HGG
Advances

From GWAS variant to function: A study of ∼148,000 variants for blood cell traits

Quan Sun,[1,12] Cheynna A. Crowley,[1,12] Le Huang,[2] Jia Wen,[3] Jiawen Chen,[1] Erik L. Bao,[9,10] Paul L. Auer,[4] Guillaume Lettre,[5,6] Alexander P. Reiner,[7,8] Vijay G. Sankaran,[9,10] Laura M. Raffield,[3] and Yun Li[1,3,11,*]

Summary

Genome-wide association studies (GWASs) have identified hundreds of thousands of genetic variants associated with complex diseases and traits. However, most variants are noncoding and not clearly linked to genes, making it challenging to interpret these GWAS signals. We present a systematic variant-to-function study, prioritizing the most likely functional elements of the genome for experimental follow-up, for >148,000 variants identified for hematological traits. Specifically, we developed VAMPIRE: Variant Annotation Method Pointing to Interesting Regulatory Effects, an interactive web application implemented in R Shiny. This tool efficiently integrates and displays information from multiple complementary sources, including epigenomic signatures from blood-cell-relevant tissues or cells, functional and conservation summary scores, variant impact on protein and gene expression, chromatin conformation information, as well as publicly available GWAS and phenome-wide association study (PheWAS) results. Leveraging data generated from independently performed functional validation experiments, we demonstrate that our prioritized variants, genes, or variant-gene links are significantly more likely to be experimentally validated. This study not only has important implications for systematic and efficient revelation of functional mechanisms underlying GWAS variants for hematological traits but also provides a prototype that can be adapted to many other complex traits, paving the path for efficient variant-to-function (V2F) analyses.

Introduction

Genome-wide association studies (GWASs) have identified thousands of genetic loci and hundreds of thousands of genetic variants associated with various complex human diseases and traits, but the underlying genetic mechanism for the vast majority of these GWAS signals remains elusive. With extensive sequencing and GWAS efforts, there is a pressing need to convert the large and ever-growing number of significant GWAS variant-trait pairs into human-interpretable functional or mechanistic knowledge.[1] Most variants identified through GWASs reside in the noncoding regions (e.g., >95% for blood cell traits[2]), and most signals include multiple highly correlated variants or variants in strong linkage disequilibrium (LD). Pinpointing the most likely causal variants within GWAS signals, and linking these variants to their target genes, is challenging, particularly as the number of GWAS loci and variants increases. For hematological traits, for instance, our recent GWAS meta-analyses[3,4] have revealed over 7,000 loci, with >148,000 variants associated with at least one blood cell index at stringent genome-wide significance threshold. Comprehensive and computationally efficient

annotation and prioritization of such GWAS findings are of ever-increasing interest.

Understanding how genetic variants contribute to a phenotype is often referred to as the variant-to-function (V2F) problem. Responding to this problem requires us to determine causal genetic variants, relative cell types/states, their target genes, and cellular/physiological functions.[5] Functional experiments are needed to fully reveal molecular mechanisms, but we cannot yet afford to perform time-, money-, and labor-consuming experimental validations of thousands of loci involving hundreds of thousands of potentially functional variants or regulatory elements controlling their nearby genes, since each gene is likely regulated by multiple variants, and each variant may regulate multiple genes. Thus, computational methods are needed to screen potential variants and their effector genes for further experiments.

In this study, we focus on hematological traits. Hematological phenotypes (red blood cell, white blood cell, and platelet counts and indices) are critical physiological intermediaries in oxygen transport, immunity, infection, thrombosis, and hemostasis and are associated with autoimmune, allergic, infectious, and cardiovascular diseases.

---

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [2]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [4]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA; [5]Montreal Heart Institute, Montreal, QC, Canada; [6]Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada; [7]Department of Epidemiology, University of Washington, Seattle, WA, USA; [8]Fred Hutchinson Cancer Research Center, University of Washington, Seattle, WA, USA; [9]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA; [10]Broad Institute of Harvard and MIT, Cambridge, MA, USA; [11]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[12]These authors contributed equally to this work
*Correspondence: yunli@med.unc.edu
https://doi.org/10.1016/j.xhgg.2021.100063

Hematological traits are highly heritable,[6] and recent large GWASs for hematological traits (including nearly 750,000 participants) identified thousands of variant-trait associations.[2,4] In addition, there are multiple large-scale functional experiments already available[2,7,8] for hematological traits, as well as fairly comprehensive functional annotation resources relevant to blood tissues. This makes hematological traits an ideal model for this type of V2F computational solution.

We have developed VAMPIRE: Variant Annotation Method Pointing to Interesting Regulatory Effects, a tool for the user to explore annotations encompassing epigenomic signatures, variant impact on protein and gene expression, chromatin conformation information from Hi-C and similar technologies, as well as publicly available GWAS and phenome-wide association study (PheWAS) results, creating a comprehensive annotation profile for variants from recent trans-ethnic blood cell trait publications[3,4] with a flexible interface for adding additional future GWAS results. This interactive web application implemented in R Shiny provides a model display mechanism for annotating GWAS variants from diverse complex traits, allowing selection of most likely causal variants and their effector genes for experimental follow-up. Importantly, we show the value of how variants and genes nominated by VAMPIRE can highlight key regulators of blood cell traits using independent functional assessment, confirming the value of this annotation tool. While blood cell traits are the focus for VAMPIRE, this framework (including our R Shiny application) is adaptable for annotation of other complex trait GWAS results and will facilitate the connection between variant and function.

Material and methods

Variant annotations

The current version of VAMPIRE (Figure 1) includes GWAS results from two studies (as detailed in the supplemental methods), including all variants in 95% credible sets for fine-mapped hematological-trait-associated loci from Chen et al.[4] (N1 = 148,019 variants) and lead variants (N2 = 2) from a TOPMed imputed GWAS meta-analysis in African American and Hispanic/Latino populations.[3] We plan to extend VAMPIRE as new trans-ethnic blood cell trait genetic analyses are released.

