

1 From GWAS Variant to Function: a Study of ~148,000 Variants for

2 Blood Cell Traits.

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

6 1. Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC,
7 USA

8 2. Curriculum in Bioinformatics and Computational Biology, University of North Carolina at
9 Chapel Hill, Chapel Hill, NC, USA

10 3. Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

11 4. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

12 5. Montreal Heart Institute, Montreal, Quebec, Canada

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

15 7. Department of Epidemiology, University of Washington, Seattle, WA, USA

16 8. Fred Hutchinson Cancer Research Center, University of Washington, Seattle, WA, USA

17 9. Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric
18 Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

19 10. Broad Institute of Harvard and MIT, Cambridge, MA, USA

20 11. Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill,
21 NC, USA

22

23 *These authors contributed equally to this work.

24 # Corresponding author: Yun Li. Department of Genetics, Biostatistics and Computer Science,
25 University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. Fax: (919) 843-4682;
26 E-mail: yunli@med.unc.edu

27

28

29

30 **Abstract**

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

49 **Keywords:** Genome-wide association studies, variant to function, functional annotations,
50 experimental validations, blood cell traits

52 Introduction

53 Genome-wide association studies (GWAS) have identified thousands of genetic loci and
54 hundreds of thousands of genetic variants associated with various complex human diseases and
55 traits, but the underlying genetic mechanism for the vast majority of these GWAS signals
56 remains elusive. With extensive sequencing and GWAS efforts, there is a pressing need to
57 convert the large and ever growing number of significant GWAS variant-trait pairs into human-
58 interpretable functional or mechanistic knowledge¹. Most variants identified through GWAS
59 reside in the noncoding regions (e.g., >95% for blood cell traits²), and most signals include
60 multiple highly correlated variants or variants in strong linkage disequilibrium (LD). Pinpointing
61 the most likely causal variants within GWAS signals, and linking these variants to their target
62 genes, is challenging, particularly as the number of GWAS loci and variants increases. For
63 hematological traits, for instance, our recent GWAS meta-analyses^{3;4} have revealed over seven
64 thousand loci, with >148,000 variants associated with at least one blood cell index at stringent
65 genome-wide significance threshold. Comprehensive and computationally efficient annotation
66 and prioritization of such GWAS findings are of ever-increasing interest.

67 Understanding how genetic variants contribute to a phenotype is often referred to as the variant-
68 to-function (V2F) problem. Responding to this problem requires us to determine causal genetic
69 variants, relative cell types/states, their target genes and cellular/physiological functions⁵.
70 Functional experiments are needed to fully reveal molecular mechanisms, but we cannot yet
71 afford to perform time-, money- and labor-consuming experimental validations of thousands of
72 loci involving hundreds of thousands of potentially functional variants or regulatory elements
73 controlling their nearby genes, since each gene is likely regulated by multiple variants and each

74 variant may regulate multiple genes. Thus, computational methods are needed to screen potential
75 variants and their effector genes for further experiments.

76 In this study, we focus on hematological traits. Hematological phenotypes (red blood cell, white
77 blood cell, and platelet counts and indices) are critical physiological intermediaries in oxygen
78 transport, immunity, infection, thrombosis, and hemostasis and are associated with autoimmune,
79 allergic, infectious, and cardiovascular diseases. Hematological traits are highly heritable ⁶, and
80 recent large GWAS for hematological traits (including nearly 750,000 participants) identified
81 thousands of variant-trait associations ^{2;4}. In addition, there are multiple large-scale functional
82 experiments already available^{2; 7; 8} for hematological traits, as well as fairly comprehensive
83 functional annotation resources relevant to blood tissues. This makes hematological traits an
84 ideal model for this type of V2F computational solution.

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

97 is adaptable for annotation of other complex trait GWAS results and will facilitate the connection
98 between variant and function.

99

100 **Methods**

101 **Variant Annotations**

102 The current version of VAMPIRE includes GWAS results from two studies (as detailed in
103 Supplemental Methods), including all variants in 95% credible sets for fine-mapped
104 hematological trait associated loci from Chen et al. (N=148,019 variants)⁴ and lead variants
105 (N=2) from a TOPMed imputed GWAS meta-analysis in African American and Hispanic/Latino
106 populations³. We plan to extend VAMPIRE as new trans-ethnic blood cell trait genetic analyses
107 are released.

108 The sources of the annotation used are stated clearly in the VAMPIRE online application, with
109 links or references to the original data sources. As a brief summary, the annotation categories are
110 trivially split into six types ("variant level", "1D", "2D", "3D", "PheWAS", "GWAS"). First,
111 "variant level" contains data on phenotypic association from the original publication or preprint
112 (such as the p-value for association with a given hematological trait, effect size, and posterior
113 probability of inclusion for fine-mapping credible sets). Second, "1D" refers to epigenomic or
114 sequence constraints features. This displays selected output from WGSA⁹ including functional
115 prediction scores, conservation scores, and epigenetic information gathered from GeneHancer¹⁰,
116 FANTOM5^{11;12}, Roadmap¹³, and ENCODE¹⁴. ATAC-seq peaks from recent studies for blood
117 cell traits^{15;16} and key histone ChIP-seq peaks such as H3K9me3, H3K36me3, H3K4me1,

