Pan-cell type continuous chromatin state annotation of all IHEC epigenomes

Habib Daneshpajouh[1], Ismail Moghul[2], Kay C Wiese[1], and Maxwell W Libbrecht[1]
[1]School of Computing Science, Simon Fraser University, Canada
[2]UCL Cancer Institute, University College London, UK

1 Abstract

Understanding the mechanistic basis of genetic disease requires annotating the regulatory elements in the human genome. To this end, the International Human Epigenome Consortium (IHEC) has generated thousands of epigenomic datasets—including ChIP-seq, DNase-seq, and ATAC-seq—that measure various biochemical activities in the genome, including transcription factor binding, histone modification, and DNA accessibility. Currently, the predominant methods for integrating these data sets to annotate regulatory elements are segmentation and genome annotation (SAGA) algorithms such as ChromHMM and Segway. The majority of annotations generated by these methods are cell type-specific. However, as the number of profiled cell types has grown into the thousands, using thousands of cell type-specific chromatin state annotations proves undesirable for many applications. Therefore, recently, researchers have sought a single unified annotation of regulatory elements across all cell types, known as a "pan-cell type" annotation. Here, we present a pan-cell type annotation that summarizes all IHEC epigenomes using the recently-developed method, epigenome-ssm. This pan-cell type annotation comprises 33 genome-wide chromatin state feature signal tracks, each of which captures a regulatory program driving genomic activity in one or more cell types. We show that these feature maps constitute an intuitive and visualizable summary of epigenomic data.

2 Background

Understanding the mechanistic basis of genetic disease requires annotating the regulatory elements in the human genome. To this end, international projects including ENCODE, Roadmap Epigenomics, BLUEPRINT, CEEHRC, DEEP, GIS, AMED-CREST, and EpiHK have generated thousands of epigenomic datasets such as ChIP-seq, DNase-seq, and ATAC-seq that measure various biochemical activities in the genome, including transcription factor binding, histone modification, and DNA accessibility. More recently, the International Human Epigenome Consortium (IHEC) unified these projects in order to re-process all these epigenomic

data sets through a uniform pipeline. The IHEC repository contains histone modification ChIP-seq, gene expression RNA-seq, and DNA methylation WGBS data across hundreds of human biosamples.

Integrative analysis of these data sets is crucial to understanding the regulatory landscape of the human genome and its role in phenotype and disease. Currently, the predominant methods for integrating epigenomic datasets to annotate regulatory elements are segmentation and genome annotation (SAGA) algorithms such as ChromHMM [1] and Segway [2]. SAGA algorithms take as input data from a given cell type, partition the genome, and assign a chromatin state label to each segment indicating the epigenetic activity at that position [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. Consequently, they can identify regulatory elements without requiring prior knowledge of known genomic elements.

As the number of profiled cell types has grown into the thousands, using thousands of cell type-specific chromatin state annotations proves undesirable for many applications. Therefore, recently, researchers have sought a single unified annotation of regulatory elements across all cell types, known as a "pan-cell type", "universal", or "stacked" annotation [28, 29, 7]. A pan-cell type annotation summarizes the activity of each genomic position across all profiled cell types and thus is ideal for understanding evolutionary conservation or disease association. However, pan-cell type annotation is challenging because the combinatorial number of patterns of activity may grow exponentially in the number of cell types.

To tackle this challenge, we recently developed epignome-ssm [30], a method that summarizes each genomic position as a vector of continuous chromatin state features, in contrast to the previous approach of assigning a discrete chromatin state label. This continuous approach can summarize complex high-dimensional datasets into a small number of interpretable chromatin state features. Unlike discrete labels, these continuous features preserve the underlying continuous nature of the epigenomic signal tracks, can easily represent varying strengths of a given element, can represent combinatorial elements with multiple types of activities, and are shown to be useful for expressive visualizations because they map complex high-dimensional datasets onto a much smaller number of dimensions.

Here, we present a chromatin state feature map summarizing all IHEC epigenomes. This feature map comprises 33 genome-wide chromatin state feature signal tracks that summarize all data sets. Each feature track captures a regulatory program that drives gene expression in one or more cell types. For example, as described below, one of the feature tracks (referred to as the Prom-Brain feature) captures the regulatory activity of the promoters active in the brain-related cell types.

We show that these feature maps constitute an intuitive and visualizable summary of epigenomic data, and that they enable accurate identification of mechanisms of disease association. We show that this is advantageous over alternative pan-cell type mapping approaches according to several metrics. Thus, we expect that this pan-cell type chromatin state feature map will provide a crucial resource to the IHEC community and enable many downstream analyses and insights.

3 Results

3.1 Pan-cell type chromatin state feature maps across thousands of human epigenomes

We present a continuous chromatin state feature map generated using epigenome-ssm on 9,539 genome-wide signal tracks from six core histone modification assays (H3K4me1, H3K27ac, H3K4me3, H3K36me3, H3K27me3, and H3K9me3) across 1,698 epigenomes (Fig. 1A, Methods). Briefly, epigenome-ssm takes as input 9,539 genome-wide signal tracks and outputs 33 pan-cell type chromatin state feature tracks (Fig. 1B-C, Methods). Each feature track summarizes a subset of the input tracks and captures a regulatory program that drives gene expression in one or more cell types.

The epigenome-ssm model is an unsupervised state-space model. It assumes that there is an unobserved vector of nonnegative chromatin state features at each genomic position, and that observed signal tracks are generated as a linear function of these chromatin state features. We trained the model using an expectation-maximization algorithm to learn parameters and active set optimization to enforce nonnegativity constraints, iterating until it reached a stable log-likelihood. After training, we normalized the feature values by their genome-wide means to enhance interpretability and standardize values for downstream analyses. This approach is similar to stacked SAGA methods like ChromHMM and Segway, with a few differences, except that existing SAGA methods generate a single discrete label per position, whereas epigenome-ssm generates a vector of 33 continuous values.

As an illustrative example, consider the ETS2 locus (Fig. 1D). This gene is on the reverse strand and has multiple transcription start sites, and is known to be expressed in several blood cell types, especially the myeloid lineage of the immune cells, such as monocytes and neutrophils (TSSs; Fig. 1D, GENCODE track). The continuous pan-cell type features provide a more complete picture of the regulatory activities at this locus than both the GENCODE gene annotation and ENCODE candidate cis-regulagory elements (cCREs). epigenome-ssm features show several blood-related promoter features that are enriched around the promoter of ETS2. The Prom-Blood1 feature captures the inner TSSs, Prom-Blood2 captures the most 5' TSS, and most importantly, Prom-Myeloid is highly enriched around all the TSSs. As shown in the ENCODE cCRE track, this gene contains several intronic enhancers; these are captured by the Enh-SG1, Enh-Fetal, Enh-Liver, and TxEnh-Myeloid, which are enriched for different subsets of those enhancers. Finally, three transcription-related features, Tx-Myeloid1, Tx-Myeloid2, and TxProm-Broad capture the pattern of transcription of this gene across different cell lineages.

We also found that epigenome-ssm features are complementary to the existing pan-cell type annotations such as the ENCODE cCREs (Fig. 1D, bottom track). The ENCODE cCRE annotations contain the position of candidate enhancer and promoter elements. ENCODE cCREs were generated based on the combination of different epigenomic signals in each region. However, such annotation may not provide a complete picture of all the epigenomic activity in a given locus. For example, the two inner TSSs are annotated as enhD

(i.e., distal enhancer), while those are likely combinatorial regions exhibiting both enhancer and promoter activities.

One of the design considerations when training unsupervised annotation models is the number of labels (or the number of features in the case of continuous chromatin state models). Here, we initially trained the model with potentially more features than needed (i.e., with 50 features). Then, we sorted all the features by their total emission value, corresponding to each feature's overall contribution to the model, and used an elbow method to drop the features with low total emission (Fig. S1). This approach is analogous to the common practice used in Principal Component Analysis (PCA), where the principal components with the largest eigenvalues explain the most variance in the data. Just as we often discard PCA components with low eigenvalues to retain only the most informative ones, we retained pan-cell type features with high emission values, ensuring that our model focuses on the most meaningful chromatin states while minimizing redundancy.

3.2 Pan-cell type features outperform alternatives according to quantitative metrics

We hypothesized that a good annotation should capture the key genomic activities involved in gene regulation, so we assessed the quality of the pan-cell annotations by examining their associations with gene expression, enhancer activity, and evolutionary conservation. Specifically, for each method and each genomic activity, we trained a linear regression model that used the annotations as predictors to estimate the corresponding activity levels.

