Genome-wide maps of transcription factor footprints identify noncoding variants rewiring gene regulatory networks with varTFBridge

Jiecong Lin[1], Wenyang Dong[1], Jiaxiang Zhang[1], Chen Xie[1,2], Xiaoxi Jing[1], Junpeng Zhao[1], Kaiyan Ma[1], Hongen Kang[1], Yilin Jiang[1,3], Xiaoliang Sunney Xie[1,2]* and Yajie Zhao[1]*

1. Changping Laboratory, Beijing 102206, China

2. Beijing Advanced Innovation Center for Genomics and Biomedical Pioneering Innovation Center, Peking University, Beijing 100871, China

3. School of Basic Medical Sciences, Peking University, Beijing, 100191, China

*Correspondence: yajiezhao@cpl.ac.cn, sunneyxie@cpl.ac.cn

Abstract

Common-variant genome-wide association studies have identified thousands of noncoding loci associated with human diseases and complex traits; however, interpreting their functional mechanisms remains a major challenge. In recent years, the decreasing cost of high-throughput sequencing has enabled large population cohorts to generate whole-genome sequencing data for hundreds of thousands of individuals, providing an opportunity to comprehensively explore the functional impact of the full spectrum of genetic variants. However, unlike common-variant GWAS or rare-variant burden testing in coding regions, robust and systematic strategies for interpreting rare variants in noncoding regions remain limited. Meanwhile, many causal variants in non-coding regions are thought to act by perturbing transcription factor (TF) binding and rewiring gene regulatory networks, but progress has been limited by the lack of accurate, high-resolution maps of TF footprints. Recent single-molecule deaminase footprinting (FOODIE) technologies enable precise genome-wide detection of TF footprints at near–single-base resolution. We discovered that K562 FOODIE footprints, although comprising less than 0.5% of the genome, exhibit approximately 70-fold enrichment for erythroid trait heritability, establishing an unprecedented resource for interrogating TF-mediated regulatory mechanisms. We further introduce varTFBridge, an integrative framework that combines both common and rare non-coding variant association analyses, footprint-gene linking models, and AlphaGenome to prioritize causal noncoding variants that rewire gene regulatory networks. Across 13 erythrocyte traits, varTFBridge linked 209 credible common and 18 driver rare variants altering TF binding affinity to modulate 207 unique target genes. We successfully recapitulated a known causal variant associated with erythrocyte traits and further elucidated its molecular mechanism, revealing how specific disruption of TF co-binding alters *CCND3* regulation to drive variation in red blood cell count. Together, our approach enables genome-wide functional predictions of trait-associated noncoding variants by linking transcription factor binding events to their target genes.

Introduction

Genome-wide association studies (GWAS) have cataloged thousands of trait-associated variants, yet the majority reside in noncoding regions and remain functionally uncharacterized, posing a major hurdle to interpreting their roles in phenotypic mechanisms[1–4]. Furthermore, GWAS typically identify variants in linkage disequilibrium (LD) rather than the causal alleles themselves[5–7]. Although fine-mapping has improved the localization of causal variants, the underlying regulatory mechanisms often remain unresolved[6,8,9]. Moreover, addressing the 'missing heritability' of complex traits requires looking beyond common variants to rare variants, which can pinpoint specific causal genes with large effect sizes[10]. However, rare variant studies have predominantly focused on coding regions due to the difficulty of interpreting non-coding effects[11]. Consequently, characterizing noncoding variants remains a critical bottleneck; establishing the functional links between these loci, their distal target genes, and the resulting epigenetic landscape constitutes a complex barrier to understanding disease etiology.

Transcription factors (TFs) are pivotal regulators of gene expression that bind to specific non-coding genomic regions[12,13]. Genetic variants located within these binding motifs can alter TF affinity by either disrupting existing motifs or creating *de novo* ones, thereby dysregulating target gene expression and potentially contributing to disease pathogenesis[14–16]. Elucidating these non-coding mechanisms is critical for therapeutic development[1,2]. A prime example is the gene-editing therapy Casgevy[17], which targets a *BCL11A* enhancer, originally identified via GWAS, to reactivate fetal hemoglobin (γ-globin) expression for the treatment of β-thalassemia[18–20]. Recently, whole-genome sequencing (WGS) data from nearly 500,000 UK Biobank participants have been released, making it one of the largest ongoing population-based studies[21]. Approximately 1.5 billion variants including single-nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants (SVs), have been identified in this cohort. This dataset constitutes a robust resource for GWAS analyses and affords an unprecedented opportunity to advance understanding of disease mechanisms, especially for noncoding regulatory variants that modulate disease risk.

In the meantime, a novel technique named FOODIE was introduced, which measures TF binding sites in the human genome at near single-base resolution using single-molecule deaminase footprinting[22] (Fig. 1a-b). This technique provides a novel avenue to understand the TF-mediated mechanisms of noncoding variants[23]. Concurrently, a new artificial intelligence model, AlphaGenome[24,25], has emerged as a leading predictor of how single variants impact gene regulatory processes[26]. Together, these tools offer the potential to decode the TF-mediated genome functions contributing to complex traits and diseases. However, there is currently no integrative framework that allows us to systematically combine this cutting-edge experimental technique with AI models for variant effect investigation.

To bridge this gap, we propose an integrative framework, varTFBridge, which adapts a broad spectrum of well-established approaches alongside AlphaGenome, to identify causal noncoding common and rare variants within FOODIE footprints and to functionally investigate their regulatory mechanisms in TF-mediated gene regulation. We first demonstrate that FOODIE footprints improve key tasks in understanding gene regulation in K562 cells, including TF binding measurement and enhancer–gene linkage identification,

while showing higher heritability enrichment for erythrocyte traits compared with DNase-seq and ATAC-seq. We then apply varTFBridge to map 209 common variants (Minor Allele Frequency, MAF $\geq$ 0.1%) and 18 rare variants (MAF < 0.1%) to TF binding sites and their target genes, using 490,640 UK Biobank genomes across 13 erythrocyte traits. This framework successfully recapitulates the causal variant rs112233623 and further infers its molecular function in regulating the *CCND3* gene: the variant disrupts a GATA1/TAL1 co-binding site, resulting in higher mean corpuscular volume (MCV) and lower red blood cell count (RBC). Together, our study demonstrates a generalizable strategy utilizing FOODIE TF footprint maps to connect genetic associations to the molecular mechanisms of complex traits and diseases.

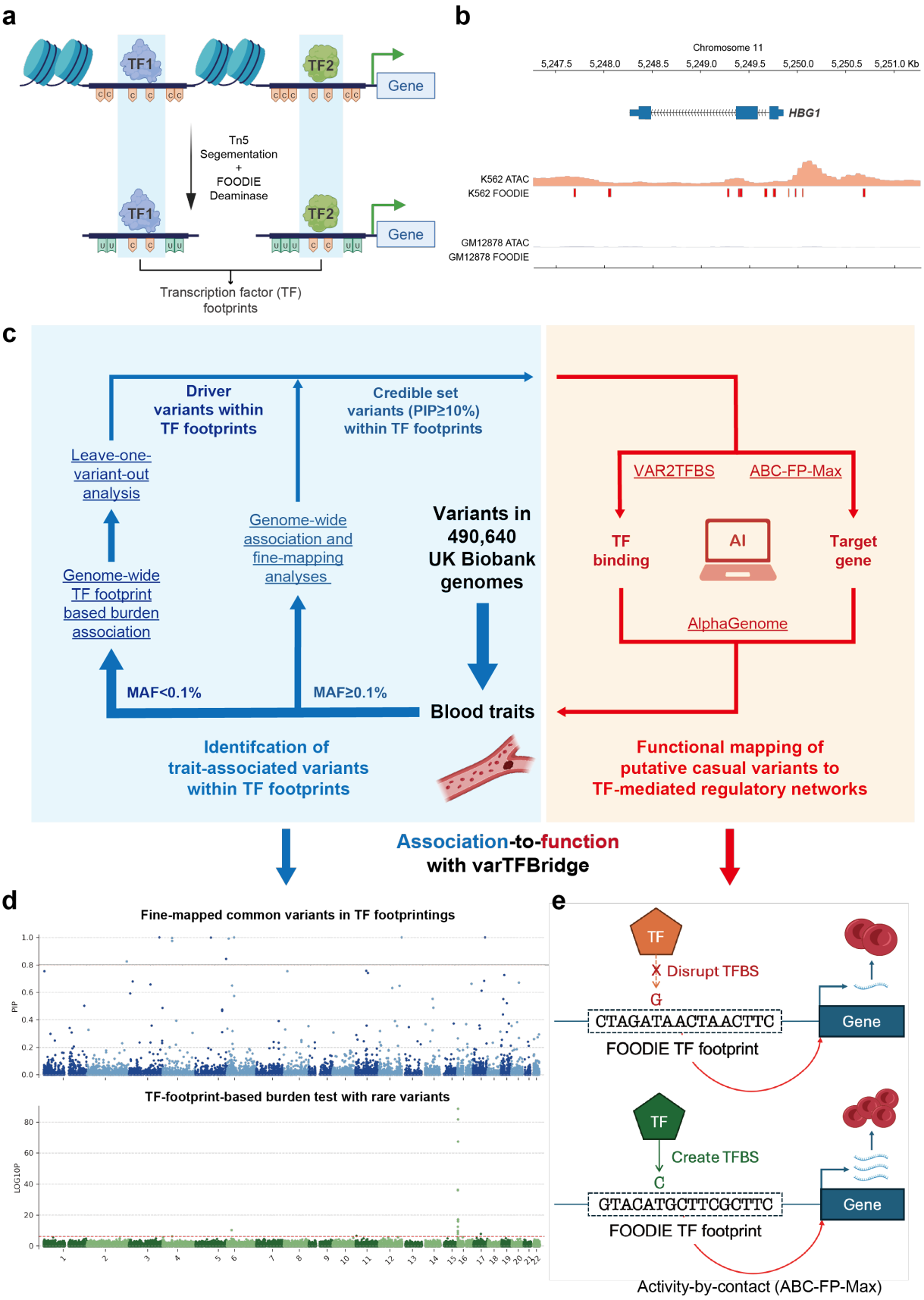

a
TF1
TF2
Gene
Tn5
Segementation
+
FOODIE
Deaminase
Transcription factor (TF)
footprints
b
Chromosome 11
5,247.5 5,248.0 5,248.5 5,249.0 5,249.5 5,250.0 5,250.5 5,251.0 Kb
HBG1
K562 ATAC
K562 FOODIE
GM12878 ATAC
GM12878 FOODIE
c
Driver
variants within
TF footprints
Credible set
variants (PIP≥10%)
within TF footprints
Leave-one-
variant-out
analysis
Genome-wide
association and
fine-mapping
analyses
Variants in
490,640
UK Biobank
genomes
Genome-wide
TF footprint
based burden
association
MAF<0.1%
MAF≥0.1%
Blood traits
Identifcation of
trait-associated variants
within TF footprints
VAR2TFBS
ABC-FP-Max
TF
binding
AI
Target
gene
AlphaGenome
Functional mapping of
putative casual variants to
TF-mediated regulatory networks
Association-to-function
with varTFBridge
d
Fine-mapped common variants in TF footprintings
PIP
TF-footprint-based burden test with rare variants
LOG10P
e
TF
Disrupt TFBS
G
CTAGATAACTAACTTC
FOODIE TF footprint
Gene
TF
Create TFBS
C
GTACATGCTTCGCTTC
FOODIE TF footprint
Gene
Activity-by-contact (ABC-FP-Max)

Fig. 1 | Overview of varTFBridge and FOODIE footprinting. a, Schematic of FOODIE, in which nuclei are treated with a cytosine deaminase following Tn5 transposition, converting cytosines to uracils in proportion to chromatin accessibility. **b,** Genome browser view of the *HBG1* locus encoding the γ-globin chain and highly expressed in K562. ATAC-seq tracks for K562 and GM12878 cells (rows 1 and 3) are shown alongside corresponding FOODIE footprint tracks (rows 2 and 4). **c,** The varTFBridge workflow. Procedures highlighted in blue identify blood-trait–associated credible variants from 490,640 UK Biobank participants, whereas procedures in red functionally map these variants to transcription factor–mediated gene regulatory networks. Underlined text denotes methods implemented within the varTFBridge framework. Arrows denote the progression from genetic association to functional interpretation. **d,** Manhattan plots for fine-mapped common variants overlapping FOODIE footprints (top) and footprint-based rare-variant burden tests (bottom) associated with blood cell traits. Posterior inclusion probabilities (PIPs) for common variants were estimated using genome-wide fine-mapping with SBayesRC. Each point represents a variant (top) or a FOODIE footprint (bottom) plotted by genomic position. **e,** Conceptual illustration of how regulatory variants that alter transcription factor binding, either disrupting or creating binding sites, to modulate gene expression and drive red blood cell phenotypic variation. TFBS: transcription factor binding site.