118 H3K4me3, and H3K27Ac generated across blood cell related tissues from Roadmap
119 Epigenomics are also included ^{13; 17}. We further include information regarding whether each
120 variant resides in any selective sweep region detected from multiple populations in the 1000
121 Genomes Project ¹⁸ using the S/HIC method ^{19; 20}. Information is displayed based on the tissue
122 relevance to the blood cell phenotype (see Supplemental Methods). All variants have 1D
123 annotation, but for prioritization purposes as described below in the five categories for
124 noncoding variant annotation, we define 1D annotation as FANTOM5_enhancer_robust =Y
125 (yes), or Genehancer_feature="Promoter" or "Enhancer" or "Promoter/Enhancer", or coreMarks
126 (for any relevant roadmap epigenomic category) = "Enhancers" or "Active TSS." Users can then
127 additionally filter by criteria such as functional prediction and conservation scores.

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

139 For the "3D" annotations, we include information on 3D genome conformation, linking blood
140 lineage specific regulatory elements to target genes from various sources. More specifically,

141 using Hi-C data we incorporated statistically significant long-range chromatin interactions
142 (LRCI)^{17; 26; 27} calculated from Fit-Hi-C²⁸, loops using the HiCCUPs methodology²⁶, and super-
143 FIREs for related tissues¹⁷. Two Promoter-Capture Hi-C (PCHi-C) data sources^{29; 30} were also
144 incorporated and matched with the 2D results to highlight consistent evidence regarding the
145 affected gene(s) across "2D" and "3D" annotations. VAMPIRE displays information on the
146 number of loops, LRCI, PCHi-C interactions, FIREs, or super-FIREs, as well as significance
147 measures such as p-values, FDR, or CHICAGO scores where applicable. This "3D" annotation
148 information can also be visualized via our HUGIn browser³¹.

149 The last two data groups present results from two PheWAS sources^{4; 32} and GWAS results of
150 blood cell traits from GWAS catalog³³, allowing the user to evaluate if hematological trait
151 associated variants may also influence other complex traits.

152 To visualize and leverage these multiple annotation categories for further analysis or
153 prioritization of experimental validations, VAMPIRE efficiently displays and integrates relevant
154 variant information, allowing the user to investigate either all the variants annotated or subsets
155 based on annotation category groupings, searching either by variant or by gene name. The
156 comprehensive annotation for the variants is summarized using a five category grouping created
157 for highlighting the most promising variants as they have various types of annotation.
158 Specifically, the five categories for noncoding variants are (1) the most restrictive category,
159 containing variants that have 1D, 2D, and 3D annotation and the genes implicated by 2D and 3D
160 evidence are consistent; (2) containing variants with 1D, 2D, and 3D evidence, but the genes
161 implicated from different resources are not consistent; (3) 2D and 3D with consistent gene
162 evidence between the 2D and 3D annotations; (4) variants with 2D and 3D information and no
163 consistent gene implied; (5) variants with 1D and 3D evidence. We also have a predicted high

164 impact coding variant category displayed, including high confidence loss of function (LoF)
165 variants and likely influential missense, in frame indels, and synonymous variants. Variants
166 without strongly compelling variant annotation are still displayed, but are not listed in these high
167 priority categories. The user can further subset results by hematological trait, hematological trait
168 category, or (for the Chen et al paper ⁴) the ancestry specific grouping in which a given credible
169 set was derived (trans-ethnic, European, East Asian, South Asian, Hispanic/Latino, or African
170 ancestry). In addition, the user can restrict the amount of information presented by selecting
171 which tables to be displayed. All tables can be exported in a csv or tab delimited format.

172 **Enrichment analysis**

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

178 At the variant level, we assessed the enrichment of variants that modify transcription factor (TF)
179 binding motif² among our annotation category 1 variants. Recently, Vuckovic et al. ²
180 characterized variants that affect erythropoiesis or hematopoiesis by modifying related TF
181 motifs, such as for KLF1, KLF6, MAFB, and GATA1. We chose these four erythroid TFs as
182 positive control TFs and two non-erythroid TFs (IRF1 and IRF8) as negative controls.

183 At the gene level, we evaluated the genes interrogated by Nandakumar et al. ⁸ with a pooled
184 short hairpin RNA (shRNA) based loss-of-function approach. Specifically, Nandakumar et al.
185 studied 389 candidate genes in the neighborhood of 75 loci associated with red blood cell traits

186 ³⁴, to identify potential causal genes underlying these GWAS signals. We assessed the
187 enrichment of genes validated by shRNA experiments among those prioritized in VAMPIRE's
188 category 1. Note that the categories were previously defined at variant level. Here we extent
189 variant category to gene category as the strongest category where a genome-wide significant
190 variant linked to this gene falls in.