The sources of the annotation used are stated clearly in the VAMPIRE online application, with links or references to the original data sources. As a brief summary, the annotation categories are trivially split into six types ("variant level," "1D," "2D," "3D," "PheWAS," and "GWAS"). First, variant level contains data on phenotypic association from the original publication or preprint (such as the p value for association with a given hematological trait, effect size, and posterior probability of inclusion for fine-mapping credible sets). Second, 1D refers to epigenomic or sequence constraint features. This displays selected output from WGSA,[9] including functional prediction scores, conservation scores, and epigenetic information gathered from GeneHancer,[10] FANTOM5,[11,12] Roadmap,[13] and ENCODE.[14] ATAC-seq peaks from recent studies for blood cell traits[15,16] and key histone chro-

matin immunoprecipation sequencing (ChIP-seq) peaks such as H3K9me3, H3K36me3, H3K4me1, H3K4me3, and H3K27Ac generated across blood-cell-related tissues from Roadmap Epigenomics are also included.[13,17] We further include information regarding whether each variant resides in any selective sweep region detected from multiple populations in the 1000 Genomes Project[18] using the S/HIC method.[19,20] Information is displayed based on the tissue relevance to the blood cell phenotype (see Supplemental methods). All variants have 1D annotation, but for prioritization purposes as described below in the five categories for noncoding variant annotation, we define 1D annotation as FANTOM5_enhancer_robust = Y (yes), or Genehancer_feature = "Promoter" or "Enhancer" or "Promoter/Enhancer," or coreMarks (for any relevant roadmap epigenomic category) = "Enhancers" or "Active TSS." Users can then additionally filter by criteria such as functional prediction and conservation scores.

For the "2D" annotations, we included impact on gene expression and splicing ratios (expression quantitative trail locus [eQTL] and splicing QTL [sQTL] information) and impact on protein abundance (protein QTL [pQTL] information[21]) from public sources relevant to blood cell traits. This includes both bulk and cell-type-specific sources from the public domain (eQTLGen,[22] CAGE,[23] BIOS[24] for whole blood, and Raj et al.[25] for purified CD4+ T cells and monocytes). Information available in these sources varies, but generally we at a minimum display the effect size estimate, p value, the allele assessed, and the gene or protein involved. Variants were matched across sources based on chromosome, position, and alleles of each variant. Only significant results (based on false discovery rate [FDR] or other publication-specific thresholds) from the respective sources are displayed in VAMPIRE; we do note that formal co-localization analyses would still need to be performed to determine if blood-cell-related and gene/protein expression QTL signals truly coincide.

For the 3D annotations, we include information on 3D genome conformation, linking blood-lineage-specific regulatory elements to target genes from various sources. More specifically, using Hi-C data we incorporated statistically significant long-range chromatin interactions (LRCI)[17,26,27] calculated from Fit-Hi-C,[28] loops using the HiCCUPs methodology,[26] and super-FIREs for related tissues.[17] Two Promoter-Capture Hi-C (PCHi-C) data sources[29,30] were also incorporated and matched with the 2D results to highlight consistent evidence regarding the affected gene(s) across 2D and 3D annotations. VAMPIRE displays information on the number of loops, LRCI, PCHi-C interactions, FIREs, or super-FIREs, as well as significance measures such as p values, FDR, or CHICAGO scores where applicable. This 3D annotation information can also be visualized via our HUGIn browser.[31]

The last two data groups present results from two PheWAS sources[4,32] and GWAS results of blood cell traits from GWAS catalog,[33] allowing the user to evaluate if hematological trait-associated variants may also influence other complex traits.

To visualize and leverage these multiple annotation categories for further analysis or prioritization of experimental validations, VAMPIRE efficiently displays and integrates relevant variant information, allowing the user to investigate either all the variants annotated or subsets based on annotation category groupings, searching either by variant or by gene name. The comprehensive annotation for the variants is summarized using a five-category grouping created for highlighting the most promising variants, as they have various types of annotation. Specifically, the five categories for noncoding variants are (1) the most restrictive category, containing variants that have 1D, 2D, and 3D annotation and the
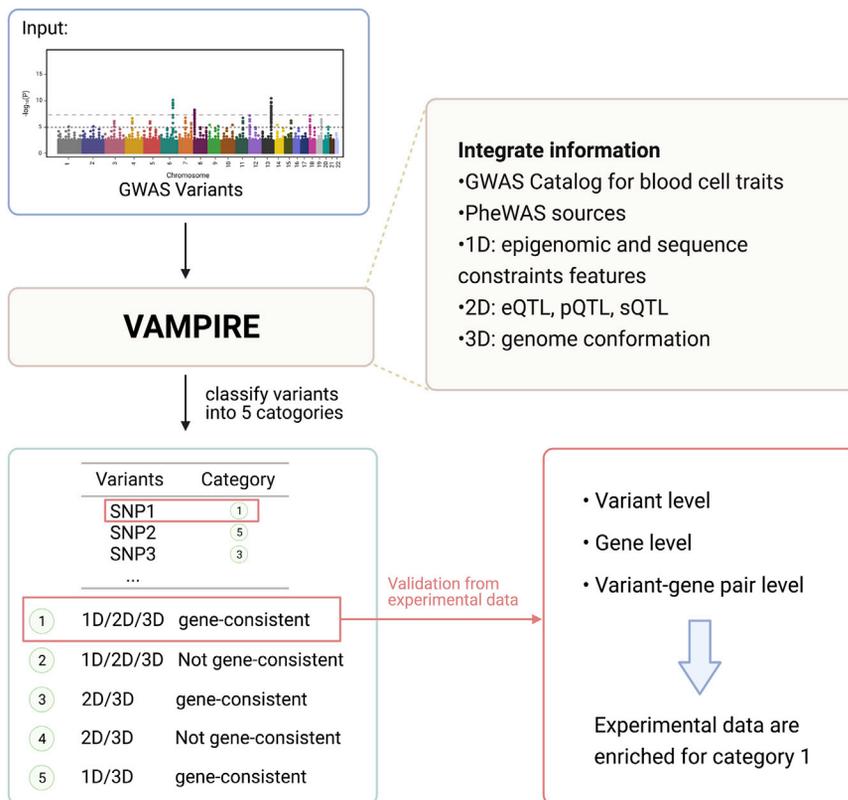
Figure 1. Overall framework of this study
VAMPIRE starts with GWAS variants in the 95% credible sets, integrates different annotations, and assigns them into different prioritization categories. We further demonstrated that our top prioritized category is enriched with variants that were experimentally validated. VAMPIRE provides a prototype that can be adapted to many other complex traits, paving the path for efficient variant-to-function (V2F) analyses.

with both uncategorized variants and variants in other categories. Recently, Vuckovic et al.[2] characterized variants that affect erythropoiesis or hematopoiesis by modifying related TF motifs, such as for KLF1, KLF6, MAFB, and GATA1. We chose these four erythroid TFs as positive control TFs and two non-erythroid TFs (IRF1 and IRF8) as negative controls.