Results

The varTFBridge framework

We developed varTFBridge, an integrated workflow leveraging whole-genome sequencing (WGS) data from 490,640 UK Biobank individuals[21] to functionally identify transcription factor (TF)-mediated noncoding regulatory variants for blood traits in a cell-type-specific context. The workflow elucidates how these variants rewire gene regulatory networks by altering TF binding sites anchored in predictive footprint-gene maps. Specifically, varTFBridge employs a two-stage analysis to elucidate the regulatory impact of these variants: 1) identification of trait-associated common and rare variants, and 2) functional dissection of putative causal variants at the level of TF binding sites and downstream target genes.

In the initial variant association stage, we implemented a comprehensive framework to analyze both common and rare variants associated with blood traits (**Fig. 1c-d**), utilizing whole-genome sequencing data from 490,640 individuals. For common variants (MAF ≥ 0.1%), we performed GWAS followed by genome-wide fine-mapping (GWFM) with SBayesRC[27] to pinpoint putative causal variants within cell-type-specific TF footprints. For rare variants (MAF < 0.1%), we conducted footprint-based burden tests by collapsing variants within each footprint to identify significant trait-associated TF footprints. As a further sensitivity analysis, we performed leave-one-variant-out analysis[28] to isolate the specific driver rare variants within each significant footprint.

In the second stage, focused on the functional dissection of putative causal variants, we adapted a series of well-established models to determine how these variants modulate TF binding sites and epigenomic states to regulate their target genes. We used FIMO[29] to identify JASPAR-based TF binding sites within TF footprints and assessed how putative causal variants affect the predicted binding affinity (termed VAR2TFBS). We then adapted the ABC-Max model[8,30] by integrating TF footprints (termed ABC-FP-Max) to improve upon

the original ABC model in predicting the variant-gene pairs. Finally, we employed AlphaGenome[25], a state-of-the-art variant effect prediction model, to assess the effects of putative causal variants across multiple layers of epigenomic states, including histone modifications, TF binding and chromatin accessibility.

Collectively, varTFBridge enables identification and mapping of regulatory variants from TF footprints to precise binding sites and their target genes. We demonstrate this utility by screening 13 erythroid-trait–associated variants to identify functional regulatory variants in K562 cells.

FOODIE footprints enhance transcription factor binding measurement and target gene predictions

Accurate TF footprint detection is foundational to Transcription Factor Binding Site (TFBS) identification and gene regulatory network inference. Therefore, we first evaluated FOODIE's capacity to capture TF binding and its utility in constructing gene regulatory networks alongside existing multi-omic assays. We applied the designated footprint-calling protocol to published FOODIE datasets for K562 cell lines[22], yielding 188,484 footprints with a median length of 19 bp. Comparison with ENCODE[31] DNase and ATAC-seq data in K562 (**Fig. 2a**) revealed that 91.8% of footprints overlap with at least one chromatin accessibility peak, while 61% are supported by both. These results align with the expectation that TF footprints predominantly reside in open chromatin regions.

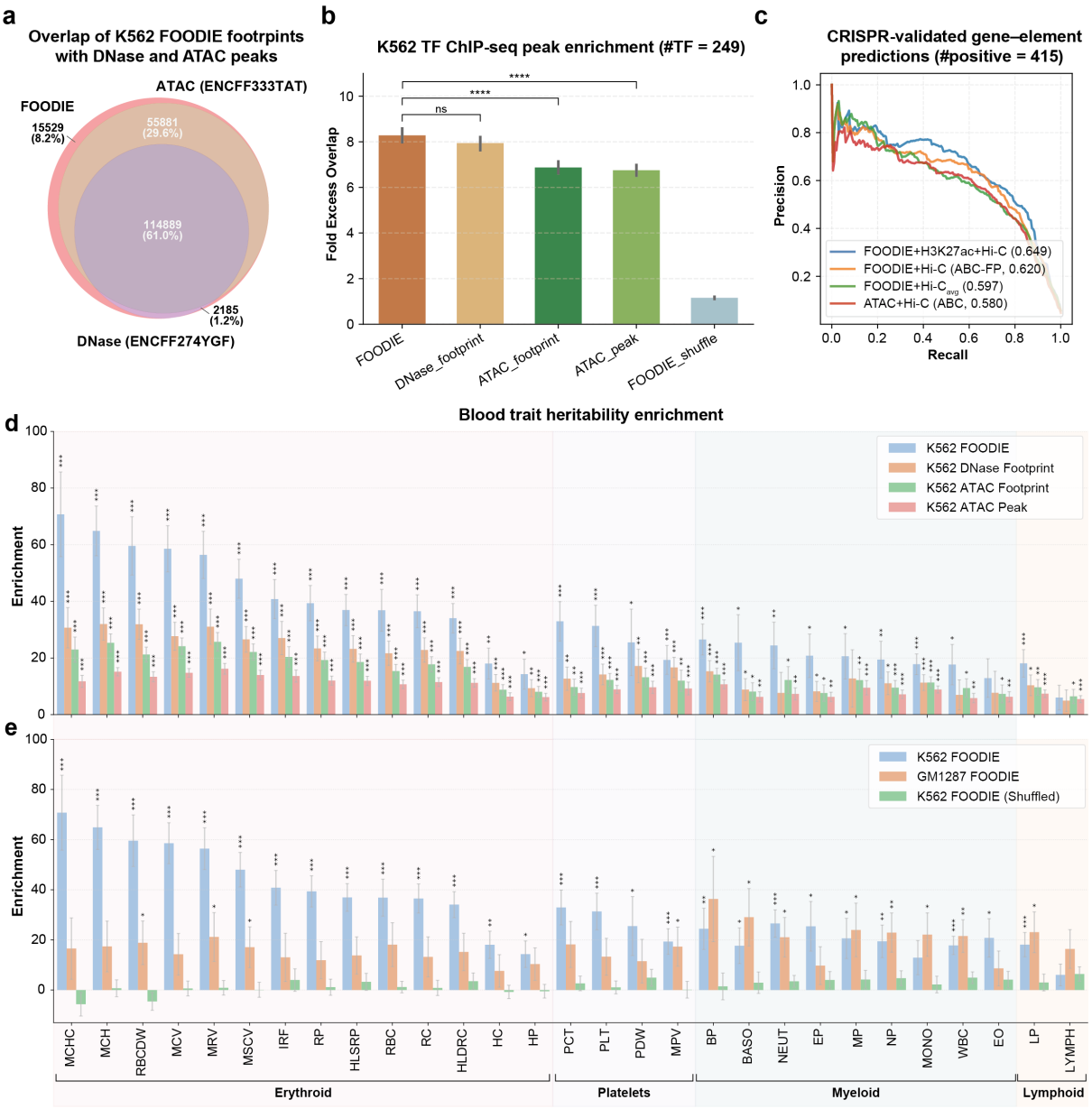
We then assessed whether FOODIE footprints accurately reflect direct measurements of TF binding in a cell-type-specific context. We compiled a K562 ChIP-seq dataset (249 TFs; 317 experiments from ENCODE[31]) and compared FOODIE with DNase- and ATAC-derived footprints using fold excess overlap[14]. The comparative footprints were obtained using the widely accepted methods footprint-tools[12] and HINT[32], respectively (**Fig. 2b**). FOODIE footprints exhibited 7.9× (SE 2.2 under the jackknife with 200 blocks dividing the whole genome) excess overlap with TF ChIP-seq, and DNase footprints exhibited comparable 7.5× (SE 2.1) excess overlap. However, ATAC footprints and ATAC peaks only achieved 6.5× (SE 1.8) and 6.4× (SE 1.7), respectively. This confirms that FOODIE footprints more accurately capture cell-type-specific TF binding.

To link noncoding regulatory variants within TF footprints to target genes, we adapted the Activity-by-Contact (ABC) framework[30] to develop ABC-FP (Activity-by-Contact for footprints). We defined putative regulatory elements by extending FOODIE footprints by 50 bp and merging overlapping regions, thereby grouping adjacent sites to capture cooperative TF binding that regulates the same target genes. This resulted in 70,670 regulatory elements with a median length of 534 bp. We then followed the ABC protocol[33] to calculate the activity (e.g., ATAC-seq and H3K27ac ChIP-seq) of these elements and the Hi-C contact frequency between them and all genes within a 100 Mb window. The resulting method, ABC-FP, quantifies a FOODIE footprint's regulatory effect by weighting its activity by its chromatin contact frequency with the target gene's promoter (**Methods**).

To evaluate the accuracy of enhancer-gene interaction predictions, we utilized a large-scale K562 CRISPR perturbation dataset[34,35], which contains 7,238 putative element-gene pairs and 418 functionally validated positives. FOODIE TF footprints showed significant enrichment within CRISPR-validated enhancer regions, achieving an odds ratio of 3.51 relative to a shuffled baseline (OR = 1; **Supplementary Figure 1**). We further performed

benchmarking by applying ABC-FP to this set using different epigenomic profiles (**Fig. 2c**). As expected, integrating FOODIE-based regulatory elements with integrative ATAC and H3K27ac activity and K562-specific Hi-C contact maps achieved the highest accuracy (AUPRC = 0.649), while using ATAC alone with Hi-C remained robust (AUPRC = 0.62). Given that cell-type-specific Hi-C data is unavailable for most cell lines, we also tested a cell-type-agnostic reference Hi-C map (averaged across 33 ENCODE cell lines) to approximate contact frequencies in K562[8]. Notably, even with this approximation, ABC-FP maintained high accuracy (AUPRC = 0.597), surpassing the original ATAC-based ABC score (AUPRC = 0.58).