191 At the variant-gene pair level, we employed the enhancer-gene connections validated via
192 CRISPRi-FlowFISH experiments by Fulco et al. ⁷ in their activity-by-contact (ABC) paper.
193 Specifically, Fulco et al. tested pairs of candidate *cis* regulatory elements (CREs, ~500bp regions)
194 and their potential effector genes via CRISPRi perturbations of the CREs, in multiple cell lines
195 including the K562 cells. Fulco et al. tested 4,124 CRE-gene pairs in total, of which 175 were
196 significant from their experiments. We overlapped their tested CREs with variants in our
197 VAMPIRE annotation database. We define a VAMPIRE variant-gene pair confirmed if the
198 variant overlaps an ABC validated CRE *and* the linked genes in VAMPIRE (from QTL and
199 chromatin capture conformation evidence) overlaps the corresponding effector gene for that CRE
200 via ABC's CRISPRi-FlowFISH experiment. We focused on ABC experiments performed on the
201 K562 cells (instead of GM12878 cells, where a very small number of CREs were tested) as the
202 number of tested CRE-gene pairs was not too small for robust statistical inference. Matching the
203 K562 cell line, we focused only on variants associated with red blood cell traits. Similar to the
204 above two sets of enrichment analyses, we focused on annotations in VAMPIRE's prioritization
205 category 1. Specifically, we tested whether variant-gene pairs prioritized in VAMPIRE's
206 category 1 are enriched within ABC's validated enhancer-gene connections. Given the CREs
207 tested in the ABC paper are rather short (~500bp), we also performed sensitivity analysis by first

208 extending the CRE regions by +/- 1kb and +/- 5kb and then overlapping variants with these
209 extended CREs, to ensure robust conclusions.

210

211 **Results**

212 **Overview of VAMPIRE annotations**

213 The overall framework of VAMPIRE is illustrated in **Figure 1**. We started with all variants in
214 95% credible sets from our recent trans-ethnic study for hematological traits (total 148,019
215 variants)⁴ and lead variants (2 variants) from Kowalski et al.³. We incorporated six types of
216 annotations (detailed in **Methods**): GWAS summary statistics and posterior probability of
217 inclusion from our previous fine-mapping analyses⁴; epigenomic or sequence constraints
218 features (1D); eQTL, sQTL and pQTL information (2D); information on 3D genome
219 conformation (3D); results from two PheWAS sources^{4;32} (PheWAS); and GWAS results from
220 blood cell traits from GWAS catalog³³ (GWAS).

221 To visualize and prioritize variants, their corresponding candidate regulatory regions, and their
222 potential effector genes, we leverage the aforementioned six types of annotation to group these
223 ~148,000 variants into various prioritization categories. Specifically, for non-coding variants,
224 we classified them into five categories (detailed in Methods). Among them, category 1 is the
225 most restrictive category, containing variants that have 1D, 2D, and 3D annotation and the genes
226 implicated by 2D and 3D evidence are consistent. Variants not falling into any of the five
227 categories are classified as uncategorized. In addition, each gene is categorized according to the

228 prioritization categories of its linked variant(s). When its linked variants fall in multiple
229 categories, the gene is assigned to the most highly prioritized category.

230 **Enrichment analysis**

231 Our enrichment analyses employing multiple previously published functional validation
232 experiments encompassing variant-level, gene-level, and variant-gene pair levels all showed
233 promising results. Specifically, at the variant level, we found significant enrichment of variants
234 affecting TF binding motifs among variants prioritized in category 1 of VAMPIRE (**Figure 2**),
235 for all the erythroid TFs ($p < 8.1E-4$) but GATA1 ($p = 0.18$) (**Table 1**), likely due a smaller
236 sample size of variants. In contrast, neither of the two negative control TFs (IRF1 and IRF8)
237 showed any significant enrichment ($p = 0.22$ and 0.62). At the gene level, we focused on two
238 statistics: (1) number of genes selected for shRNA experiments, since genes were more likely to
239 be selected for experiments when they demonstrated some prior evidence of potential causality,
240 and (2) number of genes validated ($p < 0.05$) by shRNA experiments. We compared the number
241 of genes in our annotation category 1 and all other categories, and found that both shRNA
242 candidate genes ($p = 3.5E-13$) and significant genes ($p = 3.1E-8$) show strong enrichment among
243 those in our annotation category 1 (**Table 2**), and the estimated enrichment score for significant
244 genes ($OR = 4.65$) is almost double of that for candidate genes ($OR = 2.37$). These results
245 suggest the genes prioritized by VAMPIRE's category 1 annotations are more likely to be
246 functional.

247 Finally, at the variant-gene pair level, we also observed enrichment among variants selected into
248 VAMPIRE's category 1 (**Table 3**). When restricting only to variants in category 1 and associated
249 with red blood cell traits and without extending the CRE regions, only 7 of VAMPIRE's variant-

