Minimum entropy framework identifies a novel class of genomic functional elements and reveals regulatory mechanisms at human disease loci

Michael J. Betti[1], Melinda C. Aldrich[1], Eric R. Gamazon[1,2]

[1]Vanderbilt University Medical Center, Nashville, TN
[2]Clare Hall, University of Cambridge, Cambridge, England

Correspondence to: eric.gamazon@vumc.org, michael.j.betti@vanderbilt.edu

We introduce CoRE-BED, a framework trained using 19 epigenomic features encompassing 33 major cell and tissue types to predict cell-type-specific regulatory function. CoRE-BED's interpretability facilitates causal inference and functional prioritization. CoRE-BED identifies nine functional classes *de-novo*, capturing both known and completely new regulatory categories. Notably, we describe a previously uncharacterized class termed Development Associated Elements (DAEs), which are highly enriched in stem-like cell types and distinguished by dual presence of either H3K4me2 and H3K9ac or H3K79me3 and H4K20me1. Unlike bivalent promoters, which represent a transitory state between active and silenced promoters, DAEs transition directly to or from a non-functional state during stem cell differentiation and are proximal to highly expressed genes. Across 70 GWAS traits, SNPs disrupting CoRE-BED elements explain nearly all SNP heritability, despite encompassing a fraction of all SNPs. Notably, we provide evidence that DAEs are implicated in neurodegeneration. Collectively, our results show CoRE-BED is an effective prioritization tool for post-GWAS analysis.

Introduction

Since the publication of the first genome-wide association study (GWAS) nearly two decades ago[1], there has been remarkable growth in the number of population-level genetic studies covering large swaths of the human phenome. However, the translational potential of GWAS has been hindered by a consistent bottleneck: the functional ambiguity of most single nucleotide polymorphisms (SNPs) identified through the methodology. Ninety percent of GWAS-identified variants are predominantly located in non-coding regions of the genome and do not directly impact the coding sequence of proteins.[2] Consequently, determining the functional consequence of SNPs on their associated phenotype of interest remains a challenge. Despite the unprecedented pace of genetic discoveries from GWAS, less than 1% of these associations have been functionally confirmed through follow-up studies.[3] This discrepancy underscores the need for improved *in silico* approaches to help bridge the knowledge gap, as the scale of the problem makes experimental validation of every GWAS association infeasible.

Integrating GWAS results with functional information can help assign regulatory function to non-coding variants[2,4]. In addition to the ENCODE Consortium's comprehensive registry of candidate *cis*-regulatory elements (cCREs) across over 1,500 cell types[5], recently-published prediction frameworks such as Activity-by-Contact[6] and ChromHMM[7,8] integrate rich functional genomics datasets to predict function in non-coding regions of the genome.

Here, we introduce CoRE-BED, a decision tree-based framework trained on 19 epigenomic features that classifies input coordinates into learned functional classes. Notably, in contrast to deterministic models of genome regulation, which assert that individual regulatory

classes are characterized by a single specific combination of regulatory marks, CoRE-BED regulatory elements are defined through multiple paths; there is no a priori definition for which feature combinations define a given class. Thus, the epigenomic profile of each class resembles a spectrum, each with a distinct distribution or signature of epigenomic features.

Because CoRE-BED utilizes a decision tree-based architecture, the model's functional predictions provide high interpretability[9], offering biological insights into genome structure and function. We demonstrate that the functional classes learned *de-novo* by CoRE-BED capture both well-known and previously uncharacterized categories of regulatory elements, including, for example, a regulatory class which we term Development Associated Elements (DAEs). Additionally, we show that CoRE-BED is an effective tool for GWAS variant prioritization, with the ability to functionally prioritize a set of SNPs enriched for disease heritability. In summary, the CoRE-BED framework is characterized by improved interpretability and practical applicability, providing robust prediction of genome-function at the cell type-specific level and identifying high-priority SNPs for post-GWAS analyses or downstream functional validation studies.

Results

CoRE-BED framework and regulatory predictions

We present CoRE-BED, a decision tree-based framework for predicting cell and tissue type-specific function of putative regulatory elements across the genome. The decision tree structure was learned from training data, generating a classifier (i.e., membership in a class of regulatory elements) whose generalizability was evaluated in test data (Figure 1). Each internal

node represents an observation that captures a physical property of chromatin, nucleosome positioning, or DNA accessibility; each leaf or terminal node is a class of regulatory elements; and each link is a decision. To determine the optimal tree structure, we considered the information gain, calculated using Shannon entropy[10], at each node split. Classification was performed by traversing down the tree hierarchy from the root node to the terminal node. The benefits of the decision tree framework include improved interpretability, ease of implementation (e.g., scaling and normalization not a prerequisite), and a non-parametric methodology (requiring no distributional assumptions).

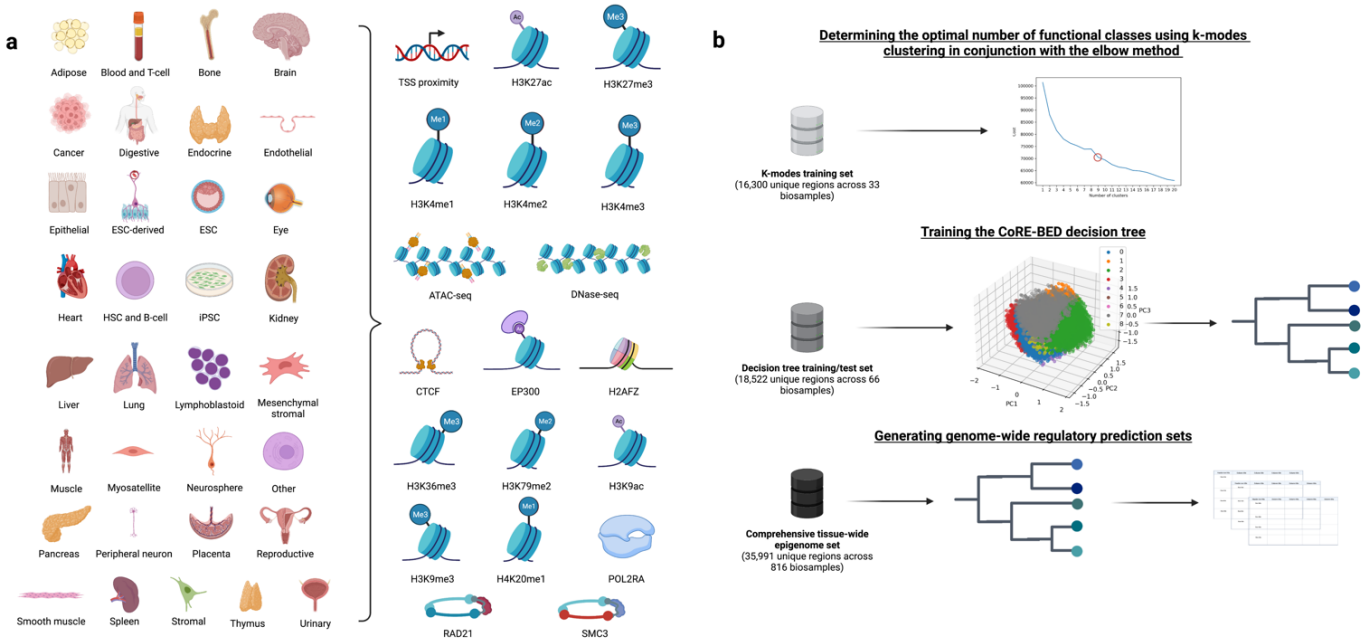


Figure 1. The CoRE-BED framework. a, *The CoRE-BED training set was composed of 19 distinct epigenomic marks representing 33 major categories of cell and tissue types.* ***b,*** *Development of the framework consisted of three steps. First, we determined the optimal number of clusters making up the data using k-modes clustering in conjunction with the elbow method. Next, using the learned* $k = 9$*, k-modes clustering was performed on an independent dataset, i.e., the training set, for the decision tree implementation. Finally, the trained decision tree was used to predict genome-wide functional classes for 816 individual biosamples encompassing the 33 major cell and tissue categories.*

The decision tree was derived from 19 epigenomic features observed across 33 cell and tissue types (see **Methods**). Since the upper size limit of a typical CRE is 1 kb[11,12], we segmented the entire human genome into approximately 3 million 1 kb bins and assessed each bin for

overlap with the 19 epigenomic features. We then applied k-modes clustering[13,14] in conjunction with the elbow method[15] to learn the optimal number of clusters representing these putative regulatory elements. We found that the data were best explained by 9 underlying clusters (Supplementary Figure 1a). Consequently, k-modes clustering using this learned k-value was again performed on an independent dataset – hereafter, the "training" dataset for the decision tree implementation (Supplementary Figure 1b). The model learned the underlying epigenomic features most highly determinate of each of the 9 regulatory classes in the training dataset. Classification performance was tested using a random 10% of the data held out as an independent test set.

The resulting optimal decision tree structure implemented a maximum depth of 15 and achieved a classification accuracy of 0.88 in the test set (Supplementary Figure 1c and Methods). The presence or absence of cohesion subunit RAD21[16] emerged as the highest-level feature determinate of regulatory state, followed by the active histone mark[17] H3K4me3. Notably, we observed that each regulatory class learned by the CoRE-BED model, rather than being determined by a single combination of overlapping features, is best characterized as a spectrum, with multiple discrete paths leading to the same regulatory classification. While each functional class is best represented as a distribution of features, we also observed that each is still most highly enriched for a specific subset of these features. Based on which genomic features are most highly determinate of each terminal class in the decision tree (Supplementary Figure 2), we assigned each of the 9 learned categories a functional label.

We term Class 0 "non-functional" due to these regions' depletion of features associated with highly active chromatin. We call members of Class 1 "distal insulators" due to their

enrichment for CTCF and cohesion subunits SMC3 and RAD21, as well as their distance from a gene transcription start site (TSS). Class 2 is termed “active promoter 1” as the class is characterized by enrichment for H3K4 methylation, proximity to a gene TSS, and chromatin accessibility[18,19], while Class 3 is “transcribed gene body” because of its enrichment for H3K36me3, H4K20me1, and H3K79me2[20]. Class 4 elements, which we term Development Associated Elements (DAEs) are enriched for either the dual presence of H3K4me2 and H3K9ac, along with proximity to a TSS, or alternatively H4K20me1 and H3K79me2. Class 5 elements are “proximal insulators” due to the strong presence of SMC3, CTCF, and RAD21[21,22], in addition to TSS proximity. Additionally, we call Class 6 elements “active promoters 2” due to their harboring of H3K4me2 and H3K4me3, as well as TSS proximity[23]. What differentiates these elements from the active promoters in Class 2 appears to be enrichment of bound RAD21. Class 7 elements are “active enhancers” due to their strong enrichment for H3K27ac, H3K4me1, and accessible chromatin regions[24,25]. Finally, Class 8 appears to capture a distinct category of enhancer-like promoters, hereafter referred to as ePromoters. Elements in this class are enriched for both signatures of active enhancers (acetyltransferase EP300 and H3K27ac)[26] and active promoters (H3K4me3, H3K9ac, and POL2RA)[23,27,28].