We compared epigenome-ssm to alternative pan-cell type annotations (Methods). Four methods are compared: DHS-NMF with 16 features, Universal ChromHMM with 100 states, and epigenome-ssm with both 16 and 50 features. Across all tasks, the epigenome-ssm models consistently achieve a high prediction performance. We found that all the methods are predictive of gene expression. However, epigenome-ssm annotations led to better prediction performance using a smaller number of features than the alternative methods. (Fig. 2A-D). Notably, even with just 16 features, epigenome-ssm rivals or exceeds the performance of ChromHMM, which uses 100 discrete states, highlighting the efficiency and predictive power of the continuous features.

We also found that the performance of epigenome-ssm is robust to changes in the number of features (K) (Fig. 2A-D). Both the K=16 and K=33 models yield similar Pearson correlation coefficients, with slight improvements when using more features. This indicates that our pan-cell type features capture the essential aspects of the epigenomic data, providing reliable predictions across different genomic activities with a relatively small set of features.

The better performance of epigenome-ssm highlights the better expressive power of continuous features compared to the discrete chromatin state labels. For example, the intensity of a promoter feature tends to correlate with the expression level of genes regulated by promoters of that type (Fig. 2E). In discrete SAGA,

the differences between low- and high-intensity regulatory elements is often represented with separate labels of e.g. "Weak Enhancer" and "Weak Transcription", which makes analysis challenging [1, 7].

Additionally, we compared epigenome-ssm with CSREP [31], a method for summarizing SAGA annotations across many cell types, and found that the pan-cell type features yield a significantly better prediction performance compared to CSREP (Fig. S2). CSREP summarizes SAGA annotations from a group of related samples into a single annotation that summarizes the group. A collection of CSREP annotations is similar to a pan-cell type annotation in the sense that both reduce thousands of experiments to a manageable size. We compared epigenome-ssm to CSREP summaries of 1,698 IHEC samples down to 56 summary annotations.

3.3 Pan-cell type features capture a variety of types of epigenomic activities

We found that epigenome-ssm chromatin state features encompass the known diversity of epigenomic activities. Each chromatin state feature captures both a type of activity and its pattern of activity across cell types. Thus, where a cell type-specific annotation of cell type X might have a mnemonic "Enh" meaning "enhancer active in cell type X", a pan-cell type model feature has mnemonics of the form "Enh-Blood", meaning "enhancer active across blood cell types". Therefore, the mnemonics that we assigned to each pan-cell type feature represent a pair of activity and cell type. We found several main categories of features (Table 1, Fig. 3) by looking at features from different perspectives: (1) Model emissions that associate features with the input tracks with respect to different characteristics of the input samples such as assay, ontology, and donor information; (2) Average feature value at annotated genomic elements such as the transcription start sites (TSSs), DNase I hypersensitive sites (DHSs), introns and so on; (3) Gene expression; and (4) Gene Ontology (GO).

"Prom" features correspond to promoter activity, characterized by the marks H3K4me3 and H3K27ac in active cell types and H3K27me3 in inactive cell types (Fig. 3C, 4B, S4-S5, Table 1). As expected, Prom features are typically present at annotated TSSs (Fig. 3B and 3G). Additionally, we found that the spatial distribution of feature values varies depending on the type of epigenomic activities that each feature captures. For instance, Prom features tend to have a punctate distribution with zero or near zero values in most regions of the genome and high values (greater than 5) in the rest of the genome (Fig. 3A).

"Enh" features correspond to enhancer activity, characterized by the marks H3K4me1 and H3K27ac in active cell types and H3K27me3 in inactive cell types (Fig. 3C, 4B, S3-S4, Table 1). These features are typically present at annotated enhancers and DHSs (Fig. 3B) and have a punctate distribution in these regions (Fig. 3A). "TxEnh" features correspond to enhancer activity of those enhancer elements that are typically found within the transcribed regions of the genome, such as the introns, characterized by the marks H3K4me1, H3K27ac, and H3K36me3 in active cell types and H3K27me3 in inactive cell types (Fig. 3C, S3-S4 and S6, Table 1).

"Tx" features are those that capture transcription activity, characterized by the mark H3K36me3 (Fig. 3C, S5, Table 1). As expected, Tx features are present at regions associated with transcription, such as the cod-

ing regions (CDS), introns, and untranslated regions (UTRs) (Fig. 3B and 3H). Unlike the Prom and Enh features, Tx features tend to have a broad distribution, with moderate values (values 1-5) over broad regions.

Epigenetic patterns at the flanking regions to regulatory elements (e.g., promoters and enhancers) belong to a category we term "Flank" (Fig. 3B), characterized by a dispersed moderate signal in H3K4me1, H3K4me3, and H3K27ac (Fig. S3-S5, Table 1).

"FaculHet" and "ConstHet" refer to Faculatative and Constitutive heterochromatin, characterized by H3K27me3 and H3K9me3, respectively. We identified two ConstHet features capturing broad H3K9me3 signal across all samples, and another ConstHet feature capturing H3K9me3 in a subset of the samples (Fig. 3C, Table 1). In contrast, most H3K27me3 signal is captured by Prom and Enh features, which typically correspond to active regulatory marks in a few samples and H3K27me3 in others (Fig. S3-S6). As expected, heterochromatin features are depleted at annotated regulatory elements (Fig. 3B). Similar to the Tx features, both types of heterochromatin features tend to have a broad distribution, with moderate values (values 1-5) over broad regions.

Apart from these canonical activity types, we identified some other rarer types: Prom-Znf, characterized by H3K4me3 at zinc-finger genes (Fig. 3B-C, Table 1). Also, we have a "TxProm" feature that is present in both promoter and transcribed regions, characterized by the marks H3K4me3 and H3K36me3 (Fig. 3B-C, Table 1).