250 gene pairs can be found in ABC's CRISPRi-FlowFISH experiments, of which 6 are not
251 significant and 1 is significant. While not significant ($p = 0.26$), the direction of enrichment is
252 nevertheless encouraging (one of seven, or 14.3%, confirmed by CRISPRi-FlowFISH
253 experiments) and 3-fold greater than that among all/background pairs from Fulco et al.⁷, where
254 175 out of 4124 variant-gene pairs (4.2%) were confirmed. Note that all the confirmed pairs
255 were linked with variants associated with red blood cell traits. Further generalizing to all
256 VAMPIRE annotation categories and to variants associated with any blood cell trait, the
257 enrichment OR increases to 8.30 with p-value 9.0E-5, indicating that variant-gene pairs
258 prioritized by VAMPIRE's five categories have much higher odds of being functional. To further
259 accommodate causal variants tagged by GWAS variants not falling into the short 500bp CREs,
260 we extended the CREs by +/- 1kb or +/- 5kb, and performed similar enrichment analysis. Our
261 conclusions remained qualitatively similar (**Table 3**), but the enrichments increased in
262 significance, thanks to larger sample size (in this context, the larger number of variant-gene pairs
263 contributing to the analysis) and suggesting that more liberal windows of *cis*-regulatory regions
264 can capture a higher rate of functional variant-gene pairs. For example, the enrichment for
265 category 1 variants associated with red blood cell (RBC) traits reached an OR of 15.77 ($p=3.8E-6$)
266 and 16.68 ($p=3.1E-15$) for 1kb and 5kb extension, respectively. We thus conclude that such
267 enrichment is significant and robust to the extension of CREs.

268 **Application example**

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

273 rs8110787 is associated with several RBC traits ⁴, including hematocrit (HCT), mean corpuscular
274 hemoglobin (MCH), mean corpuscular volume (MCV) and red blood cell counts (RBC). Based
275 on genomic distance alone, *CALR* is not the nearest gene to rs8110787, with several other closer
276 genes. However, based on H3K27ac HiChIP data in K562 cells ³⁵, rs8110787 significantly
277 interacts with *CALR* promoter region (p < 1E-120), suggesting that *CALR* is a potential target
278 gene regulated by the CRE around rs8110787. This variant is also an eQTL of *CALR* from
279 CAGE ²³ (p = 9.4E-16) and BIOS ²⁴ (p = 1.0E-25), and is an enhancer in K562 Leukemia cells
280 (E123) from Roadmap ¹³, adding additional evidence. Our VAMPIRE successfully highlights
281 this rs8110787-*CALR* pair in its category 1.

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

289 This low frequency intergenic variant rs112097551 (*GATA2-RPN1* locus, 0.15% minor allele
290 frequency in Chen et al. trans-ethnic analysis ⁴) has no close linkage disequilibrium proxies in
291 African or European populations, and thus was not compared to other highly correlated variants.

292 Based on variant frequency, particularly in European ancestry populations, we had no
293 expectation this variant would have eQTL or pQTL evidence (2D annotation), given currently
294 available sample sizes for eQTL and pQTL analysis. For low frequency variants, 1D and 3D
295 annotation would be the highest annotation category likely for a variant of interest like

296 rs112097551. The variant is ~5x more common among African versus non-African samples in
297 gnomAD version 2.1.1. It is the only variant in the credible set in fine-mapping analyses from
298 Chen et al. 1D annotation suggests this variant is highly conserved (CADD Phred score of 20.4,
299 meaning the variant is amongst the top 1% of deleterious variants in the human genome), and it
300 is rated as deleterious by FATHMM-XF (rank score 0.99169, close to the maximum score of 1).
301 It is also in open chromatin in megakaryocyte–erythroid progenitor cells, based on hematopoietic
302 ATAC-seq data³⁶. 3D annotation from PCHi-C data in erythroblasts from Javierre et al.²⁹ links
303 this variant to the gene *RUVBL1* ~500Kb away, as well as noncoding transcripts *RNU2-37P* and
304 *RUVBL1-AS1*. Based on this data, which can be quickly displayed using the VAMPIRE
305 application, we are currently working on *in vitro* follow-up of this candidate functional enhancer
306 variant³⁷.

307 Discussion

308 As genotyped sample sizes increase and meta-analysis efforts grow ever larger, more variant-
309 trait pairs are identified for complex traits than can be easily annotated on a variant by variant
310 basis. New, user-friendly applications are needed for rapid display of functional annotation
311 information and prioritization of variants for further functional follow-up to pave the V2F path.
312 Our VAMPIRE tool provides an example of how the publicly available code can be adapted to
313 accommodate other sources of annotation specific to other complex trait GWAS results or to
314 accommodate future blood cell trait GWAS and annotation resources. In addition to *a priori*
315 providing one category of coding variants and 5 categories of non-coding variants that warrant
316 prioritization consideration, VAMPIRE allows users to decide their own categories based on
317 arbitrary combinations of the annotations at adjustable thresholds (for example, prioritizing high
318 CADD score variants, or variants in open chromatin in blood cells based on ATAC-seq). Along

319 with the addition of more blood cell trait genetics papers published in the future, VAMPIRE
320 could also be used as written to annotate GWAS results for other blood related phenotypes, such
321 as recent GWAS of risk of myeloproliferative neoplasm or clonal hematopoiesis^{38,39}.

322 As we accumulate additional functional validation data, including high-throughput massively
323 parallel reporter assays (MPRA), medium-throughput CRISPRi/CRISPRa and low throughput
324 mouse xenotransplant experiments, we will provide statistics summarizing experimental
325 validation results (e.g., number of variants in the category followed-up, proportion that show
326 evidence of functional impact in their experiments) for each of the 6 VAMPIRE categories and
327 for user defined categories. Importantly, we illustrate the value of VAMPIRE using existing
328 independent functional validation and therefore illuminate the value of this type of annotation
329 tool in enabling one to go from variant to function for blood cell traits and other complex
330 phenotypes.