Enrichment of each functional category was variable across each of the 33 cell and tissue types. Class 1 (distal insulator) was most highly enriched in mesenchymal stem cells, Class 2 (active promoter 1) in the peripheral nervous system, Class 3 (transcribed gene body) in embryonic stem cells, Class 4 (DAE) in ESC-derived cell types, Class 5 (proximal insulator) and Class 6 (active promoter 2) in cancer, Class 7 (active enhancer) in embryonic stem cells, and Class 8 (ePromoter) in bone (Supplementary Figure 3).

Using the optimal CoRE-BED decision tree, we generated genome-wide functional annotations on 816 EpiMap[29] biosamples encompassing 33 distinct cell and tissue types (adipose, blood and T-cell, bone, brain, cancer, digestive, endocrine, endothelial, epithelial, embryonic stem cell-derived, embryonic stem cell, eye, heart, hematopoietic stem cell (HSC) and B-cell, induced pluripotent stem cell, kidney, liver, lung, lymphoblastoid, mesenchyme, muscle, myosatellite, neurosphere, pancreas, placenta and EMM, peripheral nervous system, reproductive, smooth muscle, spleen, stromal, thymus, and urinary). When hierarchically clustering all 816 biosamples based on similarity of genome-wide CoRE-BED functional annotations across the same loci, we observed that biosamples clustered together based on tissue of origin and regenerative potential (Figure 2a). Because each putative functional element classified by CoRE-BED incorporates 19 epigenomic features, binarized as present or absent, we would expect a total of 2^{19} = 524,288 possible feature combinations across the dataset. However, only a fraction of these expected combinations (35,991) is observed across all 816 biosamples. This small core set of epigenomic combinations relative to the number of possible combinations suggests that similar to DNA mutation, epigenomic marks are also under functional constraint[30,31]. Supporting this hypothesis is the observation that cancer cells, characterized by high level of epigenomic dysregulation[32], harbor the highest number of unique combinations of epigenomic marks in putative regulatory elements across all tissues (3,365). This is nearly three times the number of unique epigenomic signatures shared across all 33 cell type and tissue categories (1,214) (Figure 2b).

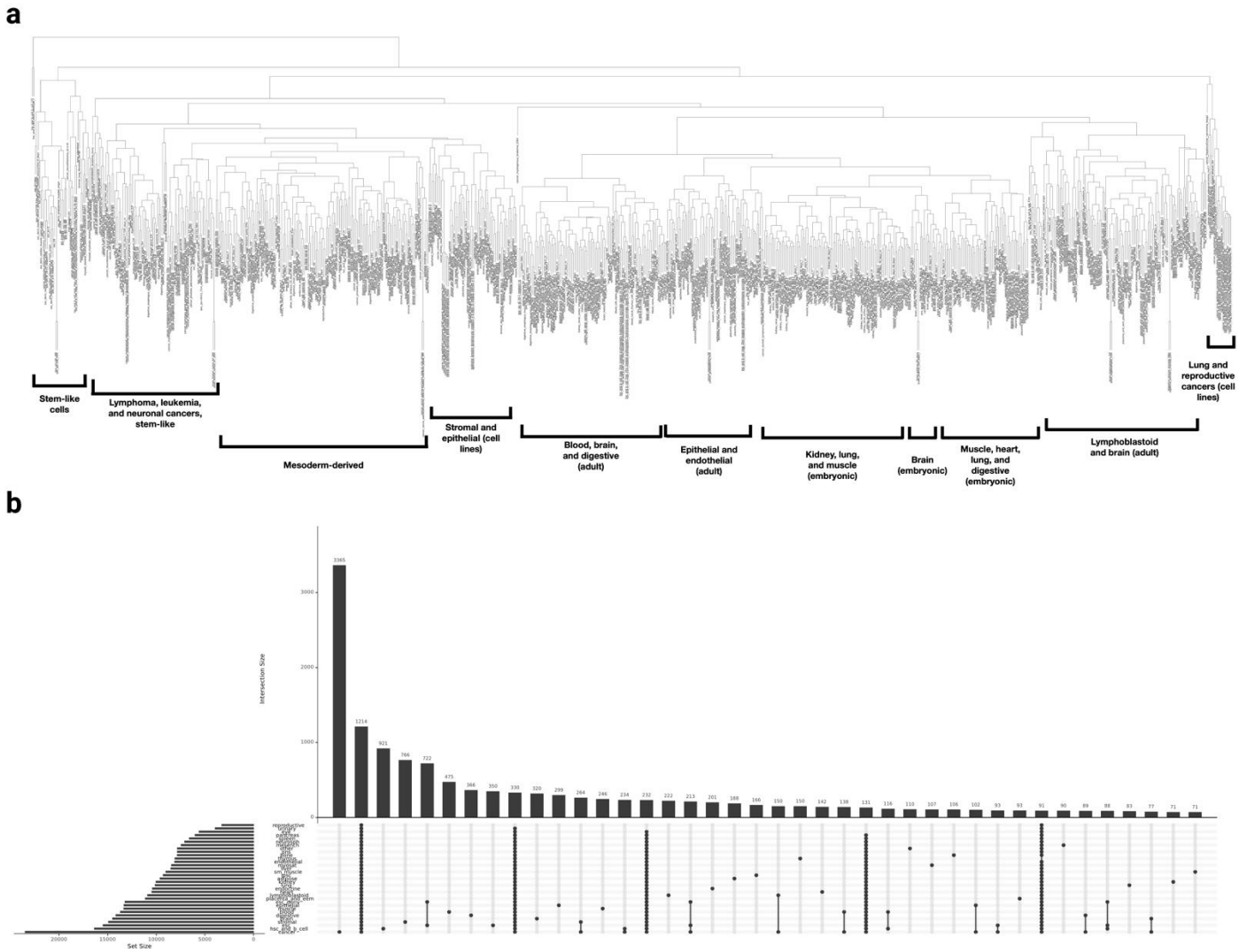


Figure 2. The distribution patterns of CoRE-BED functional classes are non-random. a, *Hierarchical clustering of 816 biospecimens based on genome-wide functional annotations at 1 kb resolution. Clustering reveals discrete groupings of samples by cell potency and differentiation pathways.* ***b,*** *Of the 524,288 (2^{19}) possible combinations of the 19 genomic features utilized by CoRE-BED, only 35,991 were observed across the 33 cell and tissue types. The largest proportion of these feature combinations was unique to cancer, suggesting disease-related epigenomic dysregulation.*

Comparison of CoRE-BED regulatory annotations with ENCODE cCREs

The CoRE-BED catalog of functional annotations presented here is not the first tissue-wide resource of functional elements. The ENCODE Consortium, for example, has published a registry of 1,063,878 cCREs across 1,518 cell types[5]. In contrast to the learned decision tree-

based approach of CoRE-BED, the ENCODE approach[5] classifies cCREs into one of five pre-defined functional categories (Promoter-like, Proximal enhancer-like, Distal enhancer-like, CTCF-only, and DNase-H3K4me3). These categories are based on the presence of pre-defined biochemical signatures, which can be applied universally to a cell or tissue type and were not learned agnostically.

Because of the high overlap between CoRE-BED's functional categories and those defined in the ENCODE cCRE registry, ENCODE cCREs present a useful dataset with which CoRE-BED's regulatory annotations can be compared. Because the size of the putative functional annotations is inherently different between the two datasets (uniform 1 kb bins for CoRE-BED versus a variable window size based on DHS signal for ENCODE cCRE), overlapping regions classified as functional in any cell or tissue type by CoRE-BED and the cCRE registry, respectively, were first merged into single regions. This resulted in a set of 290,846 putative CoRE-BED functional elements and 1,057,868 ENCODE cCREs. Notably, 186,544 of the global CoRE-BED functional regions (64.14%) were also classified as functional in the cCRE registry, while the remaining 104,302 (35.86%) were unique to CoRE-BED (Jaccard similarity index = 0.19). Among these CoRE-BED-specific annotated regions, most appeared to be Class 3 (transcribed gene body) elements, a category not included in the ENCODE cCRE registry. ENCODE cCRE registry also measured the genomic footprint of these predicted functional elements, estimating that their compendium of regulatory regions covers 7.9% of the entire genome. We quantified the coverage for our tissue-wide CoRE-BED predictions and estimated that they cover 23.72% of the entire genome.

Characterizing gene expression under CoRE-BED-predicted active promoter control in the peripheral nervous system

One should expect genes under the control of highly active promoters (Class 2) to be highly expressed. Of all 33 CoRE-BED cell and tissue types, we found that the peripheral nervous system (PNS) was most highly enriched for elements from this regulatory class (Supplementary Figure 3c). Therefore, we investigated the expression of genes proximal to PNS-specific active promoters using tibial nerve expression data from 619 individuals in GTEx. Of the 1,976 genes proximal to a Class 2 CoRE-BED active promoter in the PNS, 1,730 (87.55%) were expressed in tibial nerve. Among these 1,730 genes, the median expression was substantially higher than that of the remaining 53,755 expressed genes in tibial nerve (median TPM of genes proximal to Class 2 elements: 3,172.315, median TPM of remaining expressed genes: 1.947, $p < 2.2 \times 10^{-16}$).

Characterization of enhancer-like promoters

CoRE-BED identified enhancer-like promoters (Class 8) or ePromoters[33–35]. Notably, among the ePromoters, we observed a significantly greater proximity with an enhancer RNA (median 22.02 kb versus 64.80 kb; $p < 2.2 \times 10^{-16}$) relative to other promoters across the genome. In addition, we found an enrichment for biological pathways associated with general cell function and maintenance of homeostasis in genes closest to this regulatory class, consistent with previous studies linking this category of regulatory elements to a housekeeping function (Figure 3a). Notably, we observed enrichment for 192 highly significant (Benjamini-Hochberg-corrected $p < 0.05$) transcription factor (TF) binding motifs (Supplementary Table 1)

that were not significantly enriched throughout the rest of the genome. Among the motifs unique to the ePromoters (see **Methods**) are AP1 and members of the STAT and ATF/CREB families[33].

