearance and metabolic processes (**Supplementary Table 6**). Complement-mediate apoptotic cell clearance mechanisms are indeed important for regulating platelet count.[12]

Notably, 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., *CCDN3*[13]), 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.

We then extended our colocalization analysis to include 70 total hematopoietic, cardiovascular, autoimmune, cancer, and neuropsychiatric traits (**Supplementary Table 7**). Colocalization analysis using 7146 genome-wide significant loci from the NHGRI database[14] revealed a total of 2005 sites wherein 2 or more traits colocalized with a posterior probability > 0.7 (**Supplementary Table 8**). The average number of traits that colocalized at a given site was 3.3±2.4 (mean±SD), with 84 loci where ≥9 traits colocalized (**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 absent genetic correlation (**Fig. 2d**, r^2=0.79, y-intercept= 0.039)

Mendelian randomization analyses have established causal relationships between several trait pairs.[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 (**Supplementary Tables 9-17**). For example, whole blood QTLs related to genes known to affect asthma pathogenesis or severity (e.g. *IL18R, ZFP57, BTN3A2, NDFIP1, SMAD3, CLEC16A, TSLP*) were associated with colocalization sites for asthma and neutrophil, eosinophil, monocyte and/or lymphocyte traits (**Supplementary Tables 12-15**). Similarly, QTLs for genes linked to coronary artery disease risk (e.g. *XRCC3, SREBF1, GIT1, SKIV2L, MAP3K11/MLK3*) were associated with colocalization

sites linking coronary artery disease with lymphocyte and/or reticulocyte counts (**Supplementary Tables 9-11**). Other affected genes impacted by 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[15] and colocalized at 9 sites (**Supplementary Table 19**). These results implicate loci and genes expected to regulate eosinophil biology (*ETS1, ID2, PHF11*), autoimmune disease (*KIAA1109*, *TAGAP*) and skin inflammation (*BANF1*), as well as unexpected genes that warrant validation (*SNX32, ZNF652, KLC2*).

We also noted interesting associations outside of hematopoietic traits. For example, Type 2 diabetes and prostate cancer colocalized at 2 locations within the *HNF1b* locus (**Fig. 2b, Supplementary Fig. 1** and **Supplementary Table 20**). *HNF1b* has been associated with diabetes[16] and prostate cancer risk.[17] It is possible that regulation of *MLLT6* and *ARHGAP23* arising from this locus may underlie the epidemiologic association between T2D and prostate cancer.[16]

In conclusion, in an extensive genetic colocalization analysis, we have identified loci, genes and related pathways linking 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.

Online Methods

SNP and study selection. Human genome version hg19 was used for all analyses. GWAS summary statistics were obtained from publicly available repositories (**Supplemental Table 7**). We narrowed analysis to just those GWAS summary statistics for European populations with >1e6 sites (i.e., 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[14] (downloaded January 2019).

Colocalization analyses. We used HyPrColoc software to conduct colocalization experiments[5]. This software requires effect (beta, 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 500kb regions (250kb on either side of each site), in line with prior literature[5].

We present separate analyses from 34 hematologic traits and the composite 71 traits because the SNPs considered as input data varied. There were 29,239,447 sites analyzed for colocalization among the hematologic traits. In contrast, 1,544,383 sites were consistent among the different GWAS considered for the 71 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 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.

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[18]), as well as all sQTLs as defined by two different algorithms (sQTLseekeR[19], Altrans[20]). In the manuscript and Supplementary Tables, the quantity of QTL SNPs and pathway analyses reflect a composite number of genes impacted by a given locus. 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].

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.

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

Acknowledgements
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).

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

Competing interests
The authors declare no relevant conflicts of interest.