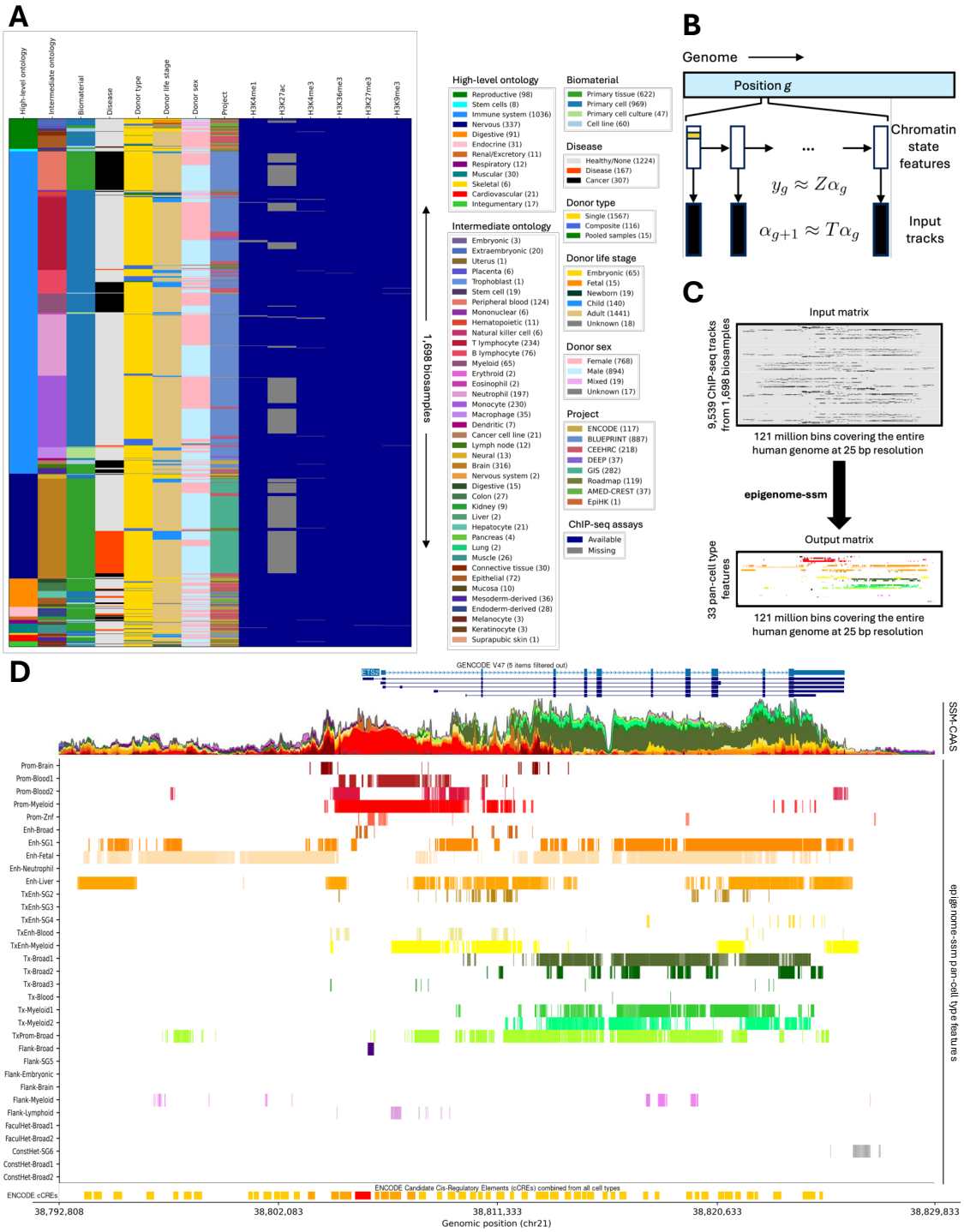

Figure 1: **We used epigenome-ssm to generate 33 pan-cell type features across 1,698 IHEC biosamples.** **A** The IHEC dataset contains tracks from six core histone modification ChIP-seq assays across 1,698 biosamples. **B** epigenome-ssm uses a state space model in which chromatin state features at each position of the genome are modeled by an emission matrix Z and a transition matrix T. **C** A total of 9,539 genome-wide tracks are passed to epigenome-ssm as input. It outputs a vector of k chromatin state features at each genomic position that represents the pan-cell type activity at that position. Here, we present a model with 33 chromatin state features. **D** UCSC Genome Browser visualization of pan-cell type features at the region around ETS2 (see Fig. 5 for color labels).

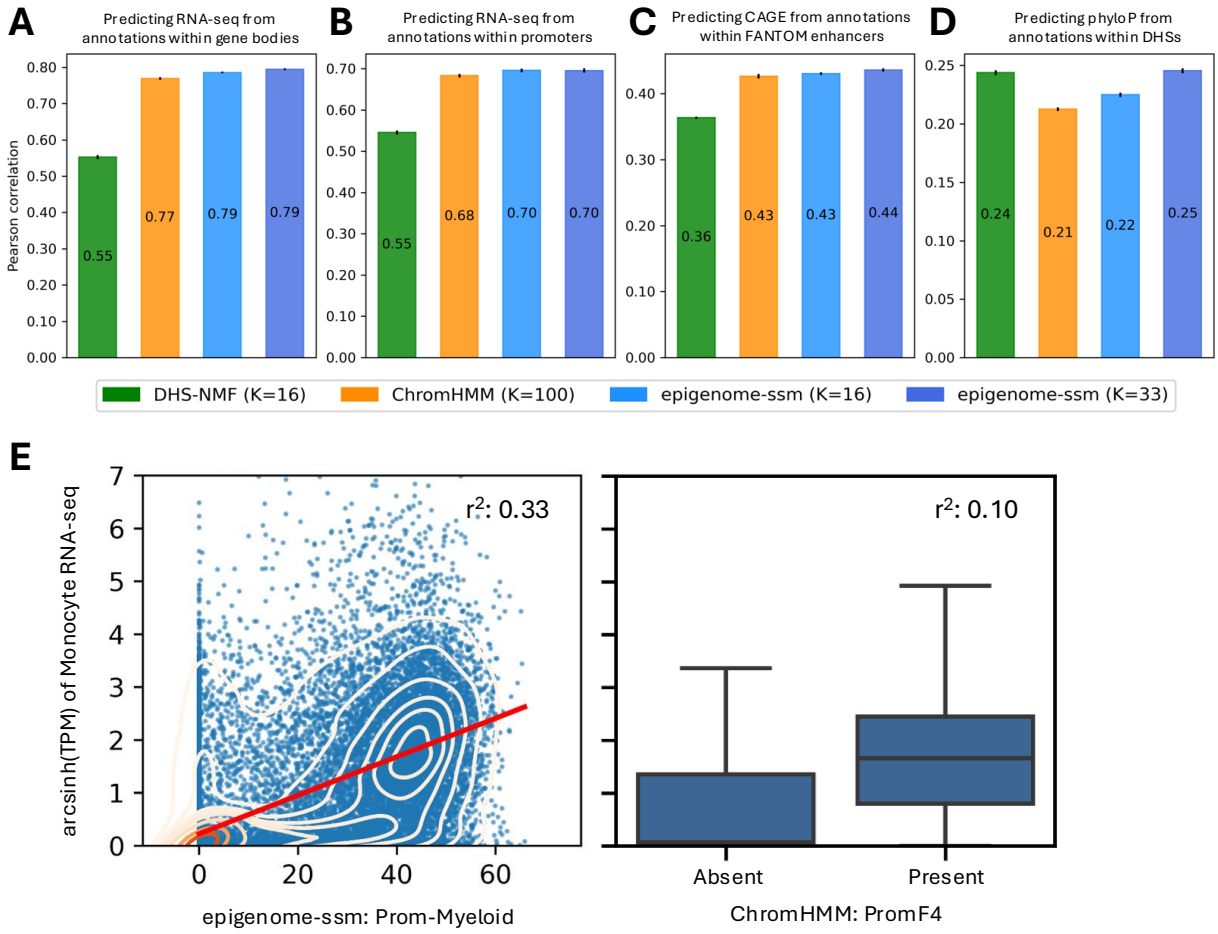

Figure 2: **Pan-cell type features are predictive of genomic activities.** (A-D) Evaluating pan-cell type annotations according to their predictive power of various genomic phenomena. K is the number of components, labels, and features for DHS-NMF, ChromHMM, and epigenome-ssm, respectively. The vertical solid lines at the top of the bars represent the standard error from a 10-fold cross-validation. **A** Using annotations from within gene bodies to predict RNA-seq TPM values across all epigenomes (see section 5.4). **B** Same as **A** but using annotations within a 2 kb window centered at TSS. **C** Using annotations from within FANTOM5 enhancers to predict CAGE TPM values across all samples (see section 5.5). **D** Using annotations from within DHSs to predict evolutionary conservation, as measured by phyloP (see section 5.6). **E** (left) predicting RNA-seq from a single ChromHMM label; (right) predicting RNA-seq from a single epigenome-ssm feature.

Table 1: Feature mnemonics. We assigned a mnemonic to each pan-cell type feature that explains the types of activity and cell types that it captures.

Feature	**Main emitted assays**	**Main emitted biosamples**	**Spatial distribution**
Prom-Brain	H3K4me3	Nervous	Punctate
Prom-Blood1	H3K4me3, H3K27ac	T-cell, Endocrine	Punctate
Prom-Blood2	H3K4me3, H3K4me1	Myeloid	Punctate
Prom-Myeloid	H3K4me3, H3K27ac	Immune system	Punctate
Prom-Znf	H3K9me3, H3K4me3	Immune system	Broad
Enh-Broad	H3K4me1, H3K27me3	Reproductive, Immune system, Respiratory, Integumentary	Broad
Enh-SG1	H3K4me1, H3K27ac	Immune system	Punctate
Enh-Fetal	H3K4me1, H3K27ac	Reproductive, Digestive, Integumentary, Endocrine	Punctate
Enh-Neutrophil	H3K27ac, H3K4me3	Immune system	Broad
Enh-Liver	H3K4me1, H3K27ac, H3K4me3	Digestive	Punctate
TxEnh-SG2	H3K4me1	Nervous	Punctate
TxEnh-SG3	H3K36me3	Immune system, Digestive	Broad
TxEnh-SG4	H3K36me3, H3K27ac	Nervous, Digestive	Punctate
TxEnh-Blood	H3K4me1, H3K27ac	Immune system	Punctate
TxEnh-Myeloid	H3K4me1, H3K27ac	Immune system	Punctate
Tx-Broad1	H3K36me3, H3K4me3, H3K27ac	Nervous, Digestive, Reproductive, Endocrine, Immune system	Punctate
Tx-Broad2	H3K36me3	Immune system, Endocrine, Reproductive, Nervous	Punctate
Tx-Broad3	H3K36me3, H3K4me3	Broad	Broad
Tx-Blood	H3K36me3	Broad	Punctate
Tx-Myeloid1	H3K36me3, H3K27ac	Myeloid	Punctate
Tx-Myeloid2	H3K36me3, H3K27me3	Myeloid	Broad
TxProm-Broad	H3K4me3, H3K36me3	Immune system, Skeletal	Punctate
Flank-Broad	H3K4me3, H3K27ac	Immune system, Reproductive	Broad
Flank-SG5	H3K4me1, H3K27ac, H3K27me3	Immune system, Reproductive	Broad
Flank-Embryonic	H3K4me3, H3K27me3	Broad	Broad
Flank-Brain	H3K4me1, H3K27ac, H3K27me3	Nervous, Reproductive	Broad
Flank-Myeloid	H3K4me1, H3K27ac	Immune system, Skeletal	Broad
Flank-Lymphoid	H3K4me1	Immune system	Broad
FaculHet-Broad1	H3K27me3	Broad	Broad
FaculHet-Broad2	H3K27me3, H3K36me3, H3K9me3	Immune system	Broad
ConstHet-SG6	H3K9me3	Nervous	Broad
ConstHet-Broad1	H3K9me3	Broad	Broad
ConstHet-Broad2	H3K9me3, H3K27me3	Immune system	Broad

