VAMPIRE: A model blood cell trait focused annotation tool for interpretation and prioritization of complex trait genome-wide association study results.

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

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, Quebec, Canada

6. Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, 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

*These authors contributed equally

Abstract

Thousands of genetic loci have been identified as associated with hematological indices (red blood cell, white blood cell, and platelet related traits), as well as other complex traits and disease. However, most loci identified are noncoding and not clearly linked to target genes, and tools are needed to prioritize the most likely functional variants for experimental follow-up. We here describe VAMPIRE: Variant Annotation Method Pointing to Interesting Regulatory Effects, an interactive web application implemented in R Shiny (http://shiny.bios.unc.edu/vampire/) for blood cell trait associated loci from recent large multi-ethnic genome-wide association studies (GWAS). This tool efficiently displays information from blood cell relevant tissues on epigenomic signatures, functional and conservation summary scores, 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. Variants are classified into multiple prioritization categories according to these functional signatures. Leveraging data generated from independent functional validation experiments, we demonstrate that our prioritized variants are enriched within experimentally validated variant sets. VAMPIRE allows rapid prioritization and interpretation of blood cell trait GWAS variants and could be easily adapted for use with other complex trait GWAS results and extended to new annotation sources.

Author Summary

Many large genome-wide association studies (GWAS) have recently been performed for blood cell traits, with thousands of associations identified. However, most of the associated variants are in noncoding regions and are often hard to interpret, link to genes, and prioritize for functional follow-up. Similar challenges exist for genetic studies of many other traits and diseases. Trying to translate knowledge of GWAS significant variants to target genes and biological insights, we here describe VAMPIRE: Variant Annotation Method Pointing to Interesting Regulatory Effects, an interactive web application implemented in R Shiny (http://shiny.bios.unc.edu/vampire/) for blood cell trait associated loci from recent large multi-ethnic GWAS. This tool displays a variety of information including epigenomic signatures, variant impact on protein and gene expression, chromatin conformation information, and publicly available GWAS and phenome-wide association study (PheWAS) results for other traits. We classified variants into annotation categories using this information, and show that variants in the highest priority categories are enriched in likely causal variant sets from previous functional experiments. We anticipate this tool will guide appropriate variants to prioritize for experimental validation for researchers studying blood cell traits, as well as providing an easily adaptable model for the creation of similar annotation tools for other complex traits and diseases.

67 **Introduction**

68 Hematological phenotypes (red blood cell, white blood cell, and platelet counts and indices) are critical
69 physiological intermediaries in oxygen transport, immunity, infection, thrombosis, and hemostasis and are
70 associated with autoimmune, allergic, infectious, and cardiovascular diseases. Hematological traits are
71 highly heritable [1], and recent large GWAS for hematological traits (including nearly 750,000
72 participants) identified thousands of variant-trait associations [2, 3]. With extensive sequencing and
73 GWAS efforts, there is a need to convert large lists of significant variant-trait pairs into human-
74 interpretable knowledge [4]. Most variants identified through GWAS are noncoding (>95% for blood cell
75 traits [2]), and most signals include multiple highly correlated variants. Pinpointing the likely causal
76 variants within GWAS signals, and linking these variants to their target genes, is challenging.

77 VAMPIRE: Variant Annotation Method Pointing to Interesting Regulatory Effects provides a tool for the
78 user to explore annotations encompassing epigenomic signatures, variant impact on protein and gene
79 expression, chromatin conformation information from Hi-C and similar technologies, as well as publicly
80 available GWAS and PheWAS results, creating a comprehensive annotation profile for variants from
81 recent trans-ethnic blood cell trait publications [3, 5] with a flexible interface for adding additional future
82 GWAS results. This interactive web application implemented in R Shiny
83 (http://shiny.bios.unc.edu/vampire/) provides a model display mechanism for annotating GWAS variants
84 from diverse complex traits, allowing selection of most likely causal variants and their effector genes for
85 experimental follow-up. Importantly, we show the value of how variants and genes nominated by
86 VAMPIRE can highlight key regulators of blood cell traits using independent functional assessment,
87 confirming the value of this annotation tool. While blood cell traits are the focus for VAMPIRE, this
88 framework is adaptable for annotation of other complex trait GWAS results and will facilitate the
89 connection between variant and function.