References

1. Ulirsch, J.C. et al. *Nat. Genet.* **51**, 683–693 (2019).
2. Astle, W.J. et al. *Cell* **167**, 1415-1429.e19 (2016).
3. Giambartolomei, C. et al. *PLoS Genet.* **10**, e1004383 (2014).
4. Giambartolomei, C. et al. *Bioinformatics* **34**, 2538–2545 (2018).
5. Foley, C.N. et al. *bioRxiv* 592238 (2019).doi:10.1101/592238
6. Bulik-Sullivan, B. et al. *Nat. Genet.* **47**, 1236–1241 (2015).
7. Giani, F.C. et al. *Cell Stem Cell* **18**, 73–78 (2016).
8. Balcerek, J. et al. *Nat. Commun.* **9**, 3915 (2018).
9. Ito, K. et al. *Nature* **431**, 997–1002 (2004).
10. Lieu, Y.K. & Reddy, E.P. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21689–94 (2009).
11. Ashburner, M. et al. *Nat. Genet.* **25**, 25–29 (2000).
12. Quach, M.E., Chen, W. & Li, R. *Blood* **131**, 1512–1521 (2018).
13. Sankaran, V.G. et al. *Genes Dev.* **26**, 2075–2087 (2012).
14. Buniello, A. et al. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
15. Leiferman, K.M. *J. Allergy Clin. Immunol.* **94**, 1310–7 (1994).
16. Stevens, V.L. et al. *Prostate* **70**, 601–7 (2010).
17. Hu, Y.L. et al. *Genet. Mol. Res.* **12**, 1327–1335 (2013).
18. Ardlie, K.G. et al. *Science (80-.).* **348**, 648–660 (2015).
19. Monlong, J., Calvo, M., Ferreira, P.G. & Guigó, R. *Nat. Commun.* **5**, 4698 (2014).
20. Ongen, H. & Dermitzakis, E.T. *Am. J. Hum. Genet.* **97**, 567–575 (2015).