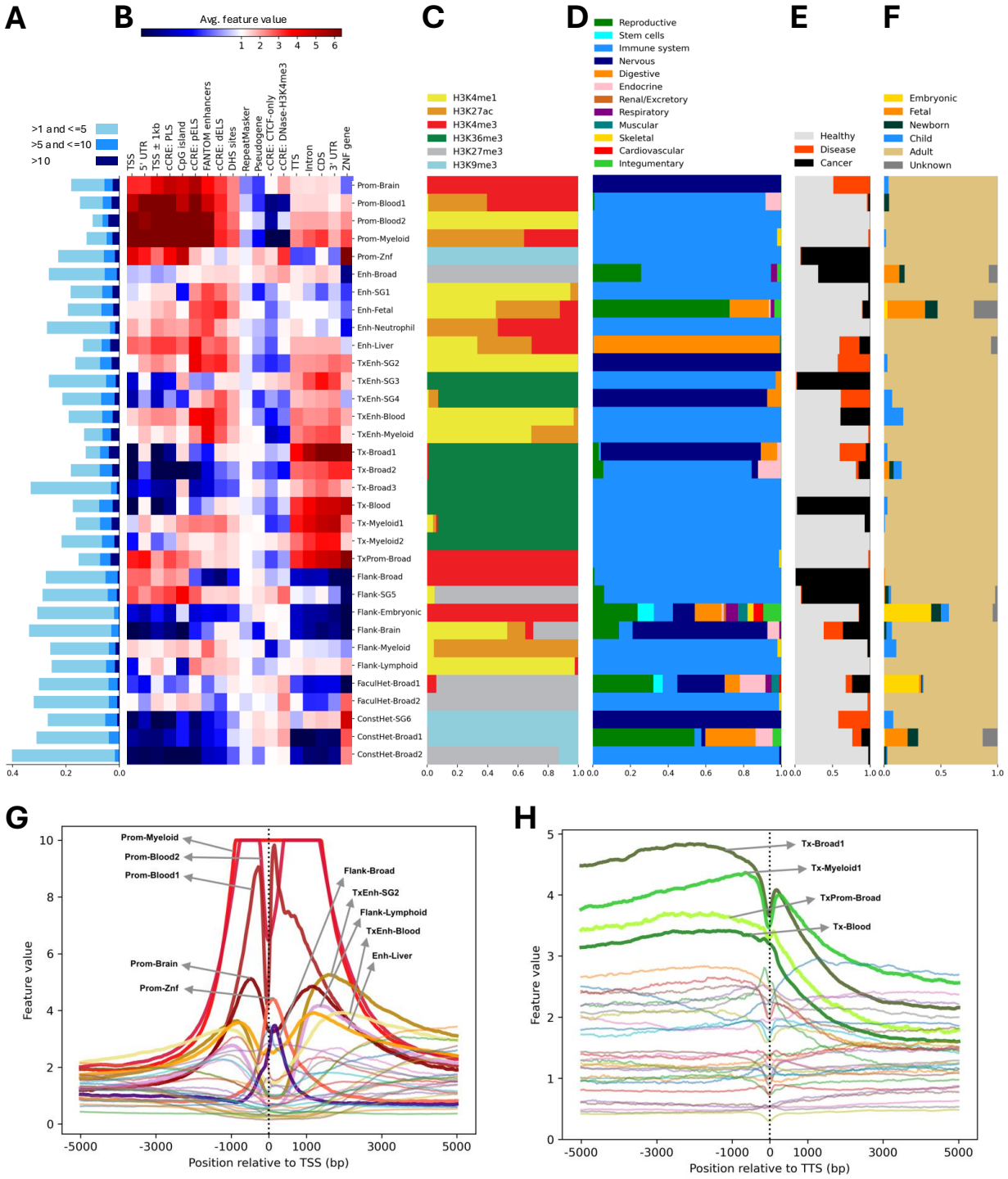

Figure 3: **Pan-cell type features capture different histone modification marks and show distinct patterns around annotated genomic elements.** **A** Distribution of feature values. Horizontal axis indicates the fraction of the genome with features of in a given value range. Features are normalized to a genome-wide average of 1, so a value of 10 indicates 10 times the genome-wide average. **B** Average value of pan-cell type features at existing genome annotations. **C-F** Of the top 1% emission values, the histone modification (C), high-level cell type ontology (D), disease state (E), and donor life stage (F). **G-H** Average value of pan-cell type features around transcription start sites (**G**) and transcription termination sites (**H**) of protein-coding genes.

3.4 Pan-cell type features capture cell type-specific regulatory programs

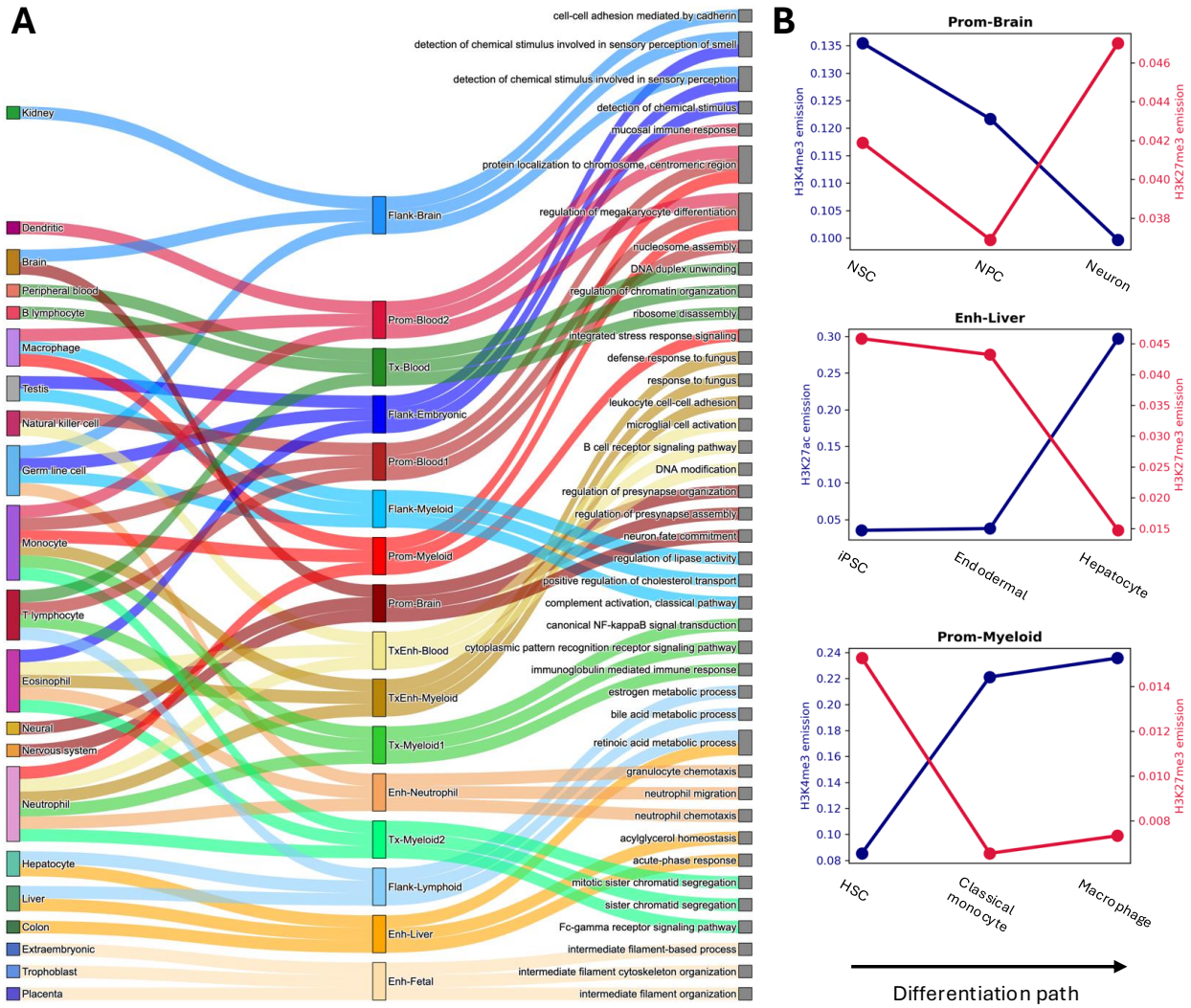

Figure 4: **Pan-cell type features capture gene regulatory programs. A** (left column) intermediate ontology; (middle column) pan-cell type features; (right column) Gene Ontology terms; The links connect each feature to the top three cell types and GO terms. Only those features with a specific cell type/sample group are shown here. **B** Average emission of each feature for active marks vs. the repressive mark, H3K27me3, across several different stem and differentiated cells.