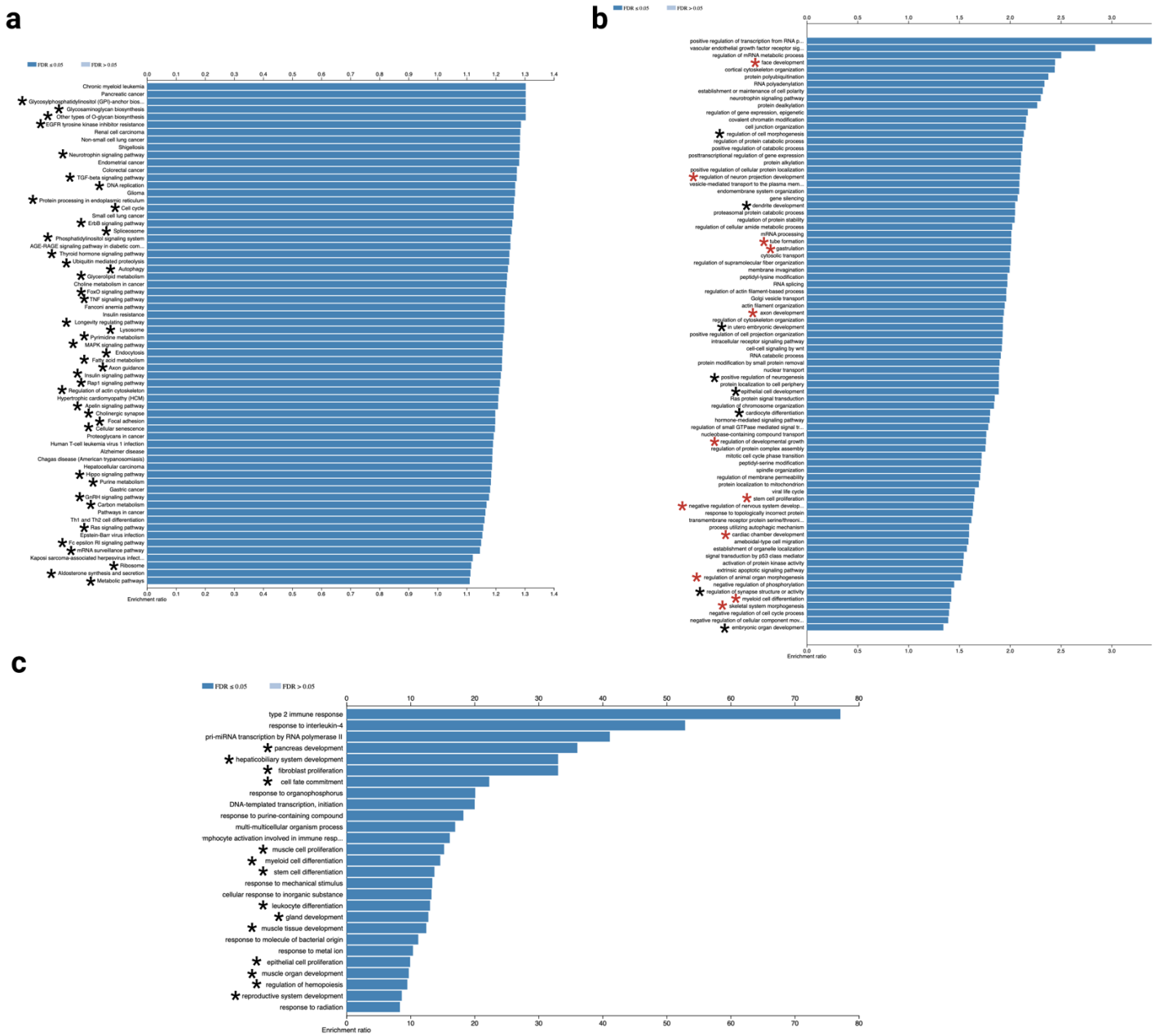


Figure 3. CoRE-BED ePromoters and DAEs are non-randomly distributed and enriched for gene pathways associated with development. a, *Genes proximal to an ePromoter in at least one cell or tissue type were found to be enriched for housekeeping pathways associated with general cell function and maintenance of homeostasis (marked with asterisk). This observation coincides with previous reports of enhancer-like promoters.* ***b,*** *Genes proximal to a DAE in ESCs (marked simultaneously by*

H3K4me2 and H3K9ac) are enriched for pathways associated with embryonic development (marked with asterisk). Pathways marked with a red asterisk are enriched ESC pathways that were not significantly enriched in thymus, the tissue with the lowest proportion of DAEs. ***c,*** *Gene ontology terms over-represented among transcription factors with binding sites at DAEs. Terms associated with developmental pathways are marked by an asterisk.*

CoRE-BED identifies class of DAEs enriched in stem-like cell types

CoRE-BED identified a previously uncharacterized class of regulatory elements, i.e., the DAEs, a unique chromatin state defined by two distinct genomic feature combinations. In the model training and test sets, which were composed of a total of 18,522 unique combinations of 19 genomic features, the 783 unique combinations representing DAEs were most highly enriched for both the active promoter mark H3K9ac[36] and H3K4me2, in addition to TSS proximity (Extended Figure 1a). While H3K4 monomethylation (H3K4me1) is most often associated with enhancers[12,24] and H3K4 trimethylation (H3K4me3) is considered a marker of actively transcribed genes[28], H3K4me2 has been linked to tissue-specific gene transcription and establishment of cellular homeostasis after DNA damage repair[37,38]. Across all 816 individual biosamples for which CoRE-BED functional annotations were generated, however, the cumulative (non-unique) set of 173,434 DAE occurrences are most highly enriched for the histone marks H4K20me1, thought to facilitate gene transcription by maintaining an open chromatin state[39], and H3K79me2, another modification associated with active transcription[40] (Extended Figure 1b). Neither of these combinations of histone modifications (H3K9ac/H3K4me2 and H4K20me1/H3K79me2) has previously been described.

Across each of the 33 major cell and tissue categories, we calculated the mean number of CoRE-BED DAEs. In concordance with expectations, three of the top four most highly enriched cell types for DAEs are stem cells with high regenerative capacity (embryonic stem cell

[ESC] derived cells = mean 687.87 DAEs per biosample, ESCs = mean 500.83 DAEs per biosample, and myosatellite = mean 444.67 DAEs per biosample). By comparison, thymus had the lowest enrichment of DAEs (mean 56.82 DAEs per biosample) (Supplementary Table 2).

Because DAEs were most highly enriched in stem-like tissues, we first characterized expression in whole blood for genes closest to a DAE in hematopoietic stem cells and B-cells. We found these 2,777 genes to be highly expressed, with a median TPM of 1,662.76, versus a median TPM of 0.53 for the remaining 52,571 genes expressed in whole blood ($p < 2.2 \times 10^{-16}$). In fact, expression levels of these genes under the putative control of a DAE were highly similar to expression levels of genes under the control of a Class 2 active promoter (median TPM of genes proximal to Class 2 active promoters: 1,396.93, median TPM of genes proximal to DAEs: 1,662.76, $p = 0.1558$). Thus, it appears that DAEs are associated with active transcription.

Given that CoRE-BED DAEs are highly enriched in stem cells, we next aimed to characterize which types of genes were under the control of a DAE in ESCs. With the 2,690 unique genes putatively regulated by a CoRE-BED DAE in one of 8 ESC lines, we ran an Over-Representation Analysis (ORA) to test for enrichment for KEGG biological pathways[41,42]. Of the 83 KEGG pathways enriched in the set of putative DAE-regulated genes in ESCs (false discovery rate [FDR] < 0.05), we observed a number related to embryonic development that were not among the pathways associated with DAE-regulated genes in thymus, the lowest-enriched tissue for DAEs. Enriched pathways include face development, regulation of neuron projectin development, tube formation, gastrulation, axon development, regulation of developmental growth, stem cell proliferation, negative regulation of nervous system development, cardiac

chamber development, regulation of animal organ morphogenesis, myeloid cell differentiation, and skeletal system morphogenesis (Figure 3b).

A similar pattern was observed for the TFs predicted to bind at DAEs. Across all 833 unique biosamples analyzed in aggregate, we identified a set of binding sites for 26 TFs enriched at DAEs. Using a subsequent gene over-representation analysis, we found these TFs to be highly enriched for 14 developmental pathways, including pancreas development, cell fate commitment, and stem cell differentiation (Figure 3c). We also subsequently characterized the tissue-specificity of genes under putative DAE control. We obtained tau indices computed across all human genes using GTEx expression data[43,44], with a tau index of 0 representing ubiquitously-expressed housekeeping genes and an index of 1 representing tissue-specific genes[45]. We found that genes under the putative control of a DAE had mean tau index of 0.373, representing moderate tissue-specificity.

Finally, we aimed to characterize how these newly defined DAEs resolve during stem cell differentiation. Over the course of embryonic development, for example, bivalent promoters represent a well-characterized “poised” regulatory state in highly pluripotent cell types that is marked by both the active H3K4me3 mark and repressive H3K27me3[46,47]. These bivalent promoters are known to transition over the course of development to either an active or permanently silenced state by losing either H3K27me3 or H3K4me3, respectively. We asked whether the DAEs identified by CoRE-BED transition to or from other regulatory classes in a similar manner. Unlike the means by which bivalent promoters resolve, however, most DAEs gained or lost during stem cell differentiation transition directly to or from Class 0 (non-functional) elements. For example, H1 ESCs have a total of 911 DAEs, which decreased to 405

when this ESC line was differentiated to neural crest cells. Of the 684 Class 4 elements lost through differentiation, 598 (87.4%) transitioned to Class 0 (non-functional) elements.

In the transition from H9 ESCs to smooth muscle cells, the number of DAEs showed a substantial increase from 176 to 3,092. Of the 2,976 DAEs gained, 2,907 (97.7%) were classified as Class 0 (non-functional) in non-differentiated H9 ESCs. Generally, Class 4 element loss during ESC differentiation appears to be the result of H4k20me1 depletion, suggesting loss of open chromatin conformation[39]. Conversely, Class 4 elements gained during ESC differentiation appear to gain either or both H3K79me2 and H4k20me1. Thus, because the DAEs primarily transition directly to or from a non-functional (Class 0) state, it is unlikely that they represent a transitional class of elements like bivalent promoters. Instead, they appear to represent a distinct class of functional elements.

A putative role for DAEs in common disease

Across 146,337 SNP associations reported within the NHGRI-EBI GWAS Catalog[48], which likely consist of causal and tag variants, 1,697 (1.16%) were found to fall within a DAE in at least one cell or tissue type (Supplementary Table 3). One of these associations is the Parkinson's disease associated variant rs823128[49,50], which is within an astrocyte-specific DAE. While this prediction was based on called H3K79me2 and H4K20me1 peaks in the immediate vicinity of this SNP, we also observed enrichment of both H3K9ac and H3K4me2, the other biochemical

signature characterizing DAEs, slightly further upstream of the variant locus (Figure 4).