90

Results

Overview of VAMPIRE annotations

We started with all variants in 95% credible sets from our recent trans-ethnic study for hematological traits (total 148,019 variants) [3] and lead variants (2 variants) from Kowalski et al. [5]. We incorporated six types of annotations (detailed in **Methods**): GWAS summary statistics and posterior probability of inclusion from our previous fine-mapping analyses [3]; epigenomic or sequence constraints features (1D); eQTL, sQTL and pQTL information (2D); information on 3D genome conformation (3D); results from two PheWAS sources [3, 28] (PheWAS); and GWAS results from blood cell traits from GWAS catalog [29] (GWAS).

To visualize and prioritize variants, their corresponding candidate regulatory regions, and their potential effector genes, we leverage the aforementioned six types of annotation to group these ~148,000 variants into various prioritization categories. Specifically, for non-coding variants, we classified them into five categories (detailed in Methods). Among them, category 1 is the most restrictive category, containing variants that have 1D, 2D, and 3D annotation and the genes implicated by 2D and 3D evidence are consistent. Variants not falling into any of the five categories are classified as uncategorized. In addition, 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.

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, we found significant enrichment of variants affecting TF binding motifs among variants prioritized in category 1 of VAMPIRE (**Fig 1**), for all the erythroid TFs ($p < 8.1E-4$) but GATA1 ($p = 0.18$) (**Table 1**), likely due a smaller sample size of variants. In contrast, neither of the two

114 negative control TFs (IRF1 and IRF8) showed any significant enrichment (p = 0.22 and 0.62). At the gene
115 level, we focused on two statistics: (1) number of genes selected for shRNA experiments, since genes
116 were more likely to be selected for experiments when they demonstrated some prior evidence of potential
117 causality, and (2) number of genes validated (p < 0.05) by shRNA experiments. We compared the number
118 of genes in our annotation category 1 and all other categories, and found that both shRNA candidate
119 genes (p = 3.5E-13) and significant genes (p = 3.1E-8) show strong enrichment among those in our
120 annotation category 1 (**Table 2**), and the estimated enrichment score for significant genes (OR = 4.65) is
121 almost double of that for candidate genes (OR = 2.37). These results suggest the genes prioritized by
122 VAMPIRE's category 1 annotations are more likely to be functional.

	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

123

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

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

131 **Table 2. Gene level enrichment analysis.** Fisher's exact test was applied to test for enrichment. Both
132 shRNA experiment candidate genes and validated genes show significant enrichment in our most
133 restrictive VAMPIRE annotation category (category 1).

134 Finally, at the variant-gene pair level, we also observed enrichment among variants selected into
135 VAMPIRE's category 1 (**Table 3)**. When restricting only to variants in category 1 and associated with red
136 blood cell traits and without extending the CRE regions, only 7 of VAMPIRE's variant-gene pairs can be
137 found in ABC's CRISPRi-FlowFISH experiments, of which 6 are not significant and 1 is significant.
138 While not significant (p = 0.26), the direction of enrichment is nevertheless encouraging (one of seven, or
139 14.3%, confirmed by CRISPRi-FlowFISH experiments) and 3-fold greater than that among
140 all/background pairs from Fulco et al. [33], where 175 out of 4124 variant-gene pairs (4.2%) were
141 confirmed. Note that all the confirmed pairs were linked with variants associated with red blood cell traits.
142 Further generalizing to all VAMPIRE annotation categories and to variants associated with any blood cell
143 trait, the enrichment OR increases to 8.30 with p-value 9.0E-5, indicating that variant-gene pairs
144 prioritized by VAMPIRE's five categories have much higher odds of being functional. To further
145 accommodate causal variants tagged by GWAS variants not falling into the short 500bp CREs, we
146 extended the CREs by +/- 1kb or +/- 5kb, and performed similar enrichment analysis. Our conclusions
147 remained qualitatively similar (**Table 3**), but the enrichments increased in significance, thanks to larger
148 sample size (in this context, the larger number of variant-gene pairs contributing to the analysis) and
149 suggesting that more liberal windows of *cis*-regulatory regions can capture a higher rate of functional
150 variant-gene pairs. For example, the enrichment for category 1 variants associated with red blood cell
151 (RBC) traits reached an OR of 15.77 (p=3.8E-6) and 16.68 (p=3.1E-15) for 1kb and 5kb extension,
152 respectively. We thus conclude that such enrichment is significant and robust to the extension of CREs.

