UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Approaches to Expand the Applications of Chromatin State Annotations

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Bioinformatics

by

Thai Ha Vu

2023

ABSTRACT OF THE DISSERTATION

Computational Approaches to Expand the Applications

of Chromatin State Annotations

by

Thai Ha Vu

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2023

Professor Jason Ernst, Chair

Genome-wide mappings of chromatin marks such as histone modifications and open chromatin sites provide valuable information for annotating the non-coding genome. Computational approaches such as ChromHMM have been applied to discover the combinatorial and spatial patterns of chromatin marks in a biosample, characterize them as chromatin states, and subsequently annotate the biosample's epigenome into chromatin states. As more biosamples' chromatin marks data are generated, it becomes more challenging to manually study biological similarities and differences in the chromatin state maps across many biosamples. We therefore have developed methods to derive epigenome annotations that incorporate data from multiple biosamples and highlight notable epigenetic properties.

First, we introduced a large-scale application of ChromHMM that generates a *universal* chromatin state map for the human genome that can be shared across cell types. In particular, we trained ChromHMM with input data from >1,000 experiments in >100 human biosamples from

Roadmap Epigenomics and ENCODE projects. We denoted the resulting chromatin state map the 'full-stack' annotation. We conducted comprehensive analyses to characterize the full-stack states' biological interpretations, and uncovered patterns of cell-type-specific and constitutive regulatory activities in each state. The full-stack annotation, along with detailed state characterizations, are useful for researchers in understanding the epigenetic contexts of genomic loci of interests.

Building on this work, we developed and analyzed an analogous universal chromatin state annotation for the mouse genome. We trained such an annotation using input data from >900 ChIP-seq/ATAC-seq or DNase-seq experiments from the Mouse ENCODE and ENCODE projects, characterized the resulting states and related them with those from the human full-stack model. Given the wide applications of mice as a model organism to study human disease mechanisms, the mouse full-stack annotation is expected to be highly useful for researchers to investigate the mouse epigenetic landscapes.

Lastly, we developed a method named CSREP to derive a genome-wide probabilistic summary chromatin state map given data from a group of biosamples with common biological properties. We validated CSREP's output summary chromatin state maps for groups of samples with shared tissue types from the Roadmap Epigenomics and EpiMap projects, and showed that CSREP can better predict genomic locations of individual chromatin states in held-out biosamples. We further showed an extension of CSREP where the summary chromatin state maps for two groups of samples are used to prioritize differential chromatin state changes between the two groups.

Overall, our work aims to derive genome-wide chromatin state annotations that can aggregate and derive the patterns of epigenetic assays within and across different cell identities. All methods we present can be widely applicable to newer and larger datasets that will be made available in the future, while the data of chromatin state annotations we provide can be useful to

the larger community in understanding the regulatory patterns across the genome of human and mouse organisms.

The dissertation of Thai Ha Vu is approved.

Jingyi Li

Matteo Pellegrini

Sriram Sankararaman

Jason Ernst, Committee Chair

University of California, Los Angeles

2023

DEDICATION

This dissertation is dedicated to my parents.

TABLE OF CONTENTS

LIST OF MAIN FIGURES AND TABLES

Chapter 2

Chapter 3

Chapter 4

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my advisor Dr. Jason Ernst. He has been an extremely reliable, patient and committed advisor. His non-compromising rigor and thoroughness in scientific research is a lesson that I will carry for the rest of my research career. He also gave me freedom to explore ideas outside of the main research trajectories. He is also very humble and kind and he treats all of us students with a great deal of respect.

I would like to thank the rest of my committee, Drs. Jessica Li, Matteo Pellegrini, and Sriram Sankararaman. I learned a great deal from rotating in labs and taking their courses. Their clarity in research and communication is admirable. I would like to thank our collaborators Drs. Steve Horvath, William Yang, Amin Haghani, Ceasar Li, Ake Lu and Nan Wang. It was a lovely journey to collaborate with Drs. Horvath and Yang's labs. Their ambition and curiosity in research all drove us forward. I thank my BIG summer mentees Zane Koch, Elijah Jones and Saiyang Liu for their hard work and enthusiasms in research to our shared goals. I am grateful to the past and present student affair officers of the UCLA Bioinformatics Interdepartmental Program for their much-needed support that helps me navigate my graduate studies. I want to thank our lab janitor Ed, who works until 2AM every day to keep the entire third flood BSRB clean and ready for more experiments the next day.

I am thankful to the former and current members of the Ernst lab for their friendship and thoughtful discussions. They have taught me invaluable skills in organization, communication and commitment to pushing our work to the finish lines. I want to thank my cohort of Bioinformatics PhD students (and some members from other departments and schools), for all the food and updates that we shared. I also want to acknowledge Dr. Ahmet Ay for his invaluable mentorship during my undergraduate; he helped me navigate down this path, and I am so glad I listened to his advice. I want to thank my friends from Vietnam for beautiful memories of growing side by side, and friends I met in America who helped me make a second home here. Their humor, artistry,

intellect, accountability, honesty and grit nurture and inspire me each and every day. I am thankful for my siblings for all the funny stories they collect about everyone and everything. Lastly, I am deeply grateful to my parents, my grandmothers and extended family for my peaceful and loving upbringing, and for trusting me with freedom to pursue my educational aspirations. I cannot thank them enough for their sacrifices.

Chapter 2 is a version of Vu, Ha, and Jason Ernst. "Universal annotation of the human genome through integration of over a thousand epigenomic datasets." *Genome Biology* 23 (2022): 1-37. We thank Adriana Arneson for helping us in collecting data of analyses involving prioritized variants, repeat classes and GWAS catalog variants. We thank members of the Ernst lab, Hector Corrada Bravo, Amin Haghani, and Steve Horvath, Caesar Li, and Ake Lu for helpful discussions related to the manuscript. We thank the ENCODE, Roadmap Epigenomics consortia for generating data and making it publicly available. We acknowledge funding from US National Institute of Health (DP1DA044371, U01MH105578, UH3NS104095); US National Science Foundation (1254200, 2125664); Kure-IT award from Kure It cancer research, a Rose Hills Innovator Award, and the UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program.

Chapter 3 is a version of Vu, Ha, and Jason Ernst. "Universal chromatin state annotation of the mouse genome." *bioRxiv* (2022): 2022-12.We thank the Mouse ENCODE consortium for generating the data and making it publicly accessible. We thank Soo Bin Kwon for helping us collect the data of per-cell-type chromatin state annotations. We thank members of the Ernst lab for helpful feedback in completing this manuscript. This research was supported by US National Institute of Health (DP1DA044371, U01MH105578, UH3NS104095, U01HG012079); US National Science Foundation (1254200, 2125664); Rose Hills Innovator Award, and the UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program.

Chapter 4 is a version of Vu, Ha, et al. "A framework for group-wise summarization and comparison of chromatin state annotations." *Bioinformatics* 39.1 (2023): btac722. We thank the ENCODE and Roadmap Epigenomics consortia for generating data and making it publicly available. We thank members of the Ernst lab for their helpful suggestions on the manuscript. This research was supported by US National Institute of Health (DP1DA044371, U01MH105578,

UH3NS104095, U01HG012079); US National Science Foundation (1254200, 1705121, 2125664); a Rose Hills Innovator Award, and the UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program.

VITA

EDUCATION

2023	Ph.D. Candidate, Bioinformatics University of California, Los Angeles
2017	B.A. Mathematics, Computer Science Colgate University, Hamilton, NY

RESEARCH EXPERIENCE

2017 – Present	Ph.D. trainee Department of Biological Chemistry University of California, Los Angeles
2021	Computational Biology Intern Genentech Inc., Early Research and Development (GRED)

RESEARCH PUBLICATIONS

Vu, Ha, et al. "A framework for group-wise summarization and comparison of chromatin state annotations." *Bioinformatics* 39.1 (2023): btac722.

Vu, Ha, and Jason Ernst. "Universal annotation of the human genome through integration of over a thousand epigenomic datasets." *Genome Biology* 23 (2022): 1-37.

Vu, Ha, and Jason Ernst. "Universal chromatin state annotation of the mouse genome." *bioRxiv* (2022): 2022-12.

Arneson, Adriana, et al. "A mammalian methylation array for profiling methylation levels at conserved sequences." *Nature communications* 13.1 (2022): 783.

Li, C. Z., et al. "Epigenetic predictors of maximum lifespan and other life history traits in mammals." *bioRxiv* (2021): 2021-05.

Lu, Ake T., et al. "Universal DNA methylation age across mammalian tissues." *BioRxiv* (2021): 2021-01.

Keskin, Sevdenur, et al. "Regulatory network of the scoliosis-associated genes establishes rostrocaudal patterning of somites in zebrafish." *Iscience* 12 (2019): 247-259.

Keskin, Sevdenur, et al. "Noise in the vertebrate segmentation clock is boosted by time delays but tamed by notch signaling." *Cell reports* 23.7 (2018): 2175-2185.

Liberman, Amanda R., et al. "Circadian clock model supports molecular link between PER3 and human anxiety." *Scientific reports* 7.1 (2017): 9893.

Hart, Evelyn L., and Ha T. Vu. "The equivalence of two methods: finding representatives of non--empty Nielsen classes." *Bulletin of the Belgian Mathematical Society-Simon Stevin* 24.4 (2017): 741-745.

Chapter 1. Introduction

Epigenomic marks such as histone modifications, open chromatin regions or DNA methylation can correspond to different categories of gene regulatory elements (Barski *et al.*, 2007; Boyle *et al.*, 2008; Thurman *et al.*, 2012; Xie *et al.*, 2013). Methods such as ChromHMM (Ernst and Kellis, 2010, 2012) or Segway (Hoffman *et al.*, 2012) were developed to learn the recurrent combinatorial patterns of multiple chromatin marks across the genome of a biosample, classify these patterns into chromatin states, and eventually generate a genome annotation for that biosample. The resulting annotation maps out various regulatory elements such as promoters, enhancers, or inactive domains, and is useful in understanding the epigenomic contexts of biological phenomena. As epigenomic data becomes more abundant and diverse in assays and profiled biosamples, new questions emerge.

The first question, which is addressed in **Chapter 2**, involves the possibility and applicability of aggregating the patterns across epigenomic mappings for multiple cell and tissue types, then learning a single chromatin state annotation shared across the input biosamples. This requires training a chromatin state discovery tool (ChromHMM) such that input datasets from different tissue types are stacked as independent tracks and the learned states are defined jointly across tissue types. This approach was previously limited due to scalability challenges and is different from the per-biosample learning that was widely used before. Here, we developed a model using data from 1032 experiments in 127 biosamples from different human cell/tissue types (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium *et al.*), denoted as the full-stack model. This results in a *universal* human genome annotation that is shared across cell types, and the states can correspond to constitutive regulatory functions or rather cell-type-specific activities. We conducted thorough analyses to characterize the states and uncover many states with different cell-type-specific and sex-specific regulatory functions such as Brain-related enhancers, embryonic stem cell (ESC)-related bivalent promoters, chromosome X-specific quiescent regions, etc. We also used the full-stack annotation to analyze the epigenetic contexts

of various classes of genetic variations such as cancer-associated somatic mutations, structural variants, rare and common variants, etc. We argue, through reasoning and quantitative analyses with external genome annotations, that the full-stack annotation offers complementary values to existing biosample-specific annotations from Roadmap Epigenomics (Roadmap Epigenomics Consortium *et al.*) and ENCODE consortia (ENCODE Project Consortium, 2012).

In **Chapter 3**, we extend the above-mentioned approach of genome annotation by training an analogous stacked ChromHMM model on 901 Chip-seq/DNase-seq/ATAC-seq datasets from 26 mouse cell types from the Mouse ENCODE and ENCODE projects (Stamatoyannopoulos *et al.*, 2012; Yue *et al.*, 2014). We conducted equivalent analyses as in the human model to characterize the biological implications of the resulting states, and generate a mouse full-stacked annotation (Vu and Ernst, 2022). We also analyzed the relationships between states from the two organisms' annotations and how the similarities and differences between the two models are also reflected in functional conservation scores between the two species (Kwon and Ernst, 2021).

The human and mouse universal chromatin state maps have been adopted in various projects where we collaborated with other labs to elucidate the epigenetic contexts of regions involved in aging (Lu *et al.*, 2022), mammalian maximum lifespan prediction (Li *et al.*, 2021) and Huntington's disease.

Another challenge that emerges with more abundant epigenomic data is capturing probabilistic summary chromatin state annotations for a group of related biosamples (such as those with shared sex, tissue/cell type, case/control status, etc.). We developed a method named CSREP for this purpose, which is presented in **Chapter 4**. CSREP trains an ensemble of multivariate logistic regression classifiers that predicts state annotation in one biosample, given the corresponding annotations in others of the same group. We can take the difference between CSREP's summary chromatin state maps for two groups of samples to derive a map of genome-wide differential chromatin scores between these two groups. We conducted leave-one-out analysis, using chromatin state annotation data from 11 cell/tissue groups from Roadmap

Epigenomics (Roadmap Epigenomics Consortium *et al.*), to evaluate how well CSREP's summary chromatin state map for a group of samples can predict the genomic locations of individual chromatin states in a held-out sample. Here, CSREP resulted in better prediction compared to a baseline approach that simply counts state-frequency across input sample. We further showed, through various analyses, that the differential chromatin scores outputted by CSREP for two groups of samples can predict external assays that distinguish the two groups. For example, CSREP differential scores between ESC and Brain sample groups better predict genomic locations of Brain-specific or ESC-specific peaks of multiple chromatin marks (H3K27ac, H3K9ac and DNase). Using CSREP, we generated the summary chromatin state maps for 11 cell/tissue groups from Roadmap Epigenomics (Roadmap Epigenomics Consortium *et al.*) and 75 groups from EpiMap (Boix *et al.*, 2021).

Chapter 2. Universal annotation of the human genome through integration of over a thousand epigenomic datasets

Abstract

Genome-wide maps of chromatin marks such as histone modifications and open chromatin sites provide valuable information for annotating the nsn-coding genome, including identifying regulatory elements. Computational approaches such as ChromHMM have been applied to discover and annotate chromatin states defined by combinatorial and spatial patterns of chromatin marks within the same cell type. An alternative 'stacked modeling' approach was previously suggested, where chromatin states are defined jointly from datasets of multiple cell types to produce a single universal genome annotation based on all datasets. Despite its potential benefits for applications that are not specific to one cell type, such an approach was previously applied only for small-scale specialized purposes. Large-scale applications of stacked modeling have previously posed scalability challenges.

Using a version of ChromHMM enhanced for large-scale applications, we apply the stacked modeling approach to produce a universal chromatin state annotation of the human genome using over 1000 datasets from more than 100 cell types, with the learned model denoted as the full-stack model. The full-stack model states show distinct enrichments for external genomic annotations, which we use in characterizing each state. Compared to per-cell-type annotations, the full-stack annotations directly differentiate constitutive from cell type specific activity and is more predictive of locations of external genomic annotations.

The full-stack ChromHMM model provides a universal chromatin state annotation of the genome and a unified global view of over 1000 datasets. We expect this to be a useful resource that complements existing per-cell-type annotations for studying the non-coding human genome.

Introduction

Genome-wide maps of histone modifications, histone variants and open chromatin provide valuable information for annotating the non-coding genome features, including various types of regulatory elements (Barski *et al.*, 2007; Boyle *et al.*, 2008; Ernst *et al.*, 2011; Thurman

et al., 2012; Xie *et al.*, 2013). These maps -- produced by assays such as chromatin immunoprecipitation followed by high-throughput sequencing to map histone modifications or DNase-seq to map open chromatin-- can facilitate our understanding of regulatory elements and genetic variants that are associated with disease (Claussnitzer *et al.*, 2015; Gjoneska *et al.*, 2015; Kheradpour *et al.*, 2013; Lay *et al.*, 2014; Lee *et al.*, 2017; Taberlay *et al.*, 2014; Varshney *et al.*, 2017). Efforts by large scale consortia as well as many individual labs have resulted in these maps for many different human cell and tissue types for multiple different chromatin marks (Barski *et al.*, 2007; Consortium, 2007; ENCODE Project Consortium, 2012; Fernández *et al.*, 2015; Kheradpour *et al.*, 2013; Meuleman *et al.*, 2015; Mikkelsen *et al.*, 2007; Stunnenberg *et al.*, 2016; Q. Wang *et al.*, 2020; Zhu *et al.*, 2013).

The availability of maps for multiple different chromatin marks in the same cell type motivated the development of methods such as ChromHMM and Segway that learn 'chromatin states' based on the combinatorial and spatial patterns of marks in such data (Ernst and Kellis, 2010, 2012; Hoffman *et al.*, 2012). These methods then annotate genomes in a per-cell-type manner based on the learned chromatin states. They have been applied to annotate more than a hundred diverse cell and tissue types (Ernst *et al.*, 2011; Meuleman *et al.*, 2015; Libbrecht *et al.*, 2019). Previously, large collections of per-cell-type chromatin state annotations have been generated using either (1) independent models that learn a different set of states in each cell type or (2) a single model that is learned across all cell types, resulting in a common set of states across cell types, yet generating per-cell-type annotations (in some cases per-tissue-type annotations are generated, but we will use the terms cell-type and tissue interchangeably for ease of presentation). This latter approach is referred to as a 'concatenated' approach (**Supplementary Fig. 2.2.1**) (Ernst and Kellis, 2012, 2017). Variants of the concatenated approach attempt to use information from related cell types to reduce the effect of noise, but still output per-cell-type annotations (Biesinger *et al.*, 2013; Zhang *et al.*, 2016). These models that produce per-cell-type annotations tend to be most appropriate in studies where researchers are interested in studying individual cell types.

A complementary approach to applying ChromHMM to data across multiple different cell types referred to as the 'stacked' modeling approach was also previously suggested (**Supplementary Fig. 2.2.1**) (Ernst and Kellis, 2012, 2017). Instead of learning per-cell-type annotations based on a limited number of datasets available in each cell type, the stacked modeling approach can learn a single universal genome annotation based on the combinatorial and spatial patterns in datasets from multiple marks across multiple cell types. This approach differs from the concatenated and independent modeling approaches as those approaches only identify combinatorial and spatial patterns present among datasets within one cell type.

Such a universal annotation from stacked modeling provides potential complementary benefits to existing concatenated and independent chromatin state annotations. First, since the model can learn patterns from signals from the same assay across cell types, a stacked model may help differentiate regions with constitutive chromatin activities from those with cell-type-specific activities. Previously, subsets of the genome assigned to individual chromatin states from 'concatenated' annotations were post-hoc clustered to analyze chromatin dynamics across cell and tissue types (Ernst *et al.*, 2011; Meuleman *et al.*, 2015). However, such an approach does not provide a view of the dynamics of all the data at once, which the stacked modeling provides. Second, the stacked modeling approach bypasses the need to pick a specific cell or tissue type when analyzing a single partitioning and annotation of the genome. Focusing on a single cell or tissue type may not be desirable for many analyses involving other data that are not inherently cell-type-specific, such as those involving conserved DNA sequence or genetic variants. For example, when studying the relationship between chromatin states and evolutionarily conserved sequences, if one uses per-cell-type chromatin state annotations from one cell type, many bases will lack an informative chromatin state assignment (e.g. many bases are in a quiescent state), while subsets of those bases will have a more informative annotation in other cell types. Third, if one tries to analyze per-cell-type annotations across cell types, one would need a post-hoc method to reason about an exponentially large number of possible combinations of chromatin

states across cell types (if each of K cell types has M states, there are M^K possible combinations of states for a genomic position) many of which would likely lack biologically meaningful distinctions. In contrast, for the stacked model, there will be a single annotation per position out of a possibly much smaller fixed number of states (compared to M^K). These states are directly informative of cross-cell type activity, though the state definitions can be more complex. Finally, annotations by the stacked modeling leverages a larger set of data for annotation, and thus has the potential to be able to identify genomic elements with greater sensitivity and specificity.

Despite the potential complementary advantages of the 'stacked' modeling approach, it has only been applied on a limited scale to combine data from a small number of cell types for highly specialized purposes (Chronis *et al.*, 2017; Mortazavi *et al.*, 2013). No large-scale application of the stacked modeling approach to many diverse cell and tissue types has been previously demonstrated. This may have in part been due to large-scale applications of stacked modeling raising scalability challenges not present in modeling approaches for concatenated and independent annotations.

Here, we present a large-scale application of the stacked modeling approach with more than a thousand human epigenomic datasets as input, using a version of ChromHMM for which we enhanced the scalability. We conduct various enrichment analyses on the states resulting from the stacked modeling and give biological interpretations to them. We show that compared to the per-cell-type annotations from independent and concatenated models, the stacked model's annotation shows greater correspondence to various external genomic annotations not used in the model learning. We analyze the states in terms of enrichment with different types of variation, and highlight specific states of the stacked model that are enriched with phenotypically associated genetic variants, cancer-associated somatic mutations, and structural variants. We expect the stacked model annotations and detailed characterization of the states that we provide will be a valuable resource for studying the epigenome and non-coding genome, complementing existing per-cell-type annotations.

Results

Annotating the human genome into universal chromatin states

We used the stacked modeling approach of ChromHMM to produce a universal chromatin state annotation of the human genome based on data from over 100 cell and tissue types from the Roadmap Epigenomics and ENCODE projects (**Fig. 2.1**) (ENCODE Project Consortium, 2012; Meuleman *et al.*, 2015). In total we applied ChromHMM to 1032 datasets for 30 histone modifications, a histone variant (H2A.Z), and DNase I hypersensitivity (**Supplementary Fig. 2.2**). The set of cell and tissue types were the same as those for which per-cell-type annotations were previously generated by applying the 'concatenated' modeling approach of ChromHMM (Ernst and Kellis, 2012, 2017; Meuleman *et al.*, 2015). We note that not all chromatin marks were profiled in all cell or tissue types, but the stacked modelling can still be applied directly.

We focused our analysis on a model with 100 states (**Methods**). The number of states is larger than typically used for models that generate per-cell-type annotations, which reflects the greater information available when defining states based on data from many cell types. This number of states was large enough to be able to capture some relatively cell-type-specific regulatory activity, while being small enough to give distinct biological interpretations to each state (**Supplementary Fig. 2.3**). We denote the model's output chromatin state annotation the 'full-stack' genome annotation.

Major groups of full-stack states

We characterized each state of the model by analyzing the model parameters (emission probabilities and transition probabilities) and state enrichments for other genome annotations (**Fig. 2.2, 3A, Supplementary Fig. 2.4-7**). The other genomic annotations include previous concatenated chromatin state annotations (**Supplementary Fig. 2.8-9**), cell-type-specific gene expression data (**Supplementary Fig. 2.10-11**), and various independent existing genomic annotations (**Fig. 2.3A**). These independent genomic annotations included annotated gene features, evolutionary constrained elements, and assembly gaps, among others (**Methods**).

These analyses led us to group the 100 full-stack states into 16 groups (**Fig. 2.2A**). One group includes states associated with assembly gaps (GapArtf1) and alignment artifacts (GapArtf2-3). Some other groups are associated with repressive or inactive states, including quiescent states (Quies1-5) (low emissions for all datasets, except possibly weak signals in H3K9me3), heterochromatin states associated with H3K9me3 (HET1-9), and polycomb repressed states associated with H3K27me3 (ReprPC1-9). There is an acetylations group marked primarily by high emission of various acetylation marks profiled only in IMR90 or ESC and ESC-derived cells, while having weaker signals for enhancer or promoter associated marks such as H3K4me1/2/3, H3K27ac, and H3K9ac (Acet1-8). We also identified weak and active candidate enhancers groups (EnhW1-8 and EnhA1-20, respectively) associated with H3K4me1, DNase, H2A.Z, and/or H3K27ac. Four groups are associated with transcriptional activities, including a group of transcribed enhancers (TxEnh1-8), two groups of weak or strong transcription (TxWk1-2, Tx1-8, respectively), and one group associated with exons and transcription (TxEx1-4). These transcriptional activities groups are associated with at least one of these marks: H3K36me3, H3K79me1, H3K79me2, or H4K20me1. Another group consists of two zinc finger (ZNF) gene states associated with H3K36me3 and H3K9me3 (znf1-2). A DNase group consists of one state (DNase1) with strong emission of *only* DNase I hypersensitivity in all profiled cell types. Three groups are associated with promoter activities, marked by emission of some promoter marks such as H3K4me3, H3K4me2, and H3K9ac. One promoter group was of bivalent states associated with promoter marks and H3K27me3 (BivProm1-4). The other two promoter groups were flanking promoter states (PromF1-7) and transcription start sites (TSS) states (TSS1-2) where the flanking promoter states also show emission of H3K4me1.

Enrichments for external annotations supported these state groupings (**Fig. 2.3A**), as well as further distinctions or characterizations among states within each group. For example, the state GapArt1 had ~8 fold enrichment for assembly gaps and contained 99.99% of all assembly gaps in hg19 (**Fig. 2.3A**). In previous concatenated models based on the Roadmap Epigenomics data

(Meuleman *et al.*, 2015), no specific state was associated with assembly gaps likely due to the limited number of input chromatin mark signals compared to the number of states, leading to assembly gaps being incorporated in the general quiescent state. The states in the zinc finger gene group, znf1-2, had 20.8 and 68.6 fold enrichment for zinc finger named genes, respectively (**Fig. 2.3A**). States in the Acet group had a lower average expression of proximal genes compared to states in the enhancer and promoter groups (**Fig. 2.3C**) while higher compared to ReprPC and HET groups. States in the transcription groups (TxEnh1-8, TxWk1-2, Tx1-8, TxEx1-4) were all at least 2.1 fold enriched for annotated gene bodies; these gene bodies covered 88.8–97.5% of the states. These states are associated with higher expression of genes across different cell types, particularly when downstream of their TSS (**Fig 3A, C, Supplementary Fig. 2.10-11**). Distinctions were seen among these states, for example, in terms of their positional enrichments relative to TES (**Fig. 2.3A, D, Supplementary Fig. 2.12**). States in the flanking promoter group (PromF1-7) showed 6.5-28 fold enrichment for being within 2kb of annotated TSS, with distinctions among states in terms of their relative distance from the TSS **(Fig. 2.3A, C, Supplementary Figure 2.12A)**. Genes whose TSS regions overlapped these states had higher average gene expression across different cell types (**Fig. 2.3A, C, Supplementary Fig. 2.10-11**). These states differed among each other in their enrichments with upstream or downstream regions of the TSS (**Fig. 2.3E, Supplementary Fig. 2.12**). The states in the transcription start site group (TSS1-2) had enrichment values that peaked at the TSS (>= 100 fold enrichment) (**Fig. 2.3A, E**). States in promoter-associated groups (TSS, PromF, BivProm), along with those in other groups, show various enriched Gene-Ontology (GO) terms based on genes overlapping or proximal to each state (**Supplementary Fig. 2.13, Supplementary Data 2.1, Methods**). For example, among biological process terms, BivProm1 is most enriched for 'embryonic organ morphogenesis' genes, while TSS1 is most enriched for 'nucleic acid metabolic process', consistent with the bivalent (Bernstein *et al.*, 2006) and the constitutively active nature of the two states, respectively. The DNase-specific state DNase1, showed distinct enrichment for CTCF-specific chromatin states

defined in a concatenated model for six cell types, compared to other full-stack states (Hoffman *et al.*, 2013) (**Fig. 2.3F, Supplementary Fig. 2.14**). These CTCF-specific states have previously been suggested to be candidate insulators and may have other roles that CTCF has been implicated in such as demarcating TAD boundaries (Dixon *et al.*, 2012; McArthur and Capra, 2021; Phillips and Corces, 2009; Wang *et al.*, 2021). The DNase1 state may correspond to similar roles, particularly where the CTCF-binding is relatively stable across cell types.

Compared to other full-stack state groups, those associated with promoters (TSS, flanking promoters, bivalent promoters) and the DNase group showed lower average DNA methylation levels across cell types (**Supplementary Fig. 2.15**). Among promoter-associated states, those showing stronger enrichments with CpG Islands also showed lower methylation levels (**Fig. 2.3A, Supplementary Fig. 2.15**), consistent with previous studies (Jones and Takai, 2001; Weber *et al.*, 2007). Some promoter-associated states (TSS1-2, PromF3-5, BivProm1-2) are among the most enriched states at the center of binding regions of polycomb repressed complex 1 and 2 (PRC1 and PRC2) sub-units. In addition, several ReprPC and BivProm states are among the most enriched states in windows surrounding binding regions of the EZH2 and SUZ12 subunits of PRC2, consistent with these states' association with H3K27me3 (**Supplementary Fig. 2.16-17**). A detailed characterization of all states in terms of associated chromatin marks, genomic elements and different associated per-cell-type chromatin states across cell groups can be found in **Supplementary Data 2.2-4.** We expect it will serve as a resource for future applications using the full-stack annotations.

We verified that enrichments computed based on hg19 were highly similar to those computed for full-stack annotations mapped to hg38 (average correlation 0.99; **Methods, Supplementary Fig. 2.18)**, ensuring the applicability of state annotations in hg38. We also confirmed that the full-stack annotations were generally more predictive of the positions of a variety of external genome annotations considered in **Fig. 2.3A** than two sets of per-cell-type annotations, a previous 18-state per-cell-type chromatin state annotation based on concatenated

models from 127 cell types (Meuleman *et al.*, 2015), denoted the concatenated annotations, and 100-state per-cell-type annotations learned independently in each cell type, denoted the independent annotations (**Methods**). As expected, since the full-stack model uses more data representing more cell types, the full-stack annotations had greater predictive performance in most cases (**Supplementary Fig. 2.19-22, Supplementary Data 2.5)**. One of the exceptions to this was lambin B1 associated domains from Tig3 human lung fibroblasts (Guelen *et al.*, 2008), where six independent annotations were more predictive, three of which are annotations for fibroblasts. We note that these evaluations were done under the assumption that a chromatin state annotation that is more predictive of well-established external genomic elements will also be more informative of less well-established classes of elements. These results suggest that the full-stack annotations will, in most cases, have greater information than any single concatenated or independent annotations about localization patterns of some target genomic elements, with likely exceptions when the target of interest is specific to a certain cell type. In such cases, the corresponding cell type's concatenated or independent chromatin state annotation may be more predictive.

Stacked Model Differentiates Cell-Type-Specific from Constitutive Activity

While the major groups of states outlined above can correspond to states from concatenated models (Ernst *et al.*, 2011; Meuleman *et al.*, 2015), the full-stack states provide additional information. For example, the states can differentiate cell-type-specific from constitutive activities. This cell type specificity in the full-stack states is reflected in the emission parameters of cell types from different tissue groups **(Fig. 2.2B,C, Supplementary Data 2.2 and 4)** and the overlap of concatenated chromatin state annotations from different cell types (Ernst and Kellis, 2015) **(Supplementary Fig. 2.8-9, Supplementary Data 2.4)**.