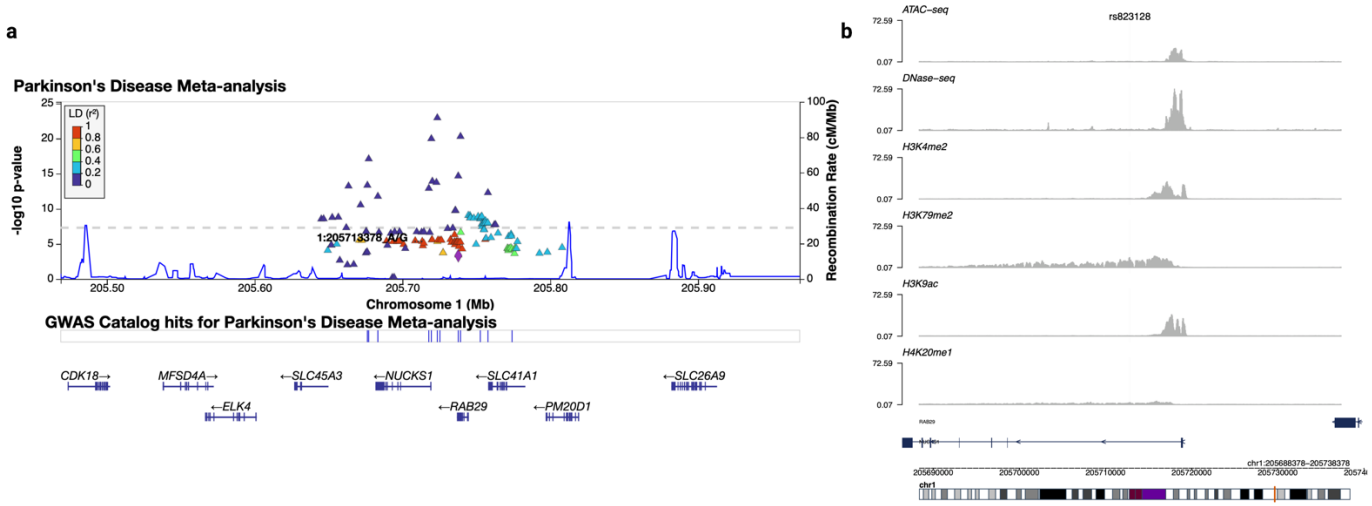


Figure 4. CoRE-BED DAEs overlap with disease-relevant variants. a, *We identified rs823128, a Parkinson's disease-associated SNP reported in the GWAS Catalog and replicated in subsequent meta-analyses. This SNP falls within a DAE in astrocytes.* ***b,*** *While this SNP region overlaps a DAE based on H3K79me2 and H4K20me1 peaks called in its vicinity, there is also notable enrichment of H3K4me2 and H3K9ac, the second never-before described biochemical signature of DAEs, further upstream near the TSS of NUCKS1.*

DAEs within brain astrocyte eQTLs

Of the 375 GWAS Catalog SNPs found to be within DAEs in brain, 286 are in astrocytes. Thus, we aimed to explore the regulatory role of DAEs in this cell type. Using individual-level germline genotyping and gene expression data from human induced pluripotent stem cell (iPSC)-derived astrocytes[51], eQTL (expression quantitative trait locus) mapping was performed. Of the 71,226 SNPs mapped as astrocyte eQTLs, 126 were within an astrocyte-specific DAE and linked to a total of 86 unique genes (Supplementary Table 4). Among these astrocyte eQTLs in a DAE is rs9408928, a SNP associated with the expression *NDUFA8*, a gene previously associated with both Alzheimer's disease and Parkinson's disease[52,53]. Interestingly, this SNP has only one immune-related phenotypic association (immune response to smallpox) reported in the NHGRI-EBI GWAS Catalog[48,54], although no associations with brain-related phenotypes have yet been

reported. Within this set of SNPs overlapping DAEs, we also identified 7 eQTLs (rs807748, rs12119677, rs57156794, rs12119878, rs72836878, rs17732024, and rs150792202) associated with the expression of four anxiety-associated genes (*ARVCF*[55], *FLII*[42,56], *HMBS*[57], and *SLC45A1*[58]). None of these SNPs have any phenotypic associations reported in the GWAS Catalog and thus could represent attractive novel targets for future functional validation experiments characterizing the potential role of DAEs in disease biology.

Practical application to GWAS variant prioritization

In addition to providing a genome-wide compendium of regulatory annotations, a potentially powerful application of CoRE-BED is functional interpretation of GWAS variants enriched for trait heritability. The underlying assumption is that SNPs within a functional region of a phenotypically associated cell or tissue type are more likely to causally affect the phenotype of interest. Focusing first on SNPs across all case-control studies in the NHGRI-EBI GWAS Catalog[48], we aimed to explore whether there was clear, observable stratification between variants within CoRE-BED regulatory elements and those that did not have a functional annotation in any cell or tissue type. Among the 25,643 GWAS Catalog SNPs (from 1,613 case-control studies), 11,760 (45.86%) had a CoRE-BED functional annotation in at least one cell or tissue type. As a negative control, we generated CoRE-BED annotations for an equally-sized set of variants randomly sampled from the over 300 million cataloged in dbSNP[59]. In this case, only 6,827 (26.62%) overlapped with a CoRE-BED functional region.

We observed a significantly higher MAF in GWAS Catalog SNPs with a CoRE-BED prediction than those without (with functional annotation median MAF 0.26; without functional

annotation median MAF 0.24; $p = 4.476 \times 10^{-13}$). We found that the GWAS Catalog variants with a CoRE-BED functional prediction had a more highly significant p-value distribution than those without (with functional annotation median GWAS p-value 6.99×10^{-9}; without functional annotation median GWAS p-value 2×10^{-8}; p-value $< 2.2 \times 10^{-16}$. We would expect this observation if genomic variants falling in a CoRE-BED functional region are enriched for causal effects on the corresponding phenotype than those without functional overlap.

We found that among the 11,760 GWAS Catalog SNPs with a CoRE-BED annotation in at least one cell or tissue type, the highest proportion (52.41%) were annotated as distal insulator variants in at least one cell type, followed by transcribed gene body variants (potentially involved in alternative transcript splicing) (52.14%), active enhancer (27.52%), ePromoter (21.84%), proximal insulator (19.48%), active promoter (type 2) (3.69%), DAEs (2.30%), and active promoter (type 1) (0.93%). Notably, we observed a high proportion (73.80%) of variants overlapping with a fine-mapped cis-eQTL mapped by GTEx. Of the individual CoRE-BED regulatory classes, promoter variants had greater overlap with fine-mapped cis-eQTLs than enhancer variants (56.33% for active promoter variants versus 49.30% for active enhancer variants).

We and others have previously shown that there is a negative correlation between the cis-defined heritability of gene expression and evolutionary rate, i.e., a depletion of cis-regulatory effects near conserved genes, where GWAS signals tend to cluster[60]. A recent study reports a systematic discontinuity between GWAS signals and eQTLs[61]. The authors argue that genes near GWAS hits are subject to strong selective constraint whereas genes near eQTLs show relaxed constraint and simpler regulatory landscapes. Our GWAS Catalog variant

comparisons support this proposed model and suggest that some alternative biological mechanism(s), which current cis-eQTL reference resources or mapping approaches do not sufficiently explain, must underlie a substantial proportion of GWAS signals.

To investigate the extent to which CoRE-BED-informed analysis improves discovery of disease-associated variants, we applied the CoRE-BED derived annotations to a GWAS meta-analysis of hypertension (performed in 126,196 cases and 325,698 controls across six populations within the UK Biobank)[62]. Because it is a cardiovascular phenotype, functional status of each variant tested in the hypertension GWAS was generated using 41 heart-derived epigenomes. Of the 28,987,534 variants tested in the GWAS, 1,487,383 (5.13%) were predicted to be functional in heart.

We compared overall p-value distributions between the variants overlapping functional annotations in heart versus the remaining set of variants. We found that there was a significantly greater enrichment of low GWAS p-values in variants with a CoRE-BED functional annotation (median p-value of 0.4597 versus median p-value of 0.4652; p-value $< 2.2 \times 10^{-16}$; Figure 5a)[63]. This suggests that CoRE-BED functional predictions are capable of enriching for trait heritability within a genome-wide set of SNPs. To test this hypothesis, we utilized partitioned linkage disequilibrium score regression (LDSC) to estimate the proportion of SNP heritability from the hypertension GWAS that is captured by heart-specific CoRE-BED regulatory elements. The broad-sense heritability of hypertension has been estimated to be between 30% and 50%[64,65]. Although only 5.13% of GWAS variants had CoRE-BED annotations in heart tissue (Figure 5b), these CoRE-BED-annotated variants were estimated to capture 100% of the total SNP heritability ($h^2 = 0.045$ overall, Figure 5c). If we consider individual regulatory categories,

Class 1 (distal insulators) account for the largest proportion of SNP heritability (0.495, h^2 = 0.022), followed by Class 7 (active enhancers) (0.442, h^2 = 0.020).