	Not significant	Significant	Significant %	p-value	Odds ratio
All pairs from Fulco et al.	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 (+/- 1kb)	10	7	41.18	3.80E-06	15.77
Confirmed pairs in category 1 for all traits (+/- 1kb)	21	9	30	3.50E-06	9.66
Confirmed pairs in all categories for all traits (+/- 1kb)	70	21	23.08	4.60E-10	6.76
Confirmed pairs in category 1 for RBC traits (+/- 5kb)	27	20	42.55	3.10E-15	16.68
Confirmed pairs in category 1 for all traits (+/- 5kb)	64	23	26.44	3.80E-12	8.1
Confirmed pairs in all categories for all traits (+/- 5kb)	160	37	18.78	3.10E-13	5.21

153 **Table 3. Variant-Gene pair level enrichment analysis.** We performed analysis for three variant
154 annotation pools (category 1, red blood cell (RBC) trait associated; category 1, any blood cell trait
155 associated; any annotation priority category (1-5), any blood cell trait associated) and three CRE lengths.
156 Fisher's exact test was applied to test for enrichment. We found enrichment for all three variant
157 annotation pools. These enrichments are also robust to the extension of CREs.

158 **Application example**

159 **Fig 2** shows one example at the *CALR* locus associated with red blood cell traits. Fulco et al. confirmed
160 by CRISPRi-FlowFISH experiment that CRE chr19:12,996,905-12,998,745 (hg19) regulates gene *CALR*
161 (adjusted p-value 1.9E-7). Annotations compiled by VAMPIRE suggest, consistently, that *CALR* is linked
162 to rs8110787 (chr19:12,999,458, hg19) in category 1. rs8110787 is associated with several RBC traits [3],

163 including hematocrit (HCT), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV)
164 and red blood cell counts (RBC). Based on genomic distance alone, *CALR* is not the nearest gene to
165 rs8110787, with several other closer genes. However, based on H3K27ac HiChIP data in K562 cells [34],
166 rs8110787 significantly interacts with *CALR* promoter region ($p < 1E-120$), suggesting that *CALR* is a
167 potential target gene regulated by the CRE around rs8110787. This variant is also an eQTL of *CALR* from
168 CAGE [19] ($p = 9.4E-16$) and BIOS [20] ($p = 1.0E-25$), and is an enhancer in K562 Leukemia cells
169 (E123) from Roadmap [10], adding additional evidence. Our VAMPIRE successfully highlights this
170 rs8110787-*CALR* pair in its category 1.

171 As a further example of the utility of the VAMPIRE application, we present the annotation results for one
172 of the lead genome-wide significant variants from recent trans-ethnic GWAS analyses from Chen et al.
173 [3] For our analysis, we were particularly interested in exploring low frequency variants, and those more
174 common in those of non-European ancestry. We were able to quickly rank and prioritize variants for
175 further examination using the annotation categories described above, including the low frequency variant
176 rs112097551 associated with mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH),
177 and red blood cell count.

178 This low frequency intergenic variant rs112097551 (*GATA2-RPN1* locus, 0.15% minor allele frequency
179 in Chen et al. trans-ethnic analysis [3]) has no close linkage disequilibrium proxies in African or
180 European populations, and thus was not compared to other highly correlated variants. Based on variant
181 frequency, particularly in European ancestry populations, we had no expectation this variant would have
182 eQTL or pQTL evidence (2D annotation), given currently available sample sizes for eQTL and pQTL
183 analysis. For low frequency variants, 1D and 3D annotation would be the highest annotation category
184 likely for a variant of interest like rs112097551. The variant is ~5x more common among African versus
185 non-African samples in gnomAD version 2.1.1. It is the only variant in the credible set in fine-mapping
186 analyses from Chen et al. 1D annotation suggests this variant is highly conserved (CADD Phred score of
187 20.4, meaning the variant is amongst the top 1% of deleterious variants in the human genome), and it is

188 rated as deleterious by FATHMM-XF (rank score 0.99169, close to the maximum score of 1). It is also in
189 open chromatin in megakaryocyte–erythroid progenitor cells, based on hematopoietic ATAC-seq data
190 [38]. 3D annotation from PCHi-C data in erythroblasts from Javierre et al.[25] links this variant to the
191 gene *RUVBL1* ~500Kb away, as well as noncoding transcripts *RNU2-37P* and *RUVBL1-AS1*. Based on
192 this data, which can be quickly displayed using the VAMPIRE application, we are currently working on
193 *in vitro* follow-up of this candidate functional enhancer variant [36].