331 We also note that there are some limitations of VAMPIRE. First, comprehensive annotations
332 specific to various cell types and cell states would further enhance classification and
333 prioritization accuracy of functional variants or regulatory elements and their target genes.
334 Although data is increasingly being generated by us^{15,16} and others^{29,35}, and has been
335 incorporated into VAMPIRE where available, interrogations in a cell-type- or state- specific
336 manner are still much needed. For instance, our recent work has demonstrated cell-type or tissue
337 specific FIREs^{17,40} and super interactive promoters (SIP)⁴¹ play key regulatory role and aid the
338 identification and prioritization of functional regulatory elements and their corresponding genes.
339 As more experimental data are generated, we will update VAMPIRE accordingly. Second, our
340 list of 148,019 variants derives primarily from fine-mapping studies, which may be inaccurate in
341 loci where more than one independent or partially independent signals exist. However, this

342 limitation cannot be resolved before more powerful methods are developed for fine-mapping
343 analysis for trans-ethnic GWAS. Finally, most of the annotations are based on analyses in
344 European ancestry individuals (e.g. eQTL, pQTL, chromatin conformation etc.). Many ongoing
345 efforts including ours are generating resources for non-European ancestry samples. For example,
346 we are involved in several recently funded efforts to generate RNA-sequencing data in non-
347 European ancestry individuals in hematopoietic cell types and anticipate relevant eQTL and
348 sQTL annotations being added to VAMPIRE in upcoming years.

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

358 **Appendix**

359 A1. Supplementary methods.

360

361 **Declaration of Interests**

362 The authors declare no competing interests.

363 **Acknowledgement**

364 This work was supported by the National Center for Advancing Translational Sciences, National
365 Institutes of Health [R01HL146500 to A.P.R., R01HL129132 to Y.L., KL2TR002490 to L.M.R.,
366 R01DK103794 to V.G.S.], and the New York Stem Cell Foundation. C.A.C. and Y.L. are also
367 partially supported by R01GM105785 and U01DA052713. V.G.S. is a New York Stem Cell
368 Foundation-Robertson Investigator. The content is solely the responsibility of the authors and
369 does not necessarily represent the official views of the NIH.