Consistent with previous findings that enhancers tend to be relatively cell-type-specific while promoters tend to be shared across cell types (Ernst *et al.*, 2011; Heintzman *et al.*, 2007), enhancer states exhibited clearer cell-type-specific associations than those of the promoter states

(**Figure 2C, Supplementary Data 2.2 and 4**). On average, states of active enhancer and weak enhancer groups (EnhA1-20, EnhW1-8) showed at least two-fold higher coefficients of variations, in terms of emission probabilities for various marks, compared to states in the TSS, flanking and bivalent promoter groups (**Supplementary Fig. 2.23**). The enhancer states differed among each other in their associations with different cell/tissue types such as brain (EnhA6), blood (EnhA7-9 and EnhWk6), digestive tissue (EnhA14-15), and embryonic stem cells (EnhA18) (**Fig. 2.1-2, Supplementary Fig. 2.24-25**). These differences in cell-type-specific activities are also associated with different gene expression levels of overlapping genes with the states. For example, some blood enhancer states (EnhA8, EnhA9, EnhWk6) overlapped genes with higher average gene expression in cell types of the blood group, while some enhancer states specific to digestive group or liver tissues (EnhA14, EnhA15) showed higher gene expression in the corresponding cell or tissue types (**Fig. 2.3C, Supplementary Fig. 2.10**).

Other groups of states besides enhancers also had individual states with cell-type-specific differences. For example, four of the nine states in the heterochromatin group (HET1-2,4,9) showed higher emission probabilities of H3K9me3 in only subsets of cell types (states HET1-2 with IMR90 and Epithelial cells; state HET4 with adipose, mesench, neurospheres, ESC, HSC&B-cells). State HET9 showed strong association as heterochromatin in ESC/iPSC groups, while being mostly quiescent in other cell types based on concatenated annotations (**Fig. 2.2C, Supplementary Fig. 2.26, Supplementary Data 2.2 and 4**). State PromF5 is associated with putative bivalent promoter chromatin states in some blood and ESC-related cell types, but with flanking promoter states in most other cell types (**Supplementary Fig. 2.27, Supplementary Data 2.2 and 4**). In addition, some quiescent states (Quies1-2, Quies4-5) show weak signals of H3K9me3 in specific groups of cell types (**Supplementary Data 2.2 and 4**). States in the polycomb repressed and bivalent promoter groups (ReprPC1-9, BivProm1-4) also showed differences in signals across cell groups, such as state ReprPC9, which showed H3K27me3 signals in only ESC/iPSC cell types (**Supplementary Data 2.2 and 4**). The ability of the stacked

modeling approach to explicitly annotate both cell-type-specific and constitutive patterns for diverse classes of chromatin states highlights a complementary advantage of this approach relative to approaches that provide per-cell-type annotations.

Full-stack states show distinct enrichments for repeat elements

As the full-stack model showed greater predictive power for repeat elements than cell-type-specific models (**Supplementary Fig. 2.19-21**), we next analyzed which states contributed most to this power. The full-stack state enrichments for bases in repeat elements ranged from 10-fold depletion to 2-fold enrichment (**Fig. 2.3A**). The top ten states most enriched with repeat elements were chromatin states associated with H3K9me3 marks and in the heterochromatin, artifact, quiescent, or ZNF genes groups (**Fig. 2.4A-B).** Repeats being consistently enriched in H3K9me3-marked states is a natural mechanisms of cells to reduce the repeats' risks to genome integrity, since H3K9me3 is characteristic of tightly-packed DNA (heterochromatin) that is physically inaccessible (Becker *et al.*, 2016).

We also observed that individual full-stack states had distinct enrichments for different repeat classes (**Fig. 2.4C, Supplementary Fig. 2.28**). For example, Acet1, a state associated with various acetylation marks and H3K9me3 had a 23-fold enrichment for simple repeats largely driven by (CA)n and (TG)n repeats which were 72 and 76 fold enriched and comprised 74% of all simple repeats in this state (**Supplementary Fig. 2.28**). As (CA)n and (TG)n repeats are known to be highly polymorphic in humans (Dib *et al.*, 1996), this suggests the possibility that signal detected in these regions may in part be due to technical issues related to deviations from the reference genome. The two states in the artifact group, GapArtf2-3, had a particularly high enrichment for satellite (181 and 145 fold, respectively) and rRNA repeat classes (75 and 580 fold, respectively) (**Fig. 2.4C, Supplementary Fig. 2.28**), likely associated with sequence mapping artifacts. States in the transcription start site group, TSS1-2, were most strongly enriched with low complexity repeat class (10.5-18 fold) and most notably GC rich repeats (195-303 fold), consistent with these states being most enriched for windows of high GC content

(**Supplementary Fig. 2.28**). Moreover, the TSS1-2 states are also most enriched with tRNA class (50-61 fold) (**Supplementary Fig. 2.28**), consistent with tRNAs being short genes (Lowe and Eddy, 1997).

We also saw specific states associated with the largest repeat classes of the genome, SINEs, LINEs, and LTRs. SINE repeats were most enriched in state Tx5 (3.7 fold) (**Fig. 2.5C**), which had high emission of H3K36me3 (**Fig. 2.2A-B, Supplementary Fig. 2.4-5**), consistent with previous studies showing that SINEs are more enriched in gene-rich regions and in transcription-related states based on concatenated annotations (Elbarbary *et al.*, 2016; Ernst and Kellis, 2010; Pehrsson *et al.*, 2019). In contrast, LINEs are depleted in most transcription-related states, reflecting the negative selection against long-sequence insertions in or near genes (Elbarbary *et al.*, 2016). LINEs are most enriched in state HET3 (3.4 folds) (**Supplementary Fig. 2.28**), and a notable property of this state is it does not show signals of H3K27me3 and acetylation marks across cell/tissue types. This property of HET3 is a pattern that would be difficult to recognize without stacked-modeling, and was only shared with HET4 and HET9 among states in the heterochromatin group. HET4 was also the second most enriched state in the heterochromatin group for LINEs (2.0-fold) while HET9 was not enriched, but is distinct in that it identifies regions where H3K9me3 is relatively specific to cell types in the embryonic and iPSC groups. LTRs are most enriched in state HET5 (4.7 fold), and this state is marked by its highest signals of H3K9me3 compared to other states in the heterochromatin group (**Fig. 2.4C, Supplementary Fig. 2.28**). LTRs showing strong enrichment with states associated with strong presence of H3K9me3 is consistent with concatenated-model chromatin state analyses (Ernst and Kellis, 2010; Pehrsson *et al.*, 2019). We also confirmed that the increased predictive power of the full-stack model over concatenated and independent annotations, which was previously seen for repeat elements overall, also held for most of the individual repeat classes (**Supplementary Fig. 2.29**). Overall, results of enrichment of full-stack state annotations for repeat classes offer further details in

stratifying the states' characteristics, and concurrently confirm and refine existing knowledge about different repeat classes chromatin state associations.

Full-stack states show distinct enrichments for constrained elements and conservation states

Sequence constrained elements are another class of genomic elements that are not cell-type-specific and for which the full-stack annotations showed greater predictive power than concatenated and independent annotations (**Supplementary Fig. 2.19-21**). We next sought to better understand the relationship between full-stack states and sequence conservation annotations. We observed 10 states that had at least a 3.4 fold enrichment for PhastCons elements (**Fig. 2.4A**). These states were associated with the TSSs or being proximal to them (TSS1-2 and PromF4-5), transcription with strong H3K36me3 signals (TxEx2 and TxEnh4), or enhancers associated with mesenchymal, muscle, heart, neurosph, adipose (EnhA2) (**Fig. 2.4A-B**). In contrast, seven states (HET3-4,6-7,9, Quies4, Gap Artf2) were more than two-fold depleted for PhastCons elements, which all had more than a 1.5 fold enrichment for repeat elements (**Fig. 2.4A**).

To gain a more refined understanding of the relationship between the full-stack chromatin states and conservation, we analyzed their enrichment using 100 previously defined conservation states by the ConsHMM method (Arneson and Ernst, 2019). These conservation states were defined based on the patterns of other species' genomes aligning to or matching the human reference genome within a 100-way vertebrate alignment. We observed 29 different conservation states maximally enriched for at least one full-stack state (**Fig. 2.3B, Supplementary Fig. 2.30**). These conservation states included, for example, ConsHMM state 1, a conservation state corresponding to bases aligning and matching through all vertebrates and hence most associated with constraint. ConsHMM state 1 had >= 10 fold enrichment for exon associated full-stack states TxEx1-4 and TxEnh4 (**Supplementary Fig. 2.30A**). Another ConsHMM state, state 28, which is associated with moderate aligning and matching through many vertebrates and strongly enriched around TSS and CpG islands, had a 44.5 and 47.8 fold enrichment for TSS-associated full-stack

states TSS1 and TSS2, respectively (**Supplementary Fig. 2.30A**). Additionally, this conservation state is consistently the most enriched conservation state for full-stack states associated with flanking and bivalent promoters (**Fig. 2.3B, Supplementary Fig. 2.30A**). ConsHMM state 2, which has high aligning and matching frequencies for most mammals and a subset of non-mammalian vertebrates and previously associated with conserved enhancer regions (Arneson and Ernst, 2019), showed >2.7 fold enrichment for some full-stack enhancer states for Brain (EnhWk4 and EnhA6), ESC & iPSC (EnhA17,19 and EnhWk8), neurosph (EnhWk4, EnhA2,17), and mesenchymal, muscle, heart, adipose (EnhA2) (**Fig. 2.3B, Supplementary Fig. 2.30A**). ConsHMM state 100, a conservation state associated with alignment artifacts, was 10.9 fold and 2.1 fold enriched for full-stack state znf2 and znf1, respectively (**Fig. 2.3A-B, Supplementary Fig. 2.30A**). This is consistent with previous analysis using concatenated annotations showing that ConsHMM state 100 was most enriched in a ZNF gene-associated chromatin state (Arneson and Ernst, 2019). State znf2 also showed a 5.4-fold enrichment for ConsHMM state 1 which contrasts with state znf1, which showed a 0.8 fold enrichment for ConsHMM state 1, suggesting a stronger association of state znf1 with newly evolving ZNF genes or those under less constraint. This difference is consistent with the znf2 state's larger fold enrichment for coding exons than znf1 (10.4 vs 1.1). The znf2 state also had a greater fold enrichment for ZNF named genes in general (68.6 vs. 20.8 fold), with those enrichment stronger and the difference greater when restricting to C2H2 annotated genes (86.8 vs. 25.1 fold) (**Fig. 2.3A-B, Supplementary Fig. 2.31**). Therefore, the full-stack annotation helped distinguish two ZNF-gene associated states, which are associated with distinct conservation states. As this example illustrates, the full-stack annotation captured conservation state enrichments that were generally consistent with those seen in concatenated annotations, but could also identify additional refined enrichment patterns.

Specific full-stack states show distinct enrichments and depletions for structural variants

We also analyzed the enrichment of the full-stack states for overlap with structural variants (SVs) mapped in 17,795 deeply sequenced human genomes (Abel *et al.*, 2020), and focused on the two largest classes of SVs, deletions and duplications. Abel et al., 2020 (Abel *et al.*, 2020) analyzed the enrichments of these deletions and duplications with concatenated-model chromatin states in 127 reference epigenomes (Meuleman *et al.*, 2015), and observed that ZNF gene and heterochromatin states were enriched for deletions and duplications, with the enrichments being stronger in regions annotated as these states (ZNF or heterochromatin) in larger number of cell or tissue types (e.g. more constitutive HET/ZNF regions). Consistent with those previous results, using the full-stack model, we observed that of the 13 states that were among the top 10 maximally enriched states for either deletions or duplications (1.18 fold or greater), seven were in the heterochromatin group (HET1-2,4-7,9) and one was the znf2 state (**Fig. 2.5A, Supplementary Fig. 2.32**). The enrichment of structural variation in HET states is consistent with the notion that potentially larger effect structural variants would less likely experience negative selection in these regions of the genome. As the znf2 state is most enriched for a conservation state associated with putative alignment artifacts, this raises the possibility that technical issues may be contributing to its SV enrichments (**Supplementary Fig. 2.30**). The other five states included two artifact states (GapArtf2-3) and three quiescent states (Quies1-2,4) (**Fig. 2.5A**). The quiescent states Quies1-2,4, despite the generally low frequencies for all marks, did have higher emission probabilities for H3K9me3 compared to other chromatin marks (**Fig. 2.5B**).

The full-stack model was also more predictive of SV than concatenated and independent ones (**Supplementary Fig. 2.33-34, Supplementary Data 2.5**). Additionally, we verified that the full-stack annotations had higher AUROC in predicting duplications and deletions compared to annotations obtained by ranking genomic bases based on the number of cell or tissue types that a state was observed, as in the approach of (Abel *et al.*, 2020) (**Methods**, **Supplementary Fig. 2.35**). These results show that the full-stack annotation can uncover enrichment patterns with SVs

that are consistent with concatenated annotations, yet highlight states with greater predictive power and offer a more refined chromatin annotation of the regions enriched with SVs.

Full-stack states gives insights into bases prioritized by different variant prioritization scores

Various scores have been proposed to prioritize deleterious variants in non-coding regions of the genome or genome-wide. These scores are based on either conservation or on integrating diverse sets of genomic annotations. Though the scores all serve to prioritize variants, they can vary substantially from each other and it is often not clear the differences among the types of bases that different scores prioritize. To better understand the epigenomic contexts of bases that each score tends to prioritize, we analyzed the full-stack state enrichment for bases they prioritize. As the scores we considered are not specific to a single cell type, the full-stack states have the potential to be more informative for this analysis than concatenated or independent annotations. We considered a set of 14 different variant prioritization scores that were previously analyzed in the context of conservation state analysis (Arneson and Ernst, 2019). The 14 scores for which we analyzed prioritized variants in non-coding regions were CADD(v1.4), CDTS, DANN, Eigen, Eigen-PC, FATHMM-XF, FIRE, fitCons, FunSeq2, GERP++, LINSIGHT, PhastCons, PhyloP, and REMM (Cooper *et al.*, 2010; Davydov *et al.*, 2010; Di Iulio *et al.*, 2018; Fu *et al.*, 2014; Gulko *et al.*, 2015; Huang *et al.*, 2017; Ioannidis *et al.*, 2017; Ionita-Laza *et al.*, 2016; Pollard *et al.*, 2010; Quang *et al.*, 2015; Rentzsch *et al.*, 2019; Rogers *et al.*, 2018; Siepel *et al.*, 2005; Smedley *et al.*, 2016). For each of these scores, we first analyzed the full-stack state enrichments for the top 1% prioritized non-coding variants relative to the background of non-coding regions on the genome (**Methods**).

In the top 1% prioritized non-coding bases, 19 states were among the top five most enriched states ranked by at least one of the 14 scores (**Fig. 2.5C, Supplementary Fig. 2.36-37**). These 19 states include nine states in promoter-associated groups, five states in enhancers-related groups, three states in the exon-associated transcription group, one polycomb repressed state, and one DNase state (**Fig. 2.5C**). Seven scores (DANN, Eigen, Eigen_PC, funSeq2, CDTS,

CADD and REMM) had their top five enriched states exclusively associated with promoter and TSS states, with enrichments ranging between 8.6 and 70 fold (PromF2-5, TSS1-2, BivProm1-2,4) (**Fig. 2.5C**). In contrast, the fitCons score showed depletions for three of these states and relatively weaker enrichment for the others. This difference might be related to fitCons' approach of prioritizing bases showing depletion of human genetic polymorphisms, potentially without sufficiently accounting for the increased mutation rates in regions with high CpG content that are observed in promoter-associated states (**Supplementary Fig. 2.42**) (Karczewski *et al.*, 2020). FIRE's prioritized variants showed depletions in the bivalent promoter states (BivProm) and PromF5, which have generally lower average gene expression across cell types (**Fig. 2.3C**). This depletion is reflective of the fact that FIRE was trained to prioritize variants in cis-expression quantitative trait loci (cis-eQTLs) in one cell type (LCL) (Ioannidis *et al.*, 2017), and few eQTLs are expected to be proximal to genes with limited or no expression in that cell type. Enhancer states EnhA2-3,17 were among the states in the top five most enriched for FATHMM, GERP++, LINSIGHT, PhastCons, and PhyloP prioritized non-coding variants. In contrast, FIRE, DANN and CDTS were depleted for prioritized variants in all these enhancer states (**Fig. 2.5C**). FIRE and fitCons showed strong enrichment for exon states (TxEx1-3), which are associated with coding regions, even though coding bases were excluded in this analysis (Fig. 2.5C). FATHMM had the greatest relative enrichment (~10 fold) for the primary DNase state (DNase1), which is associated with a CTCF state from a concatenated model defined in six cell types (Hoffman *et al.*, 2013), and was the only score for which this state was among the top five most enriched states (**Fig. 2.5C, Supplementary Fig. 2.14, 36**).

We conducted similar analyses based on top 5% and 10% prioritized non-coding variants and observed relatively similar patterns of enrichments, though there did exist some differences at these thresholds (**Supplementary Fig. 2.36, 38-39**). One difference was that alignment artifact states GapArtf2-3 were among the top two most enriched states with top non-coding bases prioritized by FATHMM-XF, while a number of other scores showed depletions for these states

(**Supplementary Fig. 2.36**). In addition, we analyzed top 1%, 5%, and 10% prioritized variants genome-wide from 12 of the scores (**Methods**) (**Supplementary Fig. 2.37-40**). Compared to the non-coding analysis, we saw a majority of scores had exon-associated transcription states (TxEx1-TxEx4) among the top five enriched states with top 1% variants genome-wide, while we saw no enhancer state among the top five enriched states with top 1% variants by any score and only one enhancer state among the top five by one score (GERP++) for top 5% and 10% variants.

Overall, this analysis showed that the scores tend to prioritize bases in different epigenetic contexts. As the scores vary in the genomic features selected as input, and the predictive model for scoring bases, it is expected that different methods may show higher scores for in different classes of genomic contexts. By analyzing the state enrichments, one can gain some expectations for what types of evaluation criteria different scores might perform better, but we note that this analysis is not trying to directly conclude one method is preferred. Also, while in general it is difficult to conclude confidently what enrichments are due to technical or biological biases, by comparing enrichments across scores and considering what else is known about the states, one can still gain insights into this. For example, the inconsistent enrichments of different methods for prioritized variants in GapArtf2-3 states (**Supplementary Fig. 2.36**), along with these states' association with sequencing artifacts, is suggestive of technical biases. Similarly, DANN's top 1% non-coding bases showing enrichments in five heterochromatin states, while not showing any enrichments in enhancer states, and no other scores showing enrichments in heterochromatin states, is also suggestive of technical biases (**Supplementary Fig. 2.37**).

We verified that the full-stack annotation showed the highest AUROC in recovering the top 1% non-coding variants compared to all 18-state concatenated annotations for all 14 scores (**Supplementary Fig. 2.41**). Compared to all 100-state independent annotations, the full-stack model showed the highest AUROC for 13 out of 14 scores in all 127 cell types (**Supplementary Fig. 2.41**).

Full-stack states show distinct enrichments and depletions for human genetic variation

We next analyzed full-stack states for their enrichment with human genetic sequence variation. We calculated enrichments of full-stack states with genetic variants sequenced in 15,708 genomes from unrelated individuals in the GNOMAD database stratified by minor allele frequencies (MAFs) (Karczewski *et al.*, 2020). Across eleven ranges of MAFs, the state enrichments ranged from a 2-fold enrichment to a 4-fold depletion (**Supplementary Fig. 2.42**). As expected, the state associated with assembly gaps (GapArtf1) is most depleted with variants, regardless of the MAF range. At the other extreme, state Acet1, which is associated with simple repeats, is the most enriched state with variants for all ten minor allele frequency (MAF) ranges that are greater than 0.0001, with fold enrichments between 1.8 and 2.0 (**Supplementary Fig. 2.42**). We verified that the high enrichment for state Acet1 was not specific to GNOMAD's calling of variants as it had a 2.0 fold enriched with common variants from dbSNP (**Methods**) (**Supplementary Fig. 2.42**). TSS and promoters associated states, PromF4 and TSS1-2, were maximally enriched for variants in the lowest range of MAF (0 < MAF <= 0.0001), 1.5-1.7 fold. The enrichment of variants for these states decreased as the MAF ranges increased, falling to 0.8-1.2 fold for variants of the highest range of MAF (0.4-0.5) (**Supplementary Fig. 2.42**). The high enrichment for states PromF4 and TSS1-2 for rare variants, despite their being the most enriched states with PhastCons conserved elements, can be explained by these states' high enrichment of CpG dinucleotides, which are associated with higher mutation rates (**Fig. 2.3A, Supplementary Fig. 2.42**) (Karczewski *et al.*, 2020)**.** At the same time, purifying selection can have a weaker impact on larger-effect rare variants than on larger-effect common variants. We also observed the pattern of decreasing enrichments for variants with increasing MAF in other states associated with transcriptional activities, enhancers, DNase, or promoters (**Supplementary Fig. 2.42**). This pattern was not observed in most states from other groups such as heterochromatin, polycomb repressed, quiescent, and acetylations only (**Supplementary Fig. 2.42**).

To better identify states with a depletion of common variants that are more likely due to selection, we ranked states based on their ratios of enrichments for the rarest variants (MAF < 0.0001) relative to the most common variants (MAF 0.4-0.5) (**Fig. 2.5D**). The states with the highest ratio included a number of flanking promoter (PromF3-4) and exon-transcription states (TxEx1,2,4) that were also associated with strong sequence conservation across species (**Fig. 2.3B, Fig. 2.5D**). These results are consistent with previous analyses supporting a depletion of common human genetic variation in evolutionary conserved regions (Lindblad-Toh et al., 2011). States associated with assembly gaps and alignment artifacts (GapArtf1-3), quiescent (Quies3), or acetylations and simple repeats (Acet1) were most depleted for rare variants relative to the common variant enrichment (**Fig. 2.5D**).

Full-stack states show enrichment for phenotype-associated genetic variants

We next analyzed the relationship between the full-stack states and phenotypic associated genetic variants. We first evaluated the enrichment of the full-stack state for variants curated into the Genome-wide Association Study (GWAS) catalog relative to a background of common variation (Welter *et al.*, 2014) (**Methods**). This revealed six states with at least a two-fold enrichment (**Supplementary Fig. 2.43**). Four of these states, TxEx1-2,4 and TxEnh4, were all transcription associated states that are >= 10-fold enriched with coding sequences and >=11 fold for ConsHMM state 1, associated with the most constraint in a sequence alignment of 100 vertebrates (**Fig. 2.3B**). This observation is consistent with previous results that GWAS catalog variants show enrichments for coding sequence and sequence constrained bases (Arneson and Ernst, 2019; Hindorff *et al.*, 2009; Lindblad-Toh *et al.*, 2011). The other two states with greater than two-fold enrichment for GWAS catalog variants relative to common variants were two promoter states, PromF2-3 (**Supplementary Fig. 2.43**). On the other hand, four states were more than two-fold depleted for GWAS catalog variants, and were associated with artifacts (GapArtf2-3), or quiescent and polycomb repressed states with weak signals of H3K9me3 (Quies5) or

H3K27me3 (ReprPC8) (**Supplementary Fig. 2.43**). Both Quies5 and ReprPC8 are highly specific to chrX, 18.3 and 19.1 fold enriched respectively (**Supplementary Fig. 2.43**).

We also analyzed the full-stack state enrichments for fine-mapped variants previously generated from a large collection of GWAS studies from the UK Biobank database and other public databases (J. Wang *et al.*, 2020). Specifically, we considered separately the fine mapped variants from two fine-mapping methods, CAVIAR (Chen *et al.*, 2015) and FINEMAP (Benner *et al.*, 2016), for 3052 traits. For each method and trait, we identified the single variants that had the greatest probability of being causal at a set of distinct loci, and computed the enrichment of these variants for the full-stack states relative to a background of common variants (**Methods**). Fold enrichment results of full-stack states for the most likely causal variants were highly consistent between fine-mapping methods (FINEMAP and CAVIAR) (**Supplementary Fig. 2.44**). The ten states maximally enriched with fine-mapped variants relative to common variants, which were the same states by both methods, included five states associated with flanking and bivalent promoter activities (PromF2-5, BivProm4), an enhancer state associated with blood and thymus (EnhA9) and an enhancer state associated with most other cell types except blood cell types (EnhA1), and three highly conserved transcription-associated states (TxEnh4,6, TxEx4) (**Fig. 2.5E**). Notably, five of 10 states maximally enriched with fine-mapped variants, PromF2-5, and BivProm4, were associated with promoter regions and also among the 19 states most enriched with top 1% prioritized variants by at least two of the 14 different variant prioritization scores (**Fig. 2.5C, E**). These results show that there are agreements in the types of full-stack states preferentially overlapped by phenotype-associated fine mapped variants and variants predicted to have greater effects based on variant prioritization scores. We also confirmed that the full-stack model consistently had higher AUROC in predicting locations of fine-mapped variants within a background of common variants, compared to the concatenated and independent annotations in all cell types (**Supplementary Fig. 2.45-46**).

Full-stack states show enrichments for cancer-associated variants

In addition to investigating germline variants, we also investigated the enrichment of full-stack states for somatic variants identified from whole genome sequencing of cancer samples. We analyzed data of variants from four cancer types that have the largest number of somatic variants in the COSMIC database (Tate *et al.*, 2019): liver, breast, pancreas and haematopoietic_and_lymphoid_tissue (**Methods**). Sixteen states were among the top 10 most enriched with at least one type of cancer's associated variants (1.2-1.4 fold in breast cancer, 1.2-5.6 fold in lymphoid cancer, 1.2-5.4 in liver cancer, 1.4-4.2 in pancreas cancer) (**Fig. 2.5F**). Among these 16 states, 15 states showed higher signals of H3K9me3 compared to most other chromatin marks, including seven states in heterochromatin group (HET1-2, 4-7,9), four states in quiescent group with weak emissions of H3K9me3 (Ques1-2,4-5), one state in the polycomb repressed group with weak signals of H3K9me3 and H3K27me3 (ReprPC8), one state in the acetylation group with signals of H3K9me3 and various acetylation marks (Acet1), two artifact-associated states with higher signals of H3K9me3 and DNase relative to other marks (GapArtf2-3) (**Fig. 2.5G**). This pattern of H3K9me3-associated states being enriched with somatic mutations in cancer was previously confirmed in multiple studies where H3K9me3 and other repressive epigenetic features showed positive association with mutation density across different types of cancer cells (Parker *et al.*, 2012; Polak *et al.*, 2015; Schuster-Böckler and Lehner, 2012). One possible explanation for this association is the more limited access of DNA mismatch repair machinery in these regions due to the tightly packed nature of the genome in heterochromatin (Martincorena and Campbell, 2015; Supek and Lehner, 2015). Notably, the GapArtf2-3 states, which had strong satellite repeats enrichments (**Fig. 2.4C, Supplementary Fig. 2.28**) were the top two most enriched states with somatic variants associated with liver, pancreas and haematopoietic and lymphoid tissue (haem-lymphoid) cancers (2.0-5.6 folds enriched) (**Fig. 2.5F, Supplementary Fig. 2.47**). We suspect that the enrichments in these putative alignment artifact states are driven at least in part by false variant calls due to sequence mapping errors associated with these regions. Similarly, enrichments of somatic mutations in haem-lymphoid cancer in state

Acet1 is also suggestive of the possibility of false calls given this state's combination of H3K9me3 and acetylation signal and enrichment for simple repeats (**Fig. 2.2, 4C, Supplementary Fig. 2.42**). We note that the presence of cancer variants is better recovered by full-stack annotation as compared to the concatenated and independent chromatin state annotations for all four cancer types (**Supplementary Fig. 2.48-49**).

Discussion

We demonstrated a large-scale application of the stacked modeling approach of ChromHMM using over a thousand epigenomic datasets to annotate the human genome. In the datasets, 32 chromatin marks and 127 reference epigenomes were represented. We note that even though not every chromatin mark was profiled in every reference epigenome, we were still able to directly apply the stacked modeling to such data. Previously, concatenated models were applied to observed and imputed data (Ernst and Kellis, 2015), however, we chose not to use imputed data as input to the full-stack model primarily since imputed data would still be based on the same observed input data used in stacked-modeling. We conducted extensive enrichment analyses of the states with various other genomic annotations and datasets, including gene features, genetic variation, repetitive elements, comparative genomic annotations, and bases prioritized by different variant prioritization scores. These analyses highlighted diverse enrichment patterns of the states. Using these enrichments along with the model parameters, we provided a detailed characterization of each of the 100 states in the model.

We grouped these 100 states into 16 groups that included promoters, enhancers, transcribed regions, polycomb repressed regions, zinc finger genes among others. We also highlighted important distinctions among states within the groups. In many cases, identifying these distinctions was enabled by the full-stack modeling using data from multiple cell types for genome annotation. For example, we identified enhancer and repressive states that were active in different subsets of cell types. We also highlighted how different states in some of the groups such as those associated with transcribed and ZNF genes showed distinct enrichments for

conservation states. Overall, the full-stack model showed enrichment patterns supporting observations based on concatenated or independent annotations, while providing a more detailed stratification of genomic regions into chromatin states with more refined associations with other genomic information. We provide extensive characterizations of full-stack states in **Supplementary Data 2.1-5** that we expect will be a resource in future applications of the full-stack annotations.

The full-stack modeling has advantages to commonly used concatenated and independent chromatin state annotations in several respects. First, the full-stack model learns patterns of signals of the same or different assays across cell types, hence can provide a unified view of all the data and directly uncover states that correspond to constitutive or cell-type-specific activities. For example, a state from the model, HET9, was associated with only the mark H3K9me3 specifically in ESCs and iPSCs even though this mark is typically associated with constitutive repression. Second, the full-stack annotation consistently showed better recovery of various genomic features compared to concatenated and independent annotations. This improvement is expected since full-stack models can leverage information from multiple cell types for genome annotations. Third, in cases where it is not desirable to focus on only one specific cell or tissue for analysis, the full-stack modeling can bypass the need to pick one such cell or tissue type or to consider a large number of different concatenated or independent chromatin state annotations simultaneously. Such cases may arise when studying other genomic information that is not inherently cell-type-specific such as genome variation and sequence conservation. Overall, the full-stack model provides a universal annotation of the genome that can be viewed as a single track in a genome browser or used for a variety of downstream bioinformatic analyses.