194

195 **Discussion**

196 As genotyped sample sizes increase and meta-analysis efforts grow ever larger, more variant-trait pairs
197 are identified for complex traits than can be easily annotated on a variant by variant basis. New, user-
198 friendly applications are needed for rapid display of functional annotation information and prioritization
199 of variants for further functional follow-up. Our VAMPIRE tool provides an example of how; the
200 publicly available code can be adapted to accommodate other sources of annotation specific to other
201 complex trait GWAS results or to accommodate future blood cell trait GWAS and annotation resources.
202 In addition to *a priori* providing one category of coding variants and 5 categories of non-coding variants
203 that warrant prioritization consideration, VAMPIRE allows users to decide their own categories based on
204 arbitrary combinations of the annotations at adjustable thresholds (for example, prioritizing high CADD
205 score variants, or variants in open chromatin in blood cells based on ATAC-seq). Along with the addition
206 of more blood cell trait genetics papers published in the future, VAMPIRE could also be used as written
207 to annotate GWAS results for other blood related phenotypes, such as recent GWAS of risk of
208 myeloproliferative neoplasm or clonal hematopoiesis [6, 7].

209 As we accumulate additional functional validation data, including high-throughput massively parallel
210 reporter assays (MPRA), medium-throughput CRISPRi/CRISPRa and low throughput mouse
211 xenotransplant experiments, we will provide statistics summarizing experimental validation results (e.g.,

212 number of variants in the category followed-up, proportion that show evidence of functional impact in
213 their experiments) for each of the 6 VAMPIRE categories and for user defined categories. Importantly,
214 we illustrate the value of VAMPIRE using existing independent functional validation and therefore
215 illuminate the value of this type of annotation tool in enabling one to go from variant to function for blood
216 cell traits and other complex phenotypes. Our results using existing functional experiments suggests
217 significant enrichment for variants and genes highlighted by VAMPIRE among functionally validated
218 regulatory elements. In conclusion, annotation tools like VAMPIRE are necessary to translate knowledge
219 of GWAS significant variants to target genes and biological insights, and to guide appropriate variants to
220 prioritize for functional validation.

221

222 **Methods**

223 **Variant Annotations**

224 The current version of VAMPIRE includes GWAS results from two studies (as detailed in Supplemental
225 Methods), including all variants in 95% credible sets for fine-mapped hematological trait associated loci
226 from Chen et al. (N=148,019 variants) [3] and lead variants (N=2) from a TOPMed imputed GWAS
227 meta-analysis in African American and Hispanic/Latino populations [5]. We plan to extend VAMPIRE as
228 new trans-ethnic blood cell trait genetic analyses are released.

229 The sources of the annotation used are stated clearly in the VAMPIRE online application, with links or
230 references to the original data sources. As a brief summary, the annotation categories are trivially split
231 into six types ("variant level", "1D","2D","3D", "PheWAS","GWAS"). First, "variant level" contains data
232 on phenotypic association from the original publication or preprint (such as the p-value for association
233 with a given hematological trait, effect size, and posterior probability of inclusion for fine-mapping
234 credible sets). Second, "1D" refers to epigenomic or sequence constraints features. This displays selected

235 output from WGSA [8] including functional prediction scores, conservation scores, and epigenetic
236 information gathered from GeneHancer [9], FANTOM5 [10, 11], Roadmap [12], and ENCODE [13].
237 ATAC-seq peaks from recent studies for blood cell traits [14, 15] and key histone ChIP-seq peaks such as
238 H3K9me3, H3K36me3, H3K4me1, H3K4me3, and H3K27Ac generated across blood cell related tissues
239 from Roadmap Epigenomics are also included [12, 16]. We further include information regarding whether
240 each variant resides in any selective sweep region detected from multiple populations in the 1000
241 Genomes Project [17] using the S/HIC method [18, 19]. Information is displayed based on the tissue
242 relevance to the blood cell phenotype (see Supplemental Methods). All variants have 1D annotation, but
243 for prioritization purposes as described below in the five categories for noncoding variant annotation, we
244 define 1D annotation as FANTOM5_enhancer_robust =Y (yes), or Genehancer_feature="Promoter" or
245 "Enhancer" or "Promoter/Enhancer", or coreMarks (for any relevant roadmap epigenomic category) =
246 "Enhancers" or "Active TSS." Users can then additionally filter by criteria such as functional prediction
247 and conservation scores.