At the gene level, we evaluated the genes interrogated by Nandakumar et al.[8] with a pooled short hairpin RNA (shRNA)-based loss-of-function approach. Specifically, Nandakumar et al.[8] assessed 389 genes in the neighborhood of 75 loci associated with red blood cell traits,[34] to identify potential causal genes underlying these GWAS signals. We assessed the enrichment of genes validated by shRNA experiments among those prioritized in VAMPIRE's category 1. Note that the categories were previously defined at the variant level. Here we extend the variant category to gene category as the strongest category where a genome-wide significant variant linked to this gene falls in. Due to the limited sample size of uncategorized genes, especially when overlapping with genes in the shRNA paper (leaving us with only two genes), we compared genes in category 1 to genes in all other categories. We also performed enrichment analyses at both variant and gene levels for categories 2–5, comparing one category to the others to see if any specifically exhibit a higher level of enrichment than the others. Specifically, we compared category 2 to categories 3–5; compared category 3 to categories 2, 4, and 5; and compared category 5 to categories 2–4.

At the variant-gene pair level, we employed the enhancer-gene connections validated via CRISPRi-FlowFISH experiments by Fulco et al.[7] in their activity-by-contact (ABC) paper. Specifically, Fulco et al.[7] tested pairs of candidate *cis* regulatory elements (CREs, ~500 bp regions) and their potential effector genes via CRISPRi perturbations of the CREs, in multiple cell lines including the K562 cells. Fulco et al.[7] tested 4,124 CRE-gene pairs in total, of which 175 were significant from their experiments. We overlapped their tested CREs with variants in our VAMPIRE annotation database. We define a VAMPIRE variant-gene pair confirmed if the variant overlaps an ABC-validated CRE *and* the linked genes in VAMPIRE (from QTL and chromatin capture conformation evidence) overlap the corresponding effector gene for that CRE via ABC's CRISPRi-FlowFISH experiment. We focused on ABC experiments performed on the K562 cells (instead of GM12878 cells, where a very small number of CREs were tested), as the number of tested CRE-gene pairs was not too small for robust statistical

genes implicated by 2D and 3D evidence are consistent; (2) containing variants with 1D, 2D, and 3D evidence, but the genes implicated from different resources are not consistent; (3) 2D and 3D with consistent gene evidence between the 2D and 3D annotations; (4) variants with 2D and 3D information and no consistent gene implied; and (5) variants with 1D and 3D evidence. We also have a predicted high-impact coding variant category displayed, including high-confidence loss-of-function (LoF) variants and likely influential missense, in-frame insertions and deletions (indels), and synonymous variants. Variants without strongly compelling variant annotation are still displayed but are not listed in these high-priority categories. The user can further subset results by hematological trait, hematological trait category, or (for the Chen et al.[4] paper) the ancestry-specific grouping in which a given credible set was derived (trans-ethnic, European, East Asian, South Asian, Hispanic/Latino, or African ancestry). In addition, the user can restrict the amount of information presented by selecting which tables to be displayed. All tables can be exported in a csv or tab delimited format.

Enrichment analysis

To assess whether the variants prioritized by VAMPIRE are more likely to be functionally impactful, we performed enrichment analysis at three different levels: variant level, gene level, and variant-gene pair level, leveraging data generated from previously published functional experiments.[2,7,8] For each set of analyses, we conducted Fisher's exact test and calculated odds ratios (ORs) and one-sided p values.

At the variant level, we assessed the enrichment of variants that modify transcription factor (TF) binding motif[2] among our annotation category 1 variants. We compared variants in category 1

inference. Matching the K562 cell line, we focused only on variants associated with red blood cell traits. Similar to the above two sets of enrichment analyses, we focused on annotations in VAMPIRE's prioritization category 1. Specifically, we tested whether variant-gene pairs prioritized in VAMPIRE's category 1 are enriched within ABC's validated enhancer-gene connections. Given the CREs tested in the ABC paper are rather short (∼500 bp), we also performed sensitivity analysis by first extending the CRE regions by ±1 kb and ±5 kb and then overlapping variants with these extended CREs, to ensure robust conclusions.

Comparison to FUMA

To further assess the capability of VAMPIRE in terms of gene prioritization, we compared the genes prioritized by VAMPIRE to genes prioritized by FUMA[35] for seven red blood cell traits, including hematocrit (HCT), hemoglobin (HGB), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), red blood cell (RBC) count, and red blood cell distribution width (RDW). We uploaded the GWAS summary statistics for each trait separately to the FUMA website with all default parameters using FUMA's SNP2GENE function. We then combined the prioritized genes for all seven red blood cell traits to compare the two methods. Similar to the gene-level enrichment analysis described above, we evaluated the number of shRNA-assessed genes and shRNA-validated genes from the shRNA experiments overlapping with the two methods. Venn diagrams were used for better illustration of the results.

Results

Overview of VAMPIRE annotations

The overall framework of VAMPIRE is illustrated in Figure 1. We started with all variants in 95% credible sets from our recent trans-ethnic study for hematological traits (total 148,019 variants)[4] and lead variants (2 variants) from Kowalski et al.[3] We incorporated six types of annotations (detailed in Material and methods): GWAS summary statistics and posterior probability of inclusion from our previous fine-mapping analyses;[4] epigenomic or sequence constraints features (1D); eQTL, sQTL, and pQTL information (2D); information on 3D genome conformation (3D); results from two PheWAS sources[4,32] (PheWAS); and GWAS results from blood cell traits from GWAS catalog[33] (GWAS).

VAMPIRE variant categories

To visualize and prioritize variants along with their corresponding candidate regulatory regions and their potential effector genes, we leverage the aforementioned six types of annotation to group these ∼148,000 blood cell traits' associated variants into various prioritization categories. Specifically, for non-coding variants, we classified them into five categories (detailed in Material and methods). Among them, category 1 is the most restrictive category, containing variants that meet all the fulling criteria: have 1D, 2D, and 3D annotation and 2D and 3D evidence supports the same effector genes (i.e., gene-consistent). Variants in category 2 also are required to have 1D, 2D, and 3D annotation simultaneously, but the genes implicated by 2D and 3D evidence are inconsistent. For example, if a variant rsXXX is an eQTL of gene A according to 2D annotations, and it also resides in a region that forms a loop with the promoter region of gene A, we say the 2D and 3D evidence is gene-consistent, and rsXXX will be classified in category 1. However, if rsYYY is an eQTL of gene B according to 2D annotations, but there is only information suggesting the rsYYY-residing region forms a loop with the promoter region of gene C, we say it is not gene-consistent and will classify rsYYY in category 2. Of course, it's possible that a SNP is an eQTL for multiple genes (e.g., gene D and gene E) and its residing region forms loops with promoters of multiple genes (e.g., gene E and gene F). As long as we can find one gene that is shared, we classify the variant as gene-consistent for the shared gene(s). In practice, we are more confident to prioritize functional experiments for rsXXX than for rsYYY, since we have consistent support from three independent sources of information for rsXXX: 1D suggesting it is regulatory, and 2D (i.e., eQTL or pQTL) and 3D (i.e., chromatin conformation) both suggesting it is regulating gene A).