Despite these advantages, there are trade-offs in using the stacked modeling approach, and we emphasize that the full-stack annotations should be considered a complement to and not a replacement of the concatenated or independent annotations. Compared to typical concatenated models, the full-stack model has increased model complexity because of the

increased number of parameters from the larger number of states and input features, which can make interpreting some model states relatively more difficult. Additionally, if one is interested in a specific cell type, then corresponding concatenated or independent annotations can have advantages in that all the annotations are directly informative about the chromatin state in the cell type of interest. An additional trade-off is that with the stacked model, it is not possible to incorporate additional data without relearning a model, while for a concatenated-model one can annotate a new cell based on an existing model, provided that the set of marks in the new cell type are the same as the existing model. We also note that post-hoc concatenated state annotations can also be used for cross-cell type analyses, particularly on a per-state basis by analyzing the frequency of a specific state across cell types or potentially in other ways. While per-state analyses using concatenated annotations can be relatively straightforward and informative, they give only partial and potentially oversimplified views of all the data, ignoring distinctions among different states. Whether to use concatenated, independent or full-stack annotations will depend on the specific application. Concatenated or independent annotations may be preferable when one is interested in studying a specific cell type, while full-stack annotations may be preferred in joint analyses of multiple cell types.

We expect many applications of the full-stack annotations that we generated here and they have already begun to be applied in other work (Arneson *et al.*, 2021; Horvath *et al.*, 2021; Li *et al.*, 2021), which we expect to further elucidate the biological significance of different states. The full-stack annotation can be used as a resource to interpret genetic variation. A possible avenue for future work is to incorporate the full-stack annotation into scoring methods to better predict genetic variants' phenotypic influences. Given the increasing availability of epigenomic datasets (Stunnenberg *et al.*, 2016), future work could also learn new stack-models to incorporate such data. The state characterizations (**Supplementary Data 2.1-5**) and analyses introduced through this work will be useful in interpreting biological implications of new models' states. Future work can also include training and deriving the full-stack annotations for key model organisms

such as mice. This work provides a new annotation resource for studying the human genome, non-coding genetic variants, and their association with diseases.

Methods

Input data and processing

We obtained coordinates of reads aligned to Human hg19 in .tagAlign format for the consolidated epigenomes as processed by the Roadmap Epigenomics Consortium from https://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/. In total we obtained data for 1032 datasets and their corresponding input control data. The datasets correspond to 127 reference epigenomes, 111 of which were generated by the Roadmap Epigenomics Consortium and 16 were generated by the ENCODE Consortium. Of the 1032 datasets, 979 were ChIP-seq data targeting 31 different epigenetic marks and 53 were of DNase-seq (**Supplementary Fig. 2.2**). For each of the 127 reference epigenomes there was a single ChIP-seq input control dataset. For the 53 reference epigenomes that had a DNase-seq dataset available, there was an additional DNase control file.

We next binarized the data at 200 base pair resolution using the BinarizeBed command of ChromHMM (v.1.18). To apply BinarizeBed in stacked mode we generated a cell_mark_file input table for ChromHMM with four tab-delimited columns. The first column had the word 'genome' for all datasets, the second column contained entries of the form '<EID>-<mark>' where 'EID' is the epigenome ID and 'mark' is the mark name, the third column specifies the name of the corresponding file with aligned reads, and the fourth column is the name of the file with the corresponding control reads. Each row in the table corresponds to one of the 1032 datasets.

In order to reduce the memory and time needed to execute BinarizeBed on a large number of datasets, we split the cell_mark_file table into 104 smaller tables with each table having at most 10 entries corresponding to at most 10 datasets to be processed. This was done with a custom script, but the same functionality has been included with the '-splitcols' and '-k' flags of BinarizedBed in ChromHMM v1.22. We then ran BinarizeBed in parallel for each of these smaller

cell_mark_file tables and generated output into separate sub-directories. We ran BinarizeBed with the option '-gzip' which generates gzipped files.

To merge data from the 104 subdirectories from the previous step into files containing binarized data of all datasets, we ran the command 'MergeBinary', which we added in v1.18 of ChromHMM. We ran the command with the options '-gzip -splitrows'. The '-splitrows' option generates multiple files of merged binarized data for each chromosome, where, under the default settings that we used, each file contains data for a genomic region of at most 1MB. Splitting each chromosome into smaller regions allows the model learning step of ChromHMM to scale in terms of memory and time to the large number of input data tracks (i.e. features) that we were using. We used chr1-22, chrX, chrY, and chrM in the binarization and model learning.

We note that we chose not to use imputed data as input to the full-stack model. A main reason is, as noted above, imputed data would still be based on observed data from the 1032 datasets used for stacked modelling. Another reason is that imputed data may have artificially high correlations across the cell types, which can particularly be the case for marks that were experimentally mapped in few cell types. This could potentially cause the stacked modeling to devote many states that correspond to heavily correlated and less informative tracks of imputed data.

Training full-stack model and generating genome-wide state annotations

We learned the full-stack chromatin state model for the 1032 datasets using the LearnModel command of ChromHMM (v1.18). This version of ChromHMM includes several options that we added to improve the scalability when training with large numbers of features. One of these features was to randomly sample different segments of the genome for training during each iteration, instead of training on the full genome. This sampling strategy was previously used by ConsHMM (Arneson and Ernst, 2019), which was built on top of ChromHMM. We note that this sampling procedure can also be applied to learn concatenated models, in which case

there would be no requirement that the same segments are sampled in each cell type. However, sampling can be unnecessary for typical instances of learning concatenated models, given that it usually involves fewer different inputs to the model, fewer number of states, and in training these models, ChromHMM is able to tolerate more parallel cores without reaching memory limits.

To learn the full-stack model with input data processed as outlined above, we used ChromHMM's LearnModel command with the options '-splitrows -holdcolumnorder -pseudo -many -p 6 -n 300 -d -1 -lowmem -gzip'. The '-splitrows' flag informs ChromHMM that binarized data for a chromosome is split into multiple files, which reduces the memory requirements and allows ChromHMM to select a subset of the genome to train on for each iteration. The '-holdcolumnorder' flag prevents ChromHMM from reordering the columns of the output emission matrix, which saves time when there are a large number of features.

The '-pseudo' flag specifies that in each update of model parameters, ChromHMM adds a pseudo count of one to the numbers of observations of transition between each pair of states, presence and absence of each mark from each state, and initial state assignments of the training chromatin state sequence. This prevents model parameters from being set to zero, which is needed for numerical stability when some features are sparse and ChromHMM does not train on the full genome in each iteration.

The '-many' flag specifies to ChromHMM to use an alternative procedure for calculating the state posterior probabilities that is more numerically stable when there are a large number of features. The procedure is designed to prevent all states from having zero posterior probability at any genomic position, which can happen due to the limits of floating-point precision. The procedure does this by leveraging the observation that only the relative product of emission probabilities across states are needed at each position to determine the posterior probabilities. Specifically, for each position, the procedure initializes the product of emission probabilities for all features, i.e. the emission product, from each state to one. For each feature, the procedure then multiplies the current emission products from each state by the emission probability of the feature

in the state, and divides all the resulting products by their maximum to obtain updated emission products. We iteratively repeat these steps of multiplication and normalization until all features have been included into the calculation of relative emission products across states.

The '-p 6' flag specifies to ChromHMM to train the model in parallel using 6 processors. The '-n 300' flag specifies to ChromHMM to randomly pick 300 files of binarized data, corresponding to 300 regions of 1 MB (or less if the last segment of the chromosome was selected) for training in each iteration. The '-d -1' option specifies to ChromHMM to not require an evaluated likelihood improvement between iterations to continue training since evaluated likelihood decreases are expected, as on each iteration the likelihood is evaluated on a different subset of data. The '-lowmem' flag has ChromHMM reduce main memory usage by not storing in main memory all the input data and instead re-loading from disk when needed. The asymptotic worst-case time and memory of the model learning is discussed in the **Supplementary Data 2.7.**

Choice of number of states

We trained full-stack models with 10-120 states, in 5 state increments, using the data and procedure outlined above. For each of these models, we calculated an estimated Akaike Information Criterion (AIC) (Sakamoto *et al.*, 1986) and Bayesian Information Criterion (BIC) (Neath and Cavanaugh, 2012) value based on a subset of the genome (**Supplementary Fig. 2.3**). AIC and BIC are calculated based on the log likelihood for 300 random 1Mb regions outputted by ChromHMM from the last training iteration. In general, both the AIC and BIC decrease as the numbers of states increase, but with diminishing improvements. We also applied the CompareModels command of ChromHMM (Ernst and Kellis, 2017) with the 100-state model as a reference model, which reports, for each state of the 100-state model, the maximum correlations of emission parameters between the state in the 100-state model and any state for each other model (**Supplementary Fig. 2.3C**). We conducted a similar analysis with the emission parameters of H3K4me1, hence for each state in each model, we obtained emission probabilities

of H3K4me1 in 127 cell types. For this analysis, for each of the 19 tissue groups previously defined (Meuleman *et al.*, 2015), we calculated the correlation of each state's H3K4me1 emission parameters with a binary vector indicating if the cell type in each parameter is in the tissue group (1) or not (0). We then report, for each tissue group, the maximum correlations among all states in each model (**Supplementary Fig. 2.3D**). These analyses showed, for instance, that a state corresponding to Brain-specific enhancers in the 100-state model, EnhA6, was well captured in models with 55-states or more (correlation of >=0.98 with states in models with >=55 states and correlation >=0.78 for H3K4me1-emissions with the Brain binary vector). A state characterized as enhancers specific to Huvec cells in the 100-state model, EnhA20, was well captured in models with 100 or more states (correlation >=1.00 based on all marks' emission parameters).

Additionally, for models with 20, 40, 60, 80, 100 and 120 states, we also produced genome annotations and then quantitatively compared the chromatin state annotations from models in terms of their power to predict locations of various other genomic annotations not used in the model training: Exon, Gene Body, TSS, TSS2kb, CpG Islands, TES, laminB1lads elements (listed in section *External Annotation Sources* section). Specifically, we evaluated the predictive power using the AUROCs that are calculated as described in a subsection below. Across different genomic contexts, as the number of full-stack states increased, the AUROC increased, but with diminishing improvements as the number of states increased (**Supplementary Fig. 2.3A**).

To balance the additional information available in models with an increased number of states, while keeping the number of states manageable for interpretation and downstream analysis, we choose to focus on a model with 100 states. We note that this choice is greater than previously used for concatenated models (Ernst *et al.*, 2011; Ernst and Kellis, 2010; Meuleman *et al.*, 2015), which reflects the additional information available for genome annotation based on the large number of datasets spanning many cell types that we are using.

Lifting chromatin state annotations to hg38

The full stack chromatin state annotations were learned directly in hg19, as this was the assembly for which uniformly processed data from the Roadmap Epigenomics integrative analysis was available. Learning full-stack annotation directly in hg19 allowed direct comparison with existing concatenated annotations. We also generated a version of full-stack annotation in hg38 by lifting over the original annotation from hg19 to hg38. To do this, we first wrote the hg19 chromatin state annotation into .bed format such that each line corresponds to a 200bp interval. We then used the liftOver tool (Kent *et al.*, 2002) with default parameters to generate the annotation in hg38. We did not annotate bases in hg38, if multiple bases in hg19 mapped to it. In total, there are 1,186,379 200-bp segments that were not mapped from hg19 to hg38, of which 98.7% fall into an assembly gap and 99.6% fall into the full-stack state primarily associated assembly gaps (GapArtf1) (**Supplementary Fig. 2.50**). In hg38 on chr1-22, X, and Y, 92.9% of bases are annotated to a state, and that number increases to 97.1% when excluding assembly gaps. We verified that we saw highly similar state fold enrichments for similar annotations between hg19 and hg38 (**Supplementary Fig. 2.18**). The sources of external annotations from hg38 are outlined in section “External annotation sources” below.

Summary sets of datasets

To construct a summary visualization of the emission parameters with a reduced set of features that approximate the annotation from the full model, we applied a greedy search over the 1032 input datasets as described in **Supplementary Data 2.7.** We applied this procedure to reduce the 1032 input datasets to 80 summary datasets.

Identifying states with differential association of marks for individual tissue groups

For each state, we tested for combinations of the 8 most profiled marks, and 19 tissue groups previously defined (Meuleman *et al.*, 2015), whether the emission probabilities of features associated with one chromatin mark and in one tissue group was significantly greater than those

of features associated with the same mark and not in the tissue group. The eight marks that we tested were H3K9me3, H3K4me1, H3K4me3, H3K27me3, H3K36me3, H3K27ac, H3K9ac, and DNase. H3K27ac, H3K9ac and DNase were profiled in 98, 62 and 53 reference epigenomes, respectively, and the remaining five marks in 127 reference epigenomes. For tests involving H3K27ac, H3K9ac, and DNase, we excluded tissue groups for which there were no datasets. In total, there were 14,200 tests among 100 states, 8 chromatin marks and 19 tissue groups. For each combination of state, chromatin mark and tissue group being tested, we applied a one-sided Mann-Whitney test to test whether the emission probabilities of the state for the features associated with the tested mark in the tested tissue group are greater than those in other tissue groups. The Bonferroni-corrected p-value threshold based on a significance level of 0.05 to declare a test significant was 3.5e-6.

Computing coefficients of variation across different tissue groups

For each state, we looked into the emission probabilities of datasets associated with six chromatin marks strongly associated with promoter and enhancer activities (DNase, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac). We grouped these datasets based on their associated chromatin mark and tissue groups, and calculated the average emission probabilities of datasets in each chromatin mark-tissue group combination. For each state and chromatin mark combination, we then calculated the coefficient of variation across different tissue groups, in terms of average emission probabilities from the previous step. For each group of states, we averaged the resulting coefficients of variation across states of the same group. The results show the average coefficients of variation of emission probabilities across different tissue groups for each state group-chromatin mark combination.

Computing fold enrichments for other annotations

All overlap enrichments for external annotations were computed using the ChromHMM OverlapEnrichment command. We used the '-b 1' flag, which specifies a binning resolution of the annotations. This '-b 1' flag is necessary when computing enrichments based on the hg38 liftOver annotations, which no longer respects the 200bp segment coordinate intervals from hg19. Including this flag gives the same results when applied to annotations from hg19 with 200bp segments, though with extra computational costs. We also included the '-lowmem' flag to specify the lower memory usage option. The ChromHMM command OverlapEnrichment computes fold enrichment between chromatin states and provided external annotations relative to a uniform genome-wide background distribution. More specifically, the fold enrichments are calculated as:

$$FE_{x,s} = \frac{\frac{\#SX}{\#X}}{\frac{\#S}{\#G}} = \frac{\frac{\#SX}{\#S}}{\frac{\#X}{\#G}} = \frac{\#SX \cdot \#G}{\#S \cdot \#X}$$

where

$FE_{x,s}$: fold enrichment of state *s* in genomic context *x*

$\#S$: number of genomic positions belonging to the state *S*

$\#X$: number of genomic positions where genomic context *X* is present

$\#SX$: number of genomic bins that overlap both state *S* and genomic context *X*

$\#G$: number of genomic positions in the entire genome

Enrichment and estimated probabilities of overlap with 25-state concatenated annotations

We obtained per-cell type chromatin state annotations based on a 25-state ChromHMM model learned using the concatenated approach for 127 reference epigenomes, which we will refer to as cell types for ease of presentation, from the Roadmap Epigenomics project (Ernst and Kellis, 2015; Meuleman *et al.*, 2015). This model was trained based on imputed data for 12-marks. We hereafter refer to this model as the CT-25-state model. As per the design of the concatenated approach of ChromHMM, CT-25-state model generates per-cell-type chromatin state annotations for each of the 127 cell types, and the 25 states' characteristics are shared across 127 cell types.

For each of these 127 cell types, we calculated overlap enrichments between the 100 full-stack states and the CT-25-states, resulting in 127 tables of size 100-by-25. We summarized this information by reporting, for each of the 100 full-stack states, and 127 cell types, the state in CT-25-state model that is maximally enriched, resulting in a 100-by-127 table (**Supplementary Fig. 2.8**). We also provided detailed comments about the patterns of maximum-enriched-states observed across 127 cell types for each full-stack state in the **Supplementary Data 2.4** to serve as a resource for future applications. We also reported, for each of the 100 full-stack states and each concatenated-25-state, the maximum and median values of fold enrichments across 127 cell types (**Supplementary Data 2.4**).

In addition, we also estimated for each combination of (1) CT-25 state, (2) cell type group and (3) full-stack state, the probability that a genomic position being annotated as the corresponding full-stack state will overlap with the corresponding per-cell type state in a cell type from the corresponding cell group. The 19 groups of cell types were previously defined by the Roadmap Epigenomics Consortium (Meuleman *et al.*, 2015). To compute the target estimate probabilities, for each full-stack state, we sampled 100 genomic bins (each of length 200bp) that are assigned to that full-stack state. Second, in each of the 127 cell types, we report the frequency that the sampled regions of each full-stack state overlapped with each CT-25-state. We repeated such a process 21 times. We then calculated the average frequencies of overlap between each full-stack state and each CT-25-state, across 21 random samplings and across the cell types in each group (Example: Blood, ESC). This results in a table of size 100 full-stack states by 475 combinations of CT-25 states and 19 cell groups, with each cell showing the values of estimated probabilities (**Supplementary Fig. 2.9**). These values, along with detailed comments about patterns of these overlap probabilities for each full-stack state, are available in **Supplementary Data 2.4**.

Receiver operator characteristic curve analysis for predicting external annotations

To evaluate how well the chromatin state annotations from different ChromHMM models can inform us about the position of external genomic annotations, we computed the Receiver Operator Characteristic (ROC) in a procedure as follows: First, we divided the genome into 200bp bins, and randomly partitioned 50% of the bins for training and the remaining 50% for testing. Second, we computed the enrichment of the target external annotation with each chromatin state on the training data, and ranked states in decreasing order of such enrichments. We used this ranking of states to iteratively add genomic bases assigned to the states as our predictions of bases that overlap the target annotation in the testing dataset. Based on the overlap of the predictions and the target annotation at each iteration, we plotted ROC curves and summarized the information by computing area under the ROC curves (AUROC).

Concatenated and independent annotations used to compare against full-stack annotations

In evaluating how predictive the full-stack model is at annotating external genomic elements, we compared the full-stack model to two sets of per-cell-type chromatin state annotations in terms of their ability to predict external annotations. One set of annotations was the 18-state ChromHMM from Roadmap Epigenomic Project (Meuleman *et al.*, 2015), which was based on a model trained using the concatenated approach and observed data of six chromatin marks (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3) in 98 cell types. In this model, we have a common set of state definitions across cell types, but unique state annotations for each cell type. The second set of annotations were based on models learned independently in each of the 127 cell types. In learning these models, we partitioned the 1032 datasets used to learn the full-stack model into 127 subsets based on their associated cell type. For each of the 127 cell types, we applied ChromHMM to learn a 100-state ChromHMM model using only the observed data in the corresponding cell type. The number of states is similar to that in the full-stack model, to control for this variable in the evaluation. This process generates 127 models, each used to generate independent annotations in one cell type. The independent

model learning approach of ChromHMM differs from the concatenated approach because the model parameters (state emission, transition, and initial probabilities) are different for different cell types, while state parameters in concatenated model are shared across cell-types. However, these two approaches both produce chromatin state annotations on a per-cell type basis. We learned these independent models with the same ChromHMM parameters as described above for the full-stack model, with the exception of using the '-init random' flag to randomly initialize models' parameters. Even when we specified the number of states to ChromHMM as 100, however, we note that due to the large number of states relative to the input tracks, for some of these models, fewer than 100 distinct states ended up being assigned to positions in the genome.

Computing fine-mapped variant enrichment

To compute enrichment of full-stack states for phenotypically associated fine-mapped variants, we downloaded data on fine-mapped variants for 3052 traits from CAUSALdb (J. Wang *et al.*, 2020). Specifically we obtained posterior probabilities of variants being causal based on two fine-mapping methods, FINEMAP (Benner *et al.*, 2016) and CAVIAR (Chen *et al.*, 2015), which do not use epigenomic annotations as part of the fine mapping procedure. For each method and trait combination, we separately partitioned the provided set of potential causal variants into distinct loci. To form the distinct loci, we merged neighboring variants into the same loci until there was at least 1MB-gap between the two closest variants from different loci. Separately for each fine-mapping method, trait, and locus combination, we selected the single variant with the highest posterior probability of being causal. For each fine-mapping method, we took the union of variants across 3052 traits, and then calculated the fold enrichments for the union of these lead variants with stacked ChromHMM states relative to the enrichment with a background set of common variants from dbSNP build 151 (hg19). To do this, we separately computed the enrichments of both of these sets relative to a genome-wide background, and then divided the enrichment of the foreground set (lead fine-mapped variants) by the enrichment of the background set (common

variants). The dbSNP variants were obtained from the UCSC genome browser (Navarro Gonzalez *et al.*, 2021).

Computing structural variant enrichments

To compute enrichment of the full-stack states for structural variant enrichments, we obtained data of structural variants from (Abel *et al.*, 2020). We used the B38 call set, which was in hg38 and used for the analysis presented in (Abel *et al.*, 2020). We filtered out structural variants that did not pass the quality control criteria of (Abel *et al.*, 2020). We then separately considered structural variants annotated as either a deletion or a duplication, for which there were, 112,328 and 28,962 sites respectively.

Since the structural variants were defined in hg38, we computed their enrichment for ChromHMM state annotations from full-stack, concatenated and independent models that were lifted over from hg19 to hg38, following the procedure outlined above. Next, we followed the enrichment analysis procedure outlined above to compare full-stack vs. concatenated and independent chromatin state annotations' power in recovering structural variants.

To compare the power of full-stack state annotations vs. concatenated state annotation frequency, we utilized the 15-state concatenated chromatin state annotation for 127 cell types (reference epigenomes) from Roadmap Epigenomics Consortium. We followed the analysis outlined in (Abel *et al.*, 2020), for each of the 15 concatenated-model states, we annotated genomic positions based on the number of cell types in which the state is present (ranging from 0 to 127), resulting in 15 state frequency annotations per genomic position. We then applied the procedure above for each concatenated-model state to compare the predictive power of the state's annotation frequency against the full-stack annotation.

Computing enrichments with cancer-associated variants

We obtained data of somatic mutations associated with different types of cancer from COSMIC non-coding variants dataset v.88 in hg38 (Tate *et al.*, 2019). We selected from this dataset variants that were from whole-genome sequencing. We filtered out variants that overlap with any of the following: the hg38 black-listed regions from the ENCODE Data Analysis Center (DAC) (Amemiya *et al.*, 2019), hg38 dbSNP (v151) set of common variants from the UCSC genome browser database, or regions annotated as coding sequence ('CDS') based on GENCODE v.30 hg38 (Harrow *et al.*, 2012) gene annotations. We decided to restrict this analysis to the four cancer types with most number of variants present in the dataset in hg38: liver (1,351,417), pancreas (500,930), haematopoietic and lymphoid tissue (354,501), and breast (323,751). We then lifted over these sets of variants from hg38 to hg19, resulting in 1,351,159, 500,798, 354,351, and 323,685, variants respectively. To obtain a background set of genomic locations for the enrichment analysis, we filtered from the genome the same set of hg38 annotations of black-listed regions, common variants, and coding sequences as we did for the foreground of COSMIC mutations. We then lifted over these remaining positions from hg38 to hg19 to obtain the background. We calculated the enrichment of chromatin states with cancer-associated variants by first calculating the enrichment values of chromatin states with filtered variants associated with each of the four cancer types, and the enrichment values with background set of genomic bases, all relative to the whole genome. We then divided the cancer-associated variant enrichment values by the background bases enrichments.

Gene ontology enrichments

We calculated the GO enrichments of genes being in proximity to each full-stack state annotation using GREAT (McLean *et al.*, 2010). For each full-stack state, we reported the 5 GO Biological Process and 5 GO Molecular Function with lowest FDR-corrected p-values, ranked by GREAT (McLean *et al.*, 2010). All bar plots showing the top GO terms and negative log 10 p-values of enrichments with full-stack states are available in **Supplementary Data 2.1.**

External annotations sources

The sources for external annotations for enrichments analyses, not given above, were as follows (all download links are listed in **Supplementary Data 2.8**):

- Annotations of CpG islands, exon, gene bodies (exons and introns), transcription start (TSS), and transcription end sites (TES), 2kb windows surrounding TSSs (TSS2kb) in hg19 and hg38 were RefSeq annotations included in ChromHMM (v1.18) and originally based on annotations obtained from the UCSC genome browser on July 26th 2015.
- Lamina associated domains were for human embryonic lung fibroblasts that were included in ChromHMM (1.18), which were lifted over to hg19 from hg18 positions originally provided by (Guelen *et al.*, 2008).
- Annotations of assembly gaps in hg19 and hg38 were obtained from the UCSC genome browser and correspond to the Gap track.
- Coordinates of zinc finger genes correspond to non-overlapping coordinates from GENCODE's hg19 gene annotation, v30 (Harrow *et al.*, 2012). ZNF named genes were those whose gene named contained 'ZNF'. The list of C2H2-type genes were from https://www.genenames.org/.
- Annotations of coding sequences in hg19 and hg38 correspond to coordinates of genes whose feature type is 'CDS' from GENCODE's hg19 and hg38 gene annotation, v30 (Harrow *et al.*, 2012).
- Annotations of pseudogenes in hg19 and hg38 correspond to coordinates of genes whose gene type or transcript type contained 'pseudogene' from GENCODE's hg19 and hg38 gene annotation, v30 (Harrow *et al.*, 2012).
- Annotations of repeat elements were obtained from UCSC genome browser RepeatMasker hg19 tracks.

- Concatenated ChromHMM chromatin state annotations were obtained from the Roadmap Epigenomics Consortium through http://compbio.mit.edu/roadmap (Meuleman *et al.*, 2015). These include data of the 18-state models based on observed data and the 25-state chromatin model based on imputed data for 98 and 127 reference epigenomes, respectively.
- CTCF- concatenated chromatin states were based on the ChromHMM chromatin state annotations for six human cell types (GM12878, H1ESC, Helas3, Hepg2, Huvec, K562) for a 25-state model from the ENCODE integrative analysis (Ernst and Kellis, 2012; Hoffman *et al.*, 2013). We extracted coordinates of regions annotated to the 'Ctcf' and 'CtcfO', both associated with CTCF signal and limited histone mark signal.
- Blacklisted regions were those provided by the ENCODE Data Analysis Center (DAC) for hg19 and hg38 (Amemiya *et al.*, 2019).
- ConsHMM conservation state annotations for human (hg19) were those from (Arneson and Ernst, 2019).
- Annotations of human genetic variants and their allele frequency were from GNOMAD v2.1.1 (Karczewski *et al.*, 2020). The dataset includes 229 million SNVs and 33 million indels from 15,708 genomes of unrelated individuals, which are aligned against the GRCh37/hg19 reference.
- GWAS catalog variants were obtained from the NHGRI-EBI Catalog, accessed on December 5th, 2016 (Welter *et al.*, 2014).
- Coordinates of CpG sites profiled across cell types were obtained from DNA Methylation data in Roadmap Epigenomic portal.
- Data of G/C content at 5bp resolution from UCSC Genome Browser, file hg19.gc5Base.txt.gz.

- Data of binding regions of proteins of the polycomb repressive complexes were downloaded from the ENCODE portal (Davis *et al.*, 2018). Download links are listed in **Supplementary Data 2.8.**

Analysis of gene expression across states

To analyze the relationship between gene expression and the full-stack states, we downloaded gene expression data from the Roadmap Epigenomics Consortium (Meuleman *et al.*, 2015). Specifically, we downloaded a matrix of gene expression values, in RPKM (Reads Per Kilobase Million), for protein coding genes for 56 reference epigenomes that were among the 127 used as part of the full-stack model. In total, we obtained expression values for 19,795 Ensembl protein coding genes.

The gene expression data was obtained from (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.exon.RPKM.pc.gz). We also obtained the corresponding genomic coordinates for these genes from (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/Ensembl_v65.Gencode_v10.ENSG.gene_info). For this analysis, we filtered out genes that are not classified as protein-coding. We transformed the gene expression values by adding a pseudo-count of 1 to the raw counts in RPKM, and taking the log of the resulting values.

For each full-stack-state and 56 reference epigenomes, we calculated the average gene expression of all genes overlapping with the state, taking into account the genes' length. For each gene g we denote its length L_g and expression E_g. We let s_i denote the state assigned at the 200-bp bin i and G_i denote the set of genes overlapping the 200bp bin i. Let B_s denote the set of 200bp bins that are assigned to state s. The average normalized expression with state s then becomes:

$$avg\ exp\ \ bp\ normalized_{\ s} = \frac{\sum_{i \in B_s} \sum_{g \in G_i} \frac{E_g}{L_g}}{\sum_{i \in B_s} \sum_{g \in G_i} \frac{1}{L_g}}$$

We also calculated for each state the average and coefficient of variation of these averages across reference epigenomes. We used the BEDTools (Quinlan and Hall, 2010) *bedtools intersect* command to obtain the chromatin state assignments for 200bp segments that totally or partially overlap with any gene. To obtain average gene expressions of a state in a cell type group as presented in **Fig. 2.3C**, we averaged the reported bp-normalized average gene expressions of the corresponding state across cell types within the group.

We also analyzed average gene expression values for each state as a function of the position of the state annotations relative to TSS, following a procedure similar to what was used previously (Ernst *et al.*, 2011). We first identified a gene's outer transcription start site (TSS) based on the reported coordinates of the gene and strand in the gene annotation file noted above. For each 200bp bin that is within 25kb upstream or downstream of an annotated TSS, including those that directly overlap with an annotated TSS, we determined the assigned full-stack state at this bin, and the position of the bin relative to those TSSs. Bins directly overlapping an annotated TSS were at position 0. If the gene was on the positive strand, the segments' genomic coordinates lower than the TSSs' correspond to upstream regions at negative points (minimum value: -250000), while genomic coordinates higher than the TSSs' correspond to downstream regions at positive points (maximum value: 25000). If the gene is on the negative strand, the upstream and downstream positions are reversed. For each state and each 200-bp bin position relative to TSS, we determined the subset of genes where there is a 200bp bin annotated to that state at that position relative to their TSSs, and calculated their average expression. This produces a 100-by-251 table for one reference epigenome, corresponding to the number of full-stack states and 200-bp segments intersecting the 50kb windows surrounding genes' TSSs and one segment directly overlapping the TSSs. We then smoothed the averaged expression data spatially by applying a sliding window with a window size of 21, i.e. each segment's smoothed gene expression is the

average of data in that segment and 21 surrounding genomic segments. Data of average gene expression in the first and last 10 segments within the 50kb window are not included in the window of smoothed data. We averaged results of 56 tables corresponding to 56 reference epigenomes as the final output from this procedure.