We hypothesize that each feature corresponds to a specific regulatory program in the genome. Because regulation is driven by a small number of regulatory programs, we need only 33 pan-cell type features to capture 9,539 tracks. As discussed, we found a few features correspond to broad activity, such as ConstHet-Broad1 and ConstHet-Broad2, corresponding to constitutive heterochromatin that is repressed in all samples, and Tx-Broad 1 to 3 corresponding to genes marked by H3K36me3 in all cell types which may play housekeeping roles.

However, such broad features are the minority. Most features learned by epigenome-ssm correspond to

activity that is specific to a certain developmental lineage (Fig. 4A). For example, Prom-Brain captures promoter activities in brain-related cell types, while Enh-Liver captures enhancers active in livers (Fig. 3C-D).

As expected, we found that genes that are expressed in a given cell type (according to RNA-seq data not used by epigenome-ssm) tend to have features representing active activity for that cell type (Fig. 4A, see Sec. 5.8). We found that for example, the Prom-Brain feature occurs at genes expressed in Brain, Neural, and Nervous system, while Prom-Myeloid occurs at genes expressed in Macrophage, Monocyte, and Neutrophil. This is consistent with the emissions for these two features, where the top emitted biosamples for Prom-Brain and Prom-Myeloid are Nervous and Immune system, respectively.

We found that regulatory programs typically involve loci switching from active to repressed—or vice versa—throughout development. We found several cases where a feature captures the cell differentiation trajectory of a particular cell type group, starting with stem cells and ending with some mature differentiated cells (Fig. 4B). For instance, Prom-Brain emission for the active mark H3K4me3 decreases when moving from neural stem cells (NSCs) to neural progenitor cells (NPCs) and mature neurons, while the emission of this feature for the repressive mark H4K27me3 is at its highest at the end of the differentiation path for mature neurons. Similarly, Enh-Liver captures the differentiation path of iPSCs to endodermal and Hepatocyte cells, while Prom-Myeloid captures the differentiation path of HSCs to monocytes and macrophages.

Furthermore, by comparing features to Gene Ontology (GO) terms, we found that regulatory programs align with cellular functions (Fig. 4A). GO terms provide insights into the downstream functions of the genes in a hierarchical manner, starting with more general terms that correspond to functions carried out by most genes to more specific terms corresponding to functions that are carried out by a small subset of the genes. We assigned GO terms to each feature according to the cellular functions of genes this feature is associated with (Methods section 5.9). The GO terms assigned to each feature are highly relevant in the cell type/sample group of that feature. For instance, the top three GO terms for Prom-Brain are "neuron fate commitment", "regulation of presynapse assembly", and "regulation of presynapse organization" (Fig. 4A).

These results provide evidence that patterns of regulatory activity correspond to patterns of gene regulation and cellular function.

3.5 Conservation-associated activity score (SSM-CAAS) captures activity relevant to phenotype and disease

The pan-cell type features generated by epigenome-ssm summarize activity across all cell types and tissues, and thus can identify activity important to organism-level phenotype and disease. It is established through genome-wide association studies (GWAS) that the majority of the disease-associated variants are in the non-coding regions of the genome, likely regions that contain some regulatory elements like enhancers and promoters; integrative annotations like epigenome-ssm are needed to identify these elements.

To enable this identification, we define SSM-CAAS, the SSM-based conservation-associated activity score

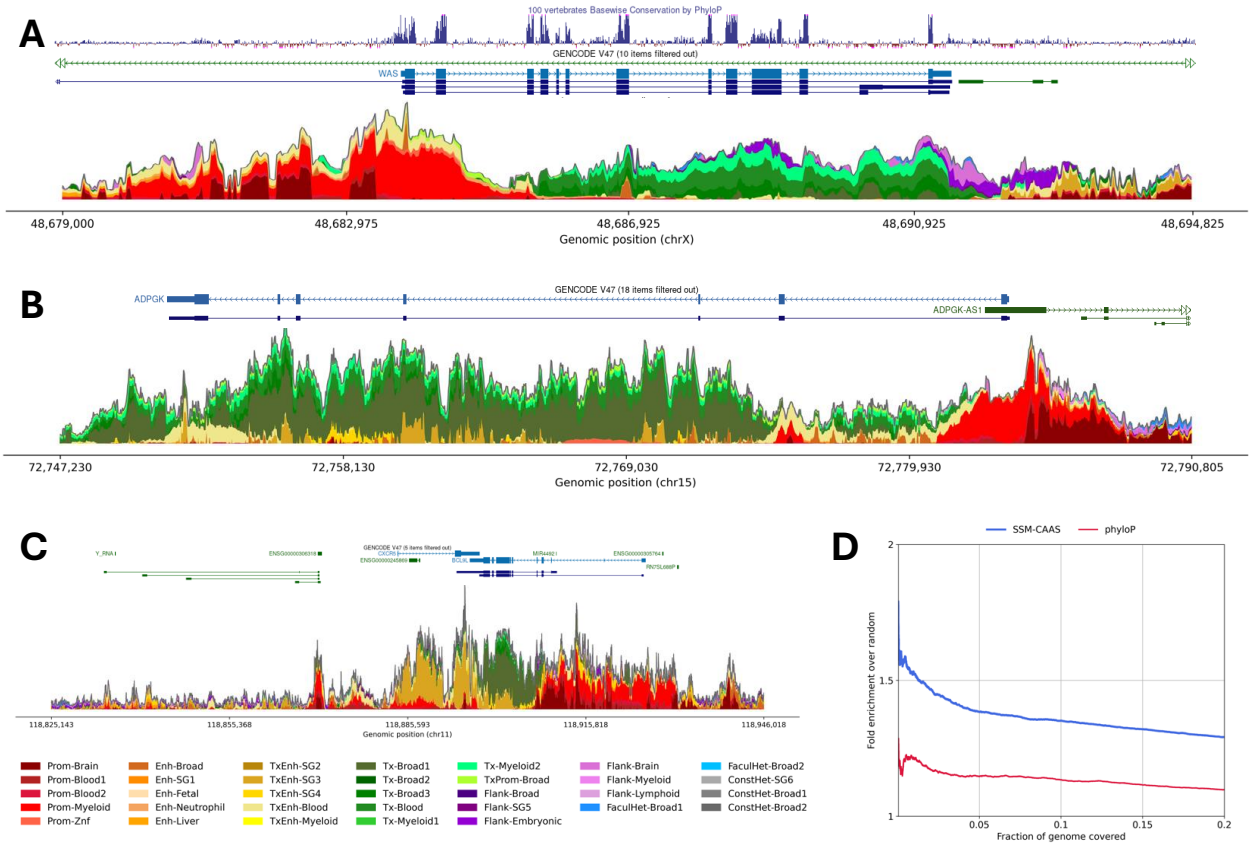

Figure 5: **SSM-CAAS plot summarizes epigenomic signals at a given locus. A** CAAS plot (see Fig. 1D) for the WAS locus. **B** CAAS plot for the ADPGK locus. **C** CAAS plot for the CXCR5 and BCL9L loci. **D** Enrichment of SSM-CAAS and phyloP scores for GWAS variants.

(CAAS), based on the previously-defined SAGA-based CAAS [32]. Briefly, we used our pan-cell type features as input to a non-negative linear regression model that predicts evolutionary conservation (measured by phyloP [33]) and defined SSM-CAAS as the predicted phyloP score by this model. Evolutionary conservation is widely used as a proxy for putatively functional regions of the genome; thus SSM-CAAS identifies positions with regulatory activity consistent with putative function.

As expected, the contribution to SSM-CAAS differs between features (Fig. S7). Features representing transcription have a high contribution to SSM-CAAS, as coding exons are generally the most conserved positions in the genome. Regulatory activity has SSM-CAAS that is comparable to transcription because, although these regions are typically less conserved than exons, epigenomic marks do not usually distinguish exons from less-conserved introns. As expected, features representing repressive or extremely cell type-specific activity have the lowest SSM-CAAS contribution.

The conservation-associated activity score offers two advantages over conservation as a measure of the importance of a given locus. First, the conservation-associated activity score is directly attributable to a specific activity in a specific set of cell types; in contrast, conservation indicates only that a position is

important, with no way to determine how it acts. SSM-CAAS has the potential to mark elements that are not conserved but have epigenetic activity similar to other conserved elements, such as those that gained regulatory or transcriptional activity only in humans. Conversely, SSM-CAAS can identify elements conserved across other species but decommissioned in humans.