248 For the "2D" annotations, we included impact on gene expression and splicing ratios (eQTL and sQTL
249 information) and impact on protein abundance (pQTL information [37]) from public sources relevant to
250 blood cell traits. This includes both bulk and cell type specific sources from the public domain (eQTLGen
251 [20], CAGE [21], BIOS [22] for whole blood, and Raj et al for purified CD4+ T cells and monocytes
252 [23]). Information available in these sources varies, but generally we at a minimum display the effect size
253 estimate, p-value, the allele assessed, and the gene or protein involved. Variants were matched across
254 sources based on chromosome, position, and alleles of each variant. Only significant results (based on
255 FDR or other publication specific thresholds) from the respective sources are displayed in VAMPIRE; we
256 do note that formal co-localization analyses would still need to be performed to determine if blood cell
257 related and gene/protein expression QTL signals truly coincide.

258 For the "3D" annotations, we include information on 3D genome conformation, linking blood lineage
259 specific regulatory elements to target genes from various sources. More specifically, using Hi-C data we

260 incorporated statistically significant long-range chromatin interactions (LRCI) [16, 24, 25] calculated
261 from Fit-Hi-C [26], loops using the HiCCUPs methodology [24], and super-FIREs for related tissues [16].
262 Two Promoter-Capture Hi-C (PCHi-C) data sources [27, 28] were also incorporated and matched with the
263 2D results to highlight consistent evidence regarding the affected gene(s) across "2D" and "3D"
264 annotations. VAMPIRE displays information on the number of loops, LRCI, PCHi-C interactions, FIREs,
265 or super-FIREs, as well as significance measures such as p-values, FDR, or CHICAGO scores where
266 applicable. This "3D" annotation information can also be visualized via our HUGIn browser [29].

267 The last two data groups present results from two PheWAS sources [3, 30] and GWAS results of blood
268 cell traits from GWAS catalog [31], allowing the user to evaluate if hematological trait associated variants
269 may also influence other complex traits.

270 To visualize and leverage these multiple annotation categories for further analysis or prioritization of
271 experimental validations, VAMPIRE efficiently displays and integrates relevant variant information,
272 allowing the user to investigate either all the variants annotated or subsets based on annotation category
273 groupings, searching either by variant or by gene name. The comprehensive annotation for the variants is
274 summarized using a five category grouping created for highlighting the most promising variants as they
275 have various types of annotation. Specifically, the five categories for noncoding variants are (1) the most
276 restrictive category, containing variants that have 1D, 2D, and 3D annotation and the genes implicated by
277 2D and 3D evidence are consistent; (2) containing variants with 1D, 2D, and 3D evidence, but the genes
278 implicated from different resources are not consistent; (3) 2D and 3D with consistent gene evidence
279 between the 2D and 3D annotations; (4) variants with 2D and 3D information and no consistent gene
280 implied; (5) variants with 1D and 3D evidence. We also have a predicted high impact coding variant
281 category displayed, including high confidence loss of function (LoF) variants and likely influential
282 missense, in frame indels, and synonymous variants. Variants without strongly compelling variant
283 annotation are still displayed, but are not listed in these high priority categories. The user can further
284 subset results by hematological trait, hematological trait category, or (for the Chen et al paper [3]) the

285 ancestry specific grouping in which a given credible set was derived (trans-ethnic, European, East Asian,
286 South Asian, Hispanic/Latino, or African ancestry). In addition, the user can restrict the amount of
287 information presented by selecting which tables to be displayed. All tables can be exported in a csv or tab
288 delimited format.

289 **Enrichment analysis**

290 To assess whether the variants prioritized by VAMPIRE are more likely to be functionally impactful, we
291 performed enrichment analysis at three different levels: variant level, gene level, and variant-gene pair
292 level, leveraging data generated from previously published functional experiments [2, 32, 33]. For each
293 set of analysis, we conducted Fisher's exact test and calculated odds ratios (OR) and one-sided p-values.