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

373 **Web Resources**

374 VAMPIRE: <http://shiny.bios.unc.edu/vampire/>

375 GWAS summary statistics from Chen et al.⁴: <http://www.mhi-human genetics.org/en/resources/>

376 GWAS Catalog: <https://www.ebi.ac.uk/gwas/>

377 PheWAS website: <http://pheweb.sph.umich.edu>

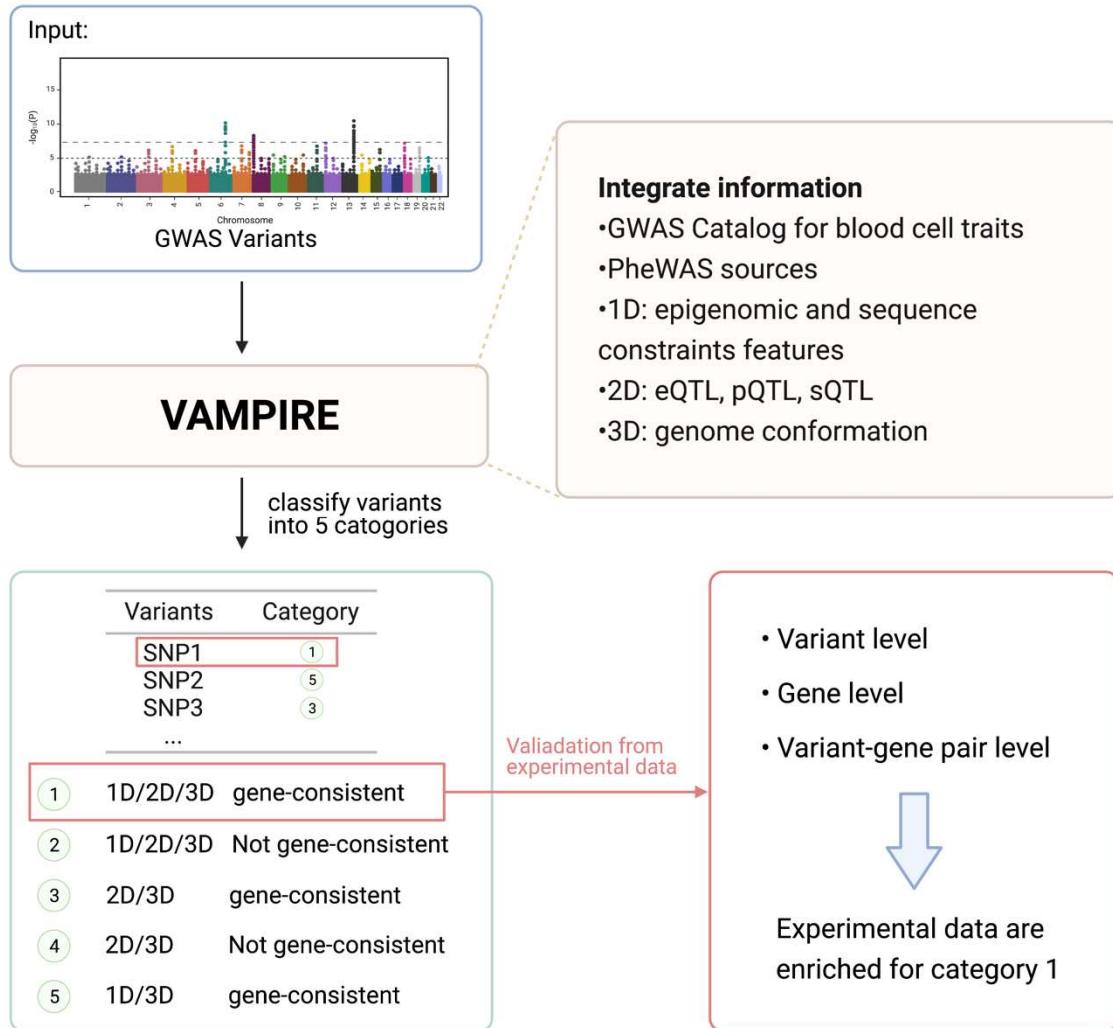
378 **Data Availability**

379 The data underlying this article are available in the article and in its online supplementary
380 material.

381

382

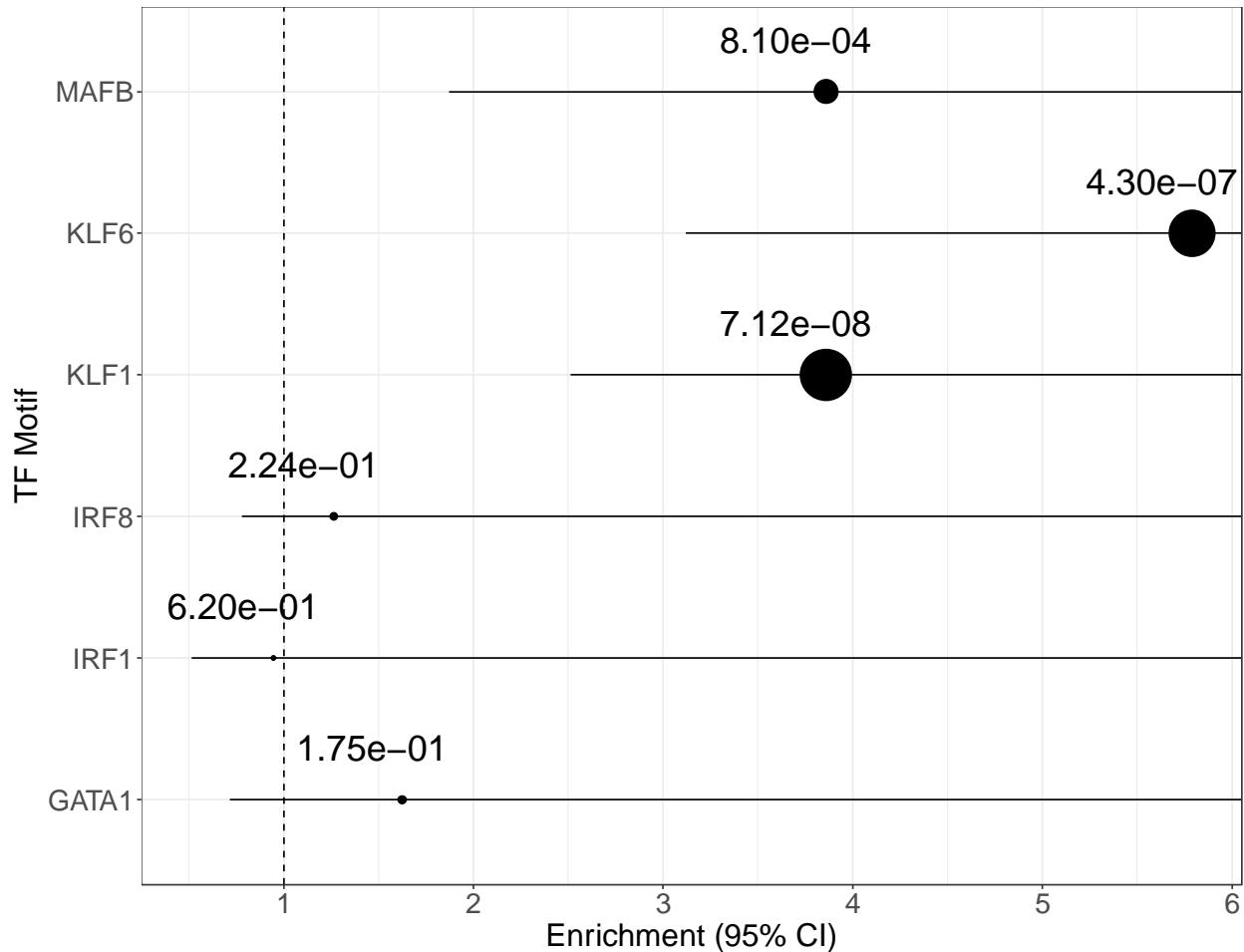
383 **Figures**



384

385 **Figure 1. Overall framework of this study.** VAMPIRE starts with GWAS variants in the 95%
386 credible sets, integrates different annotations and assigns them into different prioritization
387 categories. We further demonstrated that our top prioritized category is enriched with variants
388 that were experimentally validated. VAMPIRE provides a prototype that can be adapted to many
389 other complex traits, paving the path for efficient variant to function (V2F) analyses.