Furthermore, SSM-CAAS enables a visualization of genomic activity that emphasizes putative function. This visualization consists of a stacked ribbon plot in which the total height of the plot is the predicted conservation of that locus, and the colors of the ribbon constitute the contribution of each activity type to the prediction (Fig. 5). We found that positions with high SSM-CAAS are more enriched for genetic variants identified by genome-wide association studies (GWAS) to be involved in disease (Fig. 5D), compared to phyloP. This indicates that CAAS captures regions of the genome that are not only evolutionarily conserved but also functionally relevant to human disease, providing a valuable tool for identifying disease-associated regulatory elements.

Several examples show how epigenome-ssm features can be used to quickly understand a genomic locus. By diving deep into several different genomic loci, we found that SSM-CAAS not only captures the main genomic elements (e.g., promoters, transcribed genes, etc.), but it also captures cell type-specific activities (Fig. 5A-C). The WAS gene located on the forward strand (Fig. 5A) has a conserved promoter region, as indicated by the phyloP track, and has a high SSM-CAAS signal that is mainly derived from promoter features (in reds). The dominant promoter feature for WAS is Prom-Myeloid, and that is expected given the fact that WAS is highly expressed in the immune cells [34]. The transcription-associated features (in greens) cover the gene body while there are enhancer features (in yellows) upstream of the promoter, within the introns, and downstream of the gene.

The ADPGK gene is located on the reverse strand (Fig. 5B). Similar to WAS, the promoter features cover the region around the transcription state site while the Tx features cover the gene body. This gene also has some enhancers as indicated by the Enh and TxEnh features (yellows). The CXCR5 gene located on the forward strand is surrounded by a few other genes (Fig. 5C). The promoter features are enriched around the promoter region of CXCR5, and also the promoter regions of both the upstream and downstream genes. While the enhancer features are very high around CXCR5 (likely due to the proximal enhancers regulating the downstream genes like BCL9L), the TxBroad features are enriched for the transcribed regions of the downstream gene, BCL9L.

These examples show how visualizing epigenome-ssm features can allow a user to quickly understand the activity of a given locus across all 1,698 biosamples, both in terms of types of activity and tissue specificity.

4 Discussion

In this study, we introduced a continuous pan-cell type chromatin state annotation and demonstrated its effectiveness in summarizing the vast epigenomic data from IHEC. Our results showed that these pan-cell

type features are capable of capturing biologically meaningful patterns in a variety of genomic contexts. The ability of epigenome-ssm to provide a more nuanced representation of chromatin state activity across multiple cell types has significant implications for understanding gene regulation and disease association.

One key advantage of continuous annotations over discrete ones is their ability to capture varying levels of chromatin activity at each genomic position, allowing for a more accurate representation of the regulatory landscape. As demonstrated in our comparisons, even with fewer features (K=16), epigenome-ssm matched or exceeded the performance of ChromHMM, which used 100 discrete labels. This suggests that the continuous nature of our features allows for more efficient encoding of regulatory information, potentially reducing redundancy while still capturing the complexity of the underlying epigenetic signals.

Furthermore, the cell type-specific patterns identified by epigenome-ssm emphasize its utility in studying tissue-specific regulatory programs. Features such as Prom-Brain, Enh-Liver, and Prom-Myeloid demonstrated strong associations with their respective cell type groups, suggesting that these features effectively capture regulatory programs that drive tissue-specific gene expression. This ability to distinguish between cell type-specific activities is crucial for applications in disease association studies, as it allows researchers to pinpoint regulatory elements that may be disrupted in specific tissues or cell types.

Finally, our CAAS analysis provides further evidence of the utility of pan-cell type features in functional genomics. The enrichment of disease-associated variants within regions with high CAAS values suggests that our method can prioritize genomic regions of interest for further investigation, particularly in the context of non-coding disease-associated variants identified by GWAS. This is particularly important given the increasing recognition that many disease-associated variants lie outside of coding regions, in regulatory elements that control gene expression.

We expect that annotations from epigenome-ssm will provide an easily-interpretable summary of epigenetic activity across hundres of cell types and tissues, and thus it will be the first stop for researchers aiming to understand the epigenome.

5 Methods

5.1 Input data and processing

All the analyses are done using the human reference genome hg38. From the IHEC repository, we downloaded 9,539 bigWig tracks containing $-\log_{10}$ *p-value* signals from six core histone modification ChIP-seq assays. Those tracks contain imputed ChIP-seq signals and were generated using ChromImpute [35] at 25 bp resolution from 5,339 observed ChIP-seq tracks in the IHEC repository. To reduce the effect of outliers, we transformed the signals using the inverse hyperbolic sine function defined as $arcsinh(x) = \log(x + \sqrt{x^2 + 1})$. We trained epigenome-ssm on the ENCODE Pilot regions [36], which consist of 1% of the human genome, and used the UCSC liftOver tool [37] to convert the ENCODE Pilot coordinates from hg19 to hg38. We also removed the ENCODE exclusion list regions [38] from all the analyses. We then used the trained

epigenome-ssm model to generate annotations for the whole genome.

5.2 epigenome-ssm model, training procedure, and annotation pipeline

The epigenome-ssm model is based on the state-space model (SSM) and is an unsupervised linear sequential model that takes as input E observed epigenomic signal values for each position g of the genome and generates K non-negative chromatin state feature values that summarize epigenomic activity at position g. More specifically, this model has two main assumptions: (1) At each position g, there is a latent vector $\alpha_g \in \mathbb{R}^K$ that encodes the chromatin state features at g and the observed data $y_g \in \mathbb{R}^E$ is generated as a linear function of α_g, parameterized by an emission matrix $Z \in \mathbb{R}^{E \times K}$; and (2) α_{g+1} is generated as a linear function of α_g, parameterized by a transition matrix $T \in \mathbb{R}^{K \times K}$. To limit the model's capacity to overfit and its sensitivity to local optima, we additionally added two L_2 regularization terms to the optimization's objective function $J(Z, T)$, which encourage Z and T to have small values: $J(Z, T) = \log P(\alpha, Z, T|Y) + \lambda_1 \|Z\|_F + \lambda_2 \|T\|_F$. Additionally, we used an active set method of Lagrange multipliers to enforce non-negativity constraints on both chromatin state features and the emission parameters (see [30] for more details).

To train the model, we used an expectation maximization (EM) algorithm to maximize the model's log-likelihood as a function of its parameters, Z and T. We generated ten random initializations of the model parameters and took the one with the best log-likelihood. We did not use a fixed number of iterations—instead, we kept the training going until there was no significant improvement in log-likelihood.

As a normalization step at the end of the pipeline, we normalized each feature i such that its genome-wide mean is 1, and correspondingly multiplied its emission values $Z_{:,i}$ by its unnormalized genome-wide mean. This step does leaves the predicted observed data unchanged, but easier to interpret the features because they are all on the same scale.

5.3 Alternative pan-cell type annotations

In our quantitative evaluations, we included pan-cell type annotations generated by two other methods: DHS-NMF [28] and Universal ChromHMM [29].

DHS-NMF annotations were generated using DNase-seq data across 733 human biosamples. The DNase-seq assay measures chromatin accessibility at DNase I hypersensitive sites (DHSs). More specifically, the authors curated a set of approximately 3.6 million DHSs across all the samples and then ran non-negative matrix factorization with 16 components on a DHS-by-biosample binary matrix. Therefore, DHS-NMF annotations are only available at DHSs and not the whole genome. We used both the DHS coordinates [39] and DHS-NMF annotations [40] in our analyses.

From the IHEC repository, we downloaded Universal ChromHMM annotations generated by Ernst Lab at UCLA by running ChromHMM on six core histone modification ChIP-seq data across all IHEC epigenomes. This set of annotations is generated using a model with 100 labels. Note that these annotations are similar,

but not identical, to annotations presented in [29] that were generated using a smaller data set taken from only Roadmap Epigenomics.

We also downloaded the CSREP annotations generated by Ernst lab from the IHEC repository. Specifically, we downloaded 56 BED files corresponding to the summarized annotations of 56 sample groups in the IHEC dataset. The annotations were initially generated using a ChromHMM model with 18 states.

5.4 Gene expression prediction

We evaluated the pan-cell type annotations based on their association with gene expression, as measured by bulk RNA-seq. We downloaded RNA-seq TPM (transcripts per kilobase million) data for 1,467 epigenomes from the IHEC repository. We then compared epigenome-ssm, Universal ChromHMM, and DHS-NMF by using annotations within the gene body (i.e., the region between the transcription start site and transcription termination site) as predictors in a linear regression model to predict RNA-seq TPM values.