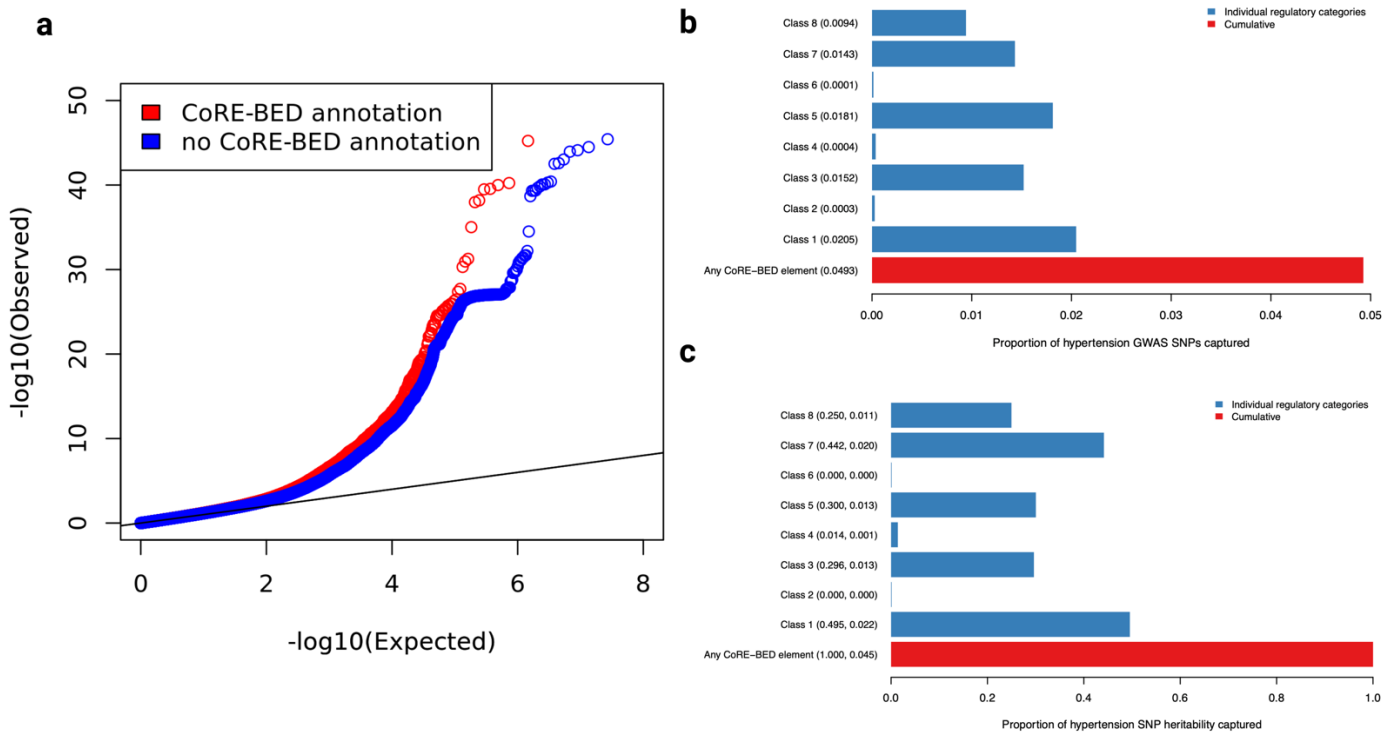


Figure 5. CoRE-BED functional annotations capture the SNP heritability from a GWAS of hypertension. a, *Across all 28,987,534 variants tested in a GWAS of hypertension, the 1,487,383 (5.13%) with CoRE-BED annotations in heart had a significantly lower p-value distribution than the remaining SNPs.* ***b,*** *The proportion of SNPs included in the partitioned heritability estimation with a CoRE-BED annotation for each functional category.* ***c,*** *The proportion of SNP heritability explained by each CoRE-BED functional category. Values in parentheses represent the proportion of SNP heritability and the overall SNP heritability, respectively.*

Post-GWAS analysis: identifying causal variants and partitioning heritability by CoRE-BED annotations

We leveraged a set of 25 experimentally validated causal SNP-gene pairs associated to disease from a recent study of a SNP-to-gene linking method, cS2G[66]. Of the 25 experimentally validated SNPs, 17 (68%) overlapped with a CoRE-BED promoter or enhancer in at least one cell

or tissue type. Of these 17 SNPs, 9 were Class 1 elements (distal insulators) in at least one cell or tissue type, 9 were classified as Class 3 (transcribed gene body), 4 as Class 5 (proximal insulators), 7 as Class 7 (active enhancers), and 4 as Class 8 (ePromoters). The same number (17) of (non-identical) SNPs overlapped an ENCODE cCRE, meaning that 21 of the 25 (84%) SNPs are cumulatively captured when using the two predictive approaches together.

Among the subset of 11 SNPs that were experimentally linked to the cS2G-predicted target gene, 7 (63.64%) overlapped with a CoRE-BED regulatory element. Of these 7 SNPs, 3 were Class 1 elements (distal insulators) in at least one cell or tissue type, 5 were Class 3 (transcribed gene body), 2 were Class 5 (proximal insulators), 4 were Class 7 (active enhancers), and 2 were Class 8 (ePromoters). Notably, this implies that some have variable functional annotations across cell types. Nine of these 11 SNPs (81.82%) overlapped with an ENCODE cCRE. Together, the two approaches can assign a predicted function to 10 of the 11 (90.91%) SNPS. Thus, through an ensemble approach with a gene prioritization method such as cS2G (which itself incorporates ABC[6] and EpiMap[29] as well as non-functionally informed approaches such as TSS) and the ENCODE cCRE catalog, CoRE-BED may be used to enhance identification of causal target genes.

One of the 11 SNPs experimentally (i.e., causally) linked to the cS2G-predicted target gene was rs16886397, an intronic variant of *MAP3K1* associated with breast cancer. Within the MCF 10A breast cancer cell line, CoRE-BED predicts this SNP to be within a proximal insulator. In concordance with this prediction, we observed distinct, overlapping ChIP-seq signals at the SNP

locus for CTCF, RAD21, and SMC3, key insulator signatures, as well as DNase-seq and ATAC-seq signals, indicating open chromatin at the variant position (Figure 6a).

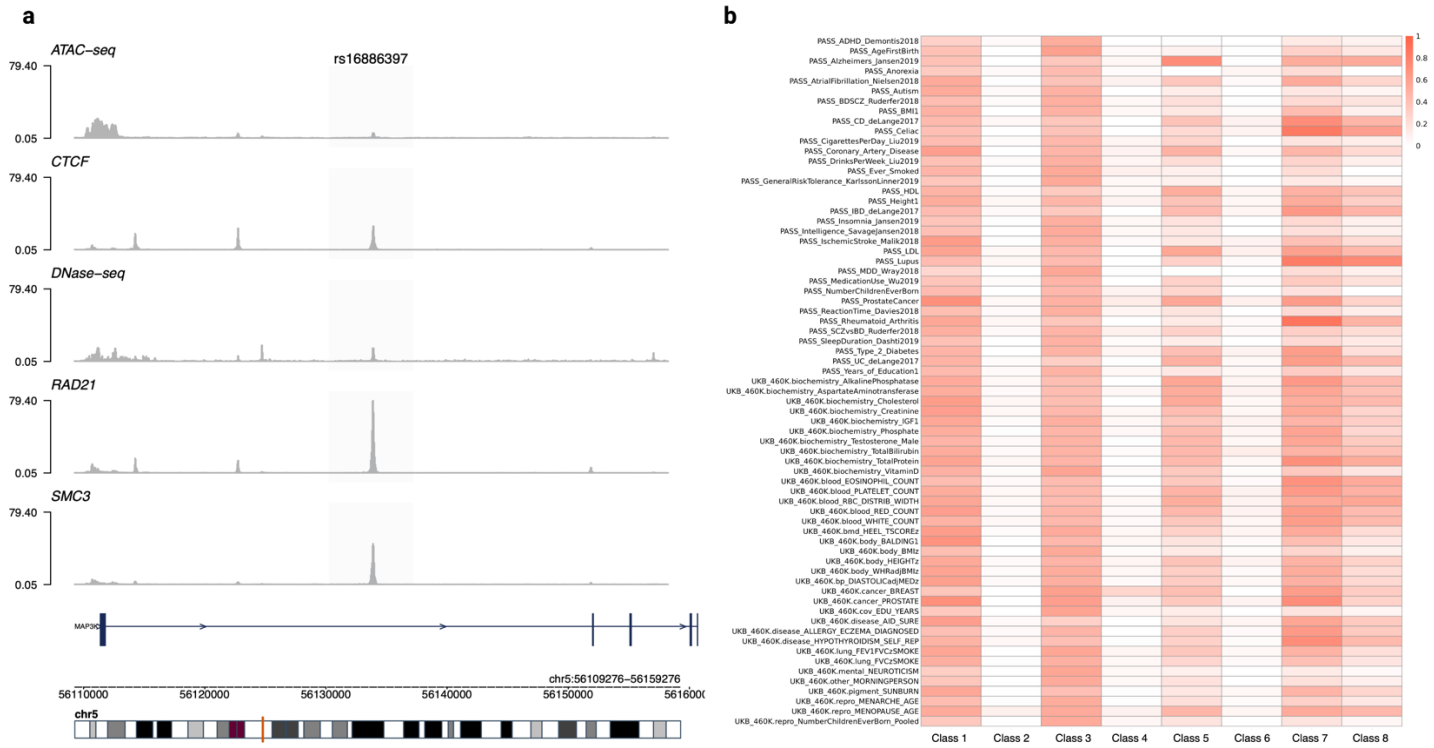


Figure 6. Applicability of CoRE-BED to post-GWAS analysis. a, *CoRE-BED accurately predicts function in gold standard list of causal variants**.*** *The breast cancer-associated SNP rs16886397 was functionally validated to confirm cS2G's prediction of MAP3K1 as the target gene. CoRE-BED predicts this SNP to be within a proximal insulator (Class 5 element). We observed an enrichment of ChIP-seq signal for CTCF, RAD21, and SMC3 at the SNP locus, all known signatures of insulator elements. Additionally, ATAC-seq and DNase-seq signal enrichment was present at the same locus, indicating open chromatin at the site. The SNP localizes to an intron of MAP3K1, supporting CoRE-BED's classification of a proximal versus distal insulator and demonstrating CoRE-BED's potential usefulness in identifying targets for functional studies.* ***b,*** *We estimated the proportion of SNP heritability attributable to CoRE-BED regulatory predictions using GWAS of 69 complex traits performed in up to 1.03 million individuals. Relative contributions varied by trait, with Class 1 (distal insulators) showing the greatest mean contribution across all traits.*

Because CoRE-BED functional predictions seemed to effectively capture all of the SNP heritability from the hypertension GWAS, we next applied this same methodology[67–69] more broadly across a total of 69 complex traits[66,70,71] to the GWAS summary statistics of the SNPs

with CoRE-BED annotations (Supplementary Table 5). Strikingly, we observed the same result across these 69 traits as we first saw for hypertension. However, the relative heritability contributions of the different classes of CoRE-BED regulatory elements were heterogeneous across phenotype (Figure 6b). Across all traits, Class 1 (distal insulator) elements explain the greatest mean proportion (0.4875) of SNP heritability (Supplementary Figure 4b), followed by Class 3 (transcribed gene body) with a mean proportion of 0.4795 (Supplementary Figure 4d) and Class 7 (active enhancer) with a mean proportion of 0.4564 (Supplementary Figure 4h).

Discussion

We present CoRE-BED, a decision tree-based framework that utilizes a broad set of genomic features to predict cell and tissue type-specific regulatory function. The decision tree was trained using 19 genomic features across 33 cell and tissue types, achieving high predictive performance in an independent test set. The decision tree implementation yields high interpretability, along with the ease of use with which functional prediction can be performed. In our software implementation, cell type-specific functional predictions from CoRE-BED can be scaled to annotate a set of genomic coordinates, including full GWAS summary statistics.