Our experiments show that FOODIE footprints outperform ATAC-seq in capturing cell-type-specific transcription factor binding and in predicting regulatory element–gene interactions. However, ChIP-seq peaks are not a true gold standard for TF-binding sites[36], since peak regions (200–500 bp) exceed canonical footprints (6–20 bp)[14], potentially underestimating enrichment. Given that binding enrichment does not necessarily imply phenotypic relevance, we therefore turn to blood cell trait heritability analyses to evaluate FOODIE footprints and investigate their potential applications.


a
Overlap of K562 FOODIE footrpints with DNase and ATAC peaks
ATAC (ENCFF333TAT)
FOODIE
15529 (8.2%)
55881 (29.6%)
114889 (61.0%)
2185 (1.2%)
DNase (ENCFF274YGF)
b
K562 TF ChIP-seq peak enrichment (#TF = 249)
Fold Excess Overlap
ns
****
****
FOODIE
DNase_footprint
ATAC_footprint
ATAC_peak
FOODIE_shuffle
c
CRISPR-validated gene–element predictions (#positive = 415)
Precision
Recall
FOODIE+H3K27ac+Hi-C (0.649)
FOODIE+Hi-C (ABC-FP, 0.620)
FOODIE+Hi-C$_{avg}$ (0.597)
ATAC+Hi-C (ABC, 0.580)
d
Blood trait heritability enrichment
Enrichment
K562 FOODIE
K562 DNase Footprint
K562 ATAC Footprint
K562 ATAC Peak
e
Enrichment
K562 FOODIE
GM1287 FOODIE
K562 FOODIE (Shuffled)
MCHC
MCH
RBCDW
MCV
MRV
MSCV
IRF
RP
HLSRP
RBC
RC
HLDRC
HC
HP
PCT
PLT
PDW
MPV
BP
BASO
NEUT
EP
MP
NP
MONO
WBC
EO
LP
LYMPH
Erythroid
Platelets
Myeloid
Lymphoid

184

Fig. 2 | Genome-wide characterization of FOODIE transcription factor footprints. a. Venn diagram showing overlap between FOODIE footprints and DNase-seq and ATAC-seq peaks in K562 cells. ENCODE sample identifiers are provided in parentheses. **b.** Comparison of transcription factor (TF) footprints and open chromatin profiles based on fold excess overlap with TF ChIP–seq peaks. Fold excess overlap with TF ChIP–seq peaks in K562 cells is shown for FOODIE, ATAC-seq and DNase-seq footprints, alongside ATAC-seq peaks and shuffled FOODIE footprints (317 ChIP–seq experiments covering 249 unique TFs). Error bars denote one standard error. **** indicates $P < 1 \times 10^{-4}$ by two-sided *t*-test; ns denotes non-significant differences ($P > 0.05$). **c.** Precision–recall curves for classifiers predicting enhancer–gene (EG) pairs. Positive EG pairs are defined as those in which CRISPR perturbation of a distal element significantly decreases expression of the target gene. In total, 415 positive EG pairs among 7,238 tested EGs were used for benchmarking. Curves show performance across different combinations of epigenomic profiles within the Activity-by-Contact (ABC) framework. ABC-FP denotes the footprint–gene linkage score proposed in this study, whereas ABC denotes the original ATAC–based ABC score. **d.** Bar plot showing heritability enrichment (proportion of heritability divided by the proportion of SNPs) for four footprint annotations and a published ENCODE ATAC-seq peak annotation in K562 cells across 29 blood cell–related traits. *** indicates $P < 1 \times 10^{-3}$ by one-sided *t*-test; ** indicates $P < 1 \times 10^{-2}$ by one-sided *t*-test; * indicates $P < 5 \times 10^{-2}$ by one-sided *t*-test. **e.** Bar plot of heritability enrichment for FOODIE footprints in K562 and K562 cells across 29 blood cell–related traits. Shuffled K562 FOODIE footprints serve as a baseline. Traits are color-coded by the respective lineage. Blood cell traits indicated in the figure are as follow: RC: Reticulocyte count; IRF: Immature fraction of reticulocytes; HP: Hematocrit; HC: Hemoglobin concentration; HLDRC: High light scatter reticulocyte count; MCHC: Mean corpuscular hemoglobin concentration; MCH: Mean corpuscular hemoglobin; HLSRP: High light scatter reticulocyte percentage; RBCDW: Red cell distribution width; MCV: Mean corpuscular volume; MRV: Mean reticulocyte volume; MSCV: Mean sphered corpuscular volume; RP: Reticulocyte fraction of red cells; RBC: Red blood cell count; LP: Lymphocyte percentage of white cells; LYMPH: Lymphocyte count; Monocyte: Monocyte count; NP: Neutrophil percentage of white cells; NEUT: Neutrophil count; EP: Eosinophil percentage of white cells; BASO: Basophil count; WBC: White blood cell count; MP: Monocyte percentage of white cells; EO: Eosinophil count; BP: Basophil percentage of white cells. PLT: Platelet count; MPV: Mean platelet volume; PCT: Plateletcrit; PDW: Platelet distribution width. MONO, NP, NEUT, EP, BASO, WBC, MP, EO, and BP are non-erythroid myeloid traits; LP and LYMPH are lymphoid traits; PLT, MPV, PDW, and PCT are megakaryocytic traits; and RC, IRF, HP, HC, HLDRC, MCHC, MCH, HLSRP, RBCDW, MCV, MRV, MSCV, RP, and RBC are erythroid traits.

FOODIE footprints are disproportionately enriched in the heritability of erythrocyte traits

Given the erythroleukemic origin and erythroid differentiation potential of K562 cells[37], we examined the enrichment of K562 FOODIE footprints for blood cell trait heritability. We analysed 29 traits from the UK Biobank (14 erythroid, 4 platelet, 9 myeloid, and 2 lymphoid phenotypes) using partitioned linkage disequilibrium score regression. We benchmarked FOODIE footprints against three other chromatin annotations: DNase footprints, ATAC footprints, and ATAC peaks as baseline. We estimated heritability enrichment for each footprint annotation using stratified LD score regression (S-LDSC) with the baseline-LD model v2.0[38] (**Methods**), conditioning on the baseline by adding each annotation individually. To ensure sufficient single-nucleotide polymorphisms (SNPs) coverage for robust estimation,

197 the three TF footprint (i.e., FOODIE, ATAC and DNase) annotations were extended by 50
198 bp[14], whereas ATAC peaks were analyzed at their original width.

199 We observed that K562 FOODIE footprints yielded the highest heritability enrichment in 28
200 of 29 blood traits, significantly exceeding ATAC and DNase footprints in erythroid and
201 myeloid traits (**Fig. 2d**). Specifically, K562 FOODIE footprints showed a median 40-fold and
202 28-fold enrichment for erythroid and myeloid traits, respectively, compared to only 12-fold for
203 lymphoid traits. This observation aligns well with the biological identity of K562 as a myeloid
204 cell line with strong erythroid characteristics. Notably, FOODIE-derived footprint regions
205 comprise <0.5% of the genome yet exhibit a 70-fold enrichment for Mean Corpuscular
206 Hemoglobin Concentration (MCHC); this small proportion of genome explains 40.27% of the
207 trait's heritability. By contrast, DNase and ATAC-derived footprints exhibit significantly lower
208 enrichment (30-fold and 21-fold, respectively). Furthermore, FOODIE footprints capture a
209 larger fraction (28.6%) of high-confidence fine-mapped variants (Posterior Inclusion
210 Probability [PIP] > 90%; **Methods**) for MCHC than DNase footprints (21.4%). These results
211 demonstrate that FOODIE footprints offer superior resolution compared to ATAC-seq and
212 DNase-seq footprints, pinpointing the functional variants driving erythroid phenotypes,
213 thereby providing a more precise explanation of trait heritability.

214 To further examine if FOODIE footprints exhibit lineage-specific enrichment in distinct blood
215 traits, we applied the same pipeline to obtain FOODIE footprints of GM12878 (a
216 lymphoblastoid cell line) and performed LDSC analysis alongside K562 (**Fig. 2e**). As
217 expected, GM12878 footprints showed higher heritability enrichment for two lymphoid traits
218 and 6 (out of 9) myeloid traits, specifically those related to white blood cell, monocyte, and
219 basophil counts. The shuffled control consistently showed negligible enrichment, confirming
220 the statistical significance of the observed signals.

221 Collectively, our results indicate that FOODIE footprints improve upon current ATAC and
222 DNase-based methods by more accurately capturing TF binding and enhancer-gene
223 linkages. We also observed that these footprints are highly enriched for heritability in blood
224 traits, particularly within the erythroid lineage. Motivated by this advantage, we integrated
225 varTFBridge with K562 FOODIE footprints to analyze 490,640 UK Biobank genomes, aiming
226 to pinpoint specific causal variants that rewire the TF-mediated gene regulatory networks
227 underlying erythroid traits.