294 At the variant level, we assessed the enrichment of variants that modify transcription factor (TF) binding
295 motif [2] among our annotation category 1 variants. Recently, Vuckovic et al. [2] characterized variants
296 that affect erythropoiesis or hematopoiesis by modifying related TF motifs, such as for KLF1, KLF6,
297 MAFB, and GATA1. We chose these four erythroid TFs as positive control TFs and two non-erythroid
298 TFs (IRF1 and IRF8) as negative controls.

299 At the gene level, we evaluated the genes interrogated by Nandakumar et al. [32] with a pooled short
300 hairpin RNA (shRNA) based loss-of-function approach. Specifically, Nandakumar et al. studied 389
301 candidate genes in the neighborhood of 75 loci associated with red blood cell traits [35], to identify
302 potential causal genes underlying these GWAS signals. We assessed the enrichment of genes validated by
303 shRNA experiments among those prioritized in VAMPIRE's category 1. Note that the categories were
304 previously defined at variant level. Here we extent variant category to gene category as the strongest
305 category where a genome-wide significant variant linked to this gene falls in.

306 At the variant-gene pair level, we employed the enhancer-gene connections validated via CRISPRi-
307 FlowFISH experiments by Fulco et al. [33] in their activity-by-contact (ABC) paper. Specifically, Fulco

et al. tested pairs of candidate *cis* regulatory elements (CREs, ~500bp regions) and their potential effector genes via CRISPRi perturbations of the CREs, in multiple cell lines including the K562 cells. Fulco et al. 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) overlaps 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 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 (~500bp), we also performed sensitivity analysis by first extending the CRE regions by +/- 1kb and +/- 5kb and then overlapping variants with these extended CREs, to ensure robust conclusions.

Acknowledgement

We would like to thank the Blood Cell Consortium (BCX) and the HemeNet investigators for comments on earlier versions of VAMPIRE. We also want to thank the Li lab members for feedback on the R Shiny app.

Financial Disclosure Statement

The project described is supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through R01HL146500 (APR), R01HL129132 (YL), KL2TR002490 (LMR),

R01DK103794 (VGS), and the New York Stem Cell Foundation. CC and YL are also partially supported by R01GM105785 and U01DA052713. VGS is a New York Stem Cell Foundation-Robertson Investigator. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Figures