Variants in category 3 and 4 have only 2D and 3D annotation. Category 3 includes those with consistent target genes suggested by 2D and 3D annotation, while variants in category 4 have 2D and 3D annotations suggesting different/inconsistent target genes. Category 5 includes those with 1D and 3D annotation but no 2D evidence. Variants not falling into any of the five categories are classified as uncategorized. Note that due to tissue or cell type specificity for some 2D (e.g., eQTL) and 3D (e.g., pcHiC) annotations, such variant-level categorization was separately performed for different traits. For instance, for white blood cell-related indices (e.g., monocyte), we considered 2D annotations from whole blood, peripheral blood mononuclear cells (PBMCs), and monocytes, while for platelet-related traits (e.g., platelet count), we only considered 2D annotations from whole blood and PBMCs. Suppose a variant has 1D regulatory evidence and forms a loop with gene A from 3D annotations. Furthermore, it is an eQTL for gene A based on monocyte, but is an eQTL only for another gene B from whole blood and PBMCs; the variant falls into category 1 for monocyte count, but category 2 for platelet count. In summary, a variant may fall into different categories for different traits. In addition, each gene is categorized according to the prioritization categories according to its linked variant(s). When its linked variants fall in multiple categories, the gene is assigned to the most highly prioritized category. The numbers of variants and genes in each category are shown in Table 1.

Enrichment analysis

Our enrichment analyses employing multiple previously published functional validation experiments encompassing variant-level, gene-level, and variant-gene pair levels all showed promising results. Specifically, at the variant level, compared to uncategorized variants, we found significant enrichment of variants affecting TF binding motifs among

Table 1. Numbers of variants and genes in each category

	Explanation	Unique variants (#)	Variant-trait pairs (#)	Genes (#)
Category 1	1D & 2D & 3D & gene-consistent	13,862	19,988	9,857
Category 2	1D & 2D & 3D & not gene-consistent	21,269	30,276	2,735
Category 3	2D & 3D & gene-consistent	14,155	20,192	1,300
Category 4	2D & 3D & not gene-consistent	33,732	48,497	1,621
Category 5	1D & 3D	11,820	14,507	1,578
Uncategorized	others	62,489	78,477	174
Total		148,215	211,937	17,265

Note that the category was defined initially at variant level, separately for each blood cell trait. One variant may fall in category 1 for one trait but in other categories for other traits. In total, we have 148,215 unique variants and 211,937 variant-trait association pairs. For gene-level category, each gene is categorized according to the prioritization categories of its linked variant(s). When its linked variants fall in multiple categories, the gene is assigned to the most highly prioritized category.

variants prioritized in category 1 of VAMPIRE (Figure 2) for all the erythroid TFs ($p < 8.1E-4$) but GATA1 ($p = 0.18$) (Table 2), likely due a smaller sample size of variants. In contrast, neither of the two negative control TFs (IRF1 and IRF8) showed any significant enrichment ($p = 0.22$ and 0.62). A similar pattern holds when comparing category 1 variants to all other categories, but the significance level decreased (Table S1), which suggests that variants in other prioritized categories, although not as enriched at TF binding sites as category 1 variants, tend to exhibit higher levels of enrichment than the uncategorized variants. At the gene level, we focused on two statistics: (1) number of genes selected for shRNA experiments, since genes were more likely to be selected for experiments when they demonstrated some prior evidence of potential causality; and (2) number of genes validated ($p < 0.05$) by shRNA experiments. We compared the number of genes in our annotation category 1 and all other categories and found that both shRNA-assessed genes ($p = 3.5E-13$) and validated genes ($p = 3.1E-8$) show strong enrichment among those in our annotation category 1 (Table 3), and the estimated enrichment score for validated genes (OR = 4.65) is almost double of that for shRNA-assessed genes (OR = 2.37).

We also conducted similar enrichment analyses at variant and gene levels to compare categories 2–5. Neither variant-level (Table S1) nor gene-level (Table S2) enrichment results is significant at all, except for category 3 (i.e., both 2D and 3D evidence exist and suggesting the same gene[s]). Category 3 is significantly ($p = 0.037$) enriched with KLF1 motif variants (OR = 1.44) and is significantly ($p = 0.027$) enriched with shRNA-assessed genes (OR = 1.70). These results suggest category 3 may be the next category most worthy of further investigation after category 1, but the evidence is not strong: the significance levels are not very high, other TF motifs are not enriched, and sample size (i.e., number of genes tested) is small. However, category 2 (1D/2D/3D but not gene-consistent) is significantly favorable over category 4 (2D/3D but not gene-consistent) (Table S3), suggesting that the additional 1D information provides more evidence.

Finally, at the variant-gene pair level, we also observed enrichment among variants selected into VAMPIRE's category 1 (Table 4). When restricting only to variants in category 1 and associated with red blood cell traits and without extending the CRE regions, only 7 of VAMPIRE's variant-gene pairs can be found in ABC's CRISPRi-FlowFISH experiments, of which 6 are not significant and 1 is significant. While not significant ($p = 0.26$), the direction of enrichment is nevertheless encouraging (one of seven, or 14.3%, confirmed by CRISPRi-FlowFISH experiments) and 3-fold greater than that among all/background pairs from Fulco et al.,[7] where 175 out of 4,124 variant-gene pairs (4.2%) were confirmed. Note that all the confirmed pairs were linked with variants associated with red blood cell traits. Further generalizing to all VAMPIRE annotation categories and to variants associated with any blood cell trait, the enrichment OR increases to 8.30 with p value $9.0E-5$, indicating that variant-gene pairs prioritized by VAMPIRE's five categories have much higher odds of being functional. To further accommodate causal variants tagged by GWAS variants not falling into the short 500 bp CREs, we extended the CREs by ± 1 kb or ± 5 kb and performed similar enrichment analysis. Our conclusions remained qualitatively similar (Table 4), but the enrichments increased in significance, thanks to larger sample size (in this context, the larger number of variant-gene pairs contributing to the analysis) and suggesting that more liberal windows of *cis*-regulatory regions can capture a higher rate of functional variant-gene pairs. For example, the enrichment for category 1 variants associated with red blood cell traits reached an OR of 15.77 ($p = 3.8E-6$) and 16.68 ($p = 3.1E-15$) for 1 kb and 5 kb extension, respectively. We thus conclude that such enrichment is significant and robust to the extension of CREs.