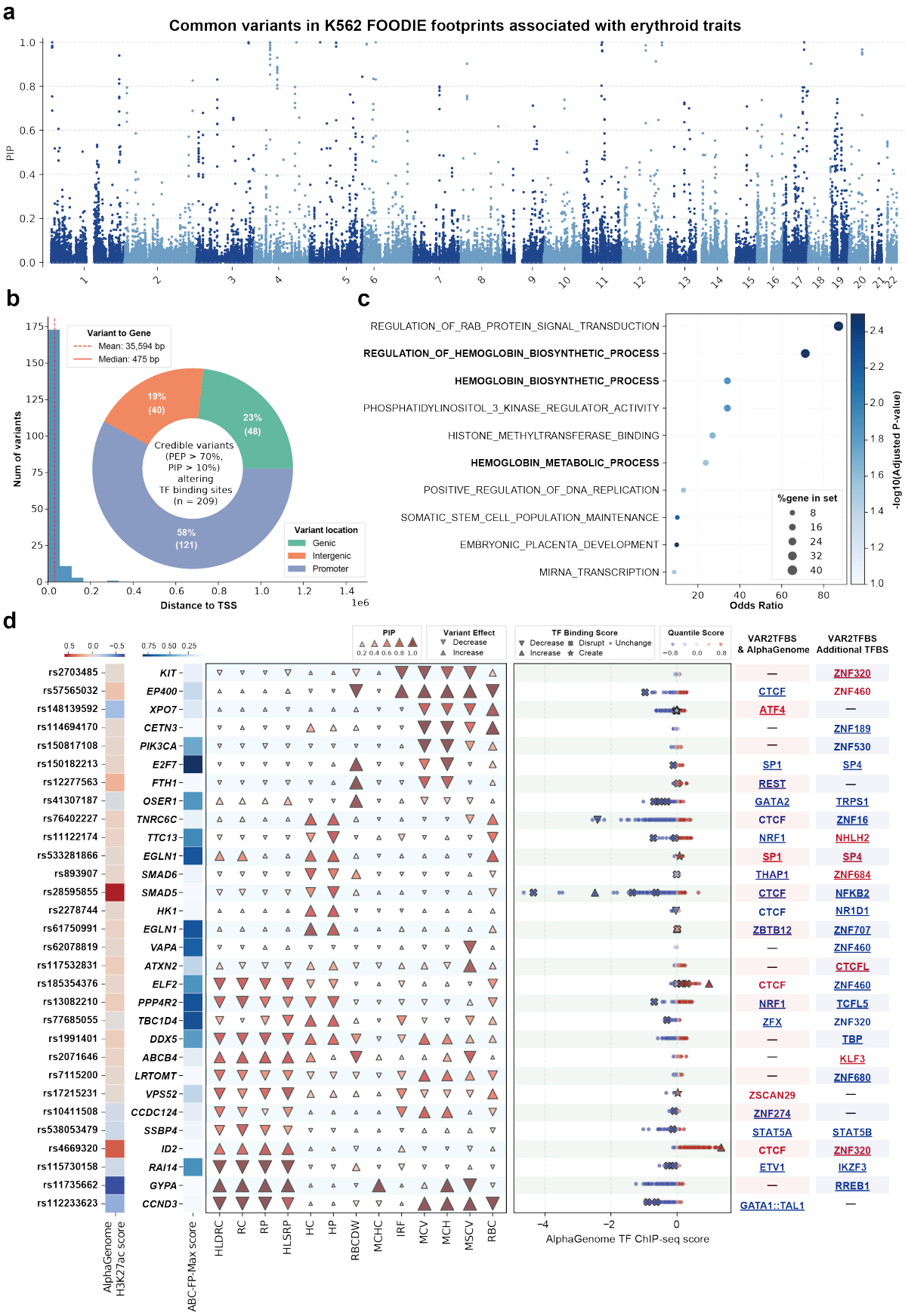

a
Common variants in K562 FOODIE footprints associated with erythroid traits
PIP
b
Variant to Gene
Mean: 35,594 bp
Median: 475 bp
Num of variants
Distance to TSS
1e6
19% (40)
23% (48)
58% (121)
Credible variants (PEP > 70%, PIP > 10%) altering TF binding sites (n = 209)
Variant location
Genic
Intergenic
Promoter
c
REGULATION_OF_RAB_PROTEIN_SIGNAL_TRANSDUCTION
REGULATION_OF_HEMOGLOBIN_BIOSYNTHETIC_PROCESS
HEMOGLOBIN_BIOSYNTHETIC_PROCESS
PHOSPHATIDYLINOSITOL_3_KINASE_REGULATOR_ACTIVITY
HISTONE_METHYLTRANSFERASE_BINDING
HEMOGLOBIN_METABOLIC_PROCESS
POSITIVE_REGULATION_OF_DNA_REPLICATION
SOMATIC_STEM_CELL_POPULATION_MAINTENANCE
EMBRYONIC_PLACENTA_DEVELOPMENT
MIRNA_TRANSCRIPTION
%gene in set
Odds Ratio
-log10(Adjusted P-value)
d
PIP
Variant Effect
Decrease
Increase
TF Binding Score
Decrease
Disrupt
Unchange
Increase
Create
Quantile Score
VAR2TFBS & AlphaGenome
VAR2TFBS Additional TFBS
rs2703485 KIT — ZNF320
rs57565032 EP400 CTCF ZNF460
rs148139592 XPO7 ATF4 —
rs114694170 CETN3 — ZNF189
rs150817108 PIK3CA — ZNF530
rs150182213 E2F7 SP1 SP4
rs12277563 FTH1 REST —
rs41307187 OSER1 GATA2 TRPS1
rs76402227 TNRC6C CTCF ZNF16
rs11122174 TTC13 NRF1 NHLH2
rs533281866 EGLN1 SP1 SP4
rs893907 SMAD6 THAP1 ZNF684
rs28595855 SMAD5 CTCF NFKB2
rs2278744 HK1 CTCF NR1D1
rs61750991 EGLN1 ZBTB12 ZNF707
rs62078819 VAPA — ZNF460
rs117532831 ATXN2 — CTCFL
rs185354376 ELF2 CTCF ZNF460
rs13082210 PPP4R2 NRF1 TCFL5
rs77685055 TBC1D4 ZFX ZNF320
rs1991401 DDX5 — TBP
rs2071646 ABCB4 — KLF3
rs7115200 LRTOMT — ZNF680
rs17215231 VPS52 ZSCAN29 —
rs10411508 CCDC124 ZNF274 —
rs538053479 SSBP4 STAT5A STAT5B
rs4669320 ID2 CTCF ZNF320
rs115730158 RAI14 ETV1 IKZF3
rs11735662 GYPA — RREB1
rs112233623 CCND3 GATA1::TAL1 —
AlphaGenome H3K27ac score
ABC-FP-Max score
HLDRC
RC
RP
HLSRP
HC
HP
RBCDW
MCHC
IRF
MCV
MCH
MSCV
RBC
AlphaGenome TF ChIP-seq score

Fig. 3 | varTFBridge deciphers the impact of common variants on transcription-factor mediated function. a, Manhattan plots of fine-mapped common variants (MAF>0.1%) overlapping K562 FOODIE footprints across 13 erythroid traits. Posterior inclusion probabilities (PIPs) were estimated by Genome-wide fine-mapping using SBayesRC. Each point represents a footprint-overlapping variant plotted by genomic position. **b,** Pie chart depicting the proportion of credible set variants (PIP > 10% and PEP > 70%) by genomic annotation. "Genic" indicates variants within a gene body (intron); "Promoter" denotes variants within 500 bp of a gene's transcription start site (TSS); "Intergenic" refers to regions without gene annotations. Gene annotations are based on ABC scores. Histogram showing the distribution of distances from variants to their predicted target genes, as determined by ABC-FP-Max. **c,** Gene Ontology (GO) enrichment of ABC-FP-Max predicted targets derived from erythroid trait–associated credible set variants (PIP > 10%). Each dot represents an enriched GO term. The x-axis indicates the odds ratio (OR) for enrichment; the y-axis lists the GO terms. Dot color signifies enrichment significance as −log10(adjusted P-value), and dot size corresponds to the percentage of genes in the GO term. **d,** Integrative demonstration of credible set variants (PIP > 70%) from trait association to TF-mediated regulatory function via VAR2TFBridge. The first column displays a heatmap indicating AlphaGenome-predicted variaant effects on K562 H3K27ac ChIP-seq (enhancer activity), with variant rsIDs on the y-axis. The second column shows a heatmap of ABC-FP-Max scores for variant-gene pairs, with the predicted target gene on the y-axis. The third column presents a heatmap of PIP values for each variant-trait pair, where triangle size denotes PIP magnitude and up/down triangles indicate positive/negative trait effects, respectively. The fifth column illustrates AlphaGenome-predicted variant variant effects on K562 TF ChIP-seq using scatter plots, where each point represents an AlphaGenome raw score for a specific TF, colored by its AlphaGenome quantile variant score. Up-triangles signify increased TF binding via VAR2TFBS, down-triangles decreased binding, crosses indicate motif disruption, and stars denote motif creation. The seventh column lists TFs whose binding sites are predicted to be altered by variants from VAR2TFBS with an AlphaGenome quantile score > 0.7. The last column identifies VAR2TFBS-predicted TFs whose binding sites are altered but are not in the AlphaGenome prediction list. TF names colored blue indicate decreased or disrupted binding (underlined for disruption), while red indicates increased or created binding (underlined for creation). If VAR2TFBS predicts multiple TFs for a variant, the one with the highest RNA-seq expression in K562 is listed; a hyphen indicates no TF was found by the specific strategy. A full list of TFs is in Supplementary Table 2, and erythroid trait names of aberration are detailed in the legend of Fig. 3.

varTFBridge functionally prioritizes common noncoding variants by elucidating TF–mediated regulatory mechanisms

To identify the credible common variants associated with 13 erythroid traits and elucidate the TF-mediated regulatory mechanism using K562 FOODIE footprint annotations, we performed genome-wide association studies (GWAS) for erythroid traits across 490,640 UK Biobank participants of White European ancestry. After quality control, we obtained an average of 21,107,428 variants per trait of 13 erythroid traits, with MAF ≥ 0.1% were available for analysis. Leveraging GWFM fine-mapping on variants, we obtained 251 fine-mapped variants with posterior inclusion probability (PIP) ≥ 10% and posterior SNP-heritability enrichment probability (PEP) ≥ 70% for at least one erythroid trait that overlap K562 FOODIE footprints (**Fig. 3a**). We then used our var2TFBridge workflow to investigate their TF-mediated regulatory mechanisms.

We first employed the ABC-FP model to construct genome-wide maps of footprint-to-gene linkage. Specifically, we calculated an ABC-FP score for each gene and FOODIE footprint within a 5-megabase pair (Mbp) window by multiplying the estimated chromatin accessibility (from ATAC-seq; **Methods**) by the three-dimensional contact frequency between the footprint and the gene promoter. Candidate footprint–gene pairs exceeding a specific threshold (0.015; **Methods**) were defined as having 'footprint–gene linkage,' and footprints predicted to regulate at least one gene were termed 'ABC-FP footprint'.

We next applied VAR2TFBS to determine how noncoding variants in the credible set, which localize to TF footprints, affect TF binding affinity. Specifically, we assessed whether these variants are predicted to disrupt, create, or alter the affinity of TF binding sites within FOODIE footprints. A variant effect on TF binding is estimated based on a position weight matrix (PWM) model for the binding specificity of the corresponding factor. We obtained 879 PWMs of vertebrate non-redundant TFs from JASPAR 2025[39]. We used FIMO[29] to identify statistically significant TFBSs ($p < 1e{-}4$; **Methods**) in both reference (REF) and alternative (ALT) sequences extracted from FOODIE footprints. For each TF, we compared the p-values between the reference and alternative sequences. A lower p-value in the alternative relative to the reference sequence indicates that the variant increases TF binding affinity, whereas a higher p-value indicates decreased binding. TFBSs with $P < 1 \times 10^{-4}$ in either the reference or alternative sequence were classified as TFBS disruptions or creations.

VAR2TFBS identified 209 credible set variants residing within ABC-FP footprints predicted to alter the binding affinity of at least one TF, in which 189 (90.4%) variants affected at least one expressed (K562 RNA-seq Transcripts Per Million [TPM] > 0.5) TF's binding site. These variants alter 1154 predicted binding sites across 330 TFs. Notably, 22% (n=254) of these are SP/KLF (Specificity Protein/Krüppel-like Factor) binding sites, a family known to play a critical role in hematopoiesis, particularly in erythropoiesis[16,40,41]. Furthermore, 3.7% (n=43) of the variants affect binding sites for CTCF/CTCFL, an essential architectural protein responsible for maintaining genome topology and mediating long-range chromatin interactions[42–44].

Subsequently, we linked these variants to target genes by assigning each variant to the gene with the highest ABC-FP score (i.e., ABC-FP-Max score). This analysis yielded 209 variant–gene pairs involving 195 unique genes (**Fig. 3b** and **Supplementary Table 1**); 58% of variants (n = 121) were in promoter regions, while 19% (n = 40) resided in distal intergenic regions. The median distance to the target gene was 475 bp (range: 5 bp–3.33 Mbp). We then performed Gene Ontology (GO) enrichment analysis[45] on the predicted target genes of TFBS-affected variants. These target genes were significantly enriched for biological processes related to hemoglobin biosynthesis and metabolism (**Fig. 3c**).

We observed that individual variants often affect multiple TF binding sites (median = 4), as identified by VAR2TFBS. This likely reflects high motif similarity across TF families. To quantify variant effects on TF binding specifically within the K562 cellular context, we leveraged AlphaGenome to generate 'quantile score'. These scores normalize the raw predicted differences between reference and alternative alleles against a background of common variants (**Methods**), thereby facilitating direct comparisons across diverse transcription factors. As such, using VAR2TFBS and AlphaGenome, we identified 28 TF binding sites with absolute quantile scores >80% (indicating effects above the 80th percentile of common variants) that are altered by a stringent credible set of 30 variants

287 (PIP > 70% for at least one erythroid trait). Notably, we observed several key TFs involved in
288 erythropoiesis such as GATA1, GATA2 and TAL1[46] (**Fig. 3d and Supplementary Data 1**).
289 Note that as AlphaGenome was trained on a limited set of 269 TF ChIP-seq profiles in K562,
290 its predictions are restricted to this specific set. Conversely, VAR2TFBS utilizes the JASPAR
291 database to cover 879 TF binding sites. This enables VAR2TFBS to predict variant effects
292 on additional key regulators of hematopoiesis, including KLF5, KLF6, and KLF16, which
293 AlphaGenome fails to identify (**Supplementary Table 2**).

294 Collectively, varTFBridge integrates a spectrum of tools and pipelines, GWAS, genome-wide
295 fine-mapping, and footprint-based burden tests, together with ABC-FP-Max, VAR2TFBS, and
296 AlphaGenome to identify and predict TF-mediated regulatory mechanisms of credible set
297 variants associated with erythroid traits. Furthermore, we proposed a visualization
298 framework (**Fig. 3d**) to demonstrate variant effects across multiple layers of regulation—from
299 chromatin states, histone modifications, and TF binding to target genes and trait effects. This
300 framework enables us to systematically elucidate how regulatory noncoding variants connect
301 to phenotypic traits, providing a roadmap for further experimental validation.