Computing average DNA methylation levels

The DNA methylation analysis was conducted based on Whole Genome Bisulfite Sequencing data from Roadmap Epigenomics (Meuleman *et al.*, 2015). The fraction DNA methylation values was obtained from https://egg2.wustl.edu/roadmap/data/byDataType/dnamethylation/WGBS/FractionalMethylation.tar.gz. For each combination of 37 reference epigenomes with DNA methylation available and 100-full-stack states, the average fractional DNA methylation in that reference epigenome was computed for all CpG bases with a non-missing DNA methylation value overlapping the full-stack state annotation.

Computing enrichment for bases prioritized by variant prioritization scores

To compute state enrichments for bases prioritized by different variant prioritization scores, we followed the approach of (Arneson and Ernst, 2019). We obtained coordinates of bases containing prioritized variants based on 14 different methods as processed and described in (Arneson and Ernst, 2019). The scores were Eigen and Eigen-PC version 1.1, funSeq2 version 2.1.6, CADD v1.4, REMM, FIRE, fitCons, CDTS, LINSIGHT, FATHMM-XF, GERP++, phastCons, phyloP and DANN (Cooper *et al.*, 2010; Di Iulio *et al.*, 2018; Fu *et al.*, 2014; Gulko *et al.*, 2015; Huang *et al.*, 2017; Ioannidis *et al.*, 2017; Ionita-Laza *et al.*, 2016; Pollard *et al.*, 2010; Quang *et al.*, 2015; Rentzsch *et al.*, 2019; Rogers *et al.*, 2018; Siepel *et al.*, 2005; Smedley *et al.*, 2016). For 12 of the 14 scores, we separately considered prioritized variants genome-wide and in non-coding regions only. Two of the variant prioritization scores, LINSIGHT and FunSeq2, were defined only in the non-coding regions, so these scores were only used in the non-coding region

analysis. As described in (Arneson and Ernst, 2019), the regions included in the non-coding analysis were defined as the bases where both LINSIGHT and FunSeq2 provided scores, which was 90.4% of the genome. For both the non-coding and whole genome analysis we computed the enrichment for bases ranked in the top 1%, 5% or 10% using the variant prioritization scores. We note that because of ties in some scores, the score-threshold above which we classified the bases as prioritized was chosen to be as close as possible to the target percentage (1%, 5% or 10%). We also note that if there were any bases with missing values for any particular score, then that base was assigned with the minimum values of such scores.

Enrichment values for the whole genome were computed as described above with the OverlapEnrichment command from ChromHMM. For computing enrichments restricted to non-coding regions, we first calculated enrichment of the non-coding prioritized variants relative to the whole genome and the enrichment of non-coding regions as defined above relative to the whole genome. We then divided these two enrichment values to obtain the enrichment of prioritized non-coding variants within non-coding regions.

Data availability

Full-stack chromatin state annotation of the genome are available at https://github.com/ernstlab/full_stack_ChromHMM_annotations. The code used to analyze the output of ChromHMM and characterize the states is provided under the open source MIT license at (Vu and Ernst, 2021b). An archival version of this code is available at (Vu and Ernst, 2021a) under the MIT license. The software ChromHMM version 1.18 is available under the GPL 3 license at (Ernst). The most up-to-date version of ChromHMM is available at https://ernstlab.biolchem.ucla.edu/ChromHMM/. All links to download publicly available data for analyses in this paper are listed in **Supplementary Data 2.8**.

Figures

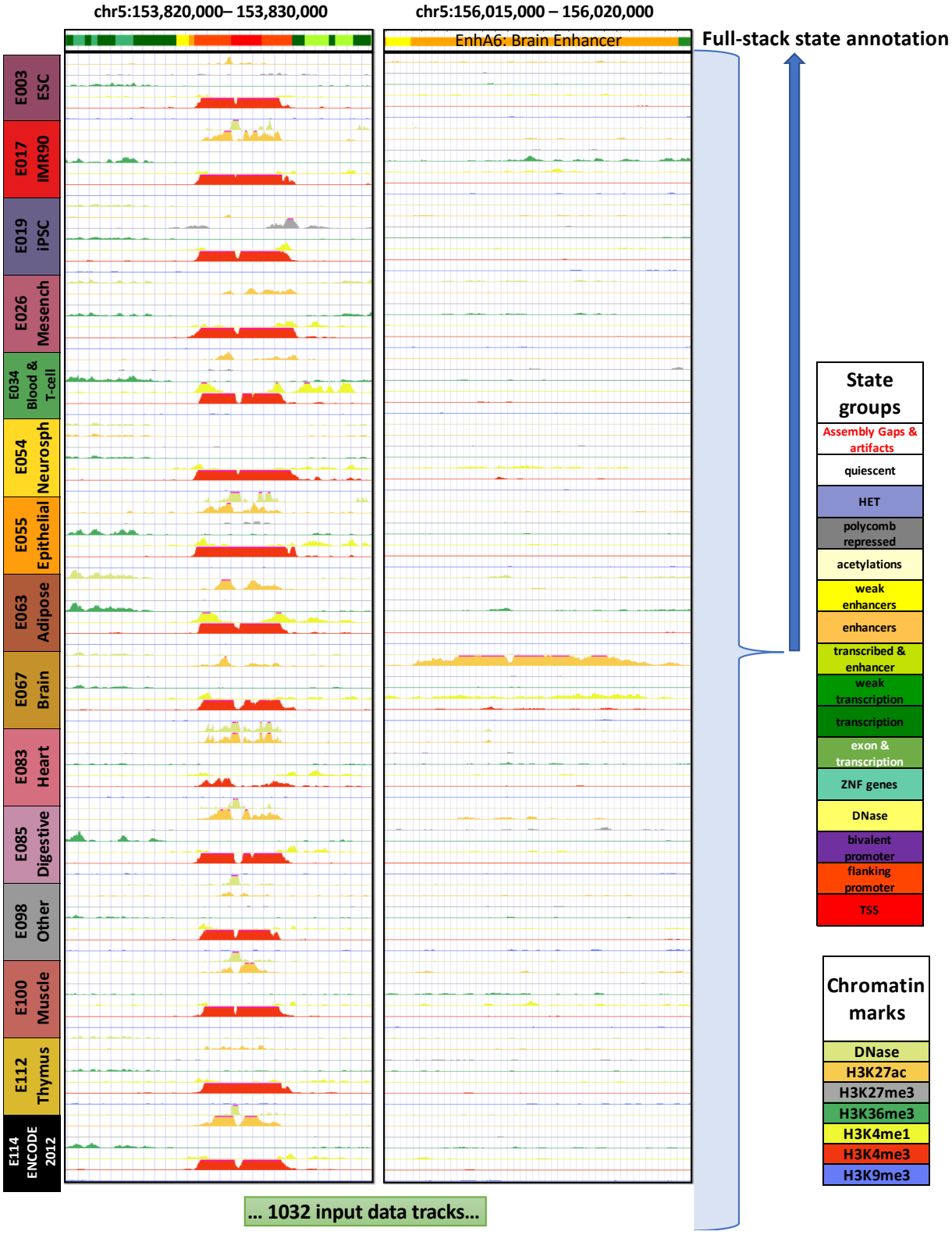


Figure 2. 1: Illustration of full-stack modeling annotations.

The figure illustrates the full-stack modeling at two loci. The top track shows chromatin state annotations from the full-stack modeling colored based on the legend at right. Below it are signal tracks for a subset of the 1032 input datasets. Data from seven (DNase I hypersensitivity, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, and H3K9me3) of the 32 chromatin marks are shown, colored based on the legend at right. These data are from 15 of the 127 reference epigenomes each representing different cell and tissue groups. The loci on left highlights a genomic region for which a portion is annotated as constitutive promoter states (TSS1-2). The loci on the right panel highlights a region for which a portion is annotated as a brain enhancer state (EnhA6), which has high signals of H3K27ac in reference epigenomes of the group Brain. The concatenated model annotations for these loci from these and additional reference epigenomes can be found in **Supplementary Figure 2.24.**

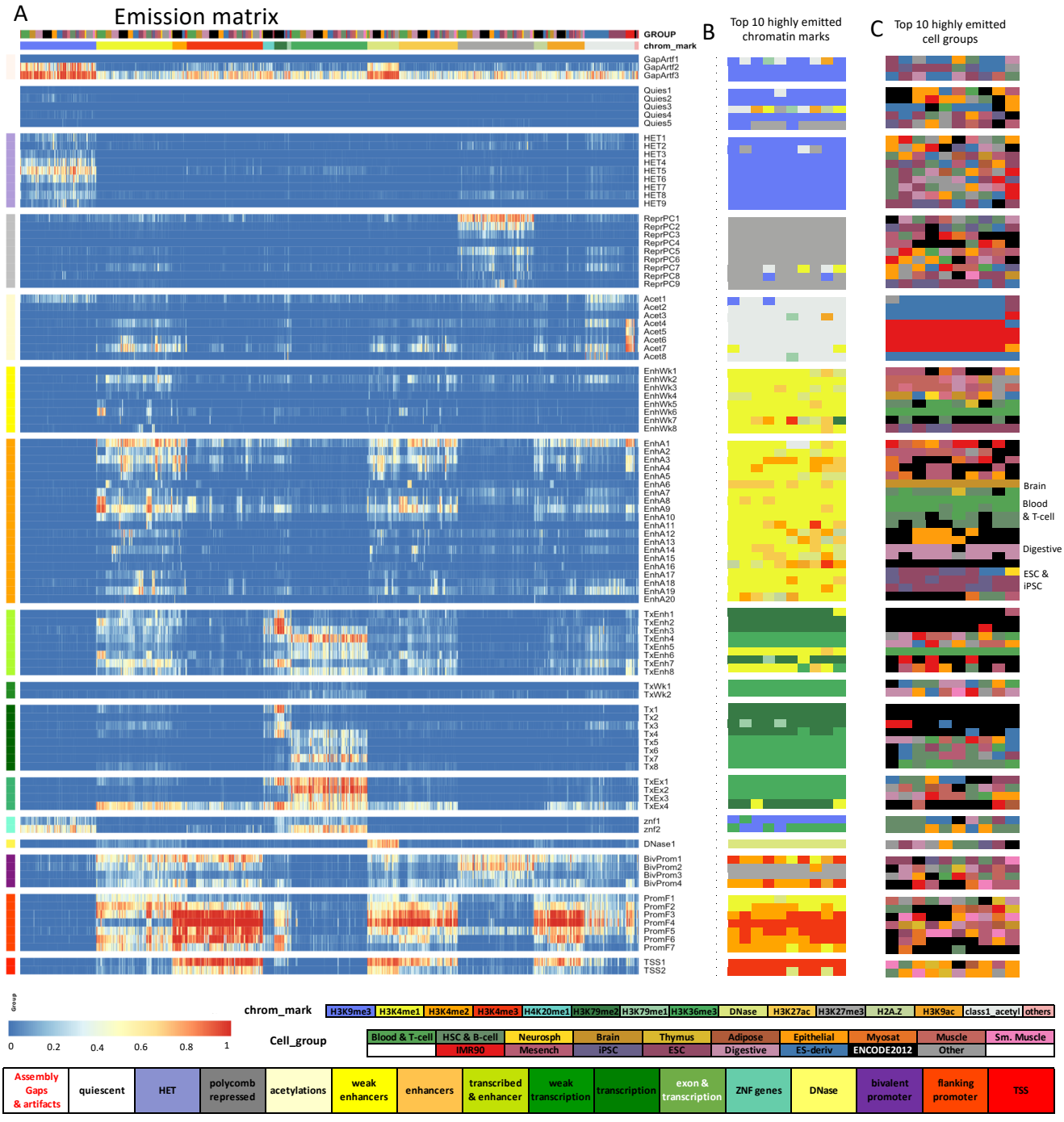


Figure 2. 2: Full-stack state emission parameters.

(A) Each of the 100 rows in the heatmap corresponds to a full-stack state. Each of the 1032 columns corresponds to one dataset. For each state and each dataset, the heatmap gives the probability within the state of observing a binary present call for the dataset's signal. Above the heatmap there are two rows, one indicating the cell or tissue type of the dataset and the other

indicating the chromatin mark. The corresponding color legends are shown towards the bottom. The states are displayed in 16 groups with white space between each group. The states were grouped based on biological interpretations indicated by the color legend at the bottom. Full characterization of states is available in **Supplementary Data 2.1-5.** The model's transition parameters between states can be found in **Supplementary Figure 2.6.** Columns are ordered such that datasets profiling the same chromatin marks are next to each other.

(B) Each row corresponds to a full-stack state as ordered in (A). The columns correspond to the top 10 datasets with the highest emission value for each state, in order of decreasing ranks, colored by their associated chromatin marks as in (A).

(C) Similar to **(B)**, but datasets are colored by the associated cell or tissue type group. On right, the cell or tissue groups primarily associated with some of the enhancer states is noted.

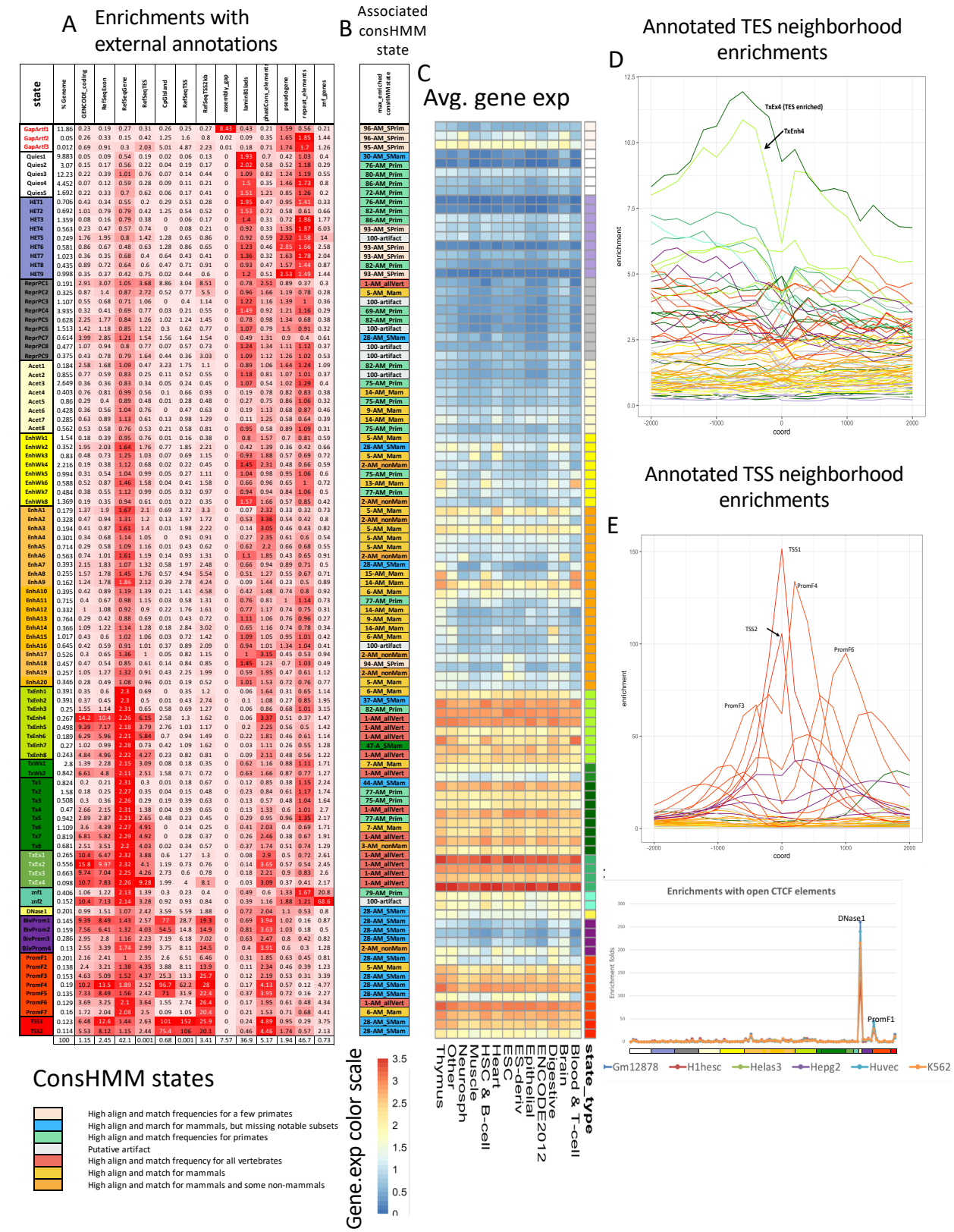


Figure 2. 3: Full-stack states enrichments for external genomic annotations.

(A) Fold enrichments of full-stack states with external genome annotations (**Methods**). Each row corresponds to a state and each column corresponds to one external genomic annotation: CpG Islands, Exons, coding sequences, gene bodies (exons and introns), transcription end sites (TES), transcription start sites (TSS), TSS and 2kb surrounding regions, lamina associated domains (laminB1lads), assembly gaps, annotated ZNF genes, repeat elements and PhastCons constrained element (**Methods**). The last row shows the percentage of the genome that each external genome annotation covers. The heatmap colors are column-normalized, i.e. within each column, the color of the cells are such that highest values are colored red and lowest values are colored white.

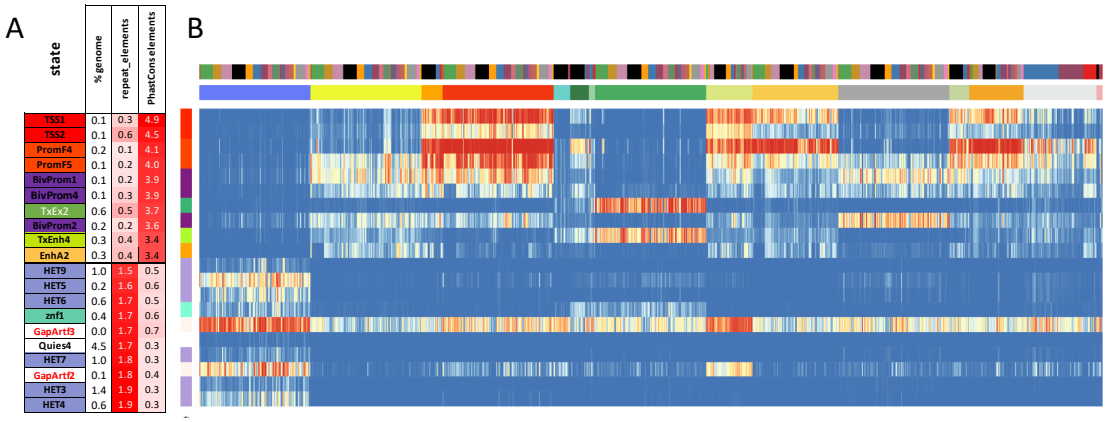
(B) Each row indicates the ConsHMM state (Arneson and Ernst, 2019) that has the highest enrichment fold in each full-stack state as ordered in **(A).** Legends of the ConsHMM state groups indicated with different colors are shown below the heatmap in **(A)**, and descriptions of select ConsHMM states curated from (Arneson and Ernst, 2019) are available in **Supplementary Data 2.7.**

(C) Average weighted expression of genes that overlap each full-stack state in different groups of cells (**Methods**). Each column corresponds to a cell group indicated at the bottom. Each row corresponds to a state, as ordered in **(A).**

(D-E) Positional enrichments of full-stack states relative to annotated **(D)** transcription end sites (TES) and **(E)** transcription start sites (TSS). Positive coordinate values represent the number of bases downstream in the 5' to 3' direction of transcription, while negative values represent the number of bases upstream. Each line shows the positional enrichments in a state. Lines are colored as indicated in **(A).**

(F) Enrichments of full-stacks states with concatenated chromatin states associated with CTCF and open chromatin, but limited histone modifications in six cell types (Hoffman *et al.*, 2013) (**Methods**). The six cell types are indicated along the bottom of the figure. States are displayed

horizontally in the same order as **(A)**. The DNase1 state showed the strongest enrichment for the concatenated chromatin states associated with CTCF and open chromatin in all six cell types.



C

state	% genome	Low_complexity	snRNA	SINE	DNA	RC	Unknown	Simple_repeat	Other	LTR	LINE	Satellite
GapArtf2	0.05	0.46	0.59	0.16	0.23	0	0	3.36	0.1	0.34	0.19	181
HET3	1.36	0.26	1.08	0.15	1.19	0.23	0.37	0.36	0.09	1.36	3.39	0.91
HET5	0.25	0.25	0.54	0.61	0.53	0.18	0.13	0.71	0.36	4.65	0.99	9.13
HET7	1.02	0.56	0.64	1.7	0.56	0.29	0.17	0.92	12.4	3.81	1.19	0.9
Acet1	0.18	0.85	0.17	0.22	0.48	0.4	0.07	23.2	0.99	1.99	0.81	0.63
EnhA2	0.33	0.59	0.56	0.49	0.74	0.48	3.18	0.57	0.02	0.55	0.27	0.01
EnhA4	0.3	0.61	1.09	0.51	1.25	2.77	2.76	0.44	0.01	0.81	0.47	0.01
Tx1	0.82	0.69	2.95	1.7	1.7	1.47	0.66	0.49	0.13	0.39	1.12	0
Tx5	0.94	0.57	1.41	3.73	0.87	0.49	0.56	0.52	0.77	0.53	0.37	0.08
TSS1	0.12	10.5	4.28	0.28	0.1	0.28	0.07	2.76	0.01	0.08	0.03	0.19
TSS2	0.11	18.4	3.55	0.5	0.19	0.35	0.35	5.58	0.17	0.3	0.09	1.04

Top enriched state with the repeat class
Enrichments >1.0

Figure 2. 4: Full-stack states enrichments with conserved elements and repeat classes.

(A) The first ten rows show the states most enriched with PhastCons elements and concurrently least enriched with RepeatMasker repeat elements, ordered by decreasing enrichments with PhastCons elements. The bottom ten rows show the states most enriched with repeat elements and concurrently least enriched with PhastCons elements, ordered by increasing enrichments with repeat elements. The columns from left to right list the state ID, the percent of the genome that each state covers, and the fold enrichments for repeat elements and PhastCons elements.

(B) Heatmap of the state emission parameters from **Fig. 2.2A** for the subset of states highlighted in panel (A). The colors are the same in **Fig. 2.2A**.

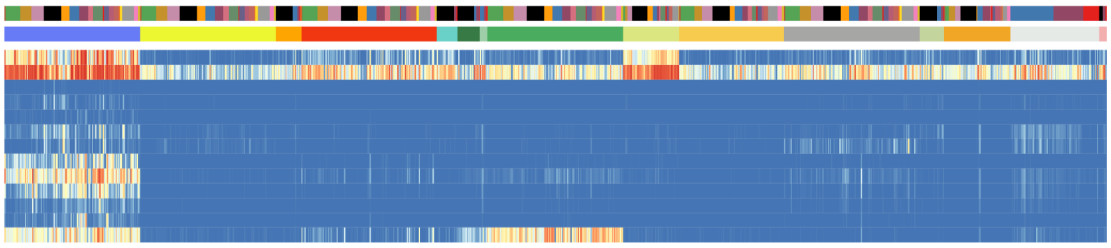
(C) Fold enrichments of full-stack states with different repeat classes (**Methods**). Rows correspond to states and columns to different repeat classes. Only states that are most enriched

with at least one repeat class are shown. Fold enrichment values that are maximal for a given are shown in dark red. Other fold enrichments greater than one are shaded light red.

A

State	Genome %	deletion	duplication
GapArtf2	0.05	1.29	1.58
GapArtf3	0.01	0.89	1.22
Quies1	10.7	1.63	1.21
Quies2	3.31	1.57	1.26
Quies4	4.8	1.23	1.14
HET1	0.76	1.58	1.3
HET2	0.75	1	1.45
HET4	0.61	1.54	1.47
HET5	0.27	1.23	1.22
HET6	0.63	1.36	1.37
HET7	1.1	1.18	1.13
HET9	1.07	1.35	1.46
znf2	0.16	1.09	1.26

B

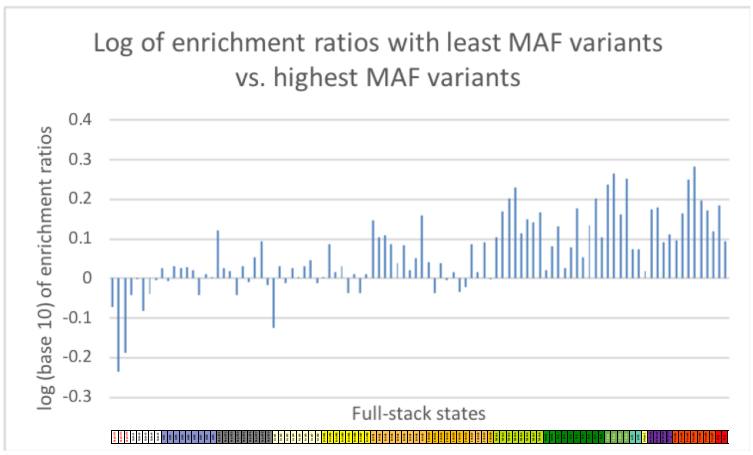


Rank of most enriched states: 1 | 2 | 3 | 4 | 5 | 6-10 | >10

C

state	% genome	FIRE	fitCons	FATHMM-XF	GERP++	LINSIGHT	PhastCons	phyloP	DANN	CDTS	CADD	REMM	Eigen	Eigen_PC	funSeq2	Comments about states
ReprPC1	0.2	0	1.1	4.4	3.9	3.9	3.6	3.8	0.7	11	4.9	5	5.2	4.3	1.5	ReprPC (except: PromBiv in ESC); H3K27me3 strong and H3K4me1 weak
EnhA2	0.36	0.3	1.9	4.5	5.6	5.2	4.8	4.2	0.2	0.8	6.3	7.4	6.2	0.8	4.6	enhancers in mesenchymal, muscle, heart, neurosph, adipose. DNase in ESC and iPSC
EnhA3	0.21	0.3	9.2	3.2	4.5	4.5	3.9	3.4	0.3	0.8	5.6	5.6	8	7.9	12	strong enhancers in most cells, weaker in blood and ESC&iPSC. Stronger signals of activate enhancer, promters and TxReg in Adipose, ENCODE, Epithelia, IMR90, Mesench and Myostat
EnhA17	0.58	0.2	1	4.8	5.7	4.8	4.7	4.3	0.3	0.6	5.9	5.8	5.5	0.1	2.3	ESC, iPSC, Neurosph enhancers and some differentiated
TxEnh4	0.25	15	17	2.9	2	2.2	1.5	1.8	0.7	2.8	2.1	2.8	2	4.1	7.8	H3K36me3 strong and some enhancers (H3K27ac, H3K4me1); exons; near TES; per-ct stateTxEnh3p all cell types
TxEnh8	0.25	4.1	9	2	2.2	2.3	1.6	1.8	0.4	0.8	1.9	2.6	2.3	2.5	7.7	Transcribed enhancers 3' in most cell types; enhancers in Myostat, IMR90, Mesench, Epithelial
TxEx1	0.26	9.9	10	2.5	2.1	1.9	1.7	2	0.6	1.2	2	2.2	1.8	0.1	3.2	H3K79me2 and H3K36me3; exon; per-ct Tx state in all cell types
TxEx2	0.5	13	14	3.2	1.9	1.8	1.4	1.8	0.8	2.5	1.6	1.9	1.1	0.3	2.1	H3K36me3 strongest; exons; per-ct Tx3p state in all cell types
TxEx3	0.65	13	8.7	1.8	1	1.1	0.9	1.2	0.7	2.1	0.9	1.1	0.7	0.8	1.1	H3K36me3 strong; exon; most enriched per-ct state Tx3p in all cell types
DNase1	0.22	0.3	3	10	2.2	3	2.4	2.5	2.2	2.3	2.8	3.2	4.3	10	7.1	DNase I only; CTCF in various cell types, Candidate Insulator
BivProm1	0.14	0.2	1.1	4.5	2.8	4.1	4.1	3.8	8.8	52	8.7	13	11	32	14	BivProm in most cell types; more balanced between H3K4me3 and H3K27me3
BivProm2	0.16	0	1.3	4.8	3.5	4.2	4.2	4.1	4.8	41	7.3	8.9	8.6	16	6.9	bivalent promoter- stronger on H3K27me3
BivProm4	0.14	0.5	1.7	5.4	6.6	5.8	5.7	5.4	0.7	7.8	8.4	9.7	9.8	6.5	8.6	PromBiv in Blood & T-cells, HSC & B-cell, ESC. Prom_D2 in Brain, muscle, Sm. Muscle, mesench, neurosph, myosate, adipose. Mix of Prom_D2, BivProm and PromP in ESC_derived, ENCODE 2012, iPSC, Digestive, Epithelia, Heath and others
PromF2	0.15	4.1	5.5	2.6	2.6	3.5	2.7	2.5	0.9	11	5	5.3	10	40	21	H3K4me1, H3K4me2, H3K4me3, DNase, acetylations promoter flank upstream bias
PromF3	0.16	11	0.8	1.2	1.4	2.6	2	1.8	2.3	35	4.2	6.7	16	64	33	H3K4me2, H3K4me3, H3K4me1(weaker than H3K4me3), DNase, acetylations - flanking tss upstream and downstream
PromF4	0.19	13	0.2	2.4	2.6	4.5	4	3.4	9.8	56	8.7	19	32	73	38	H3K4me2, H3K4me3 limited H3K4me1, heavily acetylated - flanking tss downstream bias
PromF5	0.14	0.8	1.1	3.7	3	4.2	4.4	3.9	8.9	48	9.1	17	16	52	21	flanking promoter; most enriched with per-ct state PromU in most cell types; stronger on H3K4me3; sometimes, this state can overlap with bivalent promoters in blood-related, ESC-related groups
TSS1	0.13	6	0.8	3	3.4	5.3	6.1	4.8	17	41	12	24	26	61	30	TSS more acetylated and active; most enriched per-ct state TssA in all cell types
TSS2	0.12	1.5	2	4.2	2.5	3.7	5.3	4	19	23	9.6	14	16	40	16	TSS all cell types except PromP in iPSC

D



state	enrichment ratio
GapArtf2	0.58
GapArtf3	0.65
Acet1	0.75
Quies3	0.83
GapArtf1	0.85
TxEx1	1.73
PromF3	1.78
TxEx4	1.78
TxEx2	1.83
PromF4	1.92

rank of most enriched states	
	1
	2
	3
	4
	5
	>5

E

state	% background	caviar_lead_snps	finemap_lead_snps
EnhA1	0.17	2.92	2.93
EnhA9	0.15	2.72	2.72
TxEnh4	0.23	2.93	2.92
TxEnh6	0.17	2.65	2.65
TxEx4	0.08	3.36	3.36
BivProm4	0.12	2.86	2.86
PromF2	0.12	3.05	3.04
PromF3	0.13	3	3.01
PromF4	0.19	3.08	3.08
PromF5	0.13	2.88	2.89

rank of least enriched	
	1
	2
	3
	4
	5

F

state	% genome	breast	haematopoietic and lymphoid tissue	liver	pancreas
GapArtf2	0.05	0.76	4.88	2.07	4.09
GapArtf3	0.01	1.35	5.58	5.38	4.22
Quies1	10.84	1.21	1.69	1.54	1.57
Quies2	3.36	1.26	1.41	1.49	1.75
Quies3	13.37	0.97	1.23	0.95	0.94
Quies4	4.86	1.18	1.25	1.09	1.23
Quies5	1.85	1.34	0.96	0.79	0.98
HET1	0.77	1.25	1.38	1.53	1.95
HET2	0.75	1.29	0.92	1.20	1.68
HET4	0.61	0.87	1.14	0.89	1.47
HET5	0.27	1.04	0.97	1.23	1.32
HET6	0.63	1.37	1.29	1.31	1.89
HET7	1.12	1.19	1.17	1.03	1.37
HET9	1.08	1.44	1.34	1.18	1.56
ReprPC8	0.52	1.24	0.75	0.64	0.77
Acet1	0.20	0.83	2.67	1.23	1.37

G

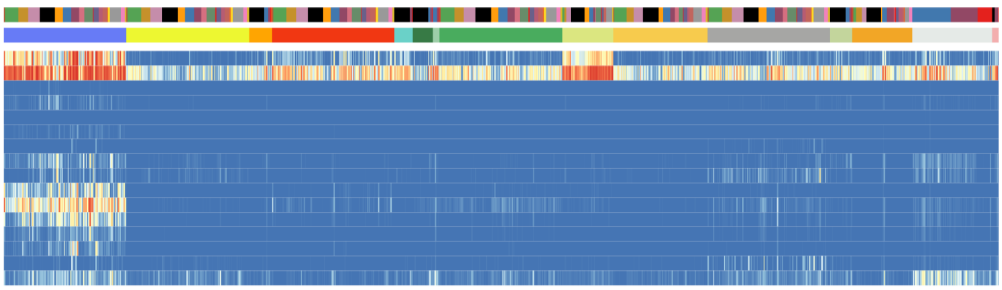


Figure 2. 5: Full-stack states' relationship with human genetic variants.