Application example

Figure 3 shows one example at the *CALR* locus associated with red blood cell traits. Fulco et al.[7] confirmed by CRISPRi-FlowFISH experiment that CRE chr19: 12,996,905–12,998,745 (hg19) regulates gene *CALR* (adjusted p value, $1.9E-7$). Annotations compiled by
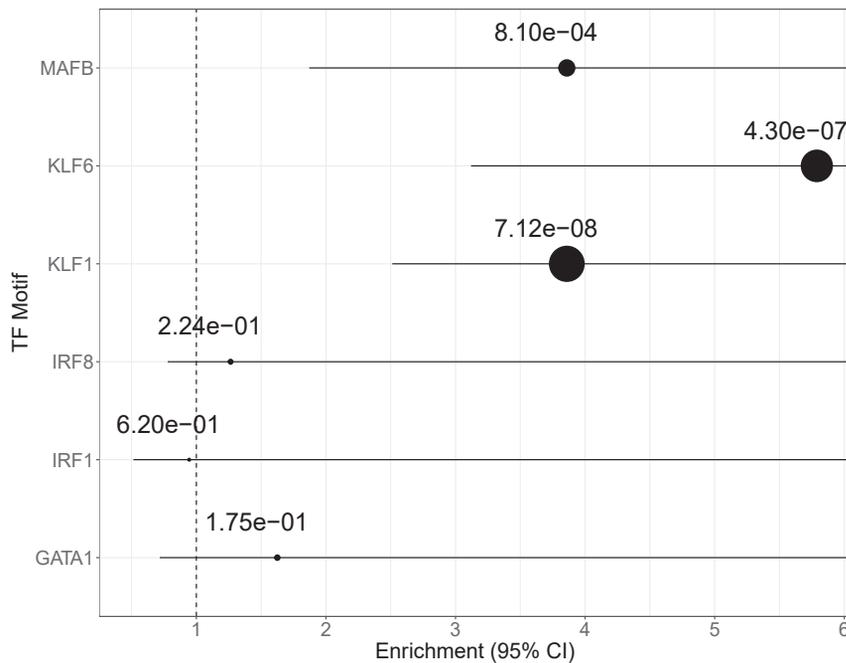
VAMPIRE suggest, consistently, that *CALR* is linked to rs8110787 (chr19: 12,999,458, hg19) in category 1. rs8110787 is associated with several red blood cell traits,[4] including HCT, MCH, MCV, and red blood cell counts. Based on genomic distance alone, *CALR* is not the nearest gene to rs8110787, with several other closer genes. However, based on H3K27ac HiChIP data in K562 cells,[36] rs8110787 significantly interacts with *CALR* promoter region (p < 1E−120), suggesting that *CALR* is a potential target gene regulated by the CRE around rs8110787. This variant is also an eQTL of *CALR* from CAGE[23] (p = 9.4E−16) and BIOS[24] (p = 1.0E−25) and is an enhancer in K562 leukemia cells (E123) from Roadmap,[13] adding additional evidence. Our VAMPIRE successfully highlights this rs8110787-*CALR* pair in its category 1.

As a further example of the utility of the VAMPIRE application, we present the annotation results for one of the lead genome-wide significant variants from recent trans-ethnic GWAS analyses from Chen et al.[4] For our analysis, we were particularly interested in exploring low-frequency variants and those more common in those of non-European ancestry. We were able to quickly rank and prioritize variants for further examination using the annotation categories described above, including the low-frequency variant rs112097551 associated with MCV, MCH, and red blood cell count.

This low-frequency intergenic variant rs112097551 (*GATA2-RPN1* locus, 0.15% minor allele frequency in Chen et al.[4] trans-ethnic analysis) has no close linkage disequilibrium proxies in African or European populations and thus was not compared to other highly correlated variants. Based on variant frequency, particularly in European ancestry populations, we had no expectation this variant would have eQTL or pQTL evidence (2D annotation), given currently available sample sizes for eQTL and pQTL analysis. For low-frequency variants, 1D and 3D annotation would be the highest annotation category likely for a variant of interest like rs112097551. The variant is ~5× more common among African versus non-African samples in gnomAD version 2.1.1. It is the only variant in the credible set in fine-mapping analyses from Chen et al.[4] 1D annotation suggests this variant is highly conserved (CADD Phred score of 20.4, meaning the variant is among the top 1% of deleterious variants in the human genome), and it is rated as deleterious by FATHMM-XF (rank score 0.99169, close to the maximum score of 1). It is also in open chromatin in megakaryocyte-erythroid progenitor cells, based on hematopoietic ATAC-seq data.[37] 3D annotation from PCHi-C data in erythroblasts from Javierre et al.[29] links this variant to the gene *RUVBL1* ~500 kb away, as well as noncoding transcripts *RNU2-37P* and *RUVBL1-AS1*. Based on these data, which can be quickly displayed using the VAMPIRE application, we most recently validated experimentally this candidate functional enhancer variant via base and nuclease editing.[38]

Table 2. Variant-level transcription factor (TF) motif enrichment analysis

	Category 1	Uncategorized	p value	Odds ratio
All RBCT variants	5,687	21,947		
KLF1	34	34	7.10E−08	3.86
KLF6	21	14	4.30E−07	5.79
MAFB	13	13	8.10E−04	3.86
GATA1	8	19	0.18	1.63
IRF1	12	49	0.62	0.95
IRF8	19	58	0.22	1.26

Four erythroid TFs and two non-erythroid TFs were selected. Fisher's exact test was applied to test for enrichment. Three erythroid TFs show enrichment for our VAMPIRE annotation category 1 (MAFB, KLF6, KLF1, p < 0.05). GATA1 motif variants also have some evidence of enrichment (odds ratio = 1.625), but this enrichment is not significant (p = 0.18), likely due to smaller sample size of variants. Two non-hematopoiesis transcription factors selected as controls do not show significant enrichment with VAMPIRE functional annotation category 1. RBCT, red blood cell trait associated.

Table 3. Gene level enrichment analysis

	Category 1	Other categories	p value	Odds ratio
All category genes	9,857	7,408		
shRNA-assessed genes	262	83	3.50E−13	2.37
shRNA-validated genes	68	11	3.10E−08	4.65

Fisher's exact test was applied to test for enrichment. Both shRNA experiment assessed genes and validated genes show significant enrichment in our most restrictive VAMPIRE annotation category (category 1).