More specifically, for a given epigenome-ssm model, we constructed a vector of K values from a given gene's body by averaging each feature in that region, where K is the number of features for each model. For ChromHMM, we formed a one-hot encoded vector of length 100 corresponding to the 100 labels in the Universal ChromHMM annotation. Position i of this vector has a value of 1 if label i is present in the region, and 0 otherwise. To account for the length of each label in the region, we multiplied each value of the vector by the proportion of the region covered by the corresponding label. Similarly, we formed a one-hot encoded vector of length 18 for CSREP. For DHS-NMF, we formed a vector of length 16 corresponding to the 16 NMF components by averaging the value of each component across DHSs within each gene's body. As the regression response value, we formed a vector of length 1,467 for each gene, corresponding to its RNA-seq TPM values across all epigenomes. We *arcsinh* transformed both predictor and response values (section 5.2).

We performed a similar analysis with all the same steps as above, except that we used annotations within the promoter region instead of the gene body. Specifically, for each gene, we centered a 2 kb window at the annotated TSS and extracted the annotations from that window.

We generated linear regression predictions in a 10-fold cross-validation setting and used Pearson correlation to measure prediction performance. We limited this analysis to protein-coding genes and used GENCODE v45 [41] gene annotations. Also, we used the *LinearRegression* model from scikit-learn [42] for this analysis, as well as for the prediction tasks in sections 5.5 and 5.6.

5.5 Enhancer activity prediction

We evaluated the pan-cell type annotations based on their association with enhancer activity, as measured by Cap Analysis of Gene Expression (CAGE). We downloaded CAGE TPM data for 1,827 samples [43] and enhancer annotations [44] from FANTOM5 and used the UCSC liftOver tool [37] to convert the coordinates from hg19 to hg38. As the regression predictor value, we extracted annotations that overlap enhancers for

each model, as described in section 5.4. As the regression response value, we formed a vector of length 1,827 for each enhancer, corresponding to its CAGE TPM values across all samples. We *arcsinh* transformed both predictor and response values (section 5.2).

5.6 Evolutionary conservation prediction

We evaluated the pan-cell type annotations based on their association with evolutionary conservation, as measured by phyloP [33]. We limited this analysis to the DHSs since DHS-NMF annotations are only available for those regions. We downloaded a phyloP track from the UCSC Genome Browser that is generated by multiple alignments of 99 vertebrate genomes to the human genome [45]. As the regression predictor value, we extracted annotations that overlap DHSs for each model, as described in section 5.4. As the regression response value, we used the absolute value of the mean phyloP score for each DHS. We *arcsinh* transformed both predictor and response values (section 5.2).

5.7 Enrichment of pan-cell type features for annotated genomic elements

We computed the enrichment of a pan-cell type feature for a given set of genomic regions as the mean feature value at those regions divided by the feature's genome-wide mean. The following annotations were used:

- We used GENCODE v45 [41] for all gene annotations and only included protein-coding genes. We took the start position of the genes as TSS and the end position of the genes as TTS. "TSS ± 1kb" includes 2 kb regions centered at TSSs. Coding sequence (CDS) regions correspond to the GENCODE elements whose feature type is "CDS". The 5' and 3' UTRs (untranslated region) correspond to the GENCODE elements whose feature type is "five_prime_UTR" "three_prime_UTR", respectively. Similarly, exons correspond to the GENCODE elements whose feature type is "exon". For introns, we took the elements whose feature type is "transcript" and removed the parts that overlapped with exons using the *bedtools subtract* command [46]. For zinc finger genes, we took those elements whose feature type is "gene" and their "gene_name" contains "ZNF". For pseudogenes, we took the elements with "gene" feature type and "pseudogene" in their "gene_type" and included both protein- and non-protein-coding genes.

- We downloaded the ENCODE candidate cis-regulatory elements from the ENCODE's SCREEN portal [47] and extracted the coordinates for the PLS (TSS-overlapping with promoter-like signatures), pELS (TSS-proximal with enhancer-like signatures), dELS (TSS-distal with enhancer-like signatures), DNase-H3K4me3 (not TSS-overlapping and with high DNase and H3K4me3 signals only), and CTCF-only (not TSS-overlapping and with high DNase and CTCF signals only) [48].

- The annotations for CpG islands [49] and RepeatMasker regions [50] were downloaded from the UCSC Genome Browser.

- Enhancer annotations were downloaded from FANTOM5 [44] (see section 5.5).

- We used the DHS coordinates [39] from [28] (see section 5.3).

5.8 Associating features with cell types through gene expression

We associated the pan-cell type features with cell type groups by looking at the feature values at protein-coding genes and the expression of those genes in different cell types. More specifically, we generated two matrices: (1) A feature-by-gene matrix that contains the average value of each feature at protein-coding genes;(2) A cell type-by-gene matrix that contains the RNA-seq expression value of each gene in different cell types. We generated the product of those two matrices over the gene axis to get a feature-by-cell type matrix. We downloaded RNA-seq data for 1,467 epigenomes from the IHEC repository for this analysis.

5.9 Associating features with Gene Ontology terms

We associated the pan-cell type features with gene ontology (GO) [51, 52] terms by looking at the feature values at protein-coding genes and the GO terms assigned to each gene in the GO database. More specifically, we generated two matrices: (1) A feature-by-gene matrix that contains the average value of each feature at protein-coding genes; (2) A GO term-by-gene binary matrix that connects each gene to the GO terms assigned to it. We generated the product of those two matrices over the gene axis to get a feature-by-GO term matrix. Since GO terms are assigned to each gene in a hierarchical manner (i.e., each gene may receive several GO terms starting from the most general term to more specific terms), we included all the terms assigned to each gene in our analysis except those terms that were assigned to less than 30 genes. For this analysis, we downloaded GO data files from [53, 54].

5.10 Conservation-associated activity score

We generated a genome-wide track of conservation-associated activity score using both the pan-cell type features as a summary of the human epigenomic data across many cell types, and evolutionary conservation as measured by phyloP [33]. We used the same phyloP track as in section 5.6 and trained a non-negative linear regression model that takes as input the pan-cell type features and predicts the absolute phyloP score for each genomic bin. We performed a 5-fold cross validation and used the average coefficients as the final model to generate a predicted phyloP track. We defined CAAS to be the predicted phyloP scores from this model.

5.11 Enrichment for GWAS variants

We evaluated the conservation-associated activity score generated from our pan-cell type features based on the enrichment for disease-associated variants. We downloaded a set of disease-associated variants from GWAS Catalog [55]. We sorted the genomic bins based on their SSM-CAAS (and similarly based on their

phyloP score), found the number of GWAS variants covered by each bin, and plotted the fold enrichment of SSM-CAAS and phyloP over the genome-wide average.

6 Code and data availability

Here is a link to a UCSC genome browser session containing the pan-cell type feature tracks along with the SSM-CAAS track: https://genome.ucsc.edu/s/habib/epigenome%2Dssm.IHEC.compact

Also, the latest code for the epigenome-ssm pipeline can be found here: https://github.com/habibdanesh/epigenome-ssm

7 Acknowledgements

We would like to thank the Ernst Lab at UCLA for generating ChromImpute imputations. We would also like to thank Guillaume Bourque, Stephan Beck, and other members of the IHEC community for their constructive feedbacks.

8 Funding

This work was funded by NSERC (RGPIN/06150-2018), Health Research BC (SCH-2021-1734), Compute Canada (kdd-445), and SFU Computing Science (N000265).

References

[1] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, 2012.

[2] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, 2012.

[3] Maxwell W Libbrecht, Rachel CW Chan, and Michael M Hoffman. Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. *PLoS Computational Biology*, 17(10):e1009423, 2021.

[4] Nathan Day, Andrew Hemmaplardh, Robert E Thurman, John A Stamatoyannopoulos, and William S Noble. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426, 2007.

[5] Guillaume J Filion, Joke G van Bemmel, Ulrich Braunschweig, Wendy Talhout, Jop Kind, Lucas D Ward, Wim Brugman, Inês J de Castro, Ron M Kerkhoven, Harmen J Bussemaker, et al. Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, 143(2):212–224, 2010.

[6] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–825, 2010.

[7] Michael M Hoffman, Jason Ernst, Steven P Wilder, Anshul Kundaje, Robert S Harris, Max Libbrecht, Belinda Giardine, Paul M Ellenbogen, Jeffrey A Bilmes, Ewan Birney, et al. Integrative annotation of chromatin elements from encode data. *Nucleic acids research*, 41(2):827–841, 2013.