Figures and Figure Legends

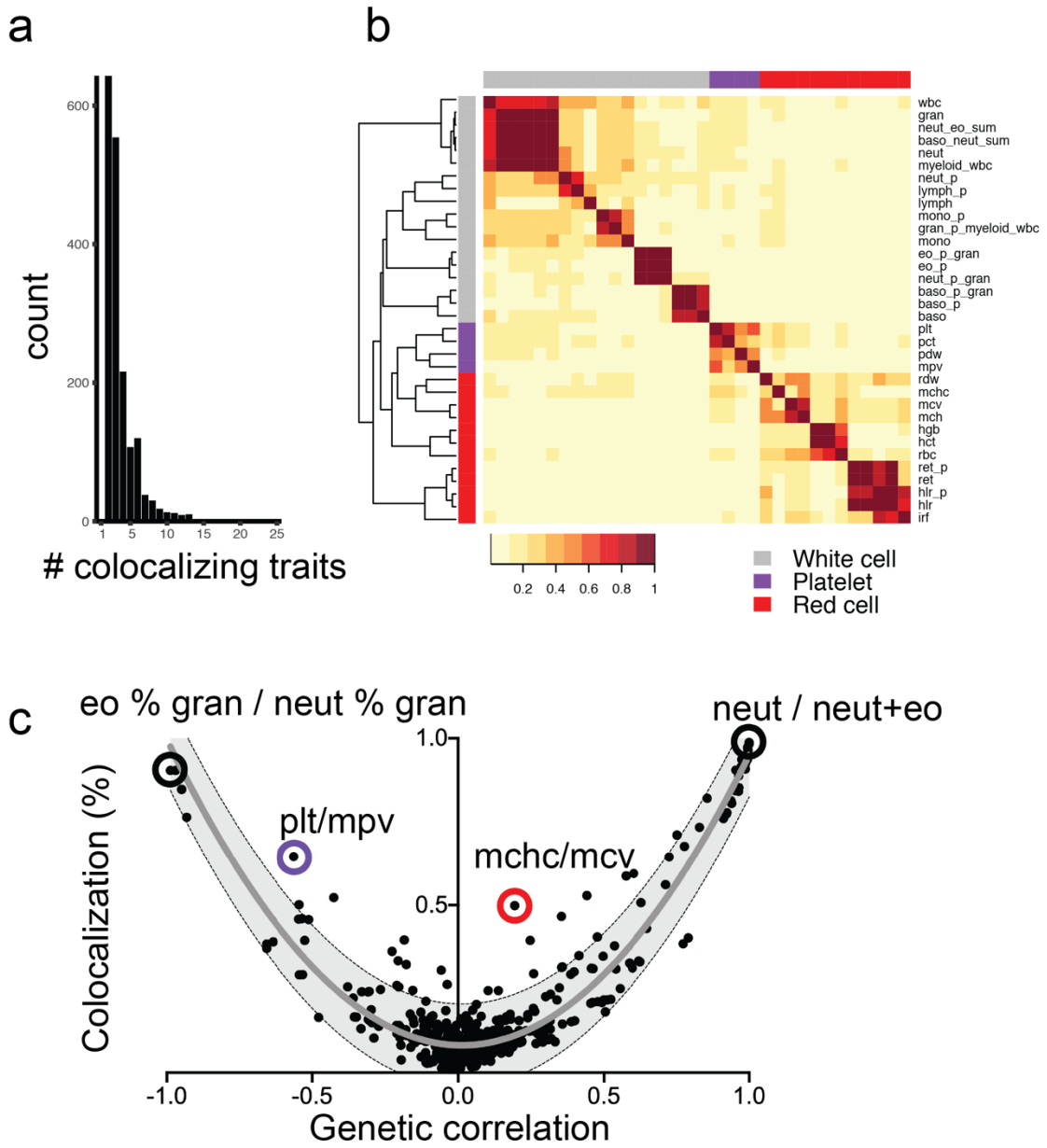


Figure 1. Genetic colocalization accurately models hematopoietic trait relationships. **a**, Number of traits colocalizing at each site (max = 25). **b**, Heat map depicting percent overlap at colocalization sites between each hematopoietic trait pair. Hierarchical clustering accurately segregates red cell, platelet, and white cell traits. **d,** Degree of colocalization (% overlap) generally reflects genetic correlation between trait pairs. Shaded area depicts 95% prediction, with gray line at mean. Exemplary trait pairs discussed in the text are highlighted.