(A) Enrichments of full-stack states with duplication and deletion structural variants from (Abel *et al.*, 2020). Only states that are in the top ten most enriched states are shown. Top five fold-enrichments for each class of structural variants are colored in increasing darker shades of red for higher ranked enrichments. Enrichment values below one, corresponding to depletions, are colored yellow. The columns from left to right are the state label, percent of genome the state covers, the fold enrichment for deletions, and fold enrichment for duplications.

(B) Emission probabilities corresponding to states in **(A).** The coloring is the same as **Fig. 2.2A**. The figure highlights how states most associated with structural variants generally had higher emission of H3K9me3 compared to other chromatin marks.

(C) Enrichments of full-stack states with top 1% prioritized bases in the non-coding genome by 14 variant prioritization scores previously analyzed (Arneson and Ernst, 2019). Only states that are among the top five most enriched states by at least one score are shown. The top five enrichment values for each score are colored in increasing darker shades of red for higher ranked enrichment values. Enrichment values below one, corresponding to depletions, are colored in yellow. The columns from left to right are the state label, percent of the genome covered, the 14 score enrichments, and a detailed description of the state.

(D) Log base 10 of ratios of states' enrichment with GNOMAD variants (Karczewski *et al.*, 2020) with the lowest MAFs (< 0.0001) vs. GNOMAD variants with the highest MAFs (0.4-0.5). States are ordered as in **Fig. 2.2A**. Top five states with the highest and lowest enrichment ratios are labeled to the right.

(E) States most enriched with fine-mapped phenotypic variants against the background of common variants. Fine-mapped phenotypic variants were identified by either CAVIAR (Chen *et al.*, 2015) or FINEMAP (Benner *et al.*, 2016) (**Methods**).

(F) State enrichments with somatic mutations associated with four cancer types in the non-coding genome. Only states that are among the ten most enriched with variants from at least one cancer

type are shown. States in the top five are colored according to their ranks. The top five enrichment values for each cancer type are colored in increasing darker shades of red for higher ranked enrichment values. The columns are the state label, the percent of the genome the state covers, and the fold enrichments of variants from breast, haematopietic and lymphoid, liver, and pancreas cancer types.

(G) Emission probabilities corresponding to states in (F), as subsetted from **Fig. 2.2A**. The coloring is the same as **Fig. 2.2A**. The figure highlights how states with the greatest enrichments for cancer-associated variants tend to have higher emission probabilities for H3K9me3 compared to other chromatin marks.

Supplementary Information

Supplementary Data 2.1: GO terms associated with each full-stack states

- Each figure shows the 5 GO Biological Process and 5 GO Molecular Function terms most significantly enriched in each full-stack state. State Quies3 did not have a list of enriched GO terms because there was no output from GREAT for regions associated with state. Therefore, there are 99 figures corresponding to 99 full-stack states with output from GREAT.

Supplementary Data 2.2: Summary characterizations of full-stack states

- Full characterization of the full-stack states, with detailed comments summarizing states' characteristics
- Emission parameters of the full-stack states.
- Top 100 highly emitted datasets associated with each full-stack state, colored by the marks associated with the datasets (similar to Fig. 2.2B)
- Top 100 highly emitted datasets associated with each full-stack state, colored by the cell groups associated with the datasets (similar to Fig. 2.2C)
- Different genome context enrichments with full-stack states: Excel version of figures 3A-B, Supplementary Figure 2.18.
- Excel version of Supplementary Figure 2.31: Enrichment of all full-stack states for ConsHMM states.
- Excel version of Supplementary Figure 2.30: Full-stack states and maximum enriched ConsHMM state.
- Excel version of Supplementary Figure 2.7: Statistically significant tissue—group specificity in full-stack states.

Supplementary Data 2.3: Full-stack states enrichments with repeats, prioritized variants, common variants, structural variants, CTCF, PRC1-PRC2, GWAS catalog variants, sex chromosomes, fine-mapped variants, cancer-associated variants.

- Excel version of Supplementary Figure 2.28: Full stack states enrichments with RepeatMasker classes of repeats and CG content.
- Excel version of Supplementary Figure 2.37: Enrichment of all full-stack states for top 1% bases prioritized by variant prioritization scores.
- Excel version of Supplementary Figure 2.32: Full-stack states enrichments with structural variants.
- Excel version of Supplementary Figure 2.42: Full-stack states enrichments with variants from GNOMAD stratified by minor allele frequencies, common variants and CG dinucleotides.
- Excel version of Supplementary Figure 2.14: Full-stack states enrichments with CTCF associated chromatin states.
- Excel version of Supplementary Figure 2.17B: Neighborhood enrichments of full-stack states with binding sites of PRC1 and PRC2 complexes.
- Excel version of Supplementary Figure 2.16: Full-stack states enrichments with Polycomb Repressive protein complexes PRC1 and PRC2.
- Excel version of Supplementary Figure 2.43: Full-stack states enrichments with GWAS catalog variants and sex chromosomes.
- Excel version of Supplementary Figure 2.44: Full-stack states enrichment values for fine-mapped variants at phenotype associated loci.
- Excel version of Supplementary Figure 2.47: Full-stack states enrichments with cancer-associated somatic mutations in the non-coding genome.

Supplementary Data 2.4: Full stack states' association with concatenated annotations for multiple cell types.

- Excel version of Supplementary Figure 2.9: Estimated probabilities of concatenated chromatin states overlapping with full-stack states. *Detailed comments about full-stack states' characteristics through this analysis are provided.*

- Excel version of Supplementary Figure 2.8: Full-stack states maximum-enrichments with annotated chromatin states in 127 reference epigenomes. *Detailed comments about full-stack states' characteristics through this analysis are provided.*
- Maximum enrichments of full-stack states and 25-state concatenated annotations for 127 cell/tissue types (reference epigenomes).
- Median enrichments of full-stack states and 25-state concatenated annotations for 127 cell/tissue types (reference epigenomes).

Supplementary Data 2.5: Data of AUROC comparison between full-stack annotation and concatenated and independent annotations in predicting different genomic contexts

- Data accompanying Supplementary Figure 2.19-21: AUROC comparison of full-stack annotations and concatenated and independent annotations in predicting external genome contexts
- Data accompanying Supplementary Figure 2.34-35: AUROC comparison of full-stack model annotations and 18-state concatenated annotations and 100-state independent annotations in predicting structural variants of type deletions and duplications.
- Data accompanying Supplementary Figure 2.29: AUROC comparison of the full-stack and concatenated and independent chromatin state annotations at predicting different classes of repeat elements.
- Data accompanying Supplementary Figure 2.45-46: AUROC comparison of full-stack model annotations and the 100-state independent annotations and 18-state concatenated annotations in predicting fine-mapped variants.
- Data accompanying Supplementary Figure 2.41: AUROC comparison of the full-stack model annotations and concatenated and independent model annotations at predicting top 1% non-coding bases prioritized by various variant prioritization scores.

Supplementary Data 2.6: Summary characteristics of ConsHMM states

- Descriptions of select ConsHMM states mentioned in the main text

- Excel version of Supplementary Figure 2.30C: Enrichment of all full-stack states for ConsHMM states.

Supplementary Data 2.7: Supplementary Information about the full-stack model analysis

- Naïve-Bayes greedy search for representative datasets for characterization of the full-stack states.
- Asymptotic worst-case time and memory usage of stacked model

Supplementary Data 2.8: Download links for annotation data used throughout the manuscript

Chapter 3. Universal chromatin state annotation of the mouse genome

Abstract
Genome-wide chromatin states learned from integrating genome-wide maps of multiple epigenetic marks within the same cell type have been widely used to generate genome annotations of individual cell types. An alternative strategy based on 'stacked modeling' can provide a single 'universal' chromatin state annotation based jointly on data from many cell types. In human, such an approach was recently demonstrated and the resulting chromatin state annotation, denoted full-stack, was shown to have complementary advantages to per-cell-type annotations. However, an analogous annotation has not been previously available in mouse. Here, we produce a chromatin state annotation for mouse based on 901 datasets assaying 14 chromatin marks in 26 different cell or tissue types. To characterize each chromatin state, we relate the states to other external annotations and compare them to analogously defined states in human. We expect the full-stack chromatin state annotation for mouse will be a useful resource for studying the genome of this key mammalian model organism.

Introduction

Mouse is widely adopted as a model organism for human for many reasons including their genetic and physiological proximity to humans, relatively short life span, and availability as test subjects for genetic manipulations (Vanhooren and Libert, 2013; Aitman *et al.*, 2011; Perlman, 2016). A wealth of epigenomic datasets in mouse, include maps of histone modifications and variants and sites of accessible DNA, has accumulated thanks to efforts from different consortia and individual labs, which can be used to annotate the mouse genome, including non-coding regions (Kazachenka *et al.*, 2018; Stamatoyannopoulos *et al.*, 2012; Yue *et al.*, 2014; Zhu *et al.*, 2021; Hon *et al.*, 2013; Tsai *et al.*, 2009; Rugg-Gunn *et al.*, 2010). This type of data has previously been integrated methods such as ChromHMM and Segway (Ernst and Kellis, 2010, 2012; Hoffman *et al.*, 2012; Libbrecht *et al.*, 2021) to generate chromatin state maps for various organisms including different mouse and human cell and tissue types (Yue *et al.*, 2014; ENCODE Project Consortium, 2012; van der Velde *et al.*, 2021; Bogu *et al.*, 2015; Sugathan and Waxman,

2013; Gorkin *et al.*, 2020). These chromatin state maps have traditionally been used to annotate genomes in a per-cell-type manner using either the 'independent' or 'concatenated' modeling approaches (for ease of presentation, we will refer to tissue types also as cell types) (Ernst and Kellis, 2017; Libbrecht *et al.*, 2021).

Recently, we applied an alternative 'stacked' modelling approach of ChromHMM to learn chromatin states from over 1000 human datasets representing more than 100 cell types, to generate a universal annotation of the human genome that can annotate all human cell types (Vu and Ernst, 2022). This modeling provided a single annotation of the genome per position based on data from all the input cell types. Such an annotation, denoted full-stack annotation, offers complementary advantages to per-cell-type annotations, such as differentiating constitutively active regions from cell-type-specific ones and simplifying genome annotations across cell types through a single annotation shared across cell types as opposed to one for each. Additionally, the full-stack annotation allows researchers to bypass picking a single cell type for analyses or conducting analyses separately for every cell type. This can be particularly useful in studies involving data that is not inherently cell-type-specific such as analyses of genetic variants or conserved DNA sequence. However, an analogous full-stack annotation has not been previously available in mouse.

To address this, we train a full-stacked model with ChromHMM using input data from >900 mouse datasets of 14 chromatin marks from 26 mouse cell type groups (**Methods**). We analyze these states with respect to their enrichments with external datasets and annotations to provide detailed characterizations for each state. We also analyze to what extent each state is conserved in human. We expect the mouse full-stack annotations along with the provided biological characterizations will be a useful resource for studying this key model organism.

Results

We learned the mouse full-stack model by applying ChromHMM to over 900 mouse epigenomic datasets, similar to how it was previously applied in human (Ernst and Kellis, 2012;

Vu and Ernst, 2022) (**Methods, Fig. 3.1, Fig. 3.S1**). We used a 100-state model for consistency with the previously analyzed human full-stack model.

We manually grouped these 100 states into 16 groups. One of the groups contains states associated with assembly gaps or alignment artifacts (mGapArtf), the latter of which are often marked by signals of both open-chromatin mark (ATAC or DNase) and heterochromatin mark H3K9me3 (**Fig. 3.1**). Another group, Quiescent group (mQuies), consists of states associated with minimal signals of any chromatin marks. We defined a Heterochromatin (mHET) group primarily associated with H3K9me3, and a Zinc finger genes (mZNF) group associated with both H3K36me3 and H3K9me3. We also defined a Polycomb repressed group (mReprPC) associated with primarily H3K27me3, and another group associated with both open chromatin marks (DNase and/or ATAC-seq) and polycomb-repressed-associated mark H3K27me3 (mReprPC_openC). We also defined a group of states associated with just open chromatin (mOpenC), based on DNase-seq and ATAC-seq signals relative to other chromatin marks.

We defined three groups of states associated with enhancers: active enhancers (mEnhA), weak enhancers (mEnhWk), and transcribed enhancers (mTxEnh). States in the mEnhA group were associated with open chromatin, H3K27ac and H3K4me1. States in the mEnhWk group (mEnhWk) also showed association with those marks, but at lower levels compared to those in mEnhA group. States in mTxEnh group showed signals of open chromatin (ATAC and/or DNase), H3K4me1, H3K27ac *and* transcription-associated marks (H3K36me3 or H3K79me2/3).

In addition to mTxEnh group, we defined three additional transcription groups: transcription (mTx), transcription and exons (mTxEx), and weak transcription (mTxWk). States in the mTx group are associated primarily with the transcription marks H3K36me3 and/or H3K79me2/3. Meanwhile, states in transcription and exon group (mTxEx) are associated with both open chromatin and transcription marks. States in the mTxWk group are associated with low levels of the transcription marks.

We also defined three promoter-associated groups: bivalent promoters (mBivProm), promoter flanking (mPromF) and transcription start sites (mTSS). States in these groups generally had relatively high levels of H3K4me2 and H3K4me3, and for some of them also H3K4me1 and/or open chromatin marks. mBivProm states were also associated with the repressive mark H3K27me3. States in the mTSS group tended to have weaker H3K4me1 levels.

Within each group, there were differences among individual states, such as the magnitude of the emission probabilities associated with specific chromatin marks, or their association with different cell type groups (**Fig. 3.1**). For example, different states in the active (mEnhA) and weak enhancer (mEnhWk) groups have enhancer associated marks that were specific to different cell type groups such as the brain, blood, immune, liver, and embryo (**Fig. 3.1C**). Detailed descriptions of each state's chromatin mark signals and cell-type-specific activities are provided in **Supplementary Data 3.1**.

We also conducted various enrichment analyses to further characterize the states (**Fig. 3.2A**). Enrichments with external annotations further highlight the distinctions among states from different groups, as well as among those within the same group. For example, the state mGapArtf1 overlapped with 99.9% of annotated assembly gaps in mm10 (6.6-fold) (**Fig. 3.2A**). States mGapArtf1 and mGapArtf3 jointly overlapped with 82.1% of the blacklisted regions from ENCODE (5.4 and 5.0-fold, respectively) (**Fig. 3.2A**). States in promoter-associated groups (mTSS, mPromF, mBivProm) showed relatively high enrichments with regions within 2kb of annotated TSSs (9.4-26.7 fold, **Fig. 3.2A**). These states vary in their enrichments with regions upstream and downstream of annotated TSSs (**Fig. 3.2D, Supplementary Figure 3.2**). Three states from the TSS group (mTSS1-3) had the strongest enrichment for TSS (59.2-159.9 fold). These three states along with mBivProm2 were strongly enriched with CpG Islands (101.1-159.2 folds, **Fig. 3.2A**). States in the transcription associated groups (mTx, mTxWk, mTxEnh, mTxEx) all had enrichments greater than 2.4-fold for annotated gene bodies. States in the transcription and exon group (mTxEx1-3) showed the highest enrichments for annotated exons (11.3-13.7

folds, **Fig. 3.2A**) and regions surrounding annotated TESs (**Fig. 3.2E, Supplementary Figure 3.2**). States mOpenC6-7, which had strong *constitutive* DNase-seq and/or ATAC-seq signal while having relatively limited histone modification signals, had the strongest enrichments with CTCF binding sites in multiple cell types (geometric mean 146- and 98- fold for states mOpenC6-7, respectively) (**Fig. 3.1, 2F, Supplementary Data 3.2**).

Additionally, we analyzed the enrichment of full-stacked states for different chromosomes. This uncovered three states in the polycomb repressed group (mReprPC4-6) that were highly enriched on chromosome X (8.9-11.4 fold, **Supplementary Figure 3.3**), likely related to H3K27me3-associated chromosome X inactivation (Wutz, 2011; Yen and Kellis, 2015). We also found chromosome Y strongly enriched for mGapArtf1 state (6.4 fold, corresponding to 96% of chrY) (**Supplementary Figure 3.3**).

We also analyzed the states' enrichments for different classes of repeat elements (Smit *et al.*, 2015). For the two largest classes of repeats, Long interspersed nuclear elements (LINE) and long tandem repeats (LTRs) (**Supplementary Figure 3.4-5**), the most enriched states were both in the HET group (mHET9,7) (2.7 and 3.3 fold). Satellite and rRNA had the strongest enrichments for the mGapArtf3 state, 22.5 and 95.5 fold, respectively.

We also related the full-stack states to average expression of overlapping genes (**Methods**). States in the transcription-associated groups (mTxEnh, mTx, mTxEx), along with those related to promoter (mPromF and mTSS groups) showed higher average gene expression across cell types compared to other groups (**Fig. 3.2C**). State mTxEx3 showed the highest gene expression of all states.

Additionally, we analyzed the mouse full-stack states' association with per-cell-type chromatin state annotations defined across 66 reference epigenomes from 12 unique cell type groups and 7 developmental stages, based on 8 marks (**Supplementary Figure 3.6-7, Supplementary Data 3.4**) (Gorkin et al., 2020). This revealed, for example, that state mEnhA17 showed the strongest enrichments with per-cell-type active enhancer states across all

developmental stages for liver (**Supplementary Figure 3.6-7, Supplementary Data 3.4**), which is consistent with this state's highest signals in enhancer-associated chromatin marks (H3K4me1, H3K27ac) for liver datasets (**Fig. 3.1**). State mTSS2 was most enriched with per-cell-type active promoter states in all reference epigenomes (**Supplementary Figure 3.6-7, Supplementary Data 3.4**), consistent with its association with individual chromatin marks (**Fig. 3.1**).

In addition, we analyzed how the mouse full-stack states correspond to those of an analogous previously defined full-stack model in human (Vu and Ernst, 2022). We evaluated the enrichments of each mouse full-stack state with each human full-stack state after mapping the human annotations to those for mouse (**Methods, Supplementary Figure 3.8-9, Supplementary Data 3.3**). Twenty-two out of 100 mouse states showed >50-fold enrichment with at least one human state (**Supplementary Figure 3.9, Methods**) (Vu and Ernst, 2022), and these states' biological implications highlight strong correspondence of states from the human and mouse models. For example, mouse state mTxEx3 showed 378.8-fold enrichment for human state TxEx4 – the largest enrichment across any pair of states— (**Supplementary Figure 3.9, Supplementary Data 3.3**). These two states showed the highest average gene expression across multiple mouse and human cell types, respectively (Vu and Ernst, 2022) (**Fig. 3.2C**). All 13 mouse states in the promoter groups (mPromF, mBivProm, mTSS states) showed strong enrichments with human full-stack states that are also promoter-associated, with 12 of these mouse states showing >90-fold enrichment (**Supplementary Figure 3.9, Supplementary Data 3.2**). Mouse states mOpenC6-7, which are associated with *constitutive* open chromatin CTCF elements (**Fig. 3.2F, Supplementary Data 3.2**), showed the strongest association with the human DNase state, which was also constitutively marked by DNase and CTCF in human (Vu and Ernst, 2022). However, there exist differences between states from the two organisms' models. For example, in the mouse model, seven states in the mOpenC group (all except mDNase6-7), which we characterized as showing cell-type-specific signals of open chromatin, did not show strong enrichment for specific human states (**Supplementary Figure 3.8-9, Supplementary Data 3.3**).

We also evaluated each full-stack state's average human-mouse LECIF score, which quantifies conservation at the functional genomics level between the two species (**Fig. 3.2B**) (Kwon and Ernst, 2021) (**Methods**), which ranged from 0.04 (mHET9) to 0.71 (mBivProm3) (**Fig. 3.2B, Supplementary Figure 3.9**). All 14 mouse states that had an average LECIF score >= 0.5 also had a >50-fold enrichment with a human full-stack state, highlighting that mouse states with high LECIF score show concordance with specific human states. In addition, we looked at each state's enrichment for sequence constraint eleements as defined by PhastCons (Siepel *et al.*, 2005). Across all states, the states' enrichments for PhastCons elements and average LECIF score showed overall consistency (Spearman correlation 0.70; p-value: 3.8e-16). We found 10 mouse states that are among the top 20 states based on average LECIF score, enrichments for PhastCons element and for a specific human full-stack state (**Supplementary Data 3.1, Fig. 3.S9**). Among these states, seven are associated with promoter activities (mBivProm1-3, mTSS1-3, mPromF1), two states are characterized by strong exon enrichments and constitutive transcriptional activities (mTxEx2-3), and one state (mEnhA3) corresponds to constitutively strong enhancers (**Supplementary Figure 3.9**). Interestingly, a few states stand out as associated with either high sequence constraint or functional conservation (LECIF score), but not in both. For example, constitutive DNase-candidate insulator states mOpenC6-7 are among top 20 with highest average LECIF scores yet had lower (Phastcons) sequence constraint enrichment (ranked 50, 59) (**Supplementary Figure 3.9**).

Discussion

We introduced the mouse full-stacked annotation to provide a single chromatin state annotation per genomic position based on over 900 epigenomic datasets representing 26 different cell type groups. The mouse full-stacked model and its characterization is analogous to the previous human full-stack model (Vu and Ernst, 2022) (**Data Availability, Supplementary Data 3.1**). As discussed previously in the context of the human genome annotation (Vu and Ernst, 2022), the full-stack model has a number of advantages, such as being able to differentiate

constitutive from cell type-specific annotations and simplifying the overall genome annotation in that there is a single genome annotation per position. However, this does come at a trade-off of a more complex set of model parameters. The full-stack annotation is not meant to replace existing per-cell-type annotations, but rather to complement them and the most appropriate annotation will likely depend on the application (Vu and Ernst, 2022). We expect the full-stack model to serve as an additional resource for work that leverages the mouse as a model organism to gain insight into human biology and disease.

Methods

Input data and processing

We obtained data of ENCODE Project Portal (The ENCODE Project Consortium *et al.*, 2020; Stamatoyannopoulos *et al.*, 2012; Yue *et al.*, 2014), and restricted the downloaded files to those with 'File analysis title' starting with 'ENCODE4' and 'File assembly' of 'mm10'. In total, we downloaded data of read alignment (.bam files) for 901 experiments, 114 of which were DNase-seq, 83 were ATAC-seq and 704 were ChIP-seq data targeting 12 chromatin marks representing 26 cell type groups (**Supplementary Data 3.1**). For each .bam file resulting from a ChIP-seq assay, we extracted the corresponding control .bam file by matching the .fastq files of reads from the ChIP-seq assay with the control reads. As the DNase-seq or ATAC-seq experiments did not have paired control .bam files, we assumed a uniform background read distribution. Links to download all input data for the stacked model are provided in **Supplementary Data 3.1**.

We then constructed the cell_mark_file input table required by ChromHMM BinarizeBam such that there are four tab-delimited columns in the table. The first column is set as 'Genome' across all rows. The second column denotes the experiment names of the form '<Biosample term name>_<Experiment target>_<Experiment accession>', where 'Biosample term name', 'Experiment target' and 'Experiment accession' correspond to the cell type, histone mark/DNase/ATAC profiled and the accession code of such experiments, respectively from the metadata from ENCODE. The third column contains the experiments' .bam file names. The last

column contains the matched control .bam file names, which is left blank for DNase-seq or ATAC-seq experiments, since we assumed a uniform background distribution for these assays.

Using this cell_mark_file input table, we next binarized the data at 200 base pair resolution using the BinarizeBam and MergeBinary commands of ChromHMM (v.1.23), following the procedures of (Vu and Ernst, 2022).

Training full-stack modeling and generating genome-wide state annotations

We learned the mouse full-stack chromatin state model for the 901 datasets using the LearnModel command of ChromHMM (v.1.23). We applied the same set of flags as in learning the human full-stack model (-splitrows -holdcolumnorder -pseudo -many -p 6 -n 300 -d -1 -lowmem -gzip), described in Vu and Ernst, 2022 (Vu and Ernst, 2022). We specified the number of states to be the same as in the human model (100 states) (Vu and Ernst, 2022).

Enrichment and estimated probabilities of overlap with per-cell-type chromatin state annotations

We obtained per-cell-type 15-chromatin state annotations for 66 reference epigenomes/cell types from Gorkin et al., 2020 (Gorkin *et al.*, 2020), with download links provided in Kwon and Ernst, 2021 (Kwon and Ernst, 2021). For simplicity, we use reference epigenome and cell types interchangeably, and we refer to the chromatin state segmentation that is used to annotate the individual reference epigenomes as per-cell-type annotations. This model was trained using the *concatenated* modeling approach from data of 8 chromatin marks measured in 12 cell type groups at up to 8 distinct stages during mouse fetal development (Gorkin *et al.*, 2020). We applied the same procedure as outlined in Vu and Ernst, 2022 (Vu and Ernst, 2022) to obtain two types of summary results of the relationship between mouse full-stack states' association with states in per-cell-type annotations. First, for each full-stack state, we report, for each of the 64 reference epigenomes, the chromatin state from the per-cell-type model that is maximally enriched in the full-stack state (Vu and Ernst, 2022). Second, for each of the 12 tissue types, we report the estimated probabilities of each full-stack state overlapping with each of the 15 states in the per-cell-type model (Vu and Ernst, 2022). These results, along with detailed comments about

the observed patterns of overlap between each full-stack state and per-cell-type state, are available in **Supplementary Data 3.4.** Data of all per-cell-type annotations are in mm10 (Gorkin *et al.*, 2020).

Average gene expression associated with each full-stack state

We obtained data of gene expression for 19 tissue types in mouse from (Shen *et al.*, 2012) (http://chromosome.sdsc.edu/mouse/download/19-tissues-expr.zip). The provided data contains two gene expression datasets for each tissue type, corresponding to two replicates. We converted the gene expression values for the 19 tissues into $log_2\,(FPKM + 1)$ values, where FPKM (Fragments per kilo base of transcript per million mapped fragments) were the provided values from the source data, and we added a pseudo count of 1 for each value.

Since the gene expression data was provided in mm9, we lifted the mouse full-stack annotation from mm10 to mm9. To do so, we first wrote the full-stack annotation in mm10 into a .bed file such that each line corresponds to a 200-bp segment. We then used the liftOver tool with default parameters to convert the 200-bp segments from mm10 to mm9. We filtered out regions in the lifted-over mm9 annotation that were mapped from >= 2 distinct segments in mm10.

For each full-stack state and each of the gene expression dataset (there are 38 of them with 2 replicates for each tissue type), we calculated the average gene expression of all genes that overlap with the state, while taking into account the genes' length. We followed the same procedure described in Vu and Ernst, 2022. In particular, within a dataset, let the length and expression of gene g be denoted L_g and E_g, respectively. Let B_s be the set of 200-bp genomic segments i's that are assigned to state s in the mouse full-stack annotation, in mm9. Let G_i denote the set of genes that overlap with genomic segment i. The gene-length-normalized average expression for state s is calculated as done previously (Vu and Ernst, 2022):

$$avg\ exp\ \ bp\ normalized_{\,s} = \frac{\sum_{i \in B_s} \sum_{g \in G_i} \frac{E_g}{L_g}}{\sum_{i \in B_s} \sum_{g \in G_i} \frac{1}{L_g}}$$

We then obtained the average gene expression for each full-stack state in each dataset. To calculate the average gene expression for the states in each of the 19 tissue types, we averaged the calculated average expression across the two replicate datasets for the same tissue type.

External annotation sources

The sources for external annotations for enrichment analyses are as follows (all download links are listed in **Supplementary Data 3.1**).

- Annotations of CpG islands, exon, gene bodies (exons and introns), transcription start (TSS), and transcription end sites (TES), 2kb windows surrounding TSSs (TSS2kb) in mm10 were RefSeq annotations included in ChromHMM (v 1.23) and originally based on annotations obtained from the UCSC genome browser (Rosenbloom *et al.*, 2015; Kent *et al.*, 2002) on July 26th, 2015.
- Annotation of coding gene regions correspond to coordinates of genes whose feature type is 'CDS' from GENCODE mm10 gene annotation, vM25 (Frankish *et al.*, 2019), accessed on February 3rd, 2022.
- Annotation of assembly gaps in mm10 were obtained from the UCSC genome browser and correspond to the Gap track (Rosenbloom *et al.*, 2015; Kent *et al.*, 2002), accessed on February 3rd, 2022.
- Annotations of pseudogenes in mm10 correspond to coordinates of genes whose gene type of transcript type contained 'pseudogene' from GENCODE's mm10 gene annotation, vM25 (Frankish *et al.*, 2019).
- Blacklisted regions were downloaded from ENCODE project portal in mm10 from (Amemiya *et al.*, 2019).
- Annotations of different repeat classes were downloaded from UCSC genome browser repeat masker track in mm10, accessed on Jan. 14th 2022 (Smit *et al.*, 2015).