Intriguingly, of the 19 genomic features utilized by CoRE-BED, we observed only 35,991 unique combinations of features across more than 2.5 billion 1 kb bins from 816 biosamples representing the 33 cell and tissue types. If these features were distributed randomly throughout the genome, we would expect to observe well over 500,000 unique combinations (2^{19}). This discrepancy between possible and observed feature combinations suggests mechanistic constraint, under which only certain epigenomic feature combinations are

compatible with cellular function, or of redundancy, under which different combinations are effectively functional equivalents. Supporting this hypothesis is our finding that a sizable proportion of the total observed feature combinations were unique to cancer cell types. Although cancer is most commonly associated with somatic mutation, it is also characterized by epigenetic dysregulation[32]. Thus, it is reasonable to hypothesize that the 3,365 feature combinations unique to cancer result in aberrant or abnormal patterns of gene regulation. Potential future work could further investigate the function of these cancer-specific epigenomic signatures by inducing them in cells and characterizing the resulting changes in gene expression or cell morphology.

After training an optimal decision tree classifier, we generated functional predictions across 816 individual biosamples encompassing the 33 cell and tissue types represented by CoRE-BED[29]. Notably, our study suggests that regulatory elements are best viewed as fluid spectra of individual epigenetic features. This dynamic model resembles the degeneracy of the genetic code itself, in which wobble base pairing allows for multiple codons to code for a single amino acid. Despite this degeneracy, however, consistent patterns emerge, with each CoRE-BED functional class being most highly enriched for a specific subset of the epigenetic features.

One of the regulatory classes learned by CoRE-BED was enhancer-like promoters, or "ePromoters," which are characterized by enrichment of EP300 and H3K27ac, signatures of active enhancers[26], as well promoter-associated marks H3K4me3, H3K9ac, and POL2RA[23,27,28]. Previous work has described a class of promoters that also exhibit enhancer activity[33,34]. While the ePromoters predicted by CoRE-BED replicated some of the features previously associated with promoters with enhancer activity, we also identified a set of nearly 200 significantly

enriched TF binding motifs which have not previously been associated with this type of regulatory class (Supplementary Table 1). Knowledge of TF binding patterns at *cis*-regulatory elements is vital to understanding the dynamics of gene regulation at those loci and how disruption of TF binding might lead to disease phenotypes[72]. Thus, the data we present here represent an important contribution toward understanding the underlying mechanisms of how promoters with enhancer activity mediate gene expression.

In addition to ePromoters, CoRE-BED identified a previously uncharacterized regulatory class which we termed Development Associated Elements (DAEs). DAEs provide an example of the degeneracy of regulatory identity, in that this state can be defined by the dual presence of histone marks H3K4me2 and H3K9ac or, alternatively, H4K20me1 and H3K79me2. An alternative model architecture predicting a larger number of functional classes might possibly resolve these two DAE sub-classes as discrete regulatory categories. However, we also observed that at some DAE sites predicted based on lone H3K79me2 and H4k20me1 marks, H3K4me2 and H3K9ac signal is present further upstream (Figure 4). Thus, the two distinct biochemical signatures marking DAEs may, in fact, be mechanistically linked. Future functional studies might further investigate the interplay between these four histone marks across different cell types and developmental states.

DAEs are most highly enriched in cells with high differentiation potential and are enriched for genes associated with key developmental pathways. At first glance, these traits make DAEs appear highly similar to the well-studied bivalent promoter class of regulatory elements. To further explore whether DAEs were, in fact, bivalent promoters, we tracked the transition of DAEs as two ESC lines were terminally differentiated. Unlike bivalent promoters,

which function as a developmental intermediate state between active and silenced promoters[47], the vast majority DAEs in ESCs transitioned directly to or from a non-functional state. Thus, we conclude that DAEs do, indeed, represent a new class of regulatory elements. Although beyond the scope of this study, future work will aim to design experimental assays to further explore the biological function of DAEs and their mediating role in developmental processes.

In addition to the novel biology captured by the CoRE-BED model itself, another promising application of the CoRE-BED framework is high-throughput interpretation and prioritization of GWAS SNPs, retaining for downstream analyses those that are most highly enriched for trait heritability. Since the publication of the very first genome-wide association study[73], more than 5,700 GWAS have been published investigating over 3,300 complex phenotypes[74]. Despite the immense progress made in terms of variant discovery, however, understanding of the underlying biological mechanisms of these identified SNP associations lags behind[75]. A major barrier to GWAS SNP interpretation is the fact that most associations fall within non-coding regions of the genome and thus do not directly disrupt the coding sequence of a gene.

We demonstrate CoRE-BED's ability to prioritize GWAS variants based on tissue-specific functional annotation, both among genome-wide significant variants reported to the NHGRI-EBI GWAS Catalog and within an entire set of GWAS summary statistics from a multi-ancestry hypertension meta-analysis conducted within the UK Biobank. For example, hypertension-associated variants with a heart-specific CoRE-BED functional prediction capture approximately

all of the total SNP heritability, despite making up only ~5% of the total SNPs in the summary statistics.

A similar phenomenon was observed when estimating the heritability enrichment of predicted CoRE-BED functional elements across 69 additional traits. Notably, distal insulators encompassed the largest proportion of heritability across these 69 traits, capturing a mean 48.75% of GWAS SNP heritability across all studies (Figure 6b). Insulators play a key role in modeling three-dimensional genomic architecture and establishing boundaries between topologically associated domains[76,77]. It has been recently reported that TAD boundaries are enriched for complex trait heritability[78]. The results reported here support this finding and provide a comprehensive reference of insulator contributions to disease heritability, predicting both proximal and distal sub-classes.

The CoRE-BED implementation has several limitations. The classifier from a decision tree structure may be highly dependent on the composition of the training data. To address this issue, averaging techniques, including bagging and random forests, may be used, reducing the problem of high variance. However, given the scale of the epigenome-wide data, these approaches (and related ones) are computationally onerous. For this reason, we tested CoRE-BED in additional epigenomic data. Another limitation is that CoRE-BED's methodological simplicity may imply loss in power. As a decision tree, CoRE-BED is scale-invariant. A split on a feature is not influenced by the other features. No feature scaling is required, and the classifier is invariant to monotonic transformations of an epigenomic feature. Decision tree classifiers are, in this regard, inherently different from model-based methods such as neural networks.

On the other hand, the methodological simplicity of CoRE-BED greatly enhances the interpretability of its functional predictions. This interpretability is one of the major strengths of the CoRE-BED model. As demonstrated here, integrating CoRE-BED functional predictions into a genomic analysis workflow enhances understanding of the functional context of regions across the genome. Utilizing the CoRE-BED framework as one pillar of an integrative, multi-method approach of investigating genotype-phenotype interactions opens completely new avenues for potential study.

Methods

CoRE-BED classifier

The CoRE-BED framework is trained using 19 genomic features (TSS proximity; histone marks H3K27ac, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me2, H3K9ac, H3K9me3, and H4K20me1; chromatin accessibility assays ATAC-seq and DNase-seq; as well as CTCF, EP300, H2AFZ, POL2RA, RAD21, and SMC3 binding) across 33 cell and tissue types (adipose, blood and T-cell, bone, brain, cancer, digestive, endocrine, endothelial, epithelial, ESC-derived, ESC, eye, heart, HSC and B-cell, iPSC, kidney, liver, lung, lymphoblastoid, mesenchymal stromal, muscle, myosatellite, neurosphere, other, pancreas, peripheral neuron, placenta, reproductive, smooth muscle, spleen, stromal, thymus, and urinary) (Figure 1). The raw data were originally published as part of the EpiMap repository[29]. Because the average size of a genomic regulatory element (i.e., promoter or enhancer) is approximately 1 kb[79,80], the entire genome was first segmented into 1 kb bins. The result was an initial genome-wide set of 3.2 million regions[81] on which to conduct the training for the classifier. Overlap was calculated

between each 1 kb bin and the 19 genomic features[29] in 33 biosamples representing each of the cell and tissue types.

The resulting training data (presence or absence of each of the 19 genomic features in each 1 kb bin) are unlabeled, meaning there is no starting "ground truth." Thus, k-modes clustering[13,14], in conjunction with the elbow method[15], was applied to the 33 biosamples, resulting in 9 clusters of putative *cis*-regulatory elements that best captured the data. K-modes clustering was then repeated using this $k = 9$ to derive regulatory class labels (Class 0 – Class 8) for an independent dataset. This dataset consisted of 66 biosamples, two for each cell or tissue type.

Using these training data and their corresponding k-modes-derived labels, a decision tree classifier was trained using sci-kit learn[82]. A series of maximum depths ranging from 1 to 19 were tested, with node splits determined based on maximal Shannon entropy reduction[10]. The Shannon entropy $H(T)$ of the tree T quantifies the uncertainty of the tree-based classifier. Given the (disjoint) classes of regulatory elements, $x_1, x_2, \ldots, x_k$, derived from T, with corresponding probability $p(x_1), p(x_2), \ldots, p(x_k)$, the Shannon entropy $H(T)$ is defined as follows:

$$H(T) = -\sum_{i=1}^{k} p(x_i) \log_2 p(x_i)$$

Shannon entropy was a natural criterion because it represents the unique family of functions, defined on the set of all k-dimensional distributions (of the k functional annotations of the

classifier), satisfying the Shannon-Khinchin axioms of continuity, symmetry, maximality, and expandability[10].

A random 10% of the training data was held out to evaluate model performance in an independent test set, with the model using a maximum depth of 15 performing best (88.29% in the test set). Based on which learned combination of genomic features contributed most highly to each class, each of the nine regulatory categories was assigned a name suggestive of putative biological function.

Generation of CoRE-BED regulatory annotations

The epigenomic features were downloaded from the EpiMap Repository (http://compbio.mit.edu/epimap/) as bigWig tracks based on GRCh37/hg19. The bigWig tracks were converted to bedGraph format before undergoing peak calling using MACS2 to generate final sets of called peaks in UCSC BED format[83]. These resulting peak calls were based on GRCh37/hg19, and an accompanying set based on genome build GRCh38/hg38 was generated using UCSC liftOver[84]. TSS coordinates based on GRCh38/hg38 were downloaded from refTSS[85], and UCSC liftOver was used to generate a set of TSS coordinates based on GRCh37/hg19. These data were binned at 1 kb resolution and fed into the trained CoRE-BED decision tree to generate genome-wide functional class across all 1 kb bins in 816 unique biosamples.