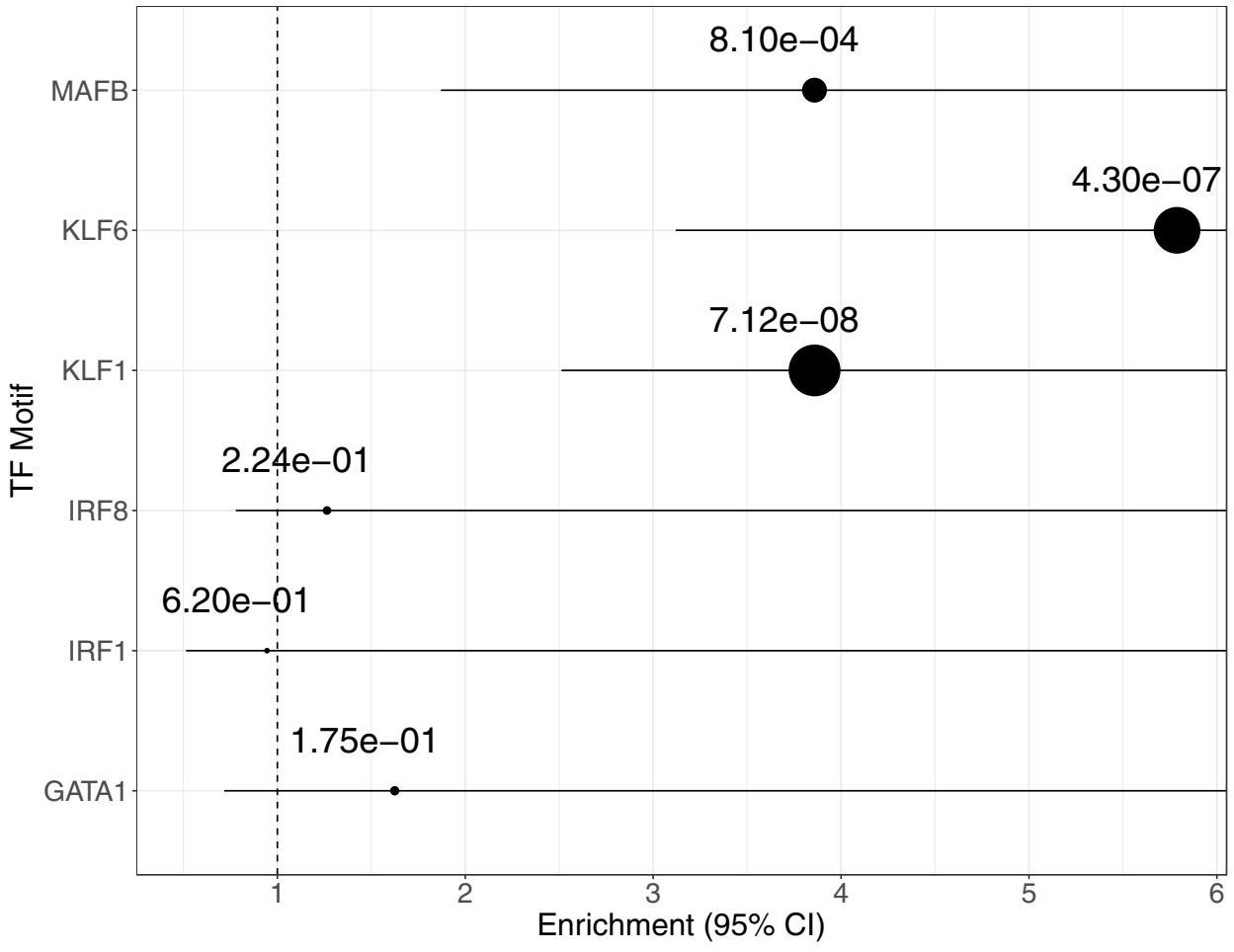

Fig 1. Variant level TF motif enrichment analysis. Each dot represents an enrichment score with the line depicting 95% confidence interval (CI). All the upper bounds of these CIs are infinity. The p-values of the enrichment are reflected by the dot size at the OR point estimate with a larger dot indicating more significant the enrichment.

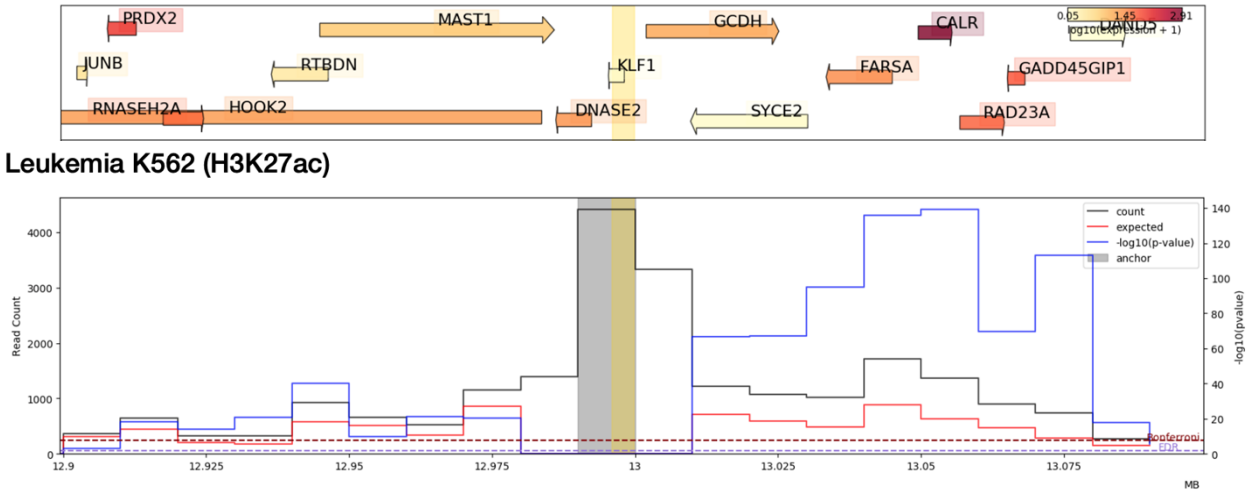

Fig 2. Variant-gene pair example (rs8110787-*CALR*) visualization from HUGIn2 [27]. Fulco et al. confirmed via CRISPRi experiments that chr19:12996905-12998745 (hg19) regulates gene *CALR* (adjusted p-value 1.9E-7). Based on annotations in VAMPIRE, *CALR* is linked to rs8110787 (chr19:12999458, hg19) in prioritization category 1, including higher than expected physical interactions with the *CALR* locus in K562 cells, as shown in the HUGIn2 plot above. rs8110787 is associated with several RBC traits (namely hematocrit, mean corpuscular hemoglobin, mean corpuscular volume, and red blood cell count) as reported in Chen et al. [3].