302 **varTFBridge elucidates the regulatory mechanisms of rs112233623 in erythroid traits**

303 To further demonstrate the capacity of varTFBridge in identifying variants that rewire
304 regulatory networks, we highlighted rs112233623, with MAF of 1% in the European
305 population, and shows a median PIP of 100% across 13 erythroid traits among all variants in
306 K562 FOODIE footprints. This variant is situated within a K562 FOODIE footprint on
307 chromosome 6, which is positioned 15 kb upstream of the CCND3 gene's TSS (Isoform ID:
308 NM_001760). varTFBridge predicts that this C-to-T variant disrupts the co-binding site of
309 GATA1 and TAL1 (predicted by VAR2TFBS), thereby affecting the regulation of CCND3
310 (linked by ABC-FP-Max). Notably, AlphaGenome's absolute quantile scores for the variant's
311 effect on GATA1 and TAL1 binding are both higher than 0.999, corroborating the VAR2TFBS
312 prediction. FIMO motif scanning demonstrates that this variant alters the most conserved
313 base within the TAL1 binding site, drastically lowering the binding affinity (Fig. 4c). This
314 variant also falls within GATA1 and TAL1 ChIP-seq peaks in K562 cells reported in
315 ENCODE[31,47], further supporting the predictions.

316 Additionally, we leveraged AlphaGenome variant scores for different histone modifications
317 and chromatin accessibility to infer that rs112233623 likely decreases chromatin accessibility
318 and enhancer activity, thereby modulating *CCND3* expression. Specifically, variant scores for
319 DNase, ATAC, and enhancer histone marks (H3K27ac and H3K4me1) were negative, while
320 the score for the repressor marker (H3K27me3) was positive (**Fig. 4b and Supplementary**
321 **Data 2**). Notably, the absolute quantile scores for these epigenomic features exceeded
322 0.999, indicating the predicted effect is significantly distinct from background common
323 variants. This inference is corroborated by a previous study[48], which demonstrated via
324 Luciferase reporter assays that rs112233623 directly decreases enhancer activity.

325 Given that *CCND3* is the predicted target gene of rs112233623, we infer that the disruption
326 of the GATA1 and TAL1 binding site at this enhancer decreases *CCND3* expression. Since
327 the *CCND3* product, Cyclin D3, is a critical regulator of terminal erythroid proliferation, its
328 reduction likely impairs erythropoiesis, leading to increased erythrocyte size and reduced red
329 blood cell (RBC) counts[49]. This aligns perfectly with our GWAS results of rs112233623,
330 which show a negative effect on RBC count and a positive effect on mean corpuscular

volume (MCV) (**Fig. 4a**). Specifically, these results indicate that carriers of this variant have larger erythrocytes, but a lower total RBC count compared to non-carriers.

In summary, varTFBridge systematically elucidates the phenotypic effects and TF-mediated gene regulatory mechanisms of rs112233623, supported by existing experimental studies. These results demonstrate the potential of our framework to accelerate functional characterization of GWAS loci, bridging genetic association and molecular mechanism.

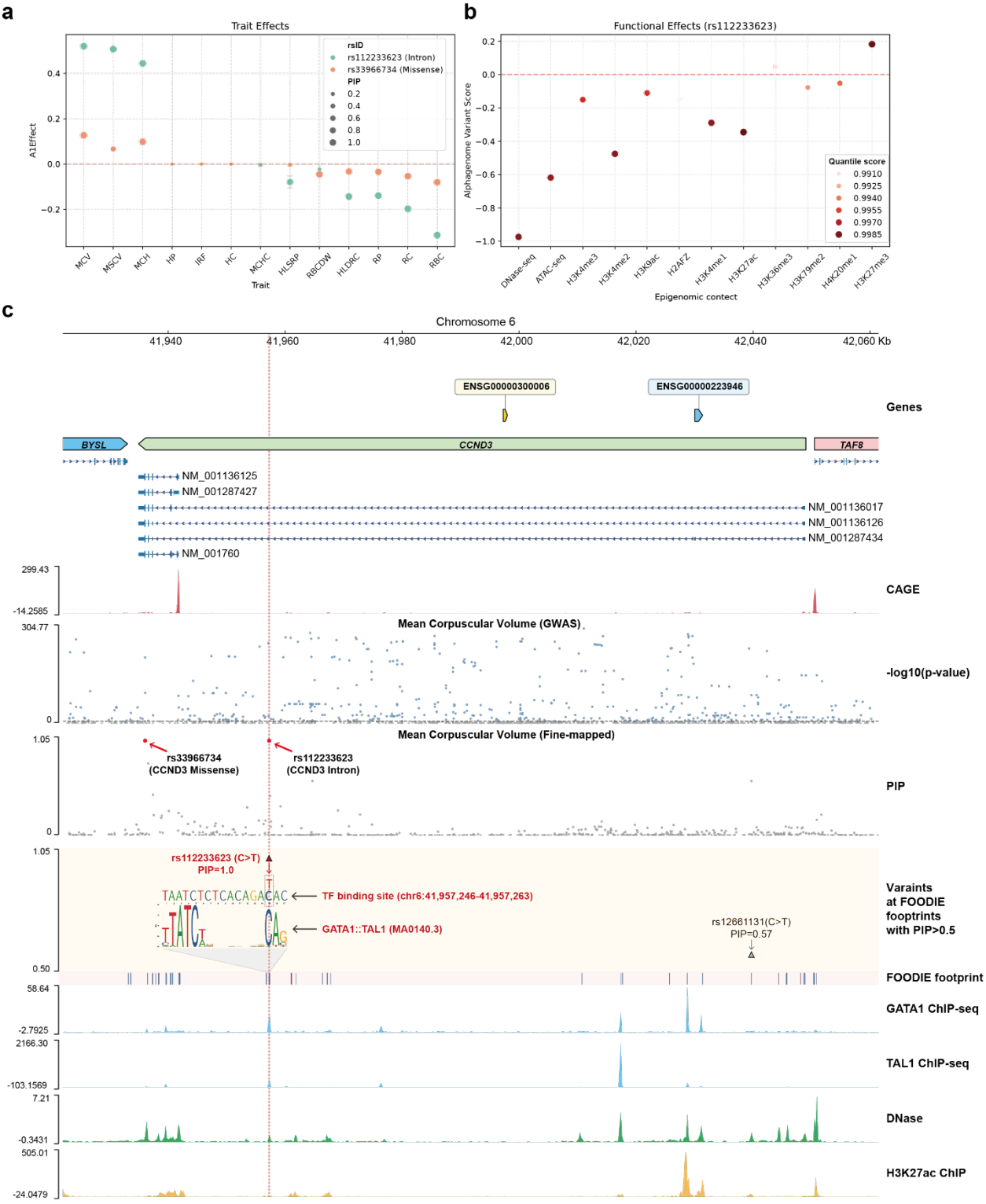


Fig. 4 | varTFBridge identifies rs112233623 as a causal regulatory variant predicted to modulate CCND3 expression by disrupting a GATA1::TAL1 binding site, thereby underlying erythroid phenotypic variation. a, Scatter plot showing the estimated effect size (BETA) and Posterior Inclusion Probability (PIP) of fine-mapped variants, rs112233623 and rs33966734, across 13 erythroid traits. A positive effect size indicates the variant increases the trait, while a negative effect decreases it. The size of each scatter point denotes its PIP value, and color differentiates between the variants. Error bars represent one standard error of the effect size, and the red dashed line indicates zero effect on the trait. **b,** Scatter plot displaying the AlphaGenome variant score of rs112233623 based on 11 histone marker ChIP-seq and 2 chromatin accessibility tracks. The color of each scatter indicates the absolute AlphaGenome quantile score, reflecting the variant's effect compared to background common variants. The red dashed line denotes a zero score from AlphaGenome, signifying no effect on the specific epigenome profiles. **d,** Genome browser view illustrating Mean Corpuscular Volume (MCV) associated variants at the *CCND3* genomic locus (chr6:41,922,000-42,061,500) with multiple epigenomic tracks. The first row displays the *CCND3* gene along with its different isoforms and two nearby genes. The second row presents GWAS signals, where each scatter represents a variant (MAF > 0.1% from 490,640 UK Biobank genomes) with-log10(p-value) from GWAS on the y-axis. Blue scatters denote significant variants (p-value > 5e-8, indicated by the grey dashed line). The third row shows fine-mapped signals for the same variants, with PIP on the y-axis. Red scatters represent the two credible set variants with PIP = 100%, namely rs112233623 (located in a distal enhancer upstream of CCND3) and rs33966734 (located in introns). The fourth row demonstrates variants at K562 FOODIE footprints (PIP > 50%) and specifically illustrates how rs112233623 (C-to-T substitution) disrupts the GATA1/TAL1 binding site. The sequence shown is extracted from the hg38 reference genome (chr6: 41,957,246-41,957,263), and the GATA1::TAL1 (JASPAR ID: MA0140.3) motif logo is sourced from the JASPAR 2025 database. The fifth row displays all K562 FOODIE footprints within this genomic region. The subsequent four rows present genomic tracks of GATA1 ChIP-seq, TAL1 ChIP-seq, DNase, and H3K27ac ChIP-seq signals from ENCODE, in order. The red vertical dashed line denotes the TSS of *CCND3*.

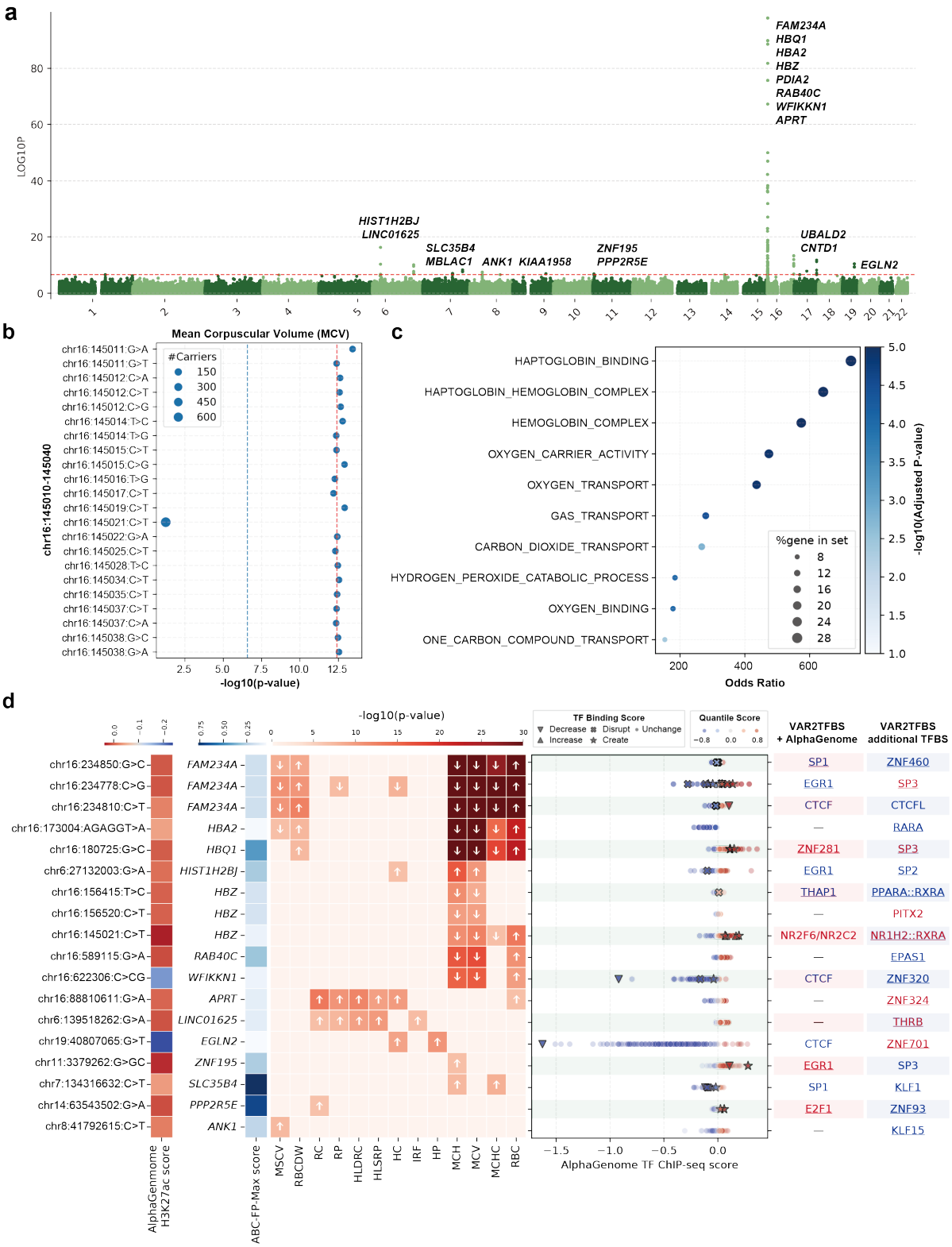
varTFBridge functionally prioritizes rare driver variants within erythroid trait–associated FOODIE footprints