Comparison of CoRE-BED functional predictions with ENCODE cCREs

GRCh38/hg38-based cCREs predicted across 1,518 cell types were downloaded from the SCREEN Registry of cCREs V3 (https://screen.encodeproject.org) and lifted over to GRCh37/hg19. For both these ENCODE cCREs and CoRE-BED-predicted functional regions,

overlapping coordinates were merged. The degree of sharing between each set of predictions was evaluated using both percent overlap and Jaccard index. We calculated the Jaccard similarity coefficient $s(x_i, E; T)$ of each class x_i of regulatory elements derived from the CoRE-BED decision tree T and the set of putative ENCODE cCREs E as follows:

$$s(x_i, E; T) = {|x_i \cap E|}/{|x_i \cup E|}$$

The Jaccard index measures the similarity between the two sets as the proportion of the number of shared elements (i.e., CoRE-BED-predicted regulatory elements in x_i overlapping a putative ENCODE cCRE) relative to the number of distinct elements across the two sets. Note the overall overlap count $\Omega(T, E)$ of the set of T-derived regulatory elements with the set of ENCODE cCREs is given by:

$$\Omega(T, E) = \sum_{i=1}^{n} |x_i \cap E| = \sum_{i=1}^{n} s(x_i, E; T) * |x_i \cup E|$$

Assessing expression of PNS genes under active promoter control

Class 2 (active promoter 1) CoRE-BED elements were most highly enriched in the PNS. In order to first identify which genes were under the control of these elements in PNS-derived tissues, a list was compiled of all unique genomic regions predicted as Class 2 elements within at least one of 10 PNS-derived biosamples. Using the closest function from bedtools[86], the closest gene to each putative active promoter was identified.

Next, gene-level expression data from tibial nerve, a PNS-derived tissue, was downloaded from the GTEx Consortium[44] (https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/gene_reads/gene_reads_2017-06-05_v8_nerve_tibial.gct.gz). Gene expression was quantified using transcripts per million (TPM). Expression levels of genes under the predicted control of a CoRE-BED active promoter were compared with the remaining genes expressed in tibial nerve using a Wilcoxon rank sum test.

Enhancer RNA (eRNA) enrichment in ePromoters

Class 8 elements learned by the CoRE-BED model were termed ePromoters, promoters with enhancer activity. A comprehensive set of eRNA coordinates was compiled using publicly-available annotation resourced compiled by ENSEMBL[87], FANTOM5[88], and Roadmap[89], in addition to the Human enhancer RNA Atlas (HeRA)[90]. Distance from an eRNA was calculated for each ePromoter and all remaining promoters across the genome using bedtools[86], and the respective distributions of distances were compared using the Wilcoxon rank sum test[86].

Gene pathway over-representation analyses for enhancer-like and DAEs

The closest gene to each CoRE-BED-predicted putative promoter was determined. The resulting sets of genes under ePromoter and DAE control, respectively, underwent over-representation analysis using WebGestalt[41] (http://www.webgestalt.org) to search for enrichment of KEGG[42] pathway terms. The Organism of Interest was set to “Homo sapiens,” Method of Interest was set to “Over-Representation Analysis (ORA),” and Functional Database was set to “pathway,” with the functional database name set to “KEGG.” The Gene ID Type

used was “Gene symbol,” while the Reference Set used was “genome.” Additionally, under “Advanced parameters,” we used an FDR of 0.05 as the significance level.

Transcription factor binding motif enrichment in enhancer-like promoters and DAEs

The ePromoters and DAEs identified by CoRE-BED underwent separate motif enrichment analysis using the *findMotifsGenome.pl* script from the HOMER suite[91]. A Benjamini-Hochberg p-value correction was used to determine the significance threshold for motif enrichment.

Computing tau indices for genes under putative DAE control

Tau scores calculated using the GTEx v8[44] dataset were downloaded from https://genomics.senescence.info/gene_expression/tau.html[43]. The closest function from bedtools[86] was used to determine which GENCODE[92] gene was closest to each predicted DAE across all 833 tested biosamples, and the corresponding tau index for each of unique gene was retrieved from the reference dataset.

Predicting GWAS Catalog variants in a DAE

All associations (v1.0) from the NHGRI-EBI GWAS Catalog[48] (https://www.ebi.ac.uk/gwas/docs/file-downloads) were downloaded, and all reported variants missing genomic coordinates, allele frequency, effect size, or GWAS p-value were filtered out. For the 146,338 remaining SNPs, CoRE-BED functional prediction was performed across all 816 supported biosamples. 1,697 were predicted to be within a DAE in at least one cell type, and

375 of these were specifically predicted to be within a DAE in a brain-derived cell type (286 in astrocytes).

LocusZoom plot for DAE locus

Once the Parkinson's disease-associated SNP rs16886397 was identified as a variant of interest, we located the GWAS summary statistics from a recent highly powered meta-analysis that replicated the association. The summary statistics were uploaded to the LocusZoom[93] web server (http://locuszoom.org) to generate the plot in Figure 4a.

Trackplot visualizations

The R package trackplot[94] (https://github.com/PoisonAlien/trackplot) was utilized to visualize bigWig signal at the genomic loci of interest in Figures 4 and 6.

Describing astrocyte eQTLs within DAEs

Astrocyte data was obtained from the Human Induced Pluripotent Stem Cell Initiative (HipSci)[51]. Using individual-level genotyping and gene expression data from astrocytes derived from differentiated iPSCs grown for 52 days, eQTL mapping was performed. Next, astrocyte-specific CoRE-BED functional predictions were generated across all 71,226 predicted eQTLs. For each of the 126 eQTLs falling within a CoRE-BED-predicted DAE, over-representation analysis (via WebGestalt[41]) using both KEGG[42] pathway and Human Phenotype Ontology[95] terms was performed on the set of genes under the predicted *cis*-regulatory control of these DAE eQTLs. Using these pathway-based approaches, genes associated with brain-related phenotypes such as Parkinson's disease, Alzheimer's disease, and anxiety were identified.

Tracking resolution of DAEs during embryonic development

Two ESC lines, H1 and H9, as well as two ESC-derived cell types differentiated from each respective ESC (H1-derived neural crest cells and H9-derived smooth muscle cells) were selected to compare differences in CoRE-BED-predicted DAEs (Class 4 elements). For each pair of cells, we assessed the number of DAEs in the ESC line that were lost after differentiation and which predicted functional categories they transitioned to. We also assessed the DAEs that were gained as a result of differentiation. The vast majority of DAEs both gained and lost transitioned to or from a predicted Class 0 (non-functional) state. Thus, at these sites, it was next observed what gain or loss of histone marks at those sites corresponded with the changes in predicted functional status.

Application to GWAS variant prioritization

Using the filtered set of GWAS Catalog variants described above (see **Predicting GWAS Catalog variants in a DAE**), additional filtering was performed to remove associations with continuous traits, retaining only those from case-control studies. CoRE-BED was used to predict functional status across the resulting variant set across all 816 supported biosamples. Annotations were based on genome build GRCh38. After annotation, variants were stratified based on whether each variant fell within a predicted functional region (Class 1-8) in at least one cell or tissue type. A Wilcoxon rank sum test with continuity correction was used to statistically compare effect size, p-value, and MAF distributions between variant sets with and without a functional annotation. Significant differences were found for both GWAS p-value and MAF between variants with and without a CoRE-BED functional prediction (see Results). We

found no statistically significant difference in odds ratio (OR) between GWAS Catalog SNPs with a CoRE-BED functional prediction and those without (with functional annotation median OR 1.1038; without functional annotation OR median 1.1039; p = 0.079).

Hypertension meta-analysis variant annotation and prioritization

The summary statistics from a multi-ancestry meta-analysis of hypertension conducted in the UK Biobank were downloaded from the Pan-UK Biobank[62] portal (https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_flat_files/categorical-20002-both_sexes-1065.tsv.bgz). Functional status was predicted for all GWAS variants, and the summary statistics were subsequently stratified based on the presence or absence of a functional prediction in at least one of 41 heart-derived biosamples. The GWAS p-values of these sets of functionally-stratified SNPs were plotted using the *qqunif* function from the gap R package[96] and tested for statistically significant divergence using a Wilcoxon rank sum test with continuity correction.

Partitioned heritability estimation of hypertension and 69 additional GWAS traits

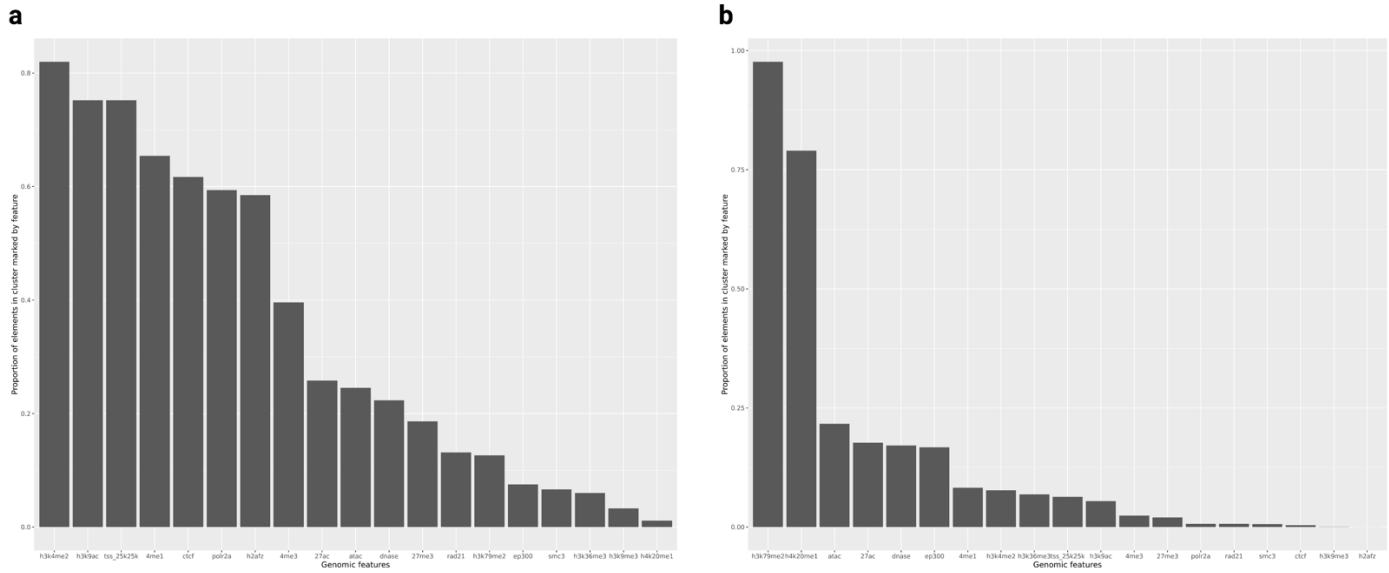
CoRE-BED functional predictions were generated for the 1000 Genomes[97] European reference set of variants. Annotation status for each functional class was used to generate partitioned linkage disequilibrium scores using LDSC[70,71]. Partitioned heritability based on CoRE-BED functional classes was first estimated for hypertension. Next, for the 69 traits assessed by Gazal et al. to evaluate the performance of cS2G[66], we leveraged these same partitioned LD scores to estimate the proportion of SNP heritability h^2 captured by variants in each of the 8 CoRE-BED derived functional (Classes 1-8) categories.