- Annotations of Zinc finger genes in the mouse genome correspond to the coordinates of genes whose name contained 'Zfp' based on GENCODE mm10 annotation vM25 (Frankish *et al.*, 2019).
- Annotations of different chromosomes' coordinates were downloaded from UCSC genome browser's data of chromosome sizes in mm10, from https://hgdownload-test.gi.ucsc.edu/goldenPath/mm10/bigZips/mm10.chrom.sizes (Kent *et al.*, 2002; Rosenbloom *et al.*, 2015).
- LECIF scores measure that human-mouse conservation at functional genomics level, and were downloaded in version 1.1 from https://github.com/ernstlab/LECIF (Kwon and Ernst, 2021). For each full-stack state, we reported the average LECIF score of overlapping genomic bases with the state
- CTCF peaks data were downloaded as bed files format from Mouse ENCODE Project (Yue *et al.*, 2014; Stamatoyannopoulos *et al.*, 2012). We only included data files that has '*File analysis title'* starting with ENCODE4 based on the metadata. In total, we obtained data of CTCF peaks for 42 ChIP-seq experiments from profiling CTCF in 28 unique biosamples. Details and download links for CTCF peaks data is available in **Supplementary Data 3.1.**
- PhastCons conserved elements (Siepel *et al.*, 2005) based on the 60-way multi-species sequence alignment were downloaded from the UCSC genome browser (https://hgdownload.soe.ucsc.edu/goldenPath/mm10/database/phastConsElements60way.txt.gz).

Data Availability

Mouse full-stack chromatin state annotation are available at https://github.com/ernstlab/mouse_fullStack_annotations in mm10. The code to analyze the full-stack states are available at https://github.com/ernstlab/mouse_fullStack_annotations. All download links for input for the mouse full-stack model are available in **Supplementary Data 3.1**.

Figures

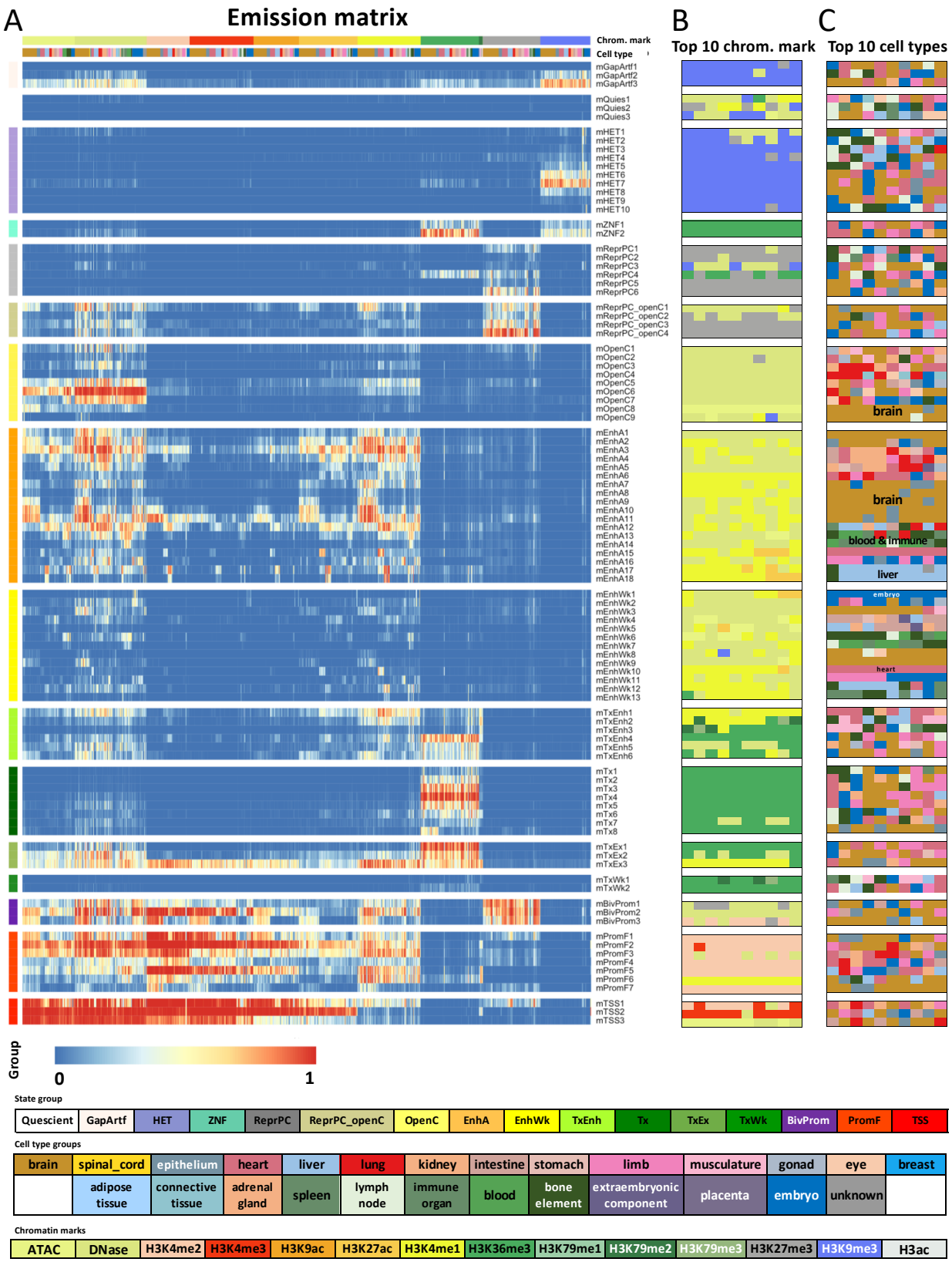


Figure 3. 1: Mouse full-stack state emission parameters.

(A) Each of the 100 rows in the heatmap corresponds to a mouse full-stack state. Each of the 901 columns corresponds to one input dataset. For each state and each dataset, the heatmap gives the probability within the state of observing a binary present call for the dataset's signal. Above the heatmap, one color bar indicates the assay/chromatin mark measured by each dataset. The other color bar shows the cell type groups associated with each dataset. The corresponding color legends are shown towards the bottom. The states are displayed in 16 groups with white space between each group, and grouped based on biological interpretations indicated by the color legend at the bottom. Full characterization of states is available in **Supplementary Data 3.1.** The model's transition parameters between states can be found in **Supplementary Figure 3.1.** Columns are ordered such that datasets profiling the same chromatin marks are next to each other.

(B) Each row corresponds to a full-stack state as ordered in (**A**). The columns correspond to the top 10 datasets with the highest emission value for each state, in order of decreasing ranks, colored by their associated chromatin marks as in (**A**).

(C) Similar to **(B)**, but datasets are colored by the associated cell type groups. The tissue groups primarily associated with some of the enhancer states are noted inside the heatmap.

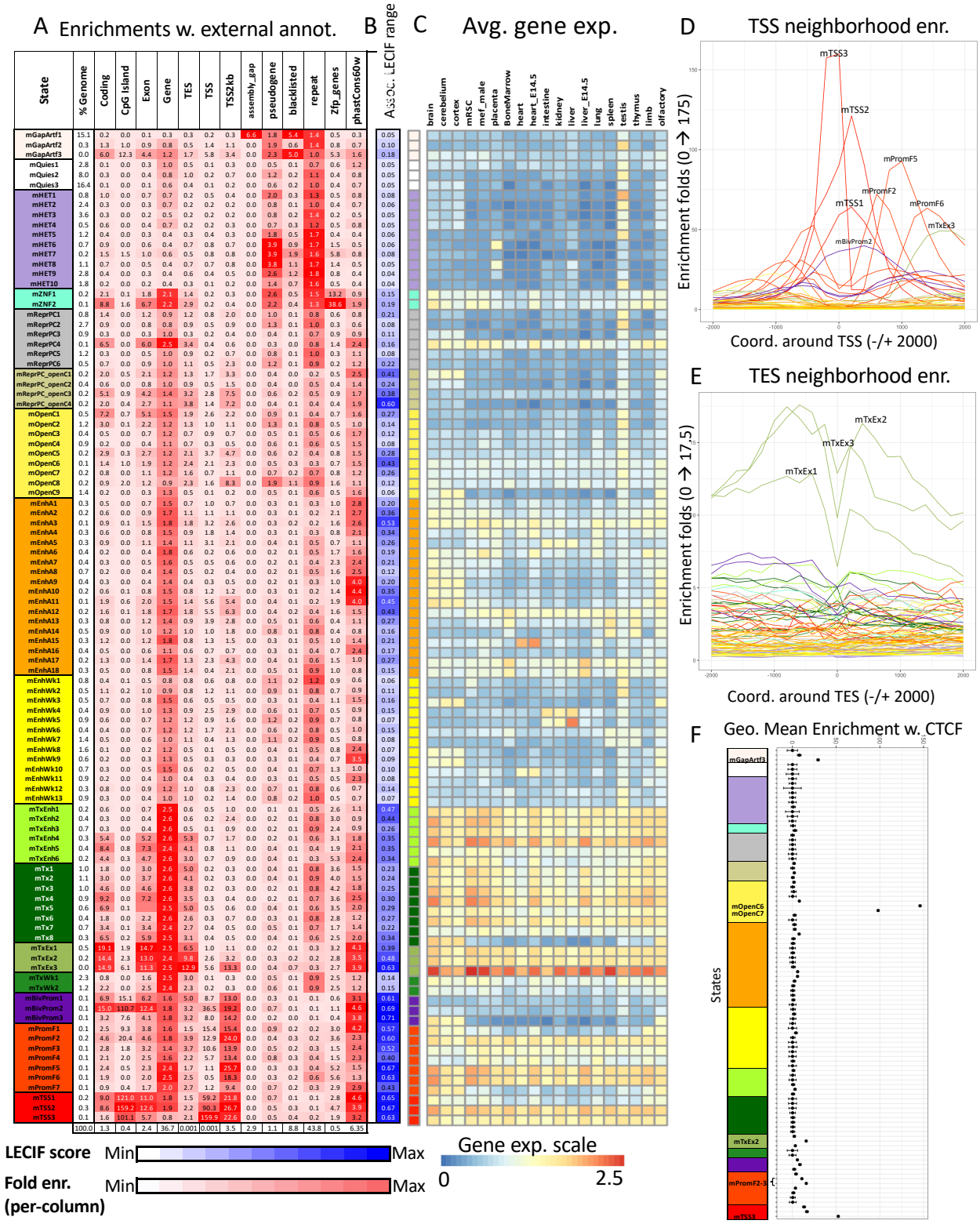


Figure 3. 2: Mouse full-stack states enrichments for external genomic annotations.

(A) Fold enrichments of mouse full-stack states with external genome annotations (**Methods**). Each row corresponds to a state and each column corresponds to one external genomic annotation: coding sequences, CpG Islands, Exons, gene bodies (exons and introns), transcription end sites (TES), transcription start sites (TSS), TSS and 2kb surrounding regions, assembly gaps, pseudogenes, blacklisted regions, repeat elements, annotated Zpf genes and PhastCons conserved elements (**Methods**). The last row shows the percentage of the genome that each external genome annotation covers. The heatmap colors are column-normalized, i.e. within each column, the colors of the cells are such that highest values are colored red and lowest values are colored white.

(B) Each row indicates the states' average LECIF score, indicating functional human-mouse conservation based on epigenetic annotations (Kwon and Ernst, 2021) (**Methods**). The list of states with top average LECIF scores and highest enrichments with PhastCons elements is in **Supplementary Data 3.1** and **Supplementary Data 3.2.**

(C) Average weighted expression of genes that overlap each full-stack state in different groups of cells (**Methods**). Each column in the heatmap corresponds to a cell group indicated at the top. Each row corresponds to a state, as ordered in **(A).**

(D-E) Positional enrichments of full-stack states relative to annotated **(D)** transcription start sites (TSS) and **(E)** transcription end sites (TES). Positive coordinate values represent the number of bases downstream in the 5' to 3' direction of transcription, while negative values represent the number of bases upstream. Each line shows the positional enrichments in a state. Lines are colored corresponding to the state group as indicated in **(A).**

(F) Geometric mean and geometric standard deviation of enrichments of full-stacks states CTCF elements across 28 cell types from ENCODE (ENCODE Project Consortium, 2012) (**Methods**). States are displayed vertically in the same order as **(A).** The DNase6-7 state showed the strongest enrichment for CTCF elements in all observed cell types. The geometric mean and standard deviation are calculated such that for each state, fold enrichment values of 0 are replaced by the

state's minimum non-zero value. The fold enrichment values accompanying this plot are available in **Supplementary Data 3.2.**

Supplementary Information

Supplementary Data 3.1: Metadata and download links for input data used for model learning, CTCF elements.

Supplementary Data 3.2: Summary characterizations of mouse full-stack states.

Supplementary Data 3.3: Mouse full-stack states' average LECIF scores, enrichments with human full-stack states, repeat classes, chromosomes and CTCF elements.

Supplementary Data 3.4: Mouse full-stack states' relationships with per-cell-type annotations (supporting supplementary figures 6-7)

Chapter 4: A framework for group-wise summarization and comparison of chromatin state annotations

Abstract

Genome-wide maps of epigenetic modifications are powerful resources for non-coding genome annotation. Maps of multiple epigenetics marks have been integrated into cell or tissue type-specific chromatin state annotations for many cell or tissue types. With the increasing availability of multiple chromatin state maps for biologically similar samples, there is a need for methods that can effectively summarize the information about chromatin state annotations within groups of samples and identify differences across groups of samples at a high resolution.

We developed CSREP, which takes as input chromatin state annotations for a group of samples. CSREP then probabilistically estimates the state at each genomic position and derives a representative chromatin state map for the group. CSREP uses an ensemble of multi-class logistic regression classifiers that predict the chromatin state assignment of each sample given the state maps from all other samples. The difference of CSREP's probability assignments for two groups can be used to identify genomic locations with differential chromatin state assignments.

Using groups of chromatin state maps of a diverse set of cell and tissue types, we demonstrate the advantages of using CSREP to summarize chromatin state maps and identify biologically relevant differences between groups at a high resolution.

The CSREP source code is openly available at http://github.com/ernstlab/csrep.

Introduction

Genome-wide maps of chromatin marks such as histone modifications and variants provide valuable information for annotating non-coding genome features (Barski *et al.*, 2007; Ernst *et al.*, 2011; Zhu *et al.*, 2013; Xie *et al.*, 2013). Efforts by large consortia and individual labs have produced chromatin state maps for many cell and tissue types (Roadmap Epigenomics Consortium *et al.*, 2015; ENCODE Project Consortium, 2012; Zhu *et al.*, 2013; Xie *et al.*, 2013). A popular representation of such data is chromatin states defined by the combinatorial and spatial

patterns of multiple marks, which are generated by methods such as ChromHMM and Segway (Libbrecht *et al.*, 2021; Ernst and Kellis, 2010, 2012; Hoffman *et al.*, 2012), and correspond to diverse classes of genomic elements including various types of enhancers and promoters.

Chromatin state maps have been produced for hundreds of different biological samples. In many cases there are multiple samples representing similar cell and tissue types (Boix *et al.*, 2021; Roadmap Epigenomics Consortium *et al.*, 2015). In such cases, to simplify analyses and visualizations, it may be desirable to have a single chromatin state annotation that summarizes the annotations for all samples in a pre-defined sample group of interest. A straightforward approach to this task is to take the most frequent chromatin state assigned at each position across samples in the group. However, when the number of samples in a group is small or the number of states is large, such an approach can be particularly vulnerable to noise. Furthermore, such an approach does not consider additional information available about the different chromatin states. For example, if a location was assigned to three different states in three samples, the summary annotation among these three states based on the frequency-based method would be arbitrary. However, by leveraging information about the co-occurrence of state assignments genome-wide, there is additional information to predict the most likely chromatin state annotation for a new sample from the group.

A related challenge is to identify differences in chromatin state annotations between two groups at a high resolution and on a per-state basis. Several methods have been developed for comparing chromatin state annotations between groups of samples, but typically either work at a coarse resolution, or do not identify differences on a per–chromatin-state basis. For instance, ChromDiff (Yen and Kellis, 2015) presents a statistical testing framework to uncover pre-defined broad regions such as gene bodies with significant differences for specific chromatin states across the two groups, but was not specifically designed for detecting differences at the resolution of the chromatin state annotations. EpiAlign (Ge *et al.*, 2019) scores the alignment patterns between two user-input sequences of chromatin state annotations in two samples, hence is also most

applicable for comparing broad domains that encompass multiple chromatin state segments. Another method, chromswitch (Jessa and Kleinman, 2018) also offers a framework to score the differential chromatin state annotations within broader user-specified input genomic locus, and is not designed for detecting chromatin state differences genome-wide at the same resolution of the annotations. EpiCompare (He and Wang, 2017) is primarily a webtool that can be used for detecting cell-type-specific chromatin state differences in terms of enhancer or promoter states, but does not support detecting differences for individual states or other types of chromatin states. SCIDDO (Ebert and Schulz, 2021) conducts fast genome-wide detection of differential chromatin domains between two groups of samples while incorporating a measure of similarity among states. However, as SCIDDO provides a single differential score per position, it does not directly answer the question of which chromatin states change at each genomic position. Another method, dPCA (Ji *et al.*, 2013), works directly on chromatin mark signals and does not quantify state differences across groups of samples.

To effectively summarize the chromatin state annotations for a group of samples and prioritize the chromatin state differences between two groups on a per-state basis, at high resolution, we introduce CSREP. CSREP leverages both the information about the input samples' chromatin states at a position, as well as information of states' co-occurrences in different samples within the same group across the genome. CSREP does this by first generating probabilistic estimates of chromatin state annotations to summarize a group of samples using an ensemble of multi-class logistic regression classifiers. These classifiers predict the state assignment in a sample at a position, given the annotations in other samples at the corresponding genomic position. From those predictions, CSREP is then able to produce a single summary state assignment per position. Furthermore, CSREP can use the difference of summary probabilistic predictions for two groups of samples to quantify the difference in state assignments between the two groups on a per-state basis, e.g. one genome-wide score track per chromatin state. CSREP's ability to summarize chromatin states for a group of samples beyond simple counting is a unique

feature of CSREP relative to existing methods for detecting differential chromatin states or domains mentioned above. CSREP is also distinguished from these existing methods by a combination of (1) considering differential chromatin state annotations at the resolution of the input annotations instead of over broad domains, (2) generating outputs genome-wide instead of at user-specified loci, and (3) providing state-specific and directionally meaningful scores for all states.

Using CSREP, we generate the summary chromatin state maps for 11 groups of tissue/cell types from Roadmap Epigenomics Project (Roadmap Epigenomics Consortium *et al.*, 2015), and for 75 groups from the EpiMap Portal (Boix *et al.*, 2021), which can be easily viewed on genome browsers (**Data Availability**). We show that CSREP can better predict chromatin state assignments in held-out samples than a counting-based baseline method. We also verify that the resulting summary chromatin maps show correspondence with the group's average gene expression profile. Additionally, we show that CSREP's differential scores can recover differential epigenetic signals on chromosome X between Male and Female samples. We also show that CSREP differential scores between samples from two different tissue groups can predict regions of differential peaks for various chromatin marks. The CSREP implementation is designed to be user-friendly and includes a detailed tutorial, available at https://github.com/ernstlab/csrep. We expect CSREP will be a useful tool for summarizing chromatin state maps within groups and finding differences across groups. Additionally, we expect the summary annotations for different tissue groups that we generated with CSREP to be a useful resource.

Results

CSREP method overview

CSREP takes as input chromatin state maps for a group of samples learned in such a way that annotations for different samples have an internally-consistent set of defined chromatin states (Ernst and Kellis, 2010, 2012). We note that the input is presented in BED file format, with each file containing the chromatin state map for one sample. CSREP then generates as output (1) a

summary probabilistic chromatin state assignment matrix and (2) a summary state map track for the group. The summary state assignment matrix represents the probabilities of each state being present at each genomic position in a new sample of that group. To generate these, CSREP takes a supervised learning approach, leveraging information about the co-occurrence of states from the different samples across the genome. Specifically, for each group of input samples, CSREP trains an ensemble of N multi-class logistic regression classifiers (Hastie *et al.*, 2009), where N is the number of samples in the group, to generate probabilistic predictions for each chromatin state at each position **(Fig. 4.1A, Methods)**. We used multi-class logistic regression classifiers since they provide well calibrated probabilities, are robust, and relatively fast to train. Each classifier is trained with *labels* based on the chromatin state assignments from one sample and *features* based on the chromatin states in other samples for the same genomic positions. Each classifier then makes a probabilistic prediction of the chromatin state assigned at each genomic position in the target sample. The chromatin state input features to each logistic regression classifier are represented with a one-hot-encoding of the chromatin states. The classifiers are trained on randomly selected genomic positions that constitute 10% of the genome, while the predictions are calculated genome-wide. The resolution of predictions is the same as that of input samples' chromatin state maps (200bp with default settings for ChromHMM). The prediction results for each sample's chromatin state map are represented in a matrix with *rows* corresponding to genomic positions and *columns* chromatin states. The values in each row sum to 1, representing the probabilities of state assignments at a genomic position. The probabilistic summary of a group is based on averaging the prediction output matrices for each sample in the group. These probabilistic predictions are then used to generate a summary chromatin state map for the group of samples by assigning the state with maximum assignment probability to each genomic position **(Fig. 4.1A, Methods).**

CSREP's summary probabilistic predictions can be directly used to generate differential chromatin state maps for two groups with multiple samples, where the input samples from both

groups share the same internally-consistent set of defined chromatin states. This is achieved by subtracting the summary chromatin state assignment matrices of one group (first group) from the other's (second group) **(Fig. 4.1B, Methods)**. At each genomic position, CSREP's chromatin differential scores for individual chromatin states are bounded between -1 and 1. A score of 1 for state S means state S was predicted to be the annotation for the first and second groups with probability 1 and 0, respectively, and vice versa for -1 (**Fig. 4.1B, C, Supplementary Figure 4.1**). Overall, in addition to summarizing the state assignments for groups of samples, CSREP can calculate scores of differential chromatin state assignments for pairs of groups at the resolution of the input chromatin state maps.

CSREP is predictive of chromatin states on held-out samples

We applied CSREP to a compendium of 18-state chromatin state maps for 64 samples (reference epigenomes) from 11 tissue groups generated by the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium *et al.*, 2015). The tissue groups include embryonic stem cells (ESCs), induced pluripotent stem cells (iPSC), ESC-derived cells, blood & T-cells, HSC & B-cells, epithelial, brain, muscle, heart, smooth muscle and digestive. The numbers of input samples for each tissue group range from 3 to 12 **(Supplementary Data 4.1)**. We provide CSREP's genome-wide summary probabilistic and state assignments for the 11 tissue groups (**Data availability**). Given our computing configuration, the run-time for CSREP to jointly preprocess input data for all 64 samples was ~40 minutes, and then the time to output the predictions for each group ranged from ~1 to 3 hours (**Supplementary Figure 4.2, Methods**).

We first visualized CSREP's summary chromatin state maps for groups of samples from digestive and heart tissue groups, which have 10 and 3 samples, respectively **(Fig. 4.2A, Supplementary Figure 4.3-6)**. We arbitrarily selected four 500-kb regions and for each group, and visualized the input chromatin state maps and CSREP's output probabilistic state estimates and summary state map at such genomic windows. We observed expected correspondences between the groups' input and output chromatin state assignment estimates **(Fig. 4.2A,**

Supplementary Figure 4.3-6). We also visualized CSREP's summary chromatin state maps at the loci of two genes that had distinctly higher expression in Digestive and Brain cell types, LGALS4 and MT3, respectively, which highlighted the corresponding groups' differences in the summary chromatin state maps (GTEx Consortium, 2020) (**Supplementary Figure 4.7-9**).

To quantitatively evaluate CSREP's summary output for a group of samples, we evaluated the accuracy of CSREP's summary probabilistic chromatin state predictions in a leave-one-out cross-validation analysis. In particular, for each chromatin state, we calculated Area Under the Receiver Operating Characteristic curve (AUROC) for predicting genomic locations assigned to the state in a held-out sample, based on the summary chromatin state maps generated from data in other samples from the group **(Methods).** We compared the performance of CSREP against a baseline method, denoted base_count (short for counting-based baseline method), which counts each state's frequency across input samples at each genomic position (**Methods**).

CSREP showed strong predictive performance for chromatin states in left-out samples with average AUROCs across 64 samples varying from 0.871 to 0.993 for the 18 states. Across the 18 states, CSREP consistently had better AUROC in recovering individual states compared to the baseline method base_count **(Fig. 4.2B)**. The average AUROC improvements by CSREP compared to base_count ranged from 0.003 (for state 18_Quies) to 0.157 (for state 4_ TssFlnkD). Larger performance improvements by CSREP relative to base_count were observed for all chromatin states when there are fewer input samples in the group (**Supplementary Figure 4.10**).

CSREP summary chromatin state maps' association with gene expression

Transcription start sites (TSS) are marked by various histone modifications and variants that can correlate with transcription (Kimura, 2013; Soboleva *et al.*, 2014). Here, we evaluated how CSREP's summary state map for a tissue group is predictive of the group's gene expression profiles at TSS of genes. First, we obtained gene expression data for available samples for the 11 tissue groups as above and calculated the average protein-coding gene expression for each group (**Methods**). Of the 11 groups, 8 had gene expression data available for at least one sample

(Methods). We then calculated the Spearman correlation between (1) the group's average expression for protein coding genes and (2) CSREP's summary state assignment probabilities for state 1_TssA (active TSS state) at the corresponding genes' TSSs. We did the same evaluation for base_count. CSREP had significantly higher correlations than base_count (**Fig. 4.2C**, paired t-test p-value < 0.0062, average 0.65 vs. 0.59, **Methods**). We next extended this analysis for a larger dataset for 552 samples in 75 groups from EpiMap repository based on state 1_TssA from the same 18-state annotations (Boix *et al.*, 2021) (**Methods**). The 75 groups were previously formed based on tissue types and developmental stages with the number of samples per group ranging from 3 to 38 (**Methods, Supplementary Data 4.1**). Of the 75 groups, 65 also had gene expression data available for at least one sample. Across these 65 groups, again CSREP had significantly higher correlations than base_count (**Fig. 4.2C**, paired t-test p-val < 2.2e-16, average 0.63 vs. 0.59, **Methods**). Overall, CSREP's summary chromatin state maps at TSS for the TssA state show significantly higher correspondence with gene expression levels compared to the base_count method.

CSREP detects differential chromatin regions associated with different sexes

We next investigated the performance of CSREP at identifying biologically meaningful chromatin state changes between groups of Male and Female samples based on its ability to prioritize chromatin state differences on chromosome X (chrX) relative to autosomal chromosomes. Specifically, we applied CSREP to calculate differential chromatin state scores between 25 Female and 44 Male samples from Roadmap Epigenomics (**Methods**) (Yen and Kellis, 2015; Ge *et al.*, 2019) by subtracting CSREP's summary state probability matrix for the Female samples from the corresponding matrix for the Male samples.

We analyzed CSREP's differential scores for all chromatin states across autosomal chromosomes and chrX (**Fig. 4.3A, Supplementary Figure 4.11-12**). Three states with the largest magnitude of difference in mean Male-Female differential scores between chrX and autosomes were states 13_Het (heterochromatin, marked by H3K9me3), 17_ReprPCWk (weak

polycomb repressive complex) and 18_Quies (quiescent). In contrast, active promoter/enhancer states showed minimal difference in the distribution of Male-Female differential scores for chrX vs. autosomes **(Fig. 4.3A, Supplementary Figure 4.11-12)**. In chrX, compared to autosomal chromosomes, the distribution of differential scores for states 13_Het and 17_ReprPCWk showed a larger tail of negative values. ChrX's average score minus the autosomes' average score values for states 13_Het and 17_ReprPCWk were -0.039 and -0.054, respectively (**Supplementary Figure 4.12**), implying that on chrX, Female samples are more often assigned to these states compared to Male samples. State 18_Quies showed the opposite trend with a difference of 0.11 (**Fig. 4.3A, Supplementary Figure 4.12**). These results are consistent with sex-specific chrX inactivation, which is used in Female mammals to achieve dosage compensation between the two sexes (Wutz, 2011; Yen and Kellis, 2015).

We next compared the performance of CSREP and other methods in recovering annotated transcription start sites (TSSs) on chrX, using the above-mentioned states, given varying numbers of input samples (**Methods**, **Fig. 4.3B**). To do this, we randomly selected 30 subsets of size n Male and n Female samples from the set of available 44 Male and 25 Female samples, where n is varied within the set of 3, 5, 9, 12 or 15 samples. Given each set of input Male and Female samples, we calculated the AUROC when using differential chromatin scores between Male and Female groups to predict locations overlapping annotated TSSs on chrX, against the background of those overlapping all annotated TSSs in the genome (**Methods**). The methods we compared CSREP against include SCIDDO, the count difference from base_count, the Mann-Whitney U test (used by ChromDiff (Yen and Kellis, 2015)), and the Fisher's exact test (used by EpiCompare (He and Wang, 2017)) (**Methods**). The Mann-Whitney U and Fisher's exact tests were applied at each genomic position, using two sample groups' chromatin state annotations at the respective position. We considered other related methods for detecting differential chromatin domains not appropriate for direct comparison against CSREP (**Methods**). We observed that CSREP showed the largest advantage over other methods, as measured by

AUROCs, when the number of input samples from Male and Female groups is relatively small, e.g. 3 samples in each group (**Fig. 4.3B**). As the number of input samples from each group increases sufficiently, the overall performance advantage of CSREP relative to base_count, Mann-Whitney U test and Fisher's exact test goes away. In all cases, CSREP showed better performance compared to SCIDDO (Ebert and Schulz, 2021) (**Fig. 4.3B**). Overall, CSREP showed the clearest advantage over other approaches when the number of samples is relatively small, which occurs frequently in practice.