390



391

392

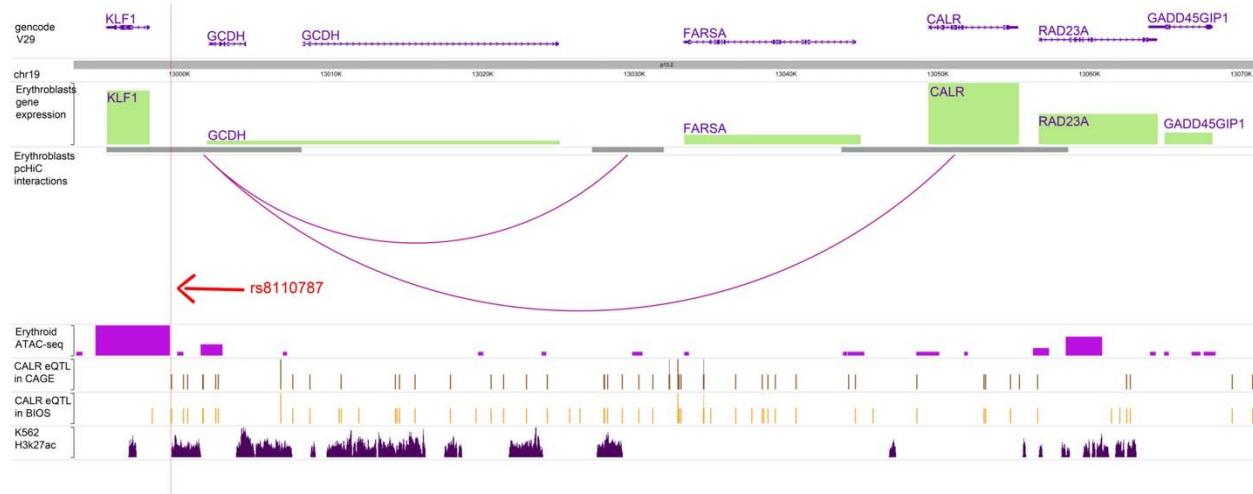
393 **Figure 2. Variant level TF motif enrichment analysis.** Each dot represents an enrichment
394 score with the line depicting 95% confidence interval (CI). All the upper bounds of these CIs are
395 infinity. The p-values of the enrichment are reflected by the dot size at the OR point estimate
396 with a larger dot indicating more significant the enrichment.

397

398

399

400



401

402

403 **Figure 3. Variant-gene pair example (rs8110787-CALR) visualization from HUGIn2³¹.**

404 Fulco et al. confirmed via CRISPRi experiments that chr19:12996905-12998745 (hg19)
405 regulates gene *CALR* (adjusted p-value 1.9E-7) which is highly expressed in Erythroblasts⁷.
406 Based on annotations in VAMPIRE, *CALR* is linked to rs8110787 (chr19:12999458, hg19) in
407 prioritization category 1, including higher than expected physical interactions with the *CALR*
408 locus from erythroblasts pcHiC data²⁹, eQTL of *CALR* in CAGE²³ and BIOS²⁴, erythroid ATAC-
409 seq peak¹⁶ and H3K27ac peak in K562 leukemia cell¹³. rs8110787 is associated with several
410 RBC traits (namely hematocrit, mean corpuscular hemoglobin, mean corpuscular volume, and
411 red blood cell count) as reported in Chen et al.⁴.

412

413

414

415 **Tables**

416

	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

417

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

426

	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

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

430

431

432

433

	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	19	7	26.92	9.00E-05	8.3

all categories for all

traits

Confirmed pairs in

category 1 for RBC	10	7	41.18	3.80E-06	15.77
--------------------	----	---	-------	----------	-------

traits (+/- 1kb)

Confirmed pairs in

category 1 for all	21	9	30	3.50E-06	9.66
--------------------	----	---	----	----------	------

traits (+/- 1kb)

Confirmed pairs in

all categories for all	70	21	23.08	4.60E-10	6.76
------------------------	----	----	-------	----------	------

traits (+/- 1kb)

Confirmed pairs in

category 1 for RBC	27	20	42.55	3.10E-15	16.68
--------------------	----	----	-------	----------	-------

traits (+/- 5kb)

Confirmed pairs in

category 1 for all	64	23	26.44	3.80E-12	8.1
--------------------	----	----	-------	----------	-----

traits (+/- 5kb)

Confirmed pairs in

all categories for all	160	37	18.78	3.10E-13	5.21
------------------------	-----	----	-------	----------	------

traits (+/- 5kb)

434 **Table 3. Variant-Gene pair level enrichment analysis.** We performed analysis for three
435 variant annotation pools (category 1, red blood cell (RBC) trait associated; category 1, any blood
436 cell trait associated; any annotation priority category (1-5), any blood cell trait associated) and
437 three CRE lengths. Fisher's exact test was applied to test for enrichment. We found enrichment

438 for all three variant annotation pools. These enrichments are also robust to the extension of
439 CREs.