Data Availability

The CoRE-BED model, command line tool, and associated code are publicly available on GitHub (https://github.com/mjbetti/CoRE-BED). Pre-compiled reference files for use with the command line tool, ChIP-seq peak calls generated from EpiMap bigWig tracks, and genome-wide sets of CoRE-BED functional predctions can be retrieved from Zenodo (10.5281/zenodo.7558115). Visualizations in Figure 1 were created with Biorender.com.

Acknowledgment

This research was supported by National Institutes of Health (NIH) grants NHGRI R35HG010718, NHGRI R01HG011138, NIA AG068026, NIGMS R01GM140287, NIMH R01MH126459, and NIH/NCI U01CA253560.



***Extended Figure 1. CoRE-BED DAEs are characterized by two previously undescribed combinations of epigenetic markings**. **a,** The decision tree training and test sets were composed of 18,522 unique combinations of 19 genomic features. The 783 feature combinations labeled Class 4 (DAEs) in this set were most highly enriched for histone marks H3K4me2, H3K9ac, in addition to TSS proximity. **b,** Across all 173,434 (non-unique) DAE occurrences in 816 individual biosamples representing 33 major cell and tissue types, elements are most highly enriched for H3K79me2 and H4K20me1.*

References

1. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
2. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).
3. Alsheikh, A. J. *et al.* The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases. *BMC Med. Genomics* **15**, 74 (2022).
4. Soskic, B. *et al.* Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nat. Genet.* **51**, 1486–1493 (2019).
5. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
6. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
7. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
8. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
9. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 22071–22080 (2019).

10. Khinchin, A. I. & Hinčin, A. J. *Mathematical Foundations of Information Theory*. (Dover Publications, 1957).

11. Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N. & Yeh, H.-Y. Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams. *Front Bioeng Biotechnol* **7**, 305 (2019).

12. Snetkova, V. & Skok, J. A. Enhancer talk. *Epigenomics* **10**, 483–498 (2018).

13. Huang, Z. & Ng, M. K. A Note on K-modes Clustering. *J. Classification* **20**, 257–261 (2003).

14. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.* **2**, 283–304 (1998).

15. Marutho, D., Hendra Handaka, S., Wijaya, E. & Muljono. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. in *2018 International Seminar on Application for Technology of Information and Communication* 533–538 (2018).

16. Cheng, H., Zhang, N. & Pati, D. Cohesin subunit RAD21: From biology to disease. *Gene* **758**, 144966 (2020).

17. Beacon, T. H. *et al.* The dynamic broad epigenetic (H3K4me3, H3K27ac) domain as a mark of essential genes. *Clin. Epigenetics* **13**, 138 (2021).

18. Cockerill, P. N. Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J.* **278**, 2182–2210 (2011).

19. Collins, B. E., Greer, C. B., Coleman, B. C. & Sweatt, J. D. Histone H3 lysine K4 methylation and its role in learning and memory. *Epigenetics Chromatin* **12**, 7 (2019).

20. Sun, Z. *et al.* H3K36me3, message from chromatin to DNA damage repair. *Cell Biosci.* **10**, 9 (2020).

21. Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife* **6**, (2017).

22. Hou, Y. *et al.* Integrative characterization of G-Quadruplexes in the three-dimensional chromatin structure. *Epigenetics* **14**, 894–911 (2019).

23. LaMere, S. A., Thompson, R. C., Komori, H. K., Mark, A. & Salomon, D. R. Promoter H3K4 methylation dynamically reinforces activation-induced pathways in human CD4 T cells. *Genes Immun.* **17**, 283–297 (2016).

24. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).

25. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295 (2013).

26. Zhang, B. *et al.* A dynamic H3K27ac signature identifies VEGFA-stimulated endothelial enhancers and requires EP300 activity. *Genome Res.* **23**, 917–927 (2013).

27. Karmodiya, K., Krebs, A. R., Oulad-Abdelghani, M., Kimura, H. & Tora, L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13**, 424 (2012).

28. Barrera, L. O. *et al.* Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.* **18**, 46–59 (2008).

29. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).

30. Putiri, E. L. & Robertson, K. D. Epigenetic mechanisms and genome stability. *Clin. Epigenetics* **2**, 299–314 (2011).

31. Sarkies, P. & Sale, J. E. Cellular epigenetic stability and cancer. *Trends Genet.* **28**, 118–127 (2012).

32. Muntean, A. G. & Hess, J. L. Epigenetic dysregulation in cancer. *Am. J. Pathol.* **175**, 1353–1361 (2009).

33. Dao, L. T. M. & Spicuglia, S. Transcriptional regulation by promoters with enhancer function. *Transcription* **9**, 307–314 (2018).

34. Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).

35. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).

36. Gates, L. A. *et al.* Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *J. Biol. Chem.* **292**, 14456–14472 (2017).

37. Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.* **20**, 1493–1502 (2010).

38. Wang, S., Meyer, D. H. & Schumacher, B. H3K4me2 regulates the recovery of protein biosynthesis and homeostasis following DNA damage. *Nat. Struct. Mol. Biol.* **27**, 1165–1177 (2020).

39. Shoaib, M. *et al.* Histone H4 lysine 20 mono-methylation directly facilitates chromatin openness and promotes transcription of housekeeping genes. *Nat. Commun.* **12**, 4800 (2021).

40. Farooq, Z., Banday, S., Pandita, T. K. & Altaf, M. The many faces of histone H3K79 methylation. *Mutat. Res. - Rev. Mut. Res.* **768**, 46–52 (2016).

41. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).

42. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457-62 (2016).

43. Palmer, D., Fabris, F., Doherty, A., Freitas, A. A. & de Magalhães, J. P. Ageing transcriptome meta-analysis reveals similarities and differences between key mammalian tissues. *Aging* **13**, 3313–3341 (2021).

44. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

45. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

46. Mas, G. *et al.* Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat. Genet.* **50**, 1452–1462 (2018).

47. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).

48. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

49. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).

50. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).

51. Streeter, I. *et al.* The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* **45**, D691–D697 (2017).

52. Wang, X.-L. & Li, L. Cell type-specific potential pathogenic genes and functional pathways in Alzheimer's Disease. *BMC Neurol.* **21**, 381 (2021).

53. Arena, G., Modjtahedi, N. & Kruger, R. Exploring the contribution of the mitochondrial disulfide relay system to Parkinson's disease: the PINK1/CHCHD4 interplay. *Neural Regeneration Res.* **16**, 2222–2224 (2021).

54. Kennedy, R. B. *et al.* Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Hum. Genet.* **131**, 1403–1421 (2012).

55. Suzuki, G. *et al.* Over-expression of a human chromosome 22q11.2 segment including TXNRD2, COMT and ARVCF developmentally affects incentive learning and working memory in mice. *Hum. Mol. Genet.* **18**, 3914–3925 (2009).

56. Stechmiller, J. K. *et al.* Biobehavioral Mechanisms Associated With Nonhealing Wounds and Psychoneurologic Symptoms (Pain, Cognitive Dysfunction, Fatigue, Depression, and

Anxiety) in Older Individuals With Chronic Venous Leg Ulcers. *Biol. Res. Nurs.* **21**, 407–419 (2019).

57. Fu, Y. *et al.* Systematically Analyzing the Pathogenic Variations for Acute Intermittent Porphyria. *Front. Pharmacol.* **10**, 1018 (2019).
58. Srour, M. *et al.* Dysfunction of the Cerebral Glucose Transporter SLC45A1 in Individuals with Intellectual Disability and Epilepsy. *Am. J. Hum. Genet.* **100**, 824–830 (2017).
59. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
60. Evans, P., Cox, N. J. & Gamazon, E. R. The regulatory genome constrains protein sequence evolution: implications for the search for disease-associated genes. *PeerJ* **8**, e9554 (2020).
61. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. *bioRxiv* 2022.05.07.491045 (2022) doi:10.1101/2022.05.07.491045.
62. Pan UKBB. https://pan.ukbb.broadinstitute.org.
63. Azevedo, T., Dimitri, G. M., Lió, P. & Gamazon, E. R. Multilayer modelling of the human transcriptome and biological mechanisms of complex diseases and traits. *npj Systems Biology and Applications* **7**, 1–13 (2021).
64. Kupper, N., Ge, D., Treiber, F. A. & Snieder, H. Emergence of novel genetic effects on blood pressure and hemodynamics in adolescence: the Georgia Cardiovascular Twin Study. *Hypertension* **47**, 948–954 (2006).
65. Shih, P.-A. B. & O'Connor, D. T. Hereditary determinants of human hypertension: strategies in the setting of genetic complexity. *Hypertension* **51**, 1456–1464 (2008).

66. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* (2022) doi:10.1038/s41588-022-01087-y.

67. Davis, L. K. *et al.* Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet.* **9**, e1003864 (2013).

68. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

69. Torres, J. M. *et al.* Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *Am. J. Hum. Genet.* **95**, 521–534 (2014).

70. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

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

72. Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor-DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43**, 110–119 (2017).

73. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).

74. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).

75. Stranger, B. E., Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367–383 (2011).

76. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genomics Hum. Genet.* **18**, 45–63 (2017).

77. Matharu, N. K. & Ahanger, S. H. Chromatin Insulators and Topological Domains: Adding New Dimensions to 3D Genome Architecture. *Genes* **6**, 790–811 (2015).

78. McArthur, E. & Capra, J. A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am. J. Hum. Genet.* **108**, 269–283 (2021).

79. Xuan, Z., Zhao, F., Wang, J., Chen, G. & Zhang, M. Q. Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biol.* **6**, R72 (2005).

80. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol.* **22**, 108 (2021).

81. Brown, T. A. *The Human Genome*. (Wiley-Liss, 2002).

82. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* (2012).

83. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).

84. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590-8 (2006).

85. Abugessaisa, I. *et al.* refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *J. Mol. Biol.* **431**, 2407–2422 (2019).

86. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

87. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).

88. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Sci Data* **4**, 170107 (2017).

89. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

90. Zhang, Z. *et al.* HeRA: an atlas of enhancer RNAs across human tissues. *Nucleic Acids Res.* **49**, D932–D938 (2021).

91. Duttke, S. H., Chang, M. W., Heinz, S. & Benner, C. Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res.* **29**, 1836–1846 (2019).

92. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

93. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

94. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618–1619 (2014).

95. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).

96. Zhao, J. H. gap: Genetic Analysis Package. *J. Stat. Softw.* **23**, 1–18 (2008).

97. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).