To investigate whether varTFBridge can identify regulatory rare variants associated with erythroid-related traits, we first performed a footprint-based burden test by aggregating rare variants (MAF < 0.1%) within each K562 FOODIE footprint (number of carriers > 30) across 13 erythroid traits in 490,640 UK Biobank participants. After applying Bonferroni correction, we identified 23 FOODIE footprints significantly associated ($P < 2.65 \times 10^{-7}$) with at least one erythroid trait (**Fig. 5a**). To identify driver variants within each significant footprint, we performed leave-one-variant-out analyses by iteratively removing each rare variant and re-applying the burden test. A loss of footprint significance upon exclusion indicated that the footprint-level association was driven by that specific variant (**Fig. 5b**). As most footprint associations were driven by a single variant (**Supplementary Fig. 2**), we designated that variant as the driver for the associated trait. In total, we identified 18 rare driver variants and subsequently applied varTFBridge to investigate the TF-mediated regulatory mechanisms.

We further linked driver rare variants to their predicted target genes using ABC-FP-Max, identifying 14 unique genes. Gene ontology analysis revealed enrichment for haptoglobin-

and hemoglobin-related functions, aligning with modulation of blood-cell traits (**Fig. 5c**). To characterize TF binding consequences, we applied VAR2TFBS and AlphaGenome, identifying 18 rare variants that altered 35 TF-binding events across 24 TFs (absolute quantile score > 0.8) (**Fig 5d** and **Supplementary Data 3**). Notable TFs include erythroid regulators (ZBTB7A, ZNF148, NR2C2), genome-architecture TFs (CTCF, SP1), and cell-fate TFs (EGR1, ETV6).

We then focused on the *HBZ* gene, a key component of primitive erythropoiesis that is highly expressed in K562 cells[50,51]. A previous study[35] performed CRISPR perturbation to identify two *HBZ* enhancers. We identified a driver rare variant (chr16:145021:C>T; **Fig. 5b**) residing in one of the enhancers (chr16:144,736-145,236), which is predicted by varTFBridge to regulate *HBZ* by altering the binding of NR2F6 (quantile score=0.981) and NR2C2 (quantile score = 0.944). This variant lies within NR2F6 ChIP-seq peaks in K562 cells from ENCODE (**Supplementary Data 3**). Both CRISPR perturbation and ChIP-seq experiments support the varTFBridge prediction for this rare variant.


a
LOG10P
FAM234A
HBQ1
HBA2
HBZ
PDIA2
RAB40C
WFIKKN1
APRT
HIST1H2BJ
LINC01625
SLC35B4
MBLAC1
ANK1
KIAA1958
ZNF195
PPP2R5E
UBALD2
CNTD1
EGLN2
b
Mean Corpuscular Volume (MCV)
#Carriers
150
300
450
600
chr16:145010-145040
-log10(p-value)
c
HAPTOGLOBIN_BINDING
HAPTOGLOBIN_HEMOGLOBIN_COMPLEX
HEMOGLOBIN_COMPLEX
OXYGEN_CARRIER_ACTIVITY
OXYGEN_TRANSPORT
GAS_TRANSPORT
CARBON_DIOXIDE_TRANSPORT
HYDROGEN_PEROXIDE_CATABOLIC_PROCESS
OXYGEN_BINDING
ONE_CARBON_COMPOUND_TRANSPORT
%gene in set
Odds Ratio
-log10(Adjusted P-value)
d
-log10(p-value)
TF Binding Score
Decrease
Disrupt
Unchange
Increase
Create
Quantile Score
VAR2TFBS + AlphaGenome
VAR2TFBS additional TFBS
AlphaGenome H3K27ac score
ABC-FP-Max score
MSCV
RBCDW
RC
RP
HLDRC
HLSRP
HC
IRF
HP
MCH
MCV
MCHC
RBC
AlphaGenome TF ChIP-seq score

369

Fig. 5 | varTFBridge deciphers the impact of driver rare variants in erythroid-associated footprints on transcription-factor mediated function. a. Manhattan plots for FOODIE footprints-based burden test results for 13 erythroid traits, showing unadjusted two-sided P values derived from footprint-burden testing conducted in REGIENE and plotted on a -log10 scale. The x-axis denotes the genomic position of the center of the footprints. Scatters above the red dashed line are significant footprints ($P < 2.65 \times 10^{-7} = \frac{0.05}{188,484}$). The ABC-FP-Max predicted target genes of the significant footprints are labeled. **b,** Scatter plot illustrating the leave-one-out analysis results on a significant footprint (chr16:145,010-145,040) associated with Mean Corpuscular Volume (MCV). Each scatter represents a rare variant (MAF < 0.1%) within this footprint, with the size of the scatter indicating the number of carriers of the variant. The blue dashed line indicates the significance threshold (6.57 = -log10(0.05/188,484)), and the red dashed line denotes the-log10(p-value) of footprint-based burden tests without variant masking. **c,** Gene Ontology (GO) enrichment of ABC-FP-Max predicted targets derived from significant footprints associated with erythroid traits. Each dot represents an enriched GO term, with the x-axis indicating the odds ratio (OR) for enrichment and the y-axis listing the GO terms. Dot color signifies enrichment significance as-log10(adjusted P-value), while dot size corresponds to the percentage of genes in the GO term. **d,** Integrative demonstration of top driver rare variants at significant FOODIE footprints, linking trait associations to transcription factor–mediated regulatory function. The first column heatmap indicates the AlphaGenome predicted effect on K562 H3K27ac ChIP-seq tracks, with the y-axis showing the genomic locus of the variants. The second column heatmap shows the ABC-FP-Max score of variant-gene pairs, with the y-axis showing the predicted target gene. The third column heatmap shows the-log10(p-value) of each variant-trait pair, where an up-arrow denotes a positive effect on the traits and a down-arrow denotes a negative effect. The fifth column shows AlphaGenome-predicted variant effects on TF ChIP-seq, where each scatter represents an AlphaGenome score of a specific TF, and the color indicates the AlphaGenome quantile variant score. Scatters marked with an up-triangle indicate the variant predicts to increase TF binding affinity via VAR2TFBS; a down-triangle indicates a decrease; a cross indicates disruption of the TF binding motif; and a star indicates creation of the TF binding motif. The seventh column lists the TF whose binding site is predicted to be altered by a variant using VAR2TFBS with an AlphaGenome quantile score > 0.85. The last column indicates the VAR2TBS identified TF whose binding site is altered by the variant but is not in the AlphaGenome prediction TF list. TF names colored in blue denote decreased or disrupted TF binding (underlined for disruption), and red denotes increased or created TF binding (underlined for creation). If VAR2TFBS predicted multiple TFs for a variant, the one with the highest RNA-seq expression in K562 is listed; a hyphen denotes no TF was found by the specific strategy. The full list of TFs is in Supplementary Table 3, and the full names of erythroid traits are detailed in the legend of Fig. 3.

Discussion

Genetic associations serve a unique role in the studies of human biology, as they can establish causal links from variants or genes to human traits and diseases. However, nearly two decades after the first genome-wide association studies (GWAS), functionally characterizing noncoding variants and mapping them to their target genes and traits remain a significant bottleneck. Compared with coding variants, noncoding variants require substantially more functional annotations to elucidate their roles across different lineages and cell types. Among noncoding regions, transcription factor binding sites (TFBSs) are

critical regulatory elements, and recent advances in TF footprinting techniques, such as FOODIE, have enabled the detection of binding events at near single-base resolution in a cell-type-specific context. Our experiments demonstrate that FOODIE improves TF binding measurement and target gene prediction and yields higher heritability enrichment of erythroid-related traits compared with widely used assays like ATAC-seq and DNase-seq in the K562 cell line. This provides an unprecedented opportunity to identify the variants that rewire TF-mediated gene regulatory networks associated with complex traits and diseases. Herein, we integrate several cutting-edge models and FOODIE footprints to develop a comprehensive framework, varTFBridge, for genome-wide mapping of FOODIE footprint-gene connections that illuminate the functions of blood cell phenotype-associated variants from 490,640 UK Biobank whole genome sequencing data.

To prioritize trait-associated causal variants across the MAF spectrum within K562 FOODIE footprints, varTFBridge integrates GWAS and GWFM fine-mapping for putative causal common variants (MAF $\geq$ 0.1%) and footprint-based burden tests with leave-one-variant-out analyses for driver rare variants (MAF < 0.1%). Consequently, we identified 224 credible common variants and 23 driver rare variants within K562 FOODIE footprints associated with erythroid traits. Equipped with VAR2TFBS, ABC-FP-Max, and AlphaGenome, varTFBridge faithfully identified 209 common and 18 rare variant–TFBS–gene trios. We demonstrate that varTFBridge successfully recapitulates the causal variant rs112233623 and further elucidates that it decreases the expression of *CCND3* by disrupting the GATA1/TAL1 co-binding sites at a distal enhancer, resulting in larger MCV and lower RBC count.

To the best of our knowledge, varTFBridge is the first framework to identify and functionally investigate both common and rare variants through FOODIE TF footprinting. The combination of VAR2TFBS, ABC-FP-Max, and AlphaGenome provides synergistic advantages in variant effect prediction. For example, while AlphaGenome falls short in predicting the effects of distal variants on target genes, ABC-FP-Max provides robust prediction of variant-to-gene links up to 5 Mb away. Furthermore, the resolution of TF binding prediction from AlphaGenome is limited to 128 bp due to the low resolution of ChIP-seq; in contrast, VAR2TFBS utilizes FIMO scans and FOODIE to achieve base-resolution variant effect predictions on TF binding affinity.

Our study has several limitations that highlight areas for future work. First, varTFBridge only considers variants residing within TF binding sites, thereby missing variants flanking the core motif that might also affect TF binding affinity. Second, some co-regulators do not bind directly to the DNA sequence; our model remains limited in predicting this type of co-regulation, though a recent model, Chromnitron[52], might help bridge this gap. Third, varTFBridge assumes a single causal gene per variant, although enhancers containing disease variants often have highly pleiotropic effects. Fourth, AlphaGenome can only predict a limited number of TFs, which hinders the extension of our framework to other tissues and cell lines. Integrating ChromBPNet[53] into our framework could potentially address this limitation, as it can be easily trained on FOODIE data from new cell lines and tissues to predict variant effects.

In summary, our approach paves the way for a comprehensive mapping of noncoding variants to TF-mediated gene regulatory networks. While this study utilizes experimental data from K562 cells to model erythroid traits underlying anemia and related disorders, we anticipate that varTFBridge will be broadly applicable to other cell lines and tissues, thereby

424 accelerating variant-to-function studies for a wide range of complex traits and diseases.