440

441

442 References

- 443 1. Liggett, L.A., and Sankaran, V.G. (2020). Unraveling Hematopoiesis through the Lens of
444 Genomics. *Cell* 182, 1384-1400.
- 445 2. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield,
446 L.M., Tardaguila, M., Huffman, J.E., et al. (2020). The Polygenic and Monogenic Basis of
447 Blood Traits and Diseases. *Cell* 182, 1214-1231.e1211.
- 448 3. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M.,
449 Arnett, D.K., Avery, C., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision
450 Medicine (TOPMed) Consortium whole genome sequences improves imputation quality
451 and detection of rare variant associations in admixed African and Hispanic/Latino
452 populations. *PLoS genetics* 15, e1008500.
- 453 4. Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B.,
454 Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific
455 Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* 182, 1198-
456 1213.e1114.
- 457 5. Nandakumar, S.K., Liao, X., and Sankaran, V.G. (2020). In The Blood: Connecting Variant to
458 Function In Human Hematopoiesis. *Trends in genetics* : TIG 36, 563-576.
- 459 6. Garner, C., Tatu, T., Reittie, J.E., Littlewood, T., Darley, J., Cervino, S., Farrall, M., Kelly, P.,
460 Spector, T.D., and Thein, S.L. (2000). Genetic influences on F cells and other
461 hematologic variables: a twin heritability study. *Blood* 95, 342-346.
- 462 7. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V.,
463 Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-
464 contact model of enhancer-promoter regulation from thousands of CRISPR
465 perturbations. *Nat Genet* 51, 1664-1669.
- 466 8. Nandakumar, S.K., McFarland, S.K., Mateyka, L.M., Lareau, C.A., Ulirsch, J.C., Ludwig, L.S.,
467 Agarwal, G., Engreitz, J.M., Przychodzen, B., McConkey, M., et al. (2019). Gene-centric
468 functional dissection of human genetic variation uncovers regulators of hematopoiesis.
469 *eLife* 8.
- 470 9. Liu, X., White, S., Peng, B., Johnson, A.D., Brody, J.A., Li, A.H., Huang, Z., Carroll, A., Wei,
471 P., Gibbs, R., et al. (2016). WGSAs: an annotation pipeline for human genome
472 sequencing studies. *Journal of medical genetics* 53, 111-112.
- 473 10. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N.,
474 Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of
475 enhancers and target genes in GeneCards. *Database : the journal of biological*
476 *databases and curation* 2017.
- 477 11. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I.,
478 Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5
479 promoter level mammalian expression atlas. *Genome Biol* 16, 22.
- 480 12. Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C.C., de Hoon, M.,
481 Severin, J., Oki, S., Hayashizaki, Y., et al. (2019). Update of the FANTOM web resource:

482 expansion to provide additional transcriptome atlases. *Nucleic Acids Res* 47, D752-
483 D758.

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

487 14. Consortium, E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*
488 306, 636-640.

489 15. Ludwig, L.S., Lareau, C.A., Bao, E.L., Nandakumar, S.K., Muus, C., Ulirsch, J.C.,
490 Chowdhary, K., Buenrostro, J.D., Mohandas, N., An, X., et al. (2019). Transcriptional
491 States and Chromatin Accessibility Underlying Human Erythropoiesis. *Cell Rep* 27,
492 3228-3240.e3227.

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

496 17. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et
497 al. (2016). A Compendium of Chromatin Contact Maps Reveal Spatially Active Regions
498 in the Human Genome. *Cell reports* 17, 2042-2059.

499 18. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M.,
500 Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global
501 reference for human genetic variation. *Nature* 526, 68-74.

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

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

506 21. Emilsson, V., Ilkov, M., Lamb, J.R., Finkel, N., Gudmundsson, E.F., Pitts, R., Hoover, H.,
507 Gudmundsdottir, V., Horman, S.R., Aspelund, T., et al. (2018). Co-regulatory networks
508 of human serum proteins link genetics to disease. *Science (New York, NY)* 361, 769-
509 773.

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

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

516 24. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W.,
517 van 't Hof, P., Mei, H., van Dijk, F., Westra, H.J., et al. (2017). Identification of context-
518 dependent expression quantitative trait loci in whole blood. *Nat Genet* 49, 139-145.

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

522 26. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T.,
523 Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the
524 human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159,
525 1665-1680.

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

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

531 29. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J.,
532 Wingett, S.W., Varnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome

Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* 167, 1369-1384 e1319.

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 phenotype-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology* 31, 1102-1111.

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, 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. 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.

36. 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.

37. Hu, Y., Stilp, A.M., McHugh, C.P., Jain, D., Zheng, X., Lane, J., Rao, S., de Bellefon, S.M., Raffield, L.M., Chen, M.-H., et al. (2020). Whole genome sequencing association analysis of quantitative red blood cell phenotypes: the NHLBI TOPMed program. *medRxiv*, 2020.2012.2009.20246736.

38. 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. (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* 586, 763-768.

39. 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. (2020). Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* 586, 769-775.

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

41. 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.