Comparison to FUMA

FUMA[35] is an integrative web-based platform using multiple different sources of biological evidence to facilitate functional annotation of GWAS results, gene prioritization, and interactive visualization. We compared our VAMPIRE and FUMA, in terms of the number of genes prioritized, shRNA-assessed genes, and validated genes of Figure 4, for red blood cell traits. FUMA prioritized 4,070 genes (A1 + A2 + A3), where 1,886 genes are also prioritized by VAMPIRE category 1 (A1) with an additional 769 genes in categories 2–5 of VAMPIRE (A2). The total number of genes prioritized by VAMPIRE category 1 (n = 4,832, A1 + A4) is similar to that by FUMA (n = 4,070), but that number is almost twice that of FUMA when considering all the categories of VAMPIRE (n = 7,922, A1 + A2 + A4 + A5). We evaluated the prioritized genes using data from the shRNA experiments. We first

checked genes assessed in the shRNA experiments (Figure 4B) and observed similar proportions of method-specific prioritized genes assessed. Comparing FUMA and VAMPIRE category 1, for example, out of the 2,184 (A2 + A3) FUMA-specific genes, 79 (B2 + B3) are assessed (3.6%); out of the 2,177 (A4) VAMPIRE category 1-specific genes, 84 (B4) are assessed (3.9%). We also found that shRNA-assessed genes exhibit a higher level of sharing than all genes prioritized. Again, comparing FUMA and VAMPIRE category 1, 178 (B1) out of 341 (B1 + B2 + B3 + B4) shRNA-assessed genes (52.2%) are shared between the two methods. In contrast, 1,886 (A1) out of 6,247 (A1 + A2 + A3 + A4) of all genes prioritized (30.2%) are shared. Finally, compared to FUMA, VAMPIRE category 1 led to a larger number (23 [C4] specific to VAMPIRE category 1 versus 16 [C2 + C3] specific to FUMA, Figure 4C) and larger proportion (27.4% [C4/B4] versus 20.3% [(C2 + C3)/(B2 + B3)], although not statistically significant due to small number of genes involved) of shRNA-validated genes (Figures 4B and 4C). These results suggest that VAMPIRE is complementary to FUMA, with VAMPIRE category 1 genes more likely being functional.

Discussion

As genotyped sample sizes increase and meta-analysis efforts grow ever larger, more variant-trait pairs are identified for complex traits than can be easily annotated on a variant-by-variant basis. New, user-friendly applications are needed for rapid display of functional annotation information and prioritization of variants for further functional follow-up to pave the V2F path. Our VAMPIRE tool

Table 4. Variant-gene pair level enrichment analysis

	Not significant	Significant	Significant (%)	p value	Odds ratio
All pairs from Fulco et al.[7]	3,949	175	4.24		
Confirmed pairs in category 1 for RBC traits	6	1	14.29	0.26	3.76
Confirmed pairs in category 1 for all traits	6	1	14.29	0.26	3.76
Confirmed pairs in all categories for all traits	19	7	26.92	9.00E−05	8.3
Confirmed pairs in category 1 for RBC traits (± 1 kb)	10	7	41.18	3.80E−06	15.77
Confirmed pairs in category 1 for all traits (± 1 kb)	21	9	30	3.50E−06	9.66
Confirmed pairs in all categories for all traits (± 1 kb)	70	21	23.08	4.60E−10	6.76
Confirmed pairs in category 1 for RBC traits (± 5 kb)	27	20	42.55	3.10E−15	16.68
Confirmed pairs in category 1 for all traits (± 5 kb)	64	23	26.44	3.80E−12	8.1
Confirmed pairs in all categories for all traits (± 5 kb)	160	37	18.78	3.10E−13	5.21

We performed analysis for three variant annotation pools (category 1, red blood cell [RBC] trait associated; category 1, any blood cell trait associated; any annotation priority category (1–5), any blood cell trait associated) and three CRE lengths. Fisher's exact test was applied to test for enrichment. We found enrichment for all three variant annotation pools. These enrichments are also robust to the extension of CREs.
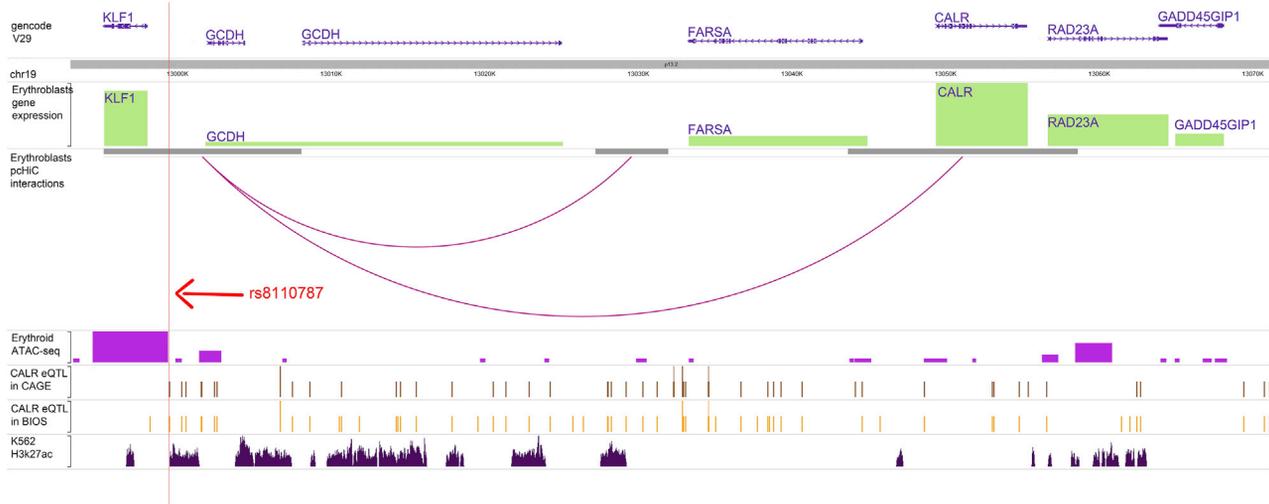
Figure 3. Variant-gene pair example (rs8110787-*CALR*) visualization from HUGIn2[31]
Fulco et al.[7] confirmed via CRISPRi experiments that chr19: 12,996,905–12,998,745 (hg19) regulates gene *CALR* (adjusted p value, 1.9E−7), which is highly expressed in erythroblasts. Based on annotations in VAMPIRE, *CALR* is linked to rs8110787 (chr19: 12,999,458, hg19) in prioritization category 1, including higher-than-expected physical interactions with the *CALR* locus from erythroblast pcHiC data,[29] eQTL of *CALR* in CAGE[23] and BIOS,[24] erythroid ATAC-seq peak,[16] and H3K27ac peak in K562 leukemia cells.[13] rs8110787 is associated with several red blood cell traits (namely hematocrit [HCT], mean corpuscular hemoglobin [MCH], mean corpuscular volume [MCV], and red blood cell count), as reported in Chen et al.[4]

provides an example of how the publicly available code can be adapted to accommodate other sources of annotation specific to other complex trait GWAS results or to accommodate future blood cell trait GWASs and annotation resources. Along with the addition of more blood cell trait genetics papers published in the future, VAMPIRE could also be used as written to annotate GWAS results for other blood-related phenotypes, such as recent GWASs of risk of myeloproliferative neoplasm or clonal hematopoiesis.[39,40]