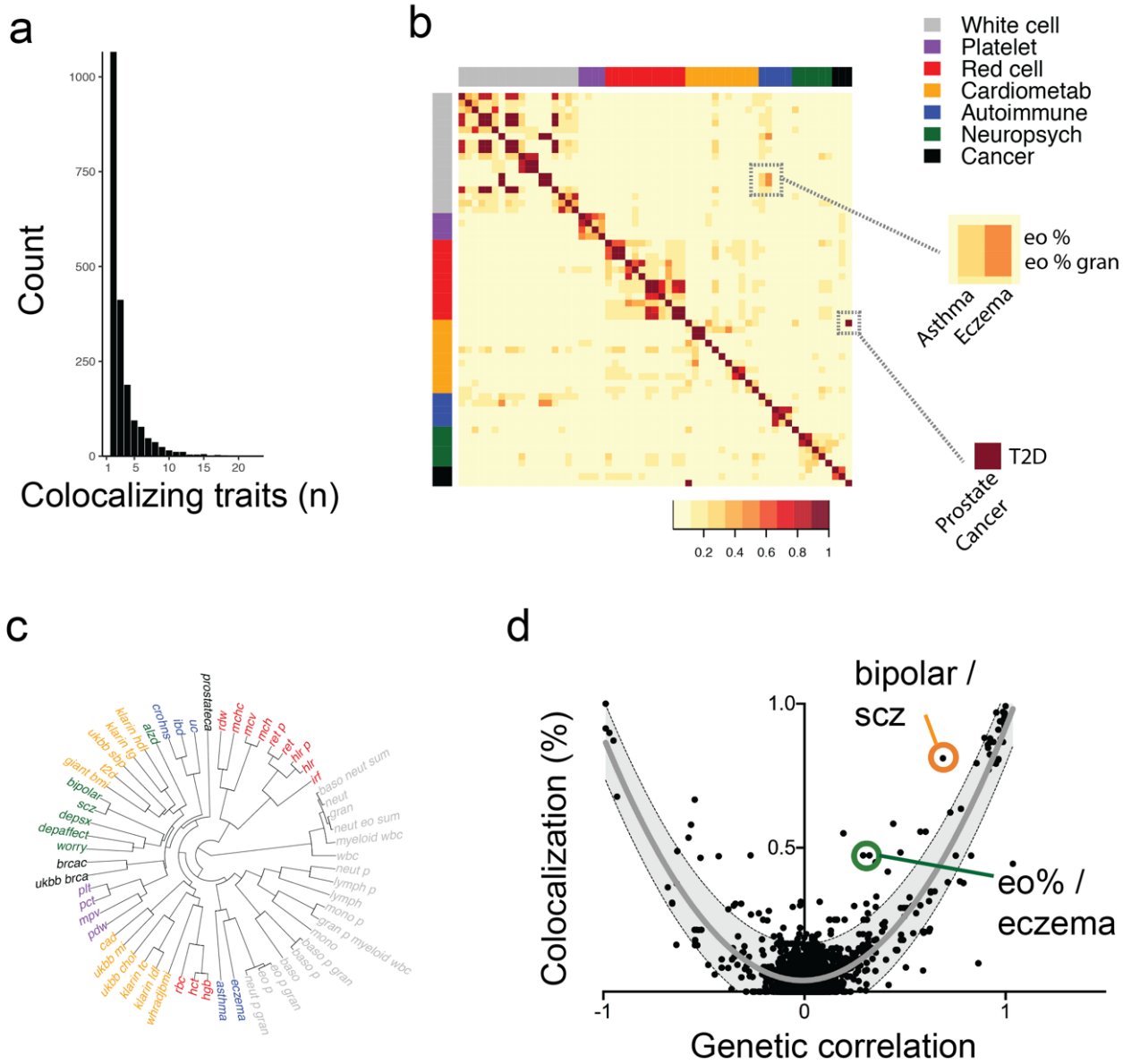
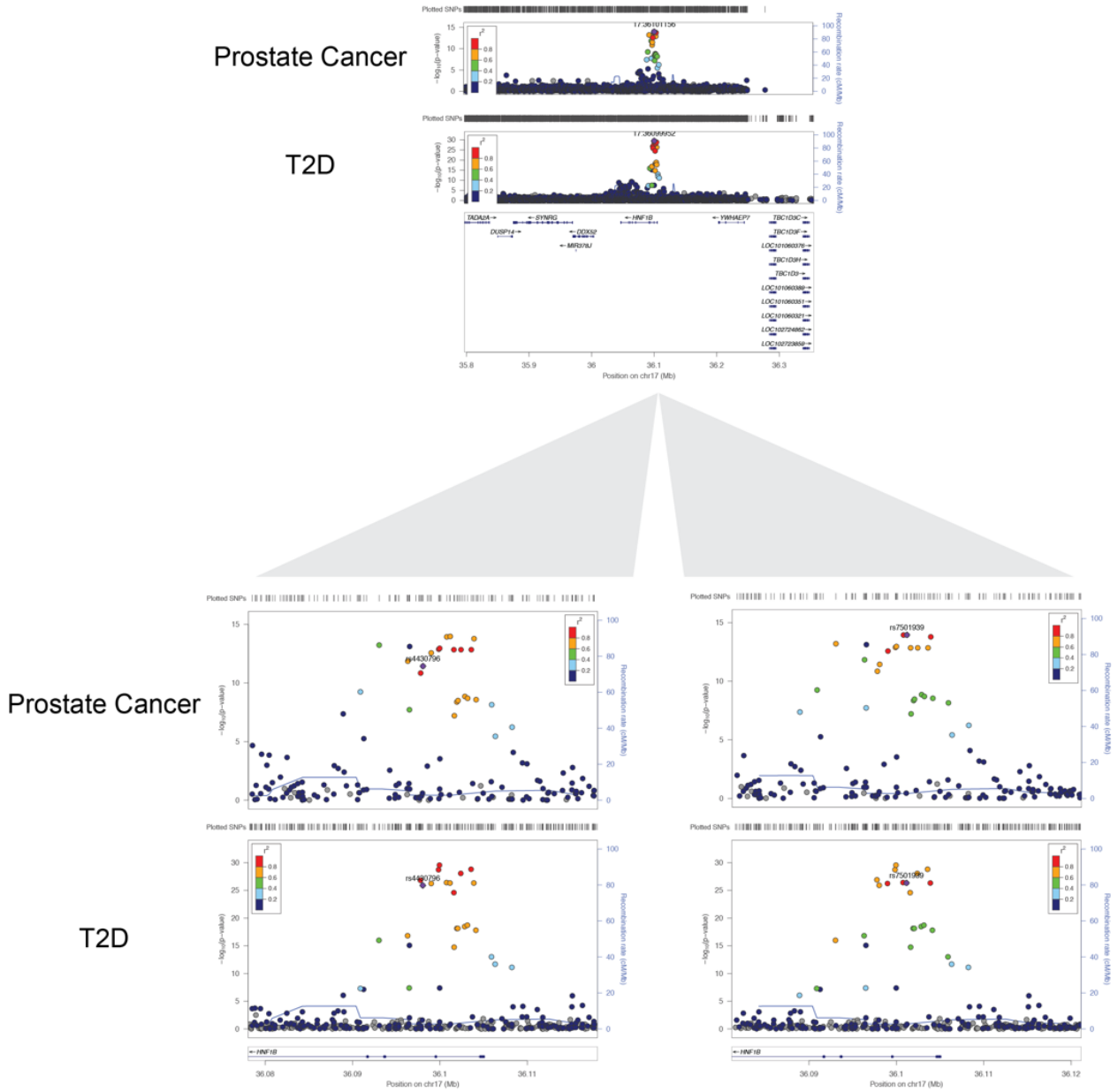