425 Methods

426 *Cell Preparation*

427 Human leukemia cell line K562 and human lymphoblastoid cell line GM12878 were grown
428 with RPMI 1640 medium (cat. no. 11875093, Gibco) containing 10% FBS (cat. no.
429 10099141C, Gibco) in a 37 °C incubator supplied with 5% CO2. Cells were rinsed once with
430 1 mL of DPBS (cat. no. 14040117, ThermoFisher) and resuspended in 1 mL DPBS before
431 use.

432 *FOODIE library preparation*

433 The FOODIE libraries were constructed using the commercial Hyperactive FOODIE Library
434 Prep Kit (cat. no. TD721, Vazyme) following the manufacturer's instruction, which is similar
435 to our previously reported protocol[22]. Briefly, approximately 50,000 cells were counted,
436 washed with TW Buffer, and centrifuged. The cell pellet was resuspended in pre-chilled Lysis
437 Buffer to facilitate permeabilization. Following permeabilization, nuclei were collected via
438 centrifugation and resuspended in Tagmentation Mix for the Tn5 transposition reaction. The
439 reaction was incubated at 37 °C for 30 min on a thermomixer with shaking at 800 rpm.
440 Subsequently, the tagmentation reaction was terminated by adding pre-chilled Wash Buffer,
441 followed by centrifugation. The isolated nuclei were then resuspended in a Deamination Mix
442 supplemented with TruePrep DddB Enzyme (cat. no. S802-01, Vazyme) and incubated at
443 37 °C for 10 min. The reaction was halted by the addition of Stop Buffer. Finally, DNA was
444 purified using the provided magnetic beads, followed by PCR amplification using the kit-
445 supplied DNA polymerase and primers.

446 *FOODIE data processing and footprint calling*

447 We followed the procedure described in the original FOODIE study[22]. Briefly, raw reads were
448 trimmed with Trim Galore (v0.6.10)[54], and aligned to the human reference genome hg38 and
449 deduplicated with Bismark (v0.24.2)[55]. Customized scripts were then applied to perform
450 deamination/conversion, followed by calculation and normalization of the conversion ratio at
451 each genomic position. The adjusted conversion ratios together with the corresponding
452 sequencing depths were used for open region identification and transcription factor footprint
453 calling. Footprints with P values $< 1 \times 10^{-7}$ were retained for downstream analyses. The
454 scripts are available at https://github.com/sunneyxie-lab/bulk-foodie-pipeline.

455 *ATAC-seq and DNase-seq footprint calling and processing*

456 We leveraged HINT[32] (v1.0.2), a Hidden Markov Models–based method, to derive
457 transcription factor footprints from ATAC-seq data in K562 cells. We downloaded the K562
458 ATAC-seq BAM file (ENCFF077FBI) and its corresponding narrow peak file (ENCFF333TAT)
459 from ENCODE and processed them with HINT's default settings. K562 DNase footprints
460 (ENCFF771QAQ) in BED format were also obtained from ENCODE.

461 *Genome-wide association testing of erythroid traits in UK Biobank*

462 We applied the same GWAS framework as described in our previous analysis[56] for UK
463 Biobank. In addition to UK Biobank's standard central quality control, we restricted analyses
464 to participants of white European genetic ancestry. This was determined using k-means

clustering of the first four principal components derived from genome-wide SNP genotypes. Individuals assigned to the white European cluster, but who self-reported a non-European ancestry were excluded. To avoid confounding due to relatedness, only unrelated individuals were included in the association analyses. After all quality control filters, a total of 490,640 participants with both genotype and phenotype data were eligible for inclusion. We further removed variants with minor allele frequency (MAF) < 0.1%. Genome-wide association analysis was performed using REGENIE v3.3[57], implemented within the UK Biobank Research Analysis Platform and following recommendations for UK Biobank analyses (https://rgcgithub.github.io/regenie/recommendations/). The association model analyzed 13 erythroid-related traits as outcome variables: hematocrit (HP), hemoglobin concentration (HC), high light scatter reticulocyte count (HLSR), high light scatter reticulocyte percentage (HLSRP), immature reticulocyte fraction (IRF), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), mean sphered corpuscular volume (MSCV), red blood cell count (RBC), red cell distribution width (RBCDW), reticulocyte count (RC), and reticulocyte fraction of red cells (RP). Covariates included genotyping array, sex, age at questionnaire completion, and the first ten principal components of genetic ancestry.

Genome-wide fine-mapping for identifying credible set variants

Credible set variants were derived from genome-wide fine-mapping (GWFM) with the SBayesRC[58] framework, which implements a Bayesian hierarchical mixture model that integrates genome-wide linkage disequilibrium (LD) patterns and functional genomic annotations to estimate posterior inclusion probabilities (PIPs) for SNPs. Genome-wide LD was captured using precomputed eigen-decomposition reference panels provided with GCTB[59], ensuring consistency with the assumed genetic architecture. Functional annotations were incorporated from Gazal et al.[38], encompassing 96 genomic features relevant to complex trait architecture, including coding and conserved regions. Fine-mapping was performed genome-wide. Local credible sets (LCS) were constructed for each candidate causal variant by aggregating variants in LD ($r^2 > 0.5$) until their cumulative PIPs reached 70%, ensuring the LCS explains more heritability than a random SNP set of the same size. In this study, variants with PIP ≥ 10% and residing in an LCS with cumulative PIP ≥ 70% were considered likely causal, following previous studies[8,27].

GWFM (v2.5.4) was downloaded from https://gctbhub.cloud.edu.au/software/gctb/#Download. Summary statistics from 13 erythroid trait GWAS were used to construct credible sets. Variants were mapped to reference SNP identifiers (rsIDs) using the dbSNP release[60] (build b157) based on the GRCh38 human genome assembly (GCF_000001405.40). To align with the GWFM database version, genomic coordinates were lifted over from hg38 to hg19 using LiftOver. The processed hg19 summary statistics were then converted to the required .ma format, and the --impute-summary option was employed for quality control and imputation of summary statistics for SNPs present in the LD reference but not in the GWAS data. Finally, GWFM analyses were run with the --gwfm RC option, implementing the SBayesRC model.

FOODIE footprint-based test and leave-one-variant-out analysis for rare variants

To identify rare germline variants (MAF < 0.1%) on FOODIE footprints associated with erythrocyte traits, we performed gene-based burden tests across chromosomes 1–22 in 490,640 UK Biobank (UKBB) participants with whole-genome sequencing data. Briefly,

490,640 participants were sequenced at an average depth of 32.5× on the Illumina NovaSeq 6000. The WGS generation and processing pipeline for UKBB participants is described in the recent study[21]. We annotated the variants with "FILTER=PASS" from ML-Corrected DRAGEN whole genome sequencing (WGS) in PLINK2 format using the ENSEMBL Variant Effect Predictor (VEP) v113.0, and analyses were restricted to rare qualifying variants (MAF < 0.1%). Footprints with <30 qualifying variant carriers were excluded from each burden mask. Accordingly, the footprint-wide significance threshold was set using Bonferroni correction at: 0.05/188,484 = 2.65×10^{-7}.

Stratified linkage disequilibrium score regression

To identify blood cell traits with heritability enrichment in K562 FOODIE footprints, we employed stratified linkage disequilibrium score regression (S-LDSC)[61]. All GWAS data were preprocessed using the 'munge_sumstats.py' script. Variants in the Major Histocompatibility Complex (MHC) region were excluded from the analysis. In addition to K562 FOODIE footprints, ATAC-seq (ENCFF333TAT) and DNase-seq narrow peaks and footprints (ENCFF274YGF, ENCFF771QAQ) were downloaded from ENCODE[31]. Linkage disequilibrium (LD) scores were calculated for each annotation using the 1000 G Phase 3 European population[62]. The heritability enrichment of each annotation for a given trait was computed by adding the annotation to the baseline LD score model (v2.0) and regression for HapMap3 SNPs. These analyses used v1.0.1 of the LDSC package.

ABC-FP-Max predictions for linking variants to the target gene

To link variants to their target genes using ABC-FP-Max, we first adapted the existing ABC model (accessible at https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction) to predict footprint–gene connections within specific cell types, thereby constructing ABC-FP models. The original ABC model was designed to link chromatin accessibility elements to genes, initially by calling peaks in chromatin accessibility data using MACS2. In contrast, for connecting FOODIE footprints to genes, we replaced the peak-calling step with FOODIE footprint calling. Following footprint identification, we extended each footprint by 50-bp on either side and merged them into a single element, recognizing that closely located footprints might co-regulate the same genes and larger windows can better count reads in ATAC-seq and H3K27ac ChIP-seq. Subsequently, we followed the original ABC models to estimate chromatin accessibility from ATAC-seq data on these footprint-based elements and to assess chromatin conformation (Hi-C) between these elements and the promoters of potential target genes, as previously described. Specifically, the activity of each footprint-based element is calculated based on ATAC-seq experiments and then quantile-normalized to the distribution observed in K562 cells. The footprint-gene contact is determined using cell-type specific Hi-C data from 4DN (4D Nucleome Data Portal). Finally, we compute the ABC-Max score for each footprint-gene pair as the product of activity and contact, normalized by the product of activity and contact for all elements within 5 Mb of that gene.

To calculate the ABC-FP-Max score for variant-gene pairs, we took the predictions from the ABC-FP model and applied the following processing steps. First, we considered all distal footprint–gene connections with an ABC-FP score ≥ 0.015 and all distal or proximal promoter–gene connections with an ABC score ≥ 0.1 (following the ABC-Max model setting). Second, we overlapped the variants with the ABC-FP nominated footprints. For each variant, the target gene was assigned as the one with the highest ABC-FP score among all target genes in the list; this top ABC-FP score is designated as the ABC-FP-Max score, quantifying

the predicted variant-gene pair.

Assessing fold excess overlap with direct measurements of TF ChIP-seq

We downloaded TF ChIP-seq data from ENCODE (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/encRegTfbsClustered/encRegTfbsClusteredWithCells.hg38.bed.gz). We selected peaks annotated with K562 cell line and created a bed file with the union of ChIP-seq peaks from all TFs assayed. We then calculated the excess overlap between an annotation (A) and the ChIP-seq peaks (B) as

$$Excess\ overlap = \frac{Fraction\ of\ SNPs\ in\ A\ and\ B}{(frac\ SNPs\ in\ A) * (frac\ SNPs\ in\ B)}$$

We calculated SEs using a block-jackknife, dividing the genome into 200 blocks of equal genomic size[14].

VAR2TFBS for predicting variants affected TF binding site

Quantifying the impact of genetic variants within FOODIE footprints necessitates assessing transcription factor (TF) binding affinity. VAR2TFBS employs a scoring procedure based on FIMO (https://github.com/jmschrei/memesuite-lite/blob/main/memelite/fimo.py), which identifies statistically significant hits by scanning Position Weight Matrix (PWM) of a certain TF against one-hot encoded DNA sequences. This process involves convolving PWMs across DNA sequences and converting the resulting scores into p-values using an exact background distribution derived from the PWM. In this study, we collected the PWMs from JASPAR 2025 core non-redundant vertebrates (https://jaspar.elixir.no/download/data/2026/CORE/JASPAR2026_CORE_vertebrates_non-redundant_pfms_meme.txt). To assess transcription factor (TF) binding within a FOODIE footprint, the sequence is first extended by 10 bp on each side. This extended reference sequence is then subjected to FIMO scanning, with TF binding sites identified by considering all Position Weight Matrices (PWMs) exhibiting a motif p-value less than 1×10^{-5}. Subsequently, each genetic variant is intersected with these identified TF binding sites, and FIMO scanning is performed on sequences incorporating the alternate allele. The resulting changes in TF binding are categorized into four types: creation, denoting a *de novo* TFBS (p-value < 1×10^{-5}) in the alternate sequence; increase, where an existing TFBS shows a decreased (more significant) p-value in the alternate sequence; disruption, defined as an existing TFBS in the reference sequence having a p-value $\geq 1 \times 10^{-5}$ in the alternate sequence; and decrease, signifying an existing TFBS with an increased (less significant) p-value in the alternate sequence.