For non-coding variants, we group them in five categories, and we have the following conclusions and observations in terms of variant prioritization. First, category 1 is the most restrictive category, and variants in category 1 are more likely to be functional than those in the other categories. Second, beyond category 1, we only found category 2 shows enhanced functional potential over category 4, while there are no strong preferences among the other categories. We have performed both variant-level (Table S1) and gene-level (Table S2) enrichment analyses comparing categories 2–5 and found no significant results, except for category 3. This may suggest that category 3 is slightly more likely to contain functional variants than categories 2, 4, and 5. However, the evidence is not strong: the significance levels are not very high, other TF motifs are not enriched, and sample size (i.e., number of genes tested) is small. Third, variant frequency information can also be helpful in interpreting eQTL/pQTL data. For a low frequency or rare variant, power is low in current eQTL/pQTL studies with small to moderate sample sizes. Thus, the absence of 2D evidence most likely reflects the power issue and should be treated as eQTL/pQTL not having been adequately assessed rather than truly not associ-

ated with the expression of gene(s) or the abundances of protein(s). Finally, different annotations have different weights depending on the trait of interest. For instance, annotations from megakaryocytes are critically important for platelet-related traits but can be rather safely ignored for red blood cell-related traits. Investigators focusing on different traits should use their discretion to up-/down-weigh various annotations.

There are several reasons that a variant does not show up in the current VAMPIRE. First, we only included variants in credible sets from the recent GWAS efforts for blood cell traits.[3,4] Variants not in those fine-mapping credible sets were not annotated. It is possible that such variants play functional roles but were not detected by GWASs and further missed by subsequent fine mapping. However, the probability tends to be low, particularly for common causal variants, given the >750,000 sample size involved in the generation of the credible sets. Second, for the included variants (i.e., credible set variants for blood cell traits), not falling in the prioritization categories (e.g., uncategorized) means that they are less likely to play functionally important roles compared to variants in categories 1–5, because no regulatory evidence or target genes are suggested based on the functional annotation information we have. Of course, it is possible that some of these uncategorized variants are indeed functional, but their functions are not reflected by the functional data we currently have.

As we accumulate additional functional validation data, including high-throughput massively parallel reporter assays (MPRAs), medium-throughput CRISPRi/CRISPRa, and low-throughput mouse xenotransplant experiments, we will provide statistics summarizing experimental validation results (e.g., number of variants in the category
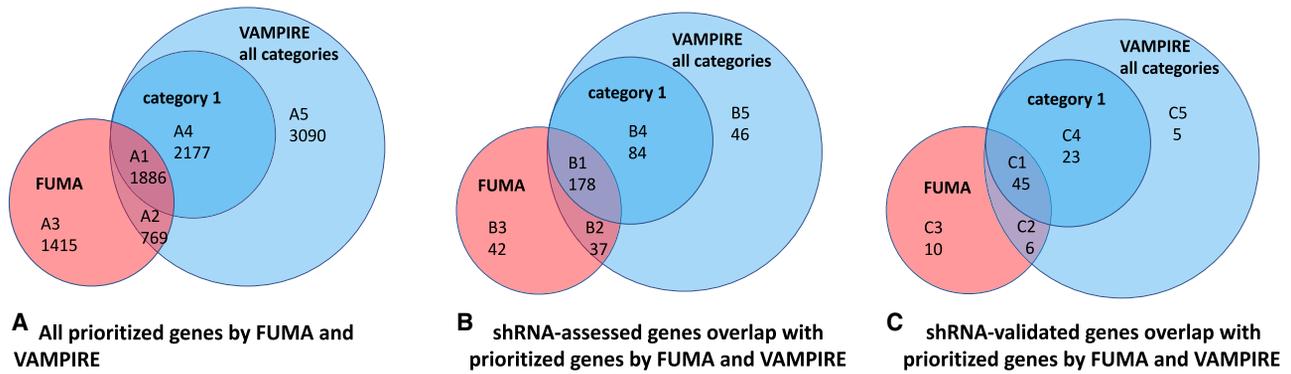
Figure 4. Venn diagrams comparing FUMA and VAMPIRE
(A) All prioritized genes by FUMA and VAMPIRE. (B) shRNA-assessed genes overlap with genes prioritized by FUMA and VAMPIRE. (C) shRNA-validated genes overlap with genes prioritized by FUMA and VAMPIRE. In each panel, numbers are the number of genes belonging to the corresponding category. A1: shared between FUMA and VAMPIRE category 1; A2: shared between FUMA and VAMPIRE other categories (categories 2–5); A3: FUMA-specific genes; A4: VAMPIRE category 1-specific genes; A5: VAMPIRE other category-specific genes. Similar interpretation for B1–B5, C1–C5.

followed up, proportion that show evidence of functional impact in their experiments) for each of the VAMPIRE categories and for user-defined categories. Importantly, we illustrate the value of VAMPIRE using existing independent functional validation and therefore illuminate the value of this type of annotation tool in enabling one to go from variant to function for blood cell traits and other complex phenotypes.

We also note that there are some limitations of VAMPIRE. First, comprehensive annotations specific to various cell types and cell states would further enhance classification and prioritization accuracy of functional variants or regulatory elements and their target genes. Although data are increasingly being generated by us[15,16] and others,[29,36] and have been incorporated into VAMPIRE where available, interrogations in a cell-type- or state-specific manner are still much needed. For instance, our recent work has demonstrated cell-type or tissue-specific FIREs[17,41] and super-interactive promoters (SIPs)[42] play key regulatory roles and aid the identification and prioritization of functional regulatory elements and their corresponding genes. As more experimental data are generated, we will update VAMPIRE accordingly. Second, our list of 148,019 variants derives primarily from fine-mapping studies, which may be inaccurate in loci where more than one independent or partially independent signal exists. However, this limitation cannot be resolved before more powerful methods are developed for fine-mapping analysis for trans-ethnic GWASs. Finally, most of the annotations are based on analyses in European ancestry individuals (e.g., eQTL, pQTL, chromatin conformation, etc.). Many ongoing efforts, including ours, are generating resources for non-European ancestry samples. For example, we are involved in several recently funded efforts to generate RNA-sequencing data in non-European ancestry individuals in hematopoietic cell types and anticipate relevant eQTL and sQTL annotations being added to VAMPIRE in upcoming years.