Figure 2. Genetic colocalization reveals shared regulatory loci and implicates causal genes underlying genetic associations between hematopoietic traits and disease end-points.

a, Number of traits colocalizing at each site (max = 23). **b**, Heat map depicting percent overlap at colocalization sites between trait pairs. Trait pairs discussed in the text are indicated with dashed lines. **c**, Hierarchical clustering based on colocalization results accurately associates related traits. **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 (schizophrenia, scz).

Supplementary Figure



Supplementary Figure 1. Type 2 diabetes and prostate cancer colocalize at 2 sites within the *HNF1b* locus.

a, Type 2 diabetes and prostate cancer share a single genetic signal within the *HNF1b* locus. **b**, The region near rs4430796 is highlighted, showing genetic association and signals for indicated traits. This SNP is an eQTL for *MLLT6*. **c**, The region near rs7501939 is highlighted, showing genetic association and signals for indicated traits. This SNP is an eQTL for *ARHGAP23*.

Supplementary Table Legends

Supplementary Table 1. 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.

Supplementary Table 2. 'Hematopoietic stem cell' sites at which white blood cell, red blood cell, and platelet counts colocalize. The rsID associated with each site is included. Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP, as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 3. 'RBC trait only' sites at which only the indicated red blood cell traits colocalized, excluding platelet or white blood cell traits. Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 4. Gene ontology pathway analysis of genes regulated by eQTLs linked to 'red blood cell trait only' sites. Shown are pathways with p<0.05 by Binomial test using Bonferroni correction for multiple testing.

Supplementary Table 5. 'Platelet trait only' sites at which only the indicated platelet traits colocalized, excluding red blood cell or white blood cell traits. Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 6. 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.

Supplemental Table 7. 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.

Supplementary Table 8. 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.

Supplementary Table 9. Colocalization sites for Lymphocyte count (LYMPH) and Asthma. Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 10. Colocalization sites for Neutrophil count (Neut) and Asthma. Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 11. Colocalization sites for Eosinophil percentage (Eo %) and Asthma. Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 12. Colocalization sites for Monocyte count (Mono) and Asthma. Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 13. Colocalization sites for Mean Platelet Volume (MPV) and Coronary Artery Disease (CAD). Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 14. Colocalization sites for Reticulocyte count (RET) and Coronary Artery Disease (CAD). Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).

Supplementary Table 15. Colocalization sites for Lymphocyte count (LYMPH) and Coronary Artery Disease (CAD). Nearest Gene, eQTL, and sQTL columns depict gene symbols related to the indicated SNP (rsID), as well as SNPs in high linkage disequilibrium (r^2>0.9).