In general, VAR2TFBS takes as input a reference genome, a variant list in BED file format, FOODIE footprints also in BED file format, and a collection of PWMs. It produces as output a comprehensive list of variants that overlap with FOODIE footprints, detailing their respective effects on transcription factor binding sites.

AlphaGenome for variant effect scoring

We leveraged AlphaGenome (v0.5.1) to predict variant effects on cell-type–specific epigenomic profiles. The approach generates modality-specific predictions for reference (REF) and alternate (ALT) sequences, which are then compared to assess the regulatory impact of each variant across modalities. The model analyzes sequences up to 1 million

bases and delivers single-base-pair resolution for ATAC-seq, DNase-seq, and TF ChIP-seq, and 128-bp resolution for histone modification ChIP-seq and TF binding. Variant scoring is performed by generating REF and ALT predictions within a modality-specific window around the variant: a 502-bp window centered on the variant for ATAC-seq, DNase-seq, and TF ChIP-seq, and a 2001-bp window for histone modification ChIP-seq. Predictions are aggregated across the window using log2-ratio of summed signals to yield a raw variant score: $\text{raw score} = log2(\frac{sum(\text{ALT})+1}{sum(\text{REF})+1})$.

Since raw scores across modalities sit on different scales, they are normalized via empirical quantiles estimated from common variants (MAF > 1% in any GnomAD v3 population). Each raw score is mapped to its percentile within this background distribution; a quantile of 0.99 corresponds to the 99th percentile among common variants. Extreme quantiles are capped near ±0.99999 due to approximately 300k variants used, so quantile scores are best for signaling unusually large scores, while raw scores convey magnitude for a given scorer/track. For signed scores with directional effects, the 0–1 quantile probability is linearly transformed to the range [-1, 1], preserving directionality: 0th percentile maps to -1, 50th to 0, and 100th to +1. The quantile score records the rank of the raw score within the background distribution as a quantile probability, enabling direct cross-track comparisons across scoring strategies that yield scores on different scales. In this study, we report both raw and quantile scores for variants across two chromatin accessibility modalities (ATAC-seq and DNase-seq) and 269 histone modification tracks in K562 cells.

The code used to develop varTFBridge, perform the analyses and generate results in this study is publicly available and has been deposited in https://github.com/JasonLinjc/varTFBridge under the MIT License.

Acknowledgments

This project is financially supported by Changping Laboratory.

Author contribution

Y.Z. and X.X. conceived and supervised the study. J.L. developed varTFBridge and led downstream computational analyses, including model benchmarking and case studies. W.D. performed FOODIE experiments; C.X. conducted FOODIE footprint calling. J.L. and J. Zhang performed genome-wide fine-mapping. J.L., J. Zhao, and K.M. conducted GWAS and footprint-based burden tests. J.L., E.H., and J. Zhang performed S-LDSC analyses. J.L., C.X., W.D., and Y.Z. drafted the manuscript, with input from all authors.

Competing interests

The authors declare no competing interests.

Reference

1. IGVF Consortium *et al.* Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).
2. Rao, S., Yao, Y. & Bauer, D. E. Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. *Genome Med.* **13**, 41 (2021).
3. Lindenhofer, D. *et al.* Functional phenotyping of genomic variants using joint multiomic single-cell DNA–RNA sequencing. *Nat. Methods* 1–10 (2025).
4. Loos, R. J. F. & Yeo, G. S. H. The genetics of obesity: from discovery to biology. *Nat. Rev. Genet.* **23**, 120–133 (2022).
5. Qi, T., Song, L., Guo, Y., Chen, C. & Yang, J. From genetic associations to genes: methods, applications, and challenges. *Trends Genet.* **40**, 642–667 (2024).
6. Li, Z. & Zhou, X. Towards improved fine-mapping of candidate causal variants. *Nat. Rev. Genet.* 1–15 (2025).
7. Southam, L. & Zeggini, E. Twenty years of genome-wide association studies. *Nature* **641**, 47–49 (2025).
8. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
9. Strober, B. J., Zhang, M. J., Amariuta, T., Rossen, J. & Price, A. L. Fine-mapping causal tissues and genes at disease-associated loci. *Nat. Genet.* 1–11 (2025).
10. Van Buren, E. *et al.* cellSTAAR: incorporating single-cell-sequencing-based functional data to boost power in rare variant association testing of noncoding regions. *Nat. Methods* 1–12 (2025).
11. Ribeiro, D. M., Hofmeister, R. J., Rubinacci, S. & Delaneau, O. Noncoding rare variant associations with blood traits in 166,740 UK Biobank genomes. *Nat. Genet.* 1–10 (2025).
12. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

13. Fu, X. *et al.* A foundation model of transcription across human cell types. *Nature* 1–9 (2025).
14. van de Geijn, B. *et al.* Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. *Hum. Mol. Genet.* **29**, 1057–1067 (2020).
15. Yao, Q. *et al.* Motif-Raptor: a cell type-specific and transcription factor centric approach for post-GWAS prioritization of causal regulators. *Bioinformatics* **37**, 2103–2111 (2021).
16. Porter, D. F. *et al.* Disease-linked regulatory DNA variants and homeostatic transcription factors in epidermis. *Nat. Commun.* **16**, 8387 (2025).
17. Frangoul, H. *et al.* Exagamglogene autotemcel for severe sickle cell disease. *N. Engl. J. Med.* **390**, 1649–1662 (2024).
18. Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
19. Levesque, S. & Bauer, D. E. CRISPR-based therapeutic genome editing for inherited blood disorders. *Nat. Rev. Drug Discov.* 1–19 (2025).
20. Orkin, S. H. & Bauer, D. E. Emerging genetic therapy for sickle cell disease. *Annu. Rev. Med.* **70**, 257–271 (2019).
21. UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* 1–10 (2025).
22. He, R. *et al.* Genome-wide single-cell and single-molecule footprinting of transcription factors with deaminase. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2423270121 (2024).
23. Chang, L. & Ren, B. Efficient, scalable, and near-nucleotide-resolution profiling of protein occupancy in the genome with deaminases. *Proceedings of the National Academy of Sciences of the United States of America* vol. 122 e2425203122 (2025).
24. Callaway, E. DeepMind's new AlphaGenome AI tackles the "dark matter" in our DNA. *Nature* **643**, 17–18 (2025).

25. Avsec, Ž. *et al.* AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. *bioRxiv* 2025.06.25.661532 (2025) doi:10.1101/2025.06.25.661532.
26. García-González, J. & Gogolewski, K. AlphaGenome: a Swiss-army knife for exploring non-coding DNA. *Trends Genet.* **0**, (2025).
27. Wu, Y. *et al.* Genome-wide fine-mapping improves identification of causal variants. *medRxiv* 2024.07.18.24310667 (2025) doi:10.1101/2024.07.18.24310667.
28. Zhao, Y. *et al.* Population-scale gene-based analysis of whole-genome sequencing provides insights into metabolic health. *Nat. Genet.* **57**, 2436–2444 (2025).
29. FIMO: scanning for occurrences of a given motif | Bioinformatics.
30. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
31. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
32. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45 (2019).
33. Fulco, C. P. *et al.* Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. *bioRxiv* 529990 (2019) doi:10.1101/529990.
34. Gschwind, A. R. *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv* 2023.11.09.563812 (2023) doi:10.1101/2023.11.09.563812.
35. Gasperini, M. *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377-390.e19 (2019).
36. Rhee, H. S. & Pugh, B. F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.* **Chapter 21**, Unit 21.24 (2012).
37. Andersson, L. C., Jokinen, M. & Gahmberg, C. G. Induction of erythroid differentiation in the human leukaemia cell line K562. *Nature* **278**, 364–365 (1979).

710 38. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits
711 shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
712 39. Ovek Baydar, D. *et al.* JASPAR 2026: expansion of transcription factor binding profiles and
713 integration of deep learning models. *Nucleic Acids Res.* **54**, D184–D193 (2026).
714 40. Presnell, J. S., Schnitzler, C. E. & Browne, W. E. KLF/SP transcription factor family
715 evolution: Expansion, diversification, and innovation in eukaryotes. *Genome Biol. Evol.* **7**,
716 2289–2309 (2015).
717 41. Bialkowska, A. B., Yang, V. W. & Mallipattu, S. K. Krüppel-like factors in mammalian stem
718 cells and development. *Development* **144**, 737–754 (2017).
719 42. Hsieh, T.-H. S. *et al.* Enhancer-promoter interactions and transcription are largely
720 maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nat. Genet.* **54**, 1919–1932
721 (2022).
722 43. Narita, T. *et al.* Disentangling the architectural and non-architectural functions of CTCF and
723 cohesin in gene regulation. *Nat. Genet.* **57**, 3137–3151 (2025).
724 44. Moore, M. M. *et al.* Multi-locus CRISPRi targeting with a single truncated guide RNA. *Nat.*
725 *Commun.* **16**, 1357 (2025).
726 45. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all.
727 *Protein Sci.* **31**, 8–22 (2022).
728 46. Lin, J., Luo, R. & Pinello, L. EPInformer: a scalable deep learning framework for gene
729 expression prediction by integrating promoter-enhancer sequences with multimodal
730 epigenomic data. *bioRxivorg* 2024.08.01.606099 (2024) doi:10.1101/2024.08.01.606099.
731 47. Moore, J. E. *et al.* An expanded registry of candidate cis-regulatory elements. *Nature* 1–10
732 (2026).
733 48. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant
734 resolution. *Nat. Genet.* **51**, 683–693 (2019).

735 49. Sankaran, V. G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to regulate
736 erythrocyte size and number. *Genes Dev.* **26**, 2075–2087 (2012).
737 50. Hardison, R. C. Mechanisms of globin gene regulation in mammals. *Annu. Rev. Genet.*
738 (2025) doi:10.1146/annurev-genet-020325-095743.
739 51. Manning, J. M., Manning, L. R., Dumoulin, A., Padovan, J. C. & Chait, B. Embryonic and
740 fetal human hemoglobins: Structures, oxygen binding, and physiological roles. *Subcell.*
741 *Biochem.* **94**, 275–296 (2020).
742 52. Tan, J. *et al.* Multimodal learning decodes the global binding landscape of chromatin-
743 associated proteins. *bioRxivorg* 2025.08.17.670761 (2025)
744 doi:10.1101/2025.08.17.670761.
745 53. Pampari, A. *et al.* ChromBPNet: bias factorized, base-resolution deep learning models of
746 chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints
747 and regulatory variants. *bioRxivorg* 2024.12.25.630221 (2025)
748 doi:10.1101/2024.12.25.630221.
749 54. Krueger, F. *et al. FelixKrueger/TrimGalore: V0.6.10 - Add Default Decompression Path*.
750 (Zenodo, 2023). doi:10.5281/ZENODO.7598955.
751 55. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-
752 Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
753 56. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood.
754 *Nature* **575**, 652–657 (2019).
755 57. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and
756 binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
757 58. Zheng, Z. *et al.* Leveraging functional genomic annotations and genome coverage to
758 improve polygenic prediction of complex traits within and between ancestries. *Nat. Genet.*
759 **56**, 767–777 (2024).

59. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).

60. Phan, L. *et al.* The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Res.* **53**, D925–D931 (2025).

61. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

62. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).