[8] Jacob Biesinger, Yuanfeng Wang, and Xiaohui Xie. Discovering and mapping chromatin states using a tree hidden markov model. In *BMC bioinformatics*, volume 14, pages 1–11. Springer, 2013.

[9] Jimin Song and Kevin C Chen. Spectacle: fast chromatin state annotation using spectral learning. *Genome biology*, 16:1–18, 2015.

[10] Kyung-Ah Sohn, Joshua WK Ho, Djordje Djordjevic, Hyun-hwan Jeong, Peter J Park, and Ju Han Kim. hihmm: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*, 31 (13):2066–2074, 2015.

[11] Alessandro Mammana and Ho-Ryun Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome biology*, 16:1–12, 2015.

[12] Maxwell Libbrecht, Michael Hoffman, Jeff Bilmes, and William Noble. Entropic graph-based posterior regularization. In *International Conference on Machine Learning*, pages 1992–2001. PMLR, 2015.

[13] Yu Zhang, Lin An, Feng Yue, and Ross C Hardison. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic acids research*, 44(14):6721–6731, 2016.

[14] Guanjue Xiang, Cheryl A Keller, Elisabeth Heuston, Belinda M Giardine, Lin An, Alexander Q Wixom, Amber Miller, April Cockburn, Michael EG Sauria, Kathryn Weaver, et al. An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis. *Genome research*, 30(3):472–484, 2020.

[15] Benedikt Zacher, Margaux Michel, Björn Schwalb, Patrick Cramer, Achim Tresch, and Julien Gagneur. Accurate promoter and enhancer identification in 127 encode and roadmap epigenomics cell types and tissues by genostan. *PLoS ONE*, 12(1):e0169249, 2017.

[16] Eugenio Marco, Wouter Meuleman, Jialiang Huang, Kimberly Glass, Luca Pinello, Jianrong Wang, Manolis Kellis, and Guo-Cheng Yuan. Multi-scale chromatin state annotation using a hierarchical hidden markov model. *Nature communications*, 8(1):15011, 2017.

[17] Senthil B Girimurugan, Yuhang Liu, Pei-Yau Lung, Daniel L Vera, Jonathan H Dennis, Hank W Bass, and Jinfeng Zhang. iseg: an efficient algorithm for segmentation of genomic and epigenomic data. *BMC bioinformatics*, 19:1–15, 2018.

[18] Simon G Coetzee, Zachary Ramjan, Huy Q Dinh, Benjamin P Berman, and Dennis J Hazelett. StateHub-StatePaintR: rapid and reproducible chromatin state evaluation for custom genome annotation. *BioRxiv*, page 127720, 2017.

[19] Adriana Arneson and Jason Ernst. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Communications biology*, 2(1):248, 2019.

[20] Axel Poulet, Ben Li, Tristan Dubos, Juan Carlos Rivera-Mulia, David M Gilbert, and Zhaohui S Qin. Rt states: systematic annotation of the human genome using cell type-specific replication timing programs. *Bioinformatics*, 35(13):2167–2176, 2019.

[21] Mickaël Mendez, FANTOM Consortium Main Contributors, Michelle S Scott, and Michael M Hoffman. Unsupervised analysis of multi-experiment transcriptomic patterns with SegRNA identifies unannotated transcripts. *bioRxiv*, pages 2020–07, 2020.

[22] Philipp Benner and Martin Vingron. ModHMM: a modular supra-Bayesian genome segmentation method. *Journal of Computational Biology*, 27(4):442–457, 2020.

[23] SPIN reveals genome-wide landscape of nuclear compartmentalization, author=Wang, Yuchuan and Zhang, Yang and Zhang, Ruochi and van Schaik, Tom and Zhang, Liguo and Sasaki, Takayo and Peric-Hupkes, Daniel and Chen, Yu and Gilbert, David M and van Steensel, Bas and others. *Genome biology*, 22:1–23, 2021.

[24] Guanjue Xiang, Belinda M Giardine, Shaun Mahony, Yu Zhang, and Ross C Hardison. S3V2-IDEAS: a package for normalizing, denoising and integrating epigenomic datasets across different cell types. *Bioinformatics*, 37(18):3011–3013, 2021.

[25] Artur Jaroszewicz and Jason Ernst. ChromGene: gene-based modeling of epigenomic data. *Genome Biology*, 24(1):203, 2023.

[26] Johanna Elena Schmitz, Nihit Aggarwal, Lukas Laufer, Jörn Walter, Abdulrahman Salhab, and Sven Rahmann. EpiSegMix: a flexible distribution hidden Markov model with duration modeling for chromatin state discovery. *Bioinformatics*, 40(4):btae178, 2024.

[27] Marjan Farahbod, Abdul Rahman Diab, Paul Sud, Meenakshi Kagda, Ian Whaling, Mehdi Foroozandeh, Ishan Goel, Habib Daneshpajouh, Benjamin Hitz, J Michael Cherry, et al. Integrative chromatin state annotation of 234 human ENCODE4 cell types using Segway reveals disease drivers. *bioRxiv*, pages 2023–10, 2023.

[28] Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, et al. Index and biological spectrum of human dnase i hypersensitive sites. *Nature*, 584(7820):244–251, 2020.

[29] Ha Vu and Jason Ernst. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biology*, 23:1–37, 2022.

[30] Habib Daneshpajouh, Bowen Chen, Neda Shokraneh, Shohre Masoumi, Kay C Wiese, and Maxwell W Libbrecht. Continuous chromatin state feature annotation of the human epigenome. *Bioinformatics*, 38 (11):3029–3036, 2022.

[31] Ha Vu, Zane Koch, Petko Fiziev, and Jason Ernst. A framework for group-wise summarization and comparison of chromatin state annotations. *Bioinformatics*, 39(1):btac722, 2023.

[32] Maxwell W Libbrecht, Oscar L Rodriguez, Zhiping Weng, Jeffrey A Bilmes, Michael M Hoffman, and William Stafford Noble. A unified encyclopedia of human functional dna elements through fully automated annotation of 164 human cell types. *Genome biology*, 20:1–14, 2019.

[33] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

[34] Adrian J Thrasher and Siobhan Burns. Wiskott-aldrich syndrome: a disorder of haematopoietic cytoskeletal regulation. *Microscopy research and technique*, 47(2):107–113, 1999.

[35] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–376, 2015.

[36] ENCODE. Encode pilot regions. http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/referenceSequences/encodePilotRegions.hg19.bed, . Accessed: 2022-10-20.

[37] Angela S Hinrichs, Donna Karolchik, Robert Baertsch, Galt P Barber, Gill Bejerano, Hiram Clawson, Mark Diekhans, Terrence S Furey, Rachel A Harte, Fan Hsu, et al. The ucsc genome browser database: update 2006. *Nucleic acids research*, 34(suppl_1):D590–D598, 2006.

[38] ENCODE. Encode exclusion list regions. https://www.encodeproject.org/files/ENCFF356LFX, . Accessed: 2022-10-20.

[39] Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, et al. Dhs-nmf loci. https://www.encodeproject.org/files/ENCFF503GCK, . Accessed: 2024-02-20.

[40] Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, et al. Dhs-nmf components. https://zenodo.org/records/3838751/files/2018-06-08NC16_NNDSVD_Mixture.npy.gz, . Accessed: 2024-02-20.

[41] GENCODE. Gencode human release 45. https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_45/gencode.v45.basic.annotation.gff3.gz. Accessed: 2024-02-20.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

[43] FANTOM5. Fantom5 human enhancers tpm matrix. https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/human_permissive_enhancers_phase_1_and_2_expression_tpm_matrix.txt.gz, . Accessed: 2024-02-20.

[44] FANTOM5. Fantom5 human enhancers annotation. https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz, . Accessed: 2024-02-20.

[45] UCSC Genome Borwser. 100 vertebrates phylop. https://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP100way/hg38.phyloP100way.bw, . Accessed: 2024-02-20.

[46] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[47] ENCODE. Encode candidate cis-regulatory elements. https://downloads.wenglab.org/V3/GRCh38-cCREs.bed, . Accessed: 2024-02-20.

[48] Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[49] UCSC Genome Borwser. Cpg islands annotation. https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/cpgIslandExt.txt.gz, . Accessed: 2024-02-20.

[50] UCSC Genome Borwser. Repeatmasker annotation. https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/nestedRepeats.txt.gz, . Accessed: 2024-02-20.

[51] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[52] Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

[53] Gene Ontology Consortium. Gene to go mapping. https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz, . Accessed: 2024-07-20.

[54] Gene Ontology Consortium. Go annotation. https://purl.obolibrary.org/obo/go/go-basic.obo, . Accessed: 2024-07-20.

[55] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The nhgri-ebi gwas catalog: knowledgebase and deposition resource. *Nucleic acids research*, 51(D1):D977–D985, 2023.