CSREP differential scores recover differential chromatin mark peaks

We next analyzed how well CSREP's, base-count's and SCIDDO's differential chromatin state scores can predict genomic regions overlapping differential signals of DNase I hypersensitivity (DNase), H3K9ac and H3K27ac between samples from embryonic stem cell (ESC) and brain. DNase and H3K9ac signals were not used for learning the 18-state model used to annotate the two groups' input samples, providing an independent validation. While H3K27ac was used in learning the input chromatin state maps, since all the methods being compared (CSREP, base_count, SCIDDO, Mann-Whitney U test based on ChromDiff and Fisher's exact test based on EpiCompare) had access to the same chromatin state maps as input, and H3K27ac is a well-established mark of cell-type specific activity (Creyghton *et al.*, 2010), we still considered H3K27ac in the evaluations of methods' performance.

For each of the three chromatin marks, we first obtained a set of bases that overlap with peaks in all samples from ESC but not in any from the Brain group and vice versa (**Methods, Supplementary Data 4.1**). We then calculated CSREP and base_count differential chromatin scores by subtracting the summary chromatin state map of Brain from that of the ESC. Additionally, we applied SCIDDO, Mann-Whitney U test (ChromDiff's approach) and Fisher's Exact test (EpiCompare's approach) to the same set of input data (**Methods**). We evaluated, in terms of AUROC, how well the methods prioritize regions overlapping bases in the ESC-/brain-specific sets of peaks (**Methods**). For CSREP, base_count, Mann-Whitney U test and Fisher's exact test,

we conducted separate evaluations for each chromatin state, but did *not* for SCIDDO since it outputs one score track that measures the overall difference across the chromatin state landscape between the two groups.

Across the different marks and groups (ESC-specific or Brain-specific peaks) we evaluated, CSREP's differential scores from either promoter- or enhancer- associated states result in the highest AUROCs, with few exceptions **(Fig. 4.4, Supplementary Figure 4.13).** For example, for identifying Brain-specific H3K9ac peaks, CSREP had an AUROC of 0.717 based on the evaluation with state 9_EnhA1, an active enhancer state, while the maximum AUROCs achieved for base_count, Mann-Whitney U test, Fisher's exact test and SCIDDO were 0.617, 0.636, 0.601 and 0.564, respectively. In total across the six evaluations, among the top-3 highest AUROCs per evaluation, 15 of the 18 AUROCs were based on CSREP's differential scores for individual chromatin states (**Fig. 4.4**). The AUROCs for states not usually associated with these marks (transcription, heterochromatin, repeats/ZNF gene, quiescent, polycomb repressed states) tended to be near 0.5 or in some cases lower (**Supplementary Figure 4.13**). These analyses suggest that CSREP differential scores tended to better correspond to locations of individual mark differences between two groups of samples genome-wide, compared to other approaches. Even though SCIDDO incorporated a measure of dissimilarity among states, it showed lower AUROCs compared to the maximum obtained by CSREP. This is potentially because SCIDDO outputs one score per genomic bin to measure the general difference across all states, while CSREP generates state-specific scores. Hence, CSREP should have better power to predict regions associated with differential signals of marks that are present in only specific states (e.g., H3K27ac is present in enhancer states but not in repressive states). Additionally, this may also be because CSREP produces scores that show the direction of differences (with positive/negative scores implying one group's higher state assignment probabilities compared to the other's) while SCIDDO's scores do not have a specific direction associated with them.

Discussion

Here, we proposed CSREP, a method for probabilistically summarizing the chromatin state maps from a group of samples. CSREP achieves this by training multi-class logistic regression models to predict the chromatin state annotations of one sample using data from others, and then averaging the prediction probabilities across all samples in the group. CSREP outputs the probabilities of each chromatin state being assigned to each genomic position, at the same resolution that chromatin states are annotated. We applied CSREP to generate summary 18-state chromatin state assignment probability matrices for 11 groups of cell and tissue types from Roadmap Epigenomics Project (Roadmap Epigenomics Consortium *et al.*, 2015), and 75 groups of samples stratified by cell and tissue types and developmental phases from EpiMap (Boix *et al.*, 2021), and have made them publicly available (**Data Availability, Supplementary Data 4.1**).

Our analyses reveal that CSREP's probabilistic summary of state assignments better predicts the chromatin states of held out samples compared to the counting-based baseline approach. We also showed that CSREP's summary assignment probabilities of state 1_TssA at TSS were well correlated with the average gene expression of the group, and significantly higher than those achieved by the counting-based baseline.

CSREP can also be used to directly quantify the difference in chromatin state maps between two groups with multiple samples, at the resolution of the input annotations. CSREP produces differential scores for each chromatin state at each genomic position, which represent the difference in probabilities that samples from two input groups are assigned to each specific state. Therefore, CSREP differential scores are bounded (-1 to 1), interpretable with respect to specific chromatin state changes, and indicative of the direction of change, which contrasts it with other approaches that provide a single score showing magnitude of difference per genomic position. We used CSREP to compare the chromatin state annotations between Male and Female samples from Roadmap Epigenomics (Roadmap Epigenomics Consortium *et al.*, 2015), and

showed that CSREP can better predict regions overlapping genes' TSS on chrX, particularly when there are few samples in each group. CSREP's differential scores for states associated with active enhancers and promoters better recovered tissue-group-specific peaks of DNase/H3K27ac/H3K9ac signals compared to alternative approaches, suggesting that CSREP provides useful additional information for analyzing epigenomic changes across tissue types.

Here, we presented applications of CSREP on samples that were grouped based on cell and tissue types and based on sex. In general, CSREP assumes the dominant signal of any variation between groups is associated with the grouping variable of interest. In cases in which the experimental design used to collect the data cannot ensure this, other known covariates can be used to detect if there are potential confounders.

CSREP works directly off of chromatin state annotations, which makes CSREP agnostic to the specific methods used to produce those annotations. Some methods for learning chromatin state annotations have the option to expose posterior probability estimates of annotations. However, in general it is not clear how well calibrated those estimates will be, and assuming accurately determined posterior probability estimates are available as input would also make CSREP less generally applicable. A possible direction for future work would be to extend CSREP to make use of posteriors or possibly other information that CSREP does not directly consider, such as the individual mark signal in each sample.

We note that CSREP's summary chromatin state maps offer complementary benefits to the recently developed universal chromatin state annotation, which provides a single integrative annotation of the genome based on a model defined from over a 1000 epigenomic datasets from over 100 cell and tissue types (denoted the full-stack model) (Vu and Ernst, 2022). The full-stack model jointly captures activity across many diverse cell and tissue types and hence can capture annotations corresponding to both constitutive and cell-type-specific activities (**Supplementary Figure 4.1**). CSREP, on the other hand, provides a more direct and focused chromatin state annotation representative specifically of the individual input samples' annotations.

To facilitate the use of CSREP, we provide an implementation of CSREP as a snakemake pipeline (Mölder *et al.*, 2021; Köster and Rahmann, 2012) with a detailed tutorial that only requires users to modify parameters in a yaml file. The program can be run either on local computers or on computing clusters, in which case snakemake will optimize the workflow for execution.

We expect CSREP to be a useful tool and the CSREP output we provided to be a valuable resource for summarizing chromatin state maps from groups of samples, and for prioritizing regions with differential chromatin state changes across pairs of groups of samples.

Methods

CSREP's summarization of a group of samples

Let G denote the number of genomic bins across the genome, S the number of chromatin states, and N the number of samples in the target group of samples. Let $C_{i,n}$ denote the chromatin state assigned to sample n at genomic position i, which can take one value of $1, 2, \ldots, S$. Let N_n denote the set of samples not including n, i.e. $N_n = \{1, \ldots, N\} - \{n\}$. In general, CSREP is an ensemble of N multi-class logistic regression classifiers such that for each sample n, CSREP trains a classifier to predict the chromatin state map of this sample based on features from the remaining samples (N_n). The predictor variables for such a model include one-hot encoding chromatin state maps of the $N - 1$ samples (all samples in the group except n) and an intercept term, resulting in $(N - 1) \times S + 1$ predictor variables. The response variable is the chromatin state of the target sample n, which can take one value of $1, 2, \ldots, S$.

In the multi-class logistic regression model, let X_i denote the vector of predictor variables at position i, which has length $(N - 1) \times S + 1$ and takes values $\{0,1\}$. The last entry of X_i is 1, corresponding to the intercept term. Let Y_i denote the value of the response variable at position i, which takes values $\{1,2, \ldots, S\}$. Since the input chromatin state maps that we used segmented the genome into 200-bp bins, we refer to each genomic position as one 200-bp window in the genome. We randomly selected genomic positions for the training data set, such that these positions constitute 10% of the genome. We chose 10% as the training proportion because increasing this parameter does not result in considerable increase in model accuracy at the cost of increased runtime (**Supplementary Figure 4.14)**. Given the training data set, for each state $s \in \{1, \ldots, S - 1\}$, the multi-class logistic regression model learns a coefficient vector β_s with length $(N - 1) \times S + 1$, corresponding to the number of predictor variables. The probability of sample n's chromatin state s being assigned at position i is then calculated as:

$$P(Y_i = s) = \frac{e^{\beta_s \times X_i}}{1 + \sum_{j=1}^{S-1} e^{\beta_j \times X_i}}$$

for $s \in \{1, .., S-1\}$, and as the following when $s = S$:

$$P(Y_i = S) = \frac{1}{1 + \sum_{j=1}^{S-1} e^{\beta_j \times X_i}}$$

The model is implemented using Python's sklearn, pybedtools package and snakemake (Dale *et al.*, 2011; Quinlan and Hall, 2010; Mölder *et al.*, 2021; Köster and Rahmann, 2012). A L2-norm penalty with the default regularization strength of 1.0 was used for training. CSREP applies the model to generate predictions of genome-wide probabilistic chromatin state map for sample n, which is presented in a matrix of size $G \times S$. The output matrices from N predictions for N samples are then averaged, so at each genomic bin, the sum of state assignment probabilities across S states is 1. In addition, the chromatin state with the maximum probability in each row is recorded to produce a single representative chromatin state map for the entire group of samples.

CSREP's application to prioritizing differential chromatin state changes between two groups of samples

To calculate differential chromatin state maps between two groups of samples, group1 and group2, CSREP first calculates the probabilistic chromatin state map matrices for each group as described above, denoted as R_1 and R_2, respectively. After this, CSREP subtracts the two matrices to represent the differential chromatin state map between group1 and group2 (denoted D_{12}), i.e. $D_{12} = R_1 - R_2$. We note that we used signed and not absolute difference here and thus the score ranges from -1 to 1. A score on row i and column s of D_{12}, denoted $D_{12,i,s}$, being -1 means group2 is estimated to have probability 1 of being assigned to state s at position i while group1 has probability of 0. Additionally, since CSREP assigns S scores of differential chromatin maps to each genomic position i, corresponding to S states, CSREP can uncover specific chromatin state changes. For example, if $D_{12,i,s} = 0.8$ when $s = 1$ while $D_{12,i,s} = -0.8$ when $s =$

2, we can infer that at position i, group1 is likely to be in state 1 while group2 is likely to be in state 2.

Primary data sources

We analyzed genome-wide 18-state chromatin state annotation for 64 reference epigenomes from the Roadmap Epigenomic Project Portal (Roadmap Epigenomics Consortium *et al.*, 2015) and 552 from the EpiMap portal (Boix *et al.*, 2021). We will refer to each reference epigenome as a sample. State annotation data for samples from Roadmap Epigenomics and EpiMap were in hg19. The 18-state model was shared between Roadmap Epigenomics and EpiMap, and was trained based on data of 6 chromatin marks: H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3 and H3K9me3. We assigned 64 samples from Roadmap Epigenomics into 11 groups based on the accompanying metadata's tissue group labels. These groups include Blood & T-cell, Brain, Digestive, embryonic stem cells (ESC), ES-deriv, Heart, induced pluripotent stem cells (iPSC), Muscle, Skin, smooth muscle (Sm_Muscle) and HSC & B-cell. We assigned the biosamples from EpiMap into 75 distinct groups based on the metadata corresponding to unique combination of: extended biosample summary (tissue and sub-tissue types) and life stage (adult or embryonic, any biosamples with samples of unknown life stage filtered out from the analyses). We only analyzed groups of samples from EpiMap with at least 3 biosamples. Among sample groups from Roadmap Epigenomics, the number of samples per group ranged from 3 to 12, while the corresponding range for samples from EpiMap is 3 to 38. Details about the samples' ID, groups and other metadata are provided in **Supplementary Data 4.1**.

Using CSREP, we generated summary chromatin state maps for chromosomes 1-22 and X for the 11 groups from Roadmap Epigenomics and 75 groups from EpiMap using input data in hg19. We ran CSREP on a high-performance compute cluster where each job was allocated 4 cores with 4 GB of memory per core. The run-time for CSREP to jointly preprocess input data for all 64 samples from Roadmap Epigenomics was ~40 minutes, and then the time to output the

predictions for each group ranged from ~1 to 3 hours (**Supplementary Figure 4.2, Methods**). We then used liftOver from the UCSC genome browser to lift the summary state maps for all groups from either Roadmap Epigenomics or EpiMap from hg19 to hg38. This procedure first finds a one-to-one mapping for a subset of 200-bins between hg19 and hg38, i.e. if there are multiple bins from hg19 that got mapped to the same bin in hg38, those bins would not be included into the annotations. Then, we map both the state assignment probabilities and the summary state annotations for each bin in hg19 to the corresponding bin in hg38. Source code for the liftOver procedure, along with a detailed tutorial, is provided at https://github.com/ernstlab/csrep.

Evaluation of CSREP in representing a group's chromatin state maps

We evaluated CSREP and an alternative baseline approach called base_count (defined below) for predicting representative chromatin state maps. We conducted this evaluation through a leave-one-out cross validation framework. Given a group with N samples, for each sample indexed n, we evaluated the prediction of chromatin state map for sample n when the state maps of the other $N-1$ samples were used as the input for generating the predictions. For these evaluations we used the data for the 64 samples from the 11 tissue groups from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium *et al.*, 2015) described above.

Base_count method: Let $C_{nis} = 1$ if in sample n, at genomic position i, the observed chromatin state is s, and $C_{nis} = 0$ otherwise. The base_count approach represents the group's chromatin state map by calculating the frequency of state s being assigned at the i position (BC_{is}) across the samples. In particular:

$$BC_{is} = \frac{\sum_{n=1}^{N} C_{nis}}{N}$$

where N is the number of samples. Similar to CSREP, the output matrix for base_count method is of size $G \times S$, with the sum of values in each row being 1, where G and S

represent the number of genomic bins and the number of states, as explained in the main Methods.

Calculating the ROC curves of prediction for a single chromatin state's location: In each round of cross-validation, one sample with index n is held-out. We then used CSREP and base_count to get the summary probabilistic chromatin state map for the group using input data from the remaining $N-1$ samples. For each state s, CSREP and base_count output the summary probability that each 200-bp genomic bin gets assigned to the state s. We divided the $[0,1]$ probability range into 500 equal-width windows with lower bounds $l \in \{0, 0.002, \ldots, 0.998\}$. Within each probability window, any genomic positions with assignment probability for the state s being no less than the window lower bound (l) will be predicted as being in state s for sample n. Given the true chromatin state map in sample n, we then calculated the cumulative true positive rates and false positive rates of the prediction at each probability threshold l to obtain the ROC curve. This analysis is repeated for each chromatin state, resulting in S ROC curves.

Evaluating CSREP's summary chromatin state maps' association with gene expression

We obtained gene expression data from the Roadmap Epigenomics Consortium (Roadmap Epigenomics Consortium *et al.*, 2015), which was available as a matrix of values in RPKM (reads per kilobase million) for genes in a subset of the samples from https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz.

We obtained the accompanying gene annotation information from https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/Ensembl_v65.Gencode_v10.ENSG.gene_info.gz.

We filtered out genes that were not annotated as protein-coding, then transformed the gene expression matrix by adding a pseudo-count of 1 to the RPKM counts, and finally log-transformed the resulting values. We also only included genes that were on chromosomes 1-22

and X in hg19 for this analysis, resulting in 20,787 distinct TSSs whose associated gene expression was available. For each of the 11 groups of samples in Roadmap Epigenomics, we obtained the group's average gene expression profile by averaging over the gene expression values across all available samples in the group, i.e. samples that are both in the group and among the samples whose expression data was available. Among the 11 groups, 8 groups (all except for groups HSC & B-cell, iPSC and Sm_Muscle) had available gene expression data from the available samples. We then calculated the Spearman correlation of the summary chromatin state assignment probabilities for the 1_TssA state at positions that overlap with the 20,787 annotated TSSs and their corresponding average gene expression for each group.

For EpiMap, we obtained data of quantile-normalized protein coding genes' expression from https://personal.broadinstitute.org/cboix/epimap/rnaseq_data/merged_qn_log2fpkm.pc.mtx.gz, which is available as values in log2(FPKM), along with data of samples' ID and genes' Ensembl ID. We utilized the same genes' annotation information as provided by Roadmap Epigenomics, and only included protein-coding genes on chromosomes 1-22 and X in hg19. This resulted in 18,543 distinct TSSs whose associated gene expression was available from EpiMap. Among the 75 groups, 10 groups (SMTH.Digestive, HSC.MPP, EYE.EMB, EPTH.BREAST.EPITH, CA.UCEC, CA.RCC, CA.MYELOMA, BRN.EMB.BRN, BRN.CAUD.NUC, BONE.EMB) had no available gene expression data. We followed the same procedure mentioned above for the remaining 65 groups to obtain the Spearman correlation between a group's average gene expression and summary state assignment probabilities for the 1_TssA state.

We used a paired t-test to compare the correlations resulting from CSREP against those from base_count, with the alternative hypothesis that CSREP's correlations with gene expression are higher than base_count's.

Applications of existing methods for detection of differential chromatin state domains across two sample groups

In this section, we denote a *chromatin state* as the learned states that represents the combinatorial patterns of chromatin mark signals, outputted by methods such as ChromHMM and Segway (Libbrecht *et al.*, 2021, Ernst and Kellis, 2010, 2012, Hoffman et al., 2012). A *chromatin state annotation* then denotes the assignment of chromatin states (e.g. active promoter, enhancer, quiescent states, etc.) to each *genomic bin (*or equivalently*, genomic position*), which are outputted by methods that discover chromatin states and conduct genome segmentation and annotation (e.g. ChromHMM and Segway) (Libbrecht *et al.*, 2021, Ernst and Kellis, 2010, 2012, Hoffman et al., 2012). Additionally, a *genomic region* or *domain* denotes a window along the genome that can span one or multiple genomic bins for which the chromatin state segmentation and annotation is defined. We note that CSREP is designed to output one differential score with respect to each chromatin state s at each genomic bin i. CSREP generates output genome-wide instead of at specific user-defined genomic region. The output of CSREP, therefore, is a matrix of size $G \times S$, where S denotes the total number of chromatin states and G denotes the total number of bins in the genome.

Similar to CSREP, the difference between base_count's summary chromatin state maps for two groups of samples can be used to calculate the base_count's differential chromatin scores to compare against those generated by CSREP. In addition, we compared CSREP's differential chromatin state scores to three additional differential scores based on the approaches used from existing methods: SCIDDO, ChromDiff, and EpiCompare (Ebert and Schulz, 2021; Yen and Kellis, 2015; He and Wang, 2017) (see below). We decided not to apply EpiAlign or chromSwitch to compare against CSREP since these methods are intended for cases where users are interested in measuring the differential chromatin state maps at a particular broad region of the genome (spanning multiple bins of chromatin state assignments) (Ge *et al.*, 2019; Jessa and Kleinman, 2018). Meanwhile, CSREP scores the genome-wide differential chromatin state maps at the same resolution as the input annotations (e.g. 200bp bins here). We did not compare CSREP against dPCA (Ji *et al.*, 2013) since it aims at scoring genomic regions' differential epigenetic patterns

directly from the chromatin signal data, while CSREP aims to uncover differential chromatin domains from input chromatin state maps.

***SCIDDO**:* To run SCIDDO, we followed the tutorial provided by the author on Github https://github.com/ptrebert/sciddo/blob/master/testdata/tutorial.md (Ebert and Schulz, 2021), and generated a list of differential chromatin domains between two conditions, where each domain can be one genomic bin of chromatin state (200-bp bin) or multiple bins. SCIDDO's output contained overlapping differential chromatin domains with different values of SCIDDO scores. For each genomic bin, we averaged the SCIDDO differential scores across overlapping differential chromatin domains, and assigned 0 to genomic bins not reported in SCIDDO's output, implying no differential signals in chromatin states between the two groups.

ChromDiff: ChromDiff's provided implementation is specifically designed to calculate differential scores for annotated genes (Yen and Kellis, 2015). Therefore, we could not directly use ChromDiff software to obtain differential scores at the same resolution as CSREP (200bp in all presented analyses). Instead, we directly implemented the same statistical test as used in the ChromDiff paper, the Mann-Whitney U-test, to determine differences in the number of samples from each group being annotated as a state s at each genomic position i (Yen and Kellis, 2015). For example, if 3 out of 5 samples in group 1 and 0 out of 7 samples in group 2 are annotated as state s at position i, then we applied the Mann-Whitney U test with two input vectors $[1,1,1,0,0]$ and $[0,0,0,0,0,0,0]$ for state s at position i. The test was implemented using Python's scipy package, and the alternative hypothesis (one-sided vs. two-sided test) was set based on the analysis purpose (see sections below about evaluating CSREP's differential scores in various analyses). The statistical test's output p-values of 1.0 imply no difference and of 0.0 imply highest difference between two groups. We converted such p-values into

differential score by the function $score = 1 - p_{value}$, so that the scores are bounded $[0,1]$, with higher value implies more differences across the two groups. The output of ChromDiff is a matrix of size $G \times S$, where G is the total number of genomic bins, and S is the number of chromatin states.

EpiCompare: As EpiCompare (He and Wang, 2017) only supports comparisons of specific groups of states (enhancer and promoter state groups only), we could not directly compare CSREP with EpiCompare. However, we did reimplement the Fisher's exact test, which is a statistical test supported by EpiCompare. Specifically, for each genomic bin i and chromatin state s, a contingency table is constructed indicating the number of samples from each group that are annotated as state s (or not as state s). We then applied Fisher's exact test using Python's scipy package to evaluate the significance of differential annotations between the two groups. The alternative hypothesis to the Fisher's exact test was set based on the application purpose (see sections below about evaluating CSREP's differential scores in various analyses). The differential scores were also obtained by the function $score = 1 - p_{value}$, to ensure that higher scores imply higher levels of difference across the two groups.

Evaluating CSREP's differential chromatin state maps between Male and Female groups in recovering chromosome X- associated genomic regions

For evaluating chromatin state differences between Male and Female groups with respect to chromosome X and autosomes, we obtained data for samples whose sex is annotated as either Male or Female (not 'Unknown' or 'Mixed'), according to the provided metadata for the samples with 18-state chromatin state maps from Roadmap Epigenomic Project (Roadmap Epigenomics Consortium *et al.*, 2015). In total, there are 44 Male samples and 25 Female samples. We then generated 30 sets of 3 male samples (randomly chosen from 44 Male samples) and 3 Female samples (randomly chosen from 25 Female samples). We calculated the differential chromatin scores for the 30 sets of samples using CSREP, base_count and SCIDDO, Mann-Whitney U test

(based on ChromDiff) and Fisher's exact test (based on EpiCompare) (see *Applications of existing methods for detection of differential chromatin state domains across two sample groups*). For each set of input samples from Male and Female groups, we used the *two-sided* statistical tests for ChromDiff and EpiCompare, and obtained the *absolute values* of CSREP and base_count differential chromatin scores. Then, we obtained a list of annotated TSSs for *protein-coding* genes from the accompanying metadata of genes' coordinates and strand provided by the Roadmap Epigenomics project. For each set of 3 male and 3 Female samples, we obtained the differential chromatin scores from each method, as outlined above, for regions that overlap these TSSs, and divided the score range window into 100 equal-width bins. The score ranges from CSREP, base_count, Mann-Whitney U test and Fisher's exact test is [0, 1] and from SCIDDO is $[0, maximum\ value]$. We then applied the same procedure as outlined in the above section (titled '*Calculating the ROC curves of prediction for a single chromatin state's location*') to obtain true-/false- positive rates and AUROCs in predictions of TSS-overlapping regions on chromosome X, among all genomic regions overlapping annotated TSSs on the autosomes and chrX. We repeated the same analysis for a total of 30 sets of n male and n Female samples, with $n \in \{3,5,9,12,15\}$. We note that not all 30 rounds of application of SCIDDO to calculate differential chromatin scores ran successfully, due to software failure. In particular, all applications of SCIDDO with 30 input sets of 15 male and 15 Female samples $(n = 15)$ failed. For such cases, we report the average AUROCs for only successful runs of SCIDDO (**Fig. 4.3B**).

Evaluating CSREP's differential chromatin state map in recovering regions associated with differential chromatin mark signals

To evaluate recovering differential chromatin mark signals, we first downloaded the available broad peaks of DNase, H3K9ac and H3K27ac for samples from the ESC and Brain groups from Roadmap Epigenomics Project at https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak. A full list of

links to data used in this analysis is provided in **Supplementary Data 4.1**. For each of the three chromatin marks and each cell group (ESC or Brain), we used $bedtools\ intersect$ (Quinlan and Hall, 2010) to obtain a set of peaks that are shared across all samples in the respective cell group. We then used $bedtools\ subtract$ function to derive peaks that are present in ESC samples and missing in Brain samples, and vice versa. We treated these ESC-specific, Brain-specific chromatin peaks as the ground-truth for this analysis. The number of base pairs overlapping peaks for each group range from 2,735,377 bp (ESC-specific H3K9ac peaks) to 85,995,111 bp (Brain-specific H3K27ac peaks) (**Supplementary Data 4.1**).

We generated differential chromatin state scores from CSREP and base_count for the ESC and Brain groups for Roadmap Epigenomics, by subtracting their probabilistic chromatin state predictions for Brain from those for ESC. We also applied the Mann-Whitney U test (based on ChromDiff) and Fisher's exact test (based on EpiCompare) for the two groups of samples for each chromatin state (see section *Applications of existing methods for detection of differential chromatin state domains across two sample groups).* For the task of predicting ESC-specific peaks, Mann-Whitney U test (ChromDiff) and Fisher's Exact test (EpiCompare) were applied such that ESC samples were used as the foreground (first) sample set, and Brain samples as the background (second) set. The foreground and background sample sets were reversed for the task of predicting Brain-specific peaks. P-values were obtained from these two methods using one-sided test, with the alternative hypothesis that a state s is more likely to be annotated at position i in the foreground samples than in the background samples. The differential scores Mann-Whitney U test (ChromDiff) and Fisher Exact test (EpiCompare) were converted as $score = 1 - pValue$, so that higher score (bounded in $[0,1]$) implies higher magnitude of difference between the two groups of samples. The differential chromatin score matrices from CSREP, base_count, Mann-Whitney U test (ChromDiff) and Fisher's exact test (EpiCompare) methods are of size $G \times S$, and denoted D_{CSREP}, D_{BC}, D_{MannW} and D_{Fisher}, respectively, where $D_{CSREP,i,s}$,

$D_{BC,i,s}$, $D_{MannW,i,s}$, $D_{Fisher,i,s}$ denote the corresponding methods' differential score for state s at genomic position i, respectively. We also obtained SCIDDO scores, which measure genome-wide differential chromatin patterns between the two groups. We denote the genome-wide SCIDDO score vector as D_{SCIDDO}, and $D_{SCIDDO,i}$ as the score at genomic position i.

To use D_{CSREP} or D_{BC} to calculate the ROC for genome-wide prediction of bases associated with ESC-specific or Brain-specific chromatin marks' peaks, we divided the score range window $[-1, 1]$ into 200 equal-width bins with lower bounds $l \in \{-1, -0.99, \ldots, 0.99\}$. To calculate ROCs for predicting bases in ESC-specific peaks, for each state s and each differential score lower-bound l, we defined genomic positions where the differential scores for state s being greater than or equal to l, denoted $\{i: D_{--,i,s} \geq l\}$. We compared such predictions with the ground-truth peaks described above to obtain the true- and false-positive rates of prediction for state s. To calculate ROCs for predicting bases associated with DNase/H3K9ac/H3K27ac Brain-specific peaks, first, we reversed the sign of D_{CSREP} and D_{BC}, resulting in differential score matrices where positive values for state s at position i implies that the respective position (i) has a higher probability of being in state s in Brain compared to ESC. Then, we applied a similar procedure as outlined above to calculate CSREP's and base_count's ROCs.

We applied a similar procedure as outlined above for CSREP and base_count to calculate ROC curves based on D_{MannW}, D_{Fisher} and D_{SCIDDO}, except using ranges of differential scores other than $[-1, 1]$. For Mann-Whitney U test and Fischer's exact test, the range was $[0,1]$. For SCIDDO, the scores range was based on the observed minimum and maximum scores across the genome. We then applied the same procedure as for CSREP and base_count scores in each state to obtain the ROCs, as mentioned above.

Data availability

The summary chromatin state maps (the chromatin state assignment matrices and the corresponding state annotation) for 11 tissue groups in Roadmap Project and 75 groups in EpiMap Portal are available for download with links from https://github.com/ernstlab/csrep, and

can be viewed on UCSC Genome Browser with the track hub link from https://github.com/ernstlab/csrep. The summary state maps for samples in Roadmap Epigenomics and EpiMap are provided for both hg38 and hg19.