In conclusion, we have built a comprehensive annotation tool, VAMPIRE, which provides characterization and prioritization of blood cell trait-related GWAS signals. Our results using existing functional experiments demonstrate that variants and genes prioritized by VAMPIRE are significantly more likely to be functionally validated at either the variant, gene, or variant-gene pair level. Annotation tools like VAMPIRE, which could be easily modified to apply to additional complex traits and diseases, are necessary to translate knowledge of GWAS-significant variants to target genes and biological insights and to guide our decisions to prioritize experimental validations of most likely functional regulatory variants/elements and their effector genes.

Data and code availability

The data underlying this article are available in the article and in its online supplemental material.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2021.100063.

Declaration of interests

The authors declare no competing interests.

Web resources

GWAS Catalog, https://www.ebi.ac.uk/gwas/
GWAS summary statistics from Chen et al.,[4] http://www.mhi-humangenetics.org/en/resources/
HUGIn, http://hugin2.genetics.unc.edu/Project/hugin/
PheWAS website, http://pheweb.sph.umich.edu
VAMPIRE, http://shiny.bios.unc.edu/vampire/

References

1. Liggett, L.A., and Sankaran, V.G. (2020). Unraveling Hematopoiesis through the Lens of Genomics. Cell *182*, 1384–1400.

2. 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.; VA Million Veteran Program (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell *182*, 1214–1231.e11.

3. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Hematology & Hemostasis Working Group (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. *15*, e1008500.

4. Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al.; VA Million Veteran Program (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198–1213.e14.

5. Nandakumar, S.K., Liao, X., and Sankaran, V.G. (2020). In The Blood: Connecting Variant to Function In Human Hematopoiesis. Trends Genet. *36*, 563–576.

6. Garner, C., Tatu, T., Reittie, J.E., Littlewood, T., Darley, J., Cervino, S., Farrall, M., Kelly, P., Spector, T.D., and Thein, S.L. (2000). Genetic influences on F cells and other hematologic variables: a twin heritability study. Blood *95*, 342–346.

7. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat. Genet. *51*, 1664–1669.

8. Nandakumar, S.K., McFarland, S.K., Mateyka, L.M., Lareau, C.A., Ulirsch, J.C., Ludwig, L.S., Agarwal, G., Engreitz, J.M., Przychodzen, B., McConkey, M., et al. (2019). Gene-centric functional dissection of human genetic variation uncovers regulators of hematopoiesis. eLife *8*, e44080.

9. Liu, X., White, S., Peng, B., Johnson, A.D., Brody, J.A., Li, A.H., Huang, Z., Carroll, A., Wei, P., Gibbs, R., et al. (2016). WGSA: an annotation pipeline for human genome sequencing studies. J. Med. Genet. *53*, 111–112.

10. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) *2017*, bax028.

11. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al.; FANTOM consortium (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22.

12. Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C.C., de Hoon, M., Severin, J., Oki, S., Hayashizaki, Y., et al. (2019). Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. Nucleic Acids Res. *47* (D1), D752–D758.

13. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045–1048.

14. Consortium, E.P.; and ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

15. Ludwig, L.S., Lareau, C.A., Bao, E.L., Nandakumar, S.K., Muus, C., Ulirsch, J.C., Chowdhary, K., Buenrostro, J.D., Mohandas, N., An, X., et al. (2019). Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis. Cell Rep. *27*, 3228–3240.e7.

16. Ulirsch, J.C., Lareau, C.A., Bao, E.L., Ludwig, L.S., Guo, M.H., Benner, C., Satpathy, A.T., Kartha, V.K., Salem, R.M., Hirschhorn, J.N., et al. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat. Genet. *51*, 683–693.

17. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., and Ren, B. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. *17*, 2042–2059.

18. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

19. Schrider, D.R., and Kern, A.D. (2016). S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. PLoS Genet. *12*, e1005928.

20. Schrider, D.R., and Kern, A.D. (2017). Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Mol. Biol. Evol. *34*, 1863–1877.

21. Emilsson, V., Ilkov, M., Lamb, J.R., Finkel, N., Gudmundsson, E.F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S.R., Aspelund, T., et al. (2018). Co-regulatory networks of human serum proteins link genetics to disease. Science *361*, 769–773.

22. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv. https://doi.org/10.1101/447367.

23. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. Am. J. Hum. Genet. *100*, 228–237.

24. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.J., et al. (2017). Identification of context-

dependent expression quantitative trait loci in whole blood. Nat. Genet. *49*, 139–145.

25. Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science *344*, 519–523.

26. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

27. Gorkin, D.U., Qiu, Y., Hu, M., Fletez-Brant, K., Liu, T., Schmitt, A.D., Noor, A., Chiou, J., Gaulton, K.J., Sebat, J., et al. (2019). Common DNA sequence variation influences 3-dimensional conformation of the human genome. Genome Biol. *20*, 255.

28. Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. *24*, 999–1011.

29. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al.; BLUEPRINT Consortium (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369–1384.e19.

30. Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. *51*, 1442–1449.

31. Martin, J.S., Xu, Z., Reiner, A.P., Mohlke, K.L., Sullivan, P., Ren, B., Hu, M., and Li, Y. (2017). HUGIn: Hi-C Unifying Genomic Interrogator. Bioinformatics *33*, 3793–3795.

32. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. *31*, 1102–1110.

33. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47* (D1), D1005–D1012.

34. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.

35. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. *8*, 1826.

36. Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat. Genet. *49*, 1602–1612.

37. 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.

38. Hu, Y., Stilp, A.M., McHugh, C.P., Rao, S., Jain, D., Zheng, X., Lane, J., Méric de Bellefon, S., Raffield, L.M., Chen, M.-H., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2021). Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. Am. J. Hum. Genet. *108*, 874–893.

39. Bick, A.G., Weinstock, J.S., Nandakumar, S.K., Fulco, C.P., Bao, E.L., Zekavat, S.M., Szeto, M.D., Liao, X., Leventhal, M.J., Nasser, J., et al.; NHLBI Trans-Omics for Precision Medicine Consortium (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature *586*, 763–768.

40. Bao, E.L., Nandakumar, S.K., Liao, X., Bick, A.G., Karjalainen, J., Tabaka, M., Gan, O.I., Havulinna, A.S., Kiiskinen, T.T.J., Lareau, C.A., et al.; FinnGen; and 23andMe Research Team (2020). Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. Nature *586*, 769–775.

41. Crowley, C., Yang, Y., Qiu, Y., Hu, B., Abnousi, A., Lipiński, J., Plewczyński, D., Wu, D., Won, H., Ren, B., et al. (2020). FIRE-caller: Detecting frequently interacting regions from Hi-C data. Comput. Struct. Biotechnol. J. *19*, 355–362.

42. Song, M., Pebworth, M.P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J.D., Choudhary, M.N.K., Cui, X., Jones, I.R., et al. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. Nature *587*, 644–649.