References

1. Garner, C., et al., *Genetic influences on F cells and other hematologic variables: a twin heritability study.* Blood, 2000. **95**(1): p. 342-6.
2. Vuckovic, D., et al., *The Polygenic and Monogenic Basis of Blood Traits and Diseases.* Cell, 2020. **182**(5): p. 1214-1231 e11.
3. Chen, M.H., et al., *Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations.* Cell, 2020. **182**(5): p. 1198-1213.e14.
4. Liggett, L.A. and V.G. Sankaran, *Unraveling Hematopoiesis through the Lens of Genomics.* Cell, 2020. **182**(6): p. 1384-1400.
5. Kowalski, M.H., et al., *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, 2019. **15**(12): p. e1008500.
6. Bick, A.G., et al., *Inherited causes of clonal haematopoiesis in 97,691 whole genomes.* Nature, 2020. **586**(7831): p. 763-768.
7. Bao, E.L., et al., *Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells.* Nature, 2020. **586**(7831): p. 769-775.
8. Liu, X., et al., *WGSA: an annotation pipeline for human genome sequencing studies.* J Med Genet, 2016. **53**(2): p. 111-2.
9. Fishilevich, S., et al., *GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.* Database (Oxford), 2017. **2017**.
10. Lizio, M., et al., *Gateways to the FANTOM5 promoter level mammalian expression atlas.* Genome Biol, 2015. **16**: p. 22.
11. Lizio, M., et al., *Update of the FANTOM web resource: expansion to provide additional transcriptome atlases.* Nucleic Acids Res, 2019. **47**(D1): p. D752-D758.
12. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.
13. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
14. Ludwig, L.S., et al., *Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis.* Cell Rep, 2019. **27**(11): p. 3228-3240.e7.
15. Ulirsch, J.C., et al., *Interrogation of human hematopoiesis at single-cell and single-variant resolution.* Nat Genet, 2019. **51**(4): p. 683-693.
16. Schmitt, A.D., et al., *A Compendium of Chromatin Contact Maps Reveal Spatially Active Regions in the Human Genome.* Cell reports, 2016. **17**(8): p. 2042-2059.
17. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
18. Schrider, D.R. and A.D. Kern, *S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning.* PLoS Genet, 2016. **12**(3): p. e1005928.
19. Schrider, D.R. and A.D. Kern, *Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome.* Mol Biol Evol, 2017. **34**(8): p. 1863-1877.
20. Võsa, U., et al., *Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis.* bioRxiv, 2018: p. 447367.

21. Lloyd-Jones, L.R., et al., *The Genetic Architecture of Gene Expression in Peripheral Blood.* Am J Hum Genet, 2017. **100**(2): p. 371.
22. Zhernakova, D.V., et al., *Identification of context-dependent expression quantitative trait loci in whole blood.* Nat Genet, 2017. **49**(1): p. 139-145.
23. Raj, T., et al., *Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes.* Science, 2014. **344**(6183): p. 519-23.
24. Rao, S.S.P., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-1680.
25. Gorkin, D.U., et al., *Common DNA sequence variation influences 3-dimensional conformation of the human genome.* Genome Biol, 2019. **20**(1): p. 255.
26. Ay, F., T.L. Bailey, and W.S. Noble, *Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.* Genome Research, 2014. **24**(6): p. 999-1011.
27. Javierre, B.M., et al., *Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters.* Cell, 2016. **167**(5): p. 1369-1384 e19.
28. Jung, I., et al., *A compendium of promoter-centered long-range chromatin interactions in the human genome.* Nat Genet, 2019. **51**(10): p. 1442-1449.
29. Martin, J.S., et al., *HUGIn: Hi-C Unifying Genomic Interrogator.* Bioinformatics (Oxford, England), 2017. **33**(23): p. 3793-3795.
30. Denny, J.C., et al., *Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data.* Nature Biotechnology, 2013. **31**(12): p. 1102-1111.
31. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D1005-d1012.

Supporting information

S1 Appendix. Supplementary methods.