Figures

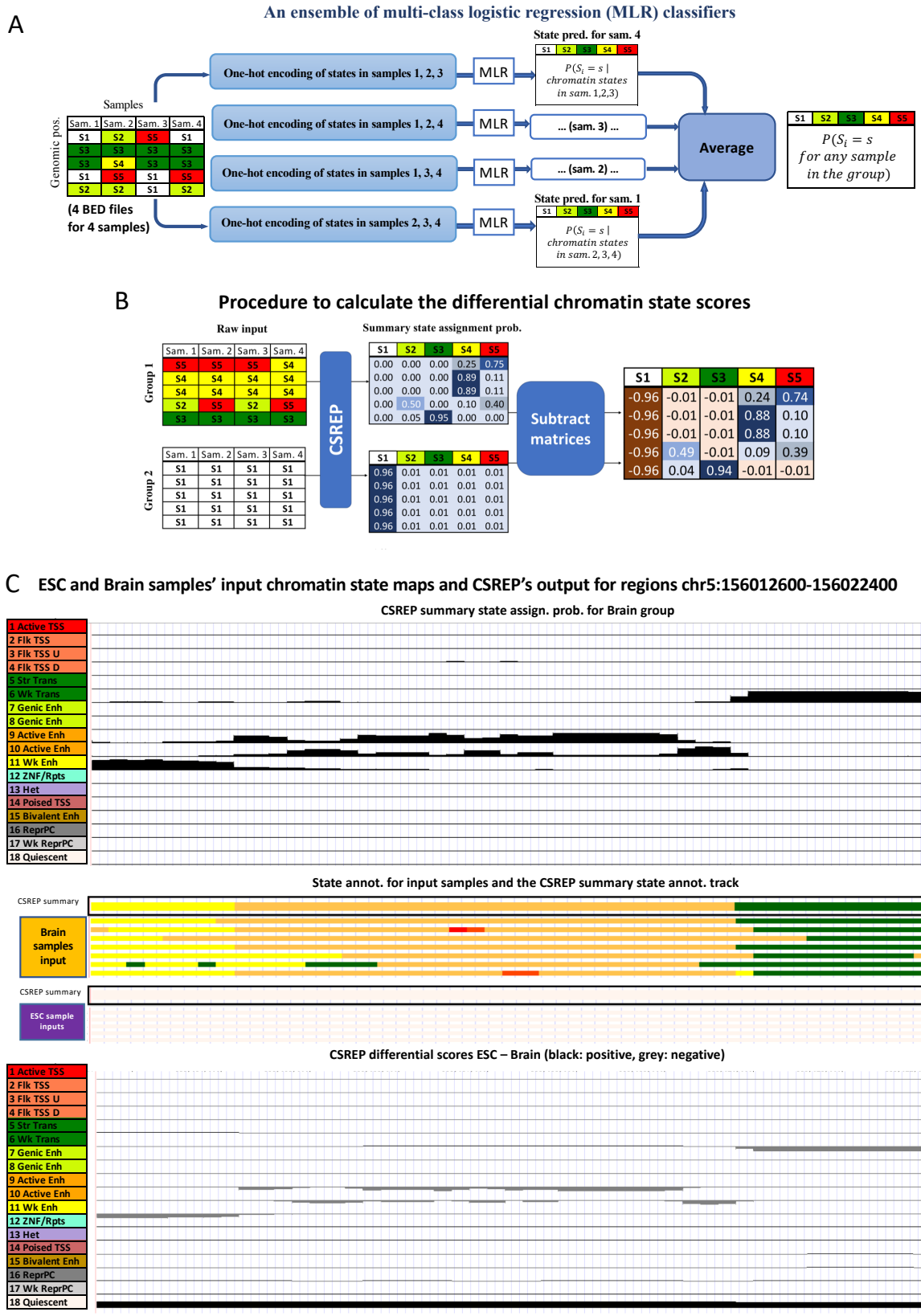


Figure 4. 1: Overview of CSREP.

(A) CSREP uses an ensemble of multi-class logistic regression models. In each model, the chromatin state map at the target sample is predicted based on the one-hot encoding of chromatin state assignments at the corresponding genomic positions in other samples. Multi-class logistic regression outputs the probabilities that each genomic position (row) in the target sample will be assigned to each state (column). CSREP averages the prediction matrices for target samples, to output the summary state assignment probability matrix. Sam.: sample; $P(S_i = s)$: probability that genomic position i is annotated as state s. **(B)** The operations to obtain differential chromatin state assignment scores between two groups with multiple samples. CSREP calculates the summary chromatin state assignment matrices for two groups and then subtracts one group's summary matrix from the other's to obtain differential chromatin scores. Differential chromatin scores are bounded between -1 (brown) and 1 (blue). **(C)** Visualization of CSREP's output in a genomic region (hg19, chr5:156,012,600-156,022,400). The top of the subpanel shows the CSREP's summary chromatin state probabilities for 18 states across seven Brain reference epigenomes. Each track shows the probabilities of assignment for one state, as named and colored on the left. The middle subpanel shows the 18-state chromatin state maps for 7 Brain samples and 5 ESC samples from Roadmap Epigenomics (Roadmap Epigenomics Consortium *et al.*, 2015), and the CSREP's output summary chromatin state maps for each group, outlined in black. States are colored as in legends at the left of this subpanel. The last subpanel shows the differential chromatin scores when ESC's summary state probabilities are subtracted from Brain's. Each track shows one state's differential scores. Scores between 0 and 1 are colored black, while those between -1 and 0 are colored grey. This region is also shown in an expanded format in **Supplementary Figure 4.1**.

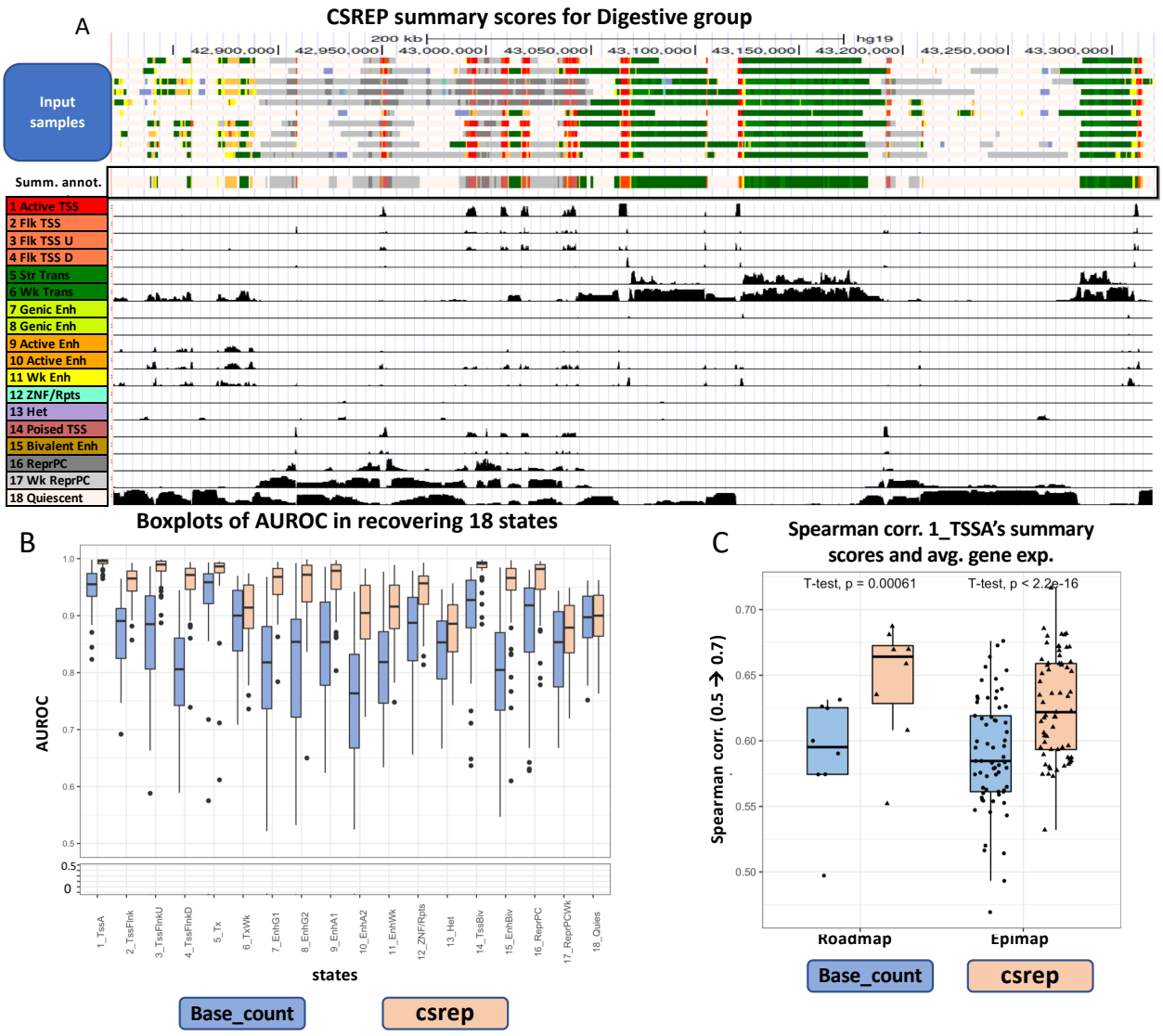


Figure 4. 2: Performance of CSREP in summarizing multiple samples' chromatin state maps from a group.

(A) Visualization of one arbitrarily selected 500-kb region (chr5: 42,821,109-43,321,109, hg19). The first 10 tracks show chromatin state maps of 10 samples of the Digestive group from the Roadmap Epigenomics Consortium, which were input to CSREP. The following track shows the summary chromatin state map from CSREP, which shows strong agreement with the input. States are colored based on the legend on the lower left. In the following 18 tracks, each track shows CSREP's probabilities of assignment for each of 18 states, with the state annotations shown in the legend on left.

(B) Boxplots showing the CSREP and base_count methods' average, range and 25, 75% quantiles of the AUROCs across 64 samples, for each of the 18 chromatin states. The AUROCs were calculated in leave-one-out cross validation analysis where we used a group's summary probabilistic chromatin state map to predict genomic locations of individual chromatin states in a left-out sample from the same cell/tissue group (**Methods**). States 1-18 (x-axis) are annotated as in **(A).**

(C) Boxplots showing the Spearman correlations between a group of samples' (1) summary probabilities of state 1_TssA (active TSS) at annotated TSSs, and (2) the corresponding group's average gene expression (**Methods**). We obtained the correlations for 8 groups of cell types from the Roadmap Epigenomics Project, and 65 groups from EpiMap. Each dot shows the Spearman correlation for data from a group of samples. Results of paired t-test to compare CSREP vs. base_count's output correlations are shown on top. The alternative hypothesis for the t-test is that correlations resulted from CSREP are higher than those from base_count (**Methods**).

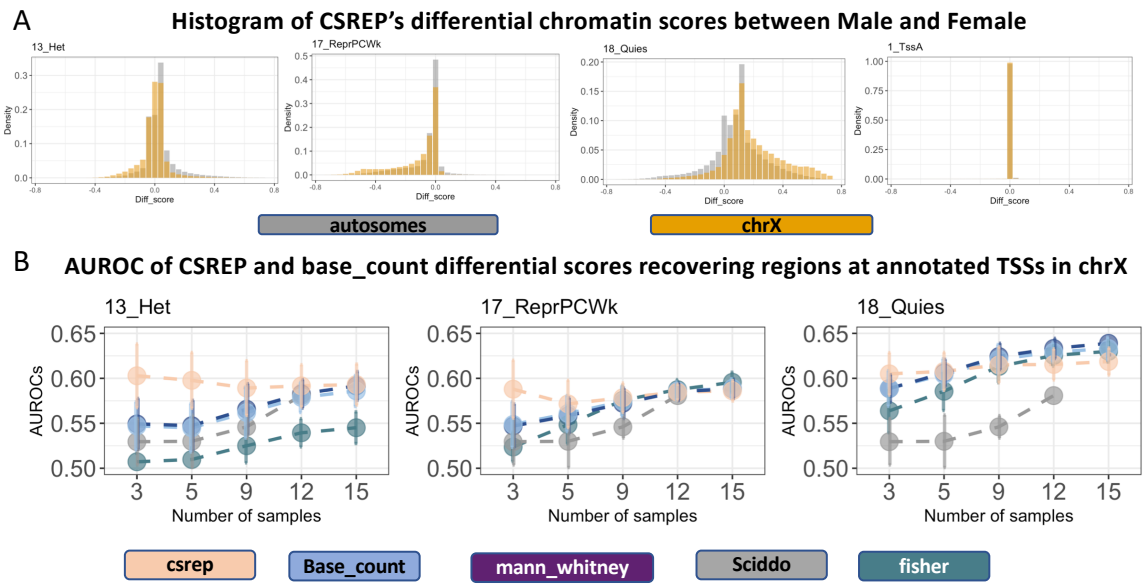


Figure 4. 3: CSREP shows signals of differential chromatin state scores in chrX when comparing Male and Female samples.

(A) Each subpanel shows the histogram of CSREP's differential scores in autosomes and chrX, for states associated with heterochromatin (13_Het), weak polycomb repressed domains (17_ReprPCWk), and quiescent regions (18_Quies), active transcription start site (1_TssA). The x-axis shows differential scores, with positive values implying Male samples have higher probabilities of being in the state compared to Female samples, and vice versa for negative values. Histograms of scores for all states are in **Supplementary Figure 4.11. (B)** AUROCs of recovering regions overlapping annotated TSSs on chrX, using differential chromatin scores of three states as in **(A)**, outputted by CSREP, base_count, SCIDDO, Mann-Whitney U test (based on ChromDiff) and Fisher's Exact test (based on EpiCompare) for Male and Female groups (**Methods**). The AUROCs based on Mann-Whitney U test showed close values with those based on base_count, hence the plotted average AUROCs from these two methods were overlapping. We calculated the AUROCs using different sets of input Male and Female samples, with varying numbers of samples in each group (x-axis). For each number of samples (x-axis), we conducted the analysis for 30 sets of Male and Female input samples (**Methods**). The plots show the average (dots) and standard deviation (error bars) of the AUROCs across the 30 sets of input samples. SCIDDO did

not successfully generate output for the case of 15 input samples, thus no results are reported for that.

AUROCs for predicting Brain-/ESC- specific peaks of chromatin marks

	DNase								H3K27ac								H3K9ac							
	Brain-specific				ESC-specific				Brain-specific				ESC-specific				Brain-specific				ESC-specific			
state	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher
1 Active TSS	0.54	0.51	0.50	0.51	0.59	0.50	0.51	0.51	0.58	0.52	0.52	0.52	0.70	0.51	0.51	0.51	0.59	0.53	0.53	0.52	**0.66**	0.55	0.55	0.55
2 Flk TSS	0.54	0.51	0.51	0.51	0.60	0.51	0.52	0.51	0.55	0.51	0.53	0.52	0.66	0.50	0.51	0.51	0.53	0.52	0.53	0.52	0.65	0.55	0.58	0.57
3 Flk TSS U	0.55	0.51	0.51	0.51	0.62	0.51	0.52	0.52	0.61	0.52	0.52	0.52	0.72	0.53	0.52	0.52	0.63	0.53	0.54	0.52	**0.68**	0.54	0.54	0.54
4 Flk TSS D	0.54	0.51	0.51	0.51	**0.65**	0.51	0.52	0.52	0.58	0.51	0.52	0.52	**0.73**	0.51	0.52	0.52	0.57	0.52	0.53	0.51	**0.69**	0.52	0.53	0.53
7 Genic Enh	0.55	0.50	0.50	0.50	0.50	0.50	0.51	0.51	**0.72**	0.53	0.53	0.53	0.55	0.52	0.52	0.52	0.68	0.53	0.54	0.53	0.50	0.51	0.52	0.51
8 Genic Enh	**0.57**	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.72	0.51	0.51	0.51	0.56	0.51	0.50	0.51	**0.71**	0.52	0.52	0.51	0.53	0.51	0.51	0.51
9 Active Enh	**0.57**	0.53	0.53	0.53	0.58	0.55	0.57	0.56	**0.74**	0.58	0.58	0.58	0.71	0.62	0.62	0.62	**0.72**	0.62	0.64	0.60	0.61	0.58	0.59	0.58
10 Active Enh	0.56	0.54	0.54	0.54	0.59	0.52	0.57	0.55	**0.73**	0.64	0.64	0.65	**0.74**	0.61	0.61	0.61	**0.70**	0.61	0.63	0.58	0.62	0.53	0.56	0.55
11 Wk Enh	0.54	0.52	**0.57**	0.56	0.64	0.61	**0.65**	**0.64**	0.69	0.58	0.62	0.61	**0.73**	0.66	0.69	0.68	0.62	0.54	0.64	0.59	0.58	0.57	0.60	0.58
sciddo	0.52				0.55				0.56				0.53				0.56				0.59			

Within each comparison category: max min

Figure 4. 4: Evaluation of recovery of differential chromatin marks signals between ESC and Brain.

The table shows AUROCs for differential scores' predictions of genomic regions associated with differential peak signals for one chromatin mark, from left to right: DNase, H3K27ac and H3K9ac. For each chromatin mark, it shows the AUROCs of predicting signal peaks observed in Brain and ESC exclusively (Brain-spec and ESC-spec, respectively). Differential scores outputted by CSREP, base-count, Mann-Whitney U test (used by ChromDiff) and Fisher's exact test (used by EpiCompare) are shown for active promoter and enhancer associated chromatin states (rows). In each category of comparisons (a chromatin mark in either ESC or Brain), the top three scores that show the highest AUROCs are in bold and underlined. Along the bottom is the AUROC for SCIDDO. Only active promoter and enhancer states are expected to be associated with differential DNase, H3K27ac and H3K9ac signals, but the AUROCs corresponding to all states are shown in **Supplementary Figure 4.13.**

Supplementary Information

Supplementary File 4.1: Metadata and download links of data used in the chapter.

Appendix—Supplementary figures

List of supplementary figures

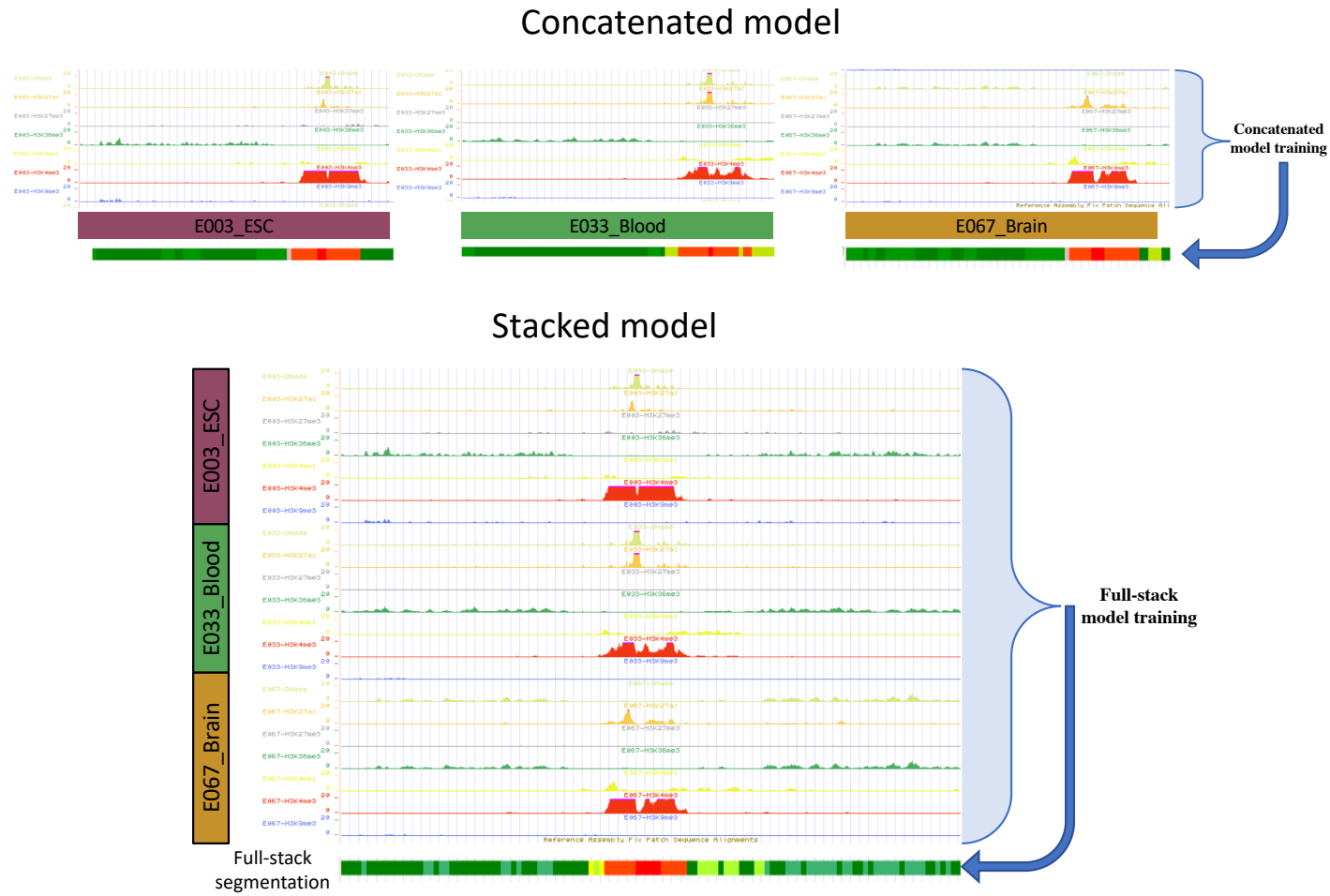
Chapter 2

Chapter 3:

Chapter 4:

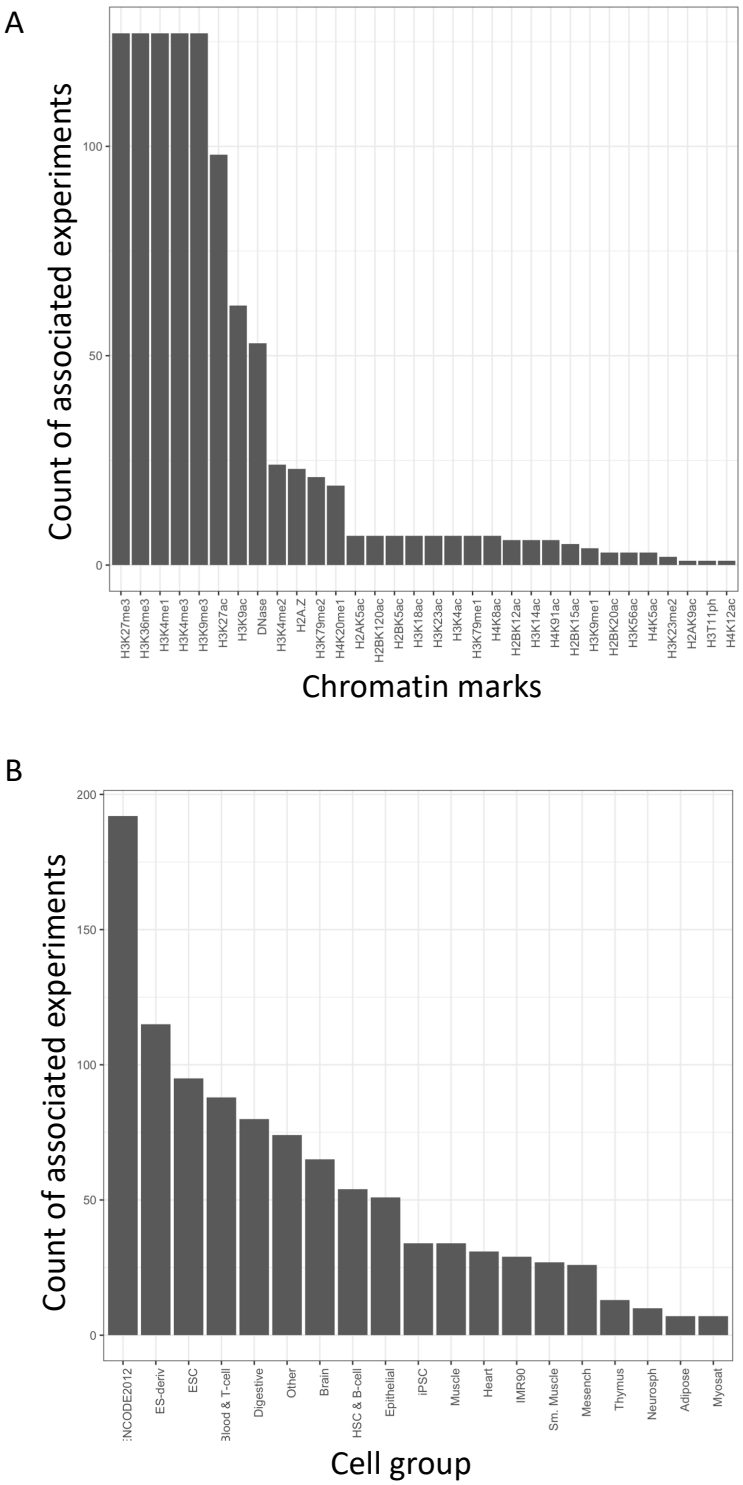
Supplementary figures for chapter 2

Universal annotation of the human genome through integration of over a thousand epigenomic datasets



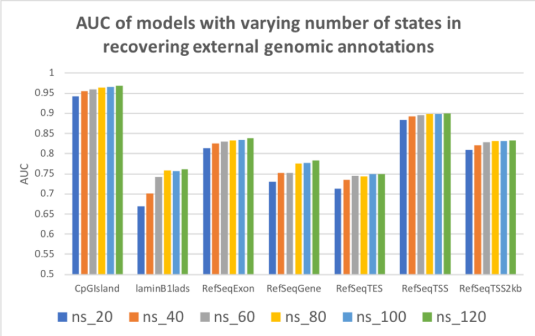
Supplementary Figure 2. 1: Illustration of concatenated model training vs. stacked model training.

The top of the figure illustrates the concatenated modeling approach where a chromatin state annotation is produced for each cell type based on the data in that cell type using a common set of chromatin state definitions. In contrast, the stacked modeling approach produces a single chromatin annotation of the genome based on all the data.

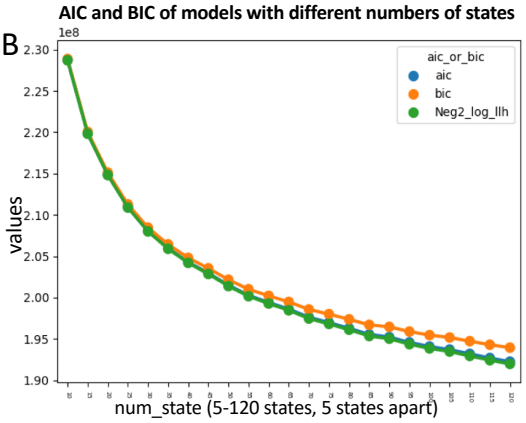


Supplementary Figure 2. 2: Mark and tissue group distribution of the input data tracks. **(A)** Counts of input tracks associated with different chromatin marks. There are five marks that were profiled in all 127 reference epigenomes, while some marks, largely acetylation marks, were profiled in few reference epigenomes. In total there were 1032 input tracks, including 53 DNase-seq datasets and 979 ChiP-seq datasets. **(B)** Count of input tracks associated with different tissue groups previously defined (Meuleman *et al.*, 2015).

A



B



D

Max correlations of emission parameters associated with H3K4me1 with indicators of cell groups

Number of states

Tissue groups	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120	max_corr
Adipose	0.17	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.17	0.18	0.19	0.18	0.18	0.18	0.19	0.18	0.18	0.18	0.19	0.21	0.19	0.19	0.19	0.21
Blood & T-cell	0.67	0.67	0.68	0.70	0.69	0.71	0.71	0.76	0.84	0.84	0.86	0.86	0.86	0.86	0.86	0.86	0.85	0.86	0.86	0.84	0.85	0.85	0.86	0.86
Brain	0.01	0.06	0.11	0.10	0.56	0.58	0.57	0.61	0.63	0.80	0.80	0.80	0.80	0.80	0.80	0.77	0.78	0.81	0.81	0.79	0.79	0.80	0.80	0.81
Digestive	-0.05	-0.02	0.07	0.06	0.23	0.33	0.45	0.55	0.54	0.50	0.61	0.54	0.55	0.63	0.64	0.65	0.64	0.65	0.64	0.64	0.64	0.65	0.64	0.65
ENCODE2012	0.39	0.44	0.43	0.45	0.45	0.45	0.46	0.44	0.45	0.44	0.47	0.49	0.48	0.48	0.48	0.50	0.52	0.50	0.52	0.49	0.49	0.52	0.53	0.53
Epithelial	0.34	0.34	0.35	0.36	0.38	0.38	0.38	0.45	0.45	0.45	0.46	0.45	0.45	0.48	0.47	0.48	0.48	0.47	0.48	0.48	0.49	0.49	0.49	0.49
ESC	0.38	0.40	0.62	0.62	0.67	0.67	0.65	0.67	0.68	0.69	0.68	0.69	0.68	0.68	0.68	0.68	0.70	0.70	0.72	0.70	0.71	0.72	0.71	0.72
ES-deriv	0.17	0.32	0.34	0.29	0.29	0.34	0.41	0.45	0.42	0.38	0.42	0.46	0.41	0.43	0.42	0.40	0.43	0.43	0.45	0.43	0.43	0.48	0.45	0.48
Heart	0.13	0.12	0.22	0.20	0.20	0.24	0.24	0.25	0.27	0.28	0.32	0.35	0.36	0.34	0.35	0.37	0.33	0.40	0.39	0.29	0.41	0.39	0.42	0.42
HSC & B-cell	0.52	0.57	0.55	0.57	0.58	0.61	0.60	0.74	0.78	0.75	0.80	0.80	0.80	0.79	0.80	0.79	0.71	0.80	0.65	0.66	0.81	0.79	0.76	0.81
IMR90	0.19	0.19	0.20	0.47	0.48	0.46	0.50	0.52	0.51	0.55	0.52	0.57	0.57	0.58	0.56	0.56	0.60	0.56	0.63	0.56	0.67	0.60	0.62	0.67
iPSC	0.37	0.32	0.44	0.45	0.48	0.48	0.48	0.48	0.49	0.48	0.48	0.48	0.48	0.48	0.48	0.50	0.51	0.52	0.51	0.51	0.51	0.51	0.52	0.52
Mesench	0.45	0.48	0.49	0.48	0.49	0.58	0.53	0.58	0.68	0.62	0.69	0.68	0.71	0.74	0.72	0.65	0.69	0.62	0.76	0.75	0.73	0.77	0.68	0.77
Muscle	0.28	0.28	0.41	0.37	0.35	0.40	0.43	0.41	0.40	0.53	0.50	0.55	0.53	0.55	0.50	0.54	0.52	0.63	0.53	0.54	0.53	0.63	0.60	0.63
Myosat	0.24	0.28	0.28	0.27	0.27	0.33	0.32	0.33	0.32	0.35	0.33	0.34	0.34	0.33	0.36	0.37	0.36	0.38	0.36	0.34	0.33	0.36	0.36	0.38
Neurosph	0.12	0.16	0.15	0.17	0.25	0.28	0.28	0.28	0.29	0.27	0.26	0.27	0.36	0.32	0.31	0.32	0.29	0.47	0.40	0.37	0.34	0.38	0.46	0.47
Other	0.15	0.17	0.17	0.18	0.19	0.23	0.23	0.23	0.24	0.29	0.28	0.29	0.29	0.37	0.35	0.36	0.31	0.29	0.32	0.42	0.45	0.32	0.38	0.45
Sm. Muscle	0.05	0.07	0.06	0.07	0.11	0.10	0.11	0.13	0.11	0.21	0.21	0.20	0.20	0.20	0.21	0.22	0.22	0.23	0.23	0.22	0.23	0.24	0.23	0.24
Thymus	0.06	0.13	0.13	0.13	0.13	0.16	0.16	0.17	0.16	0.17	0.17	0.17	0.17	0.17	0.16	0.18	0.19	0.18	0.20	0.21	0.21	0.22	0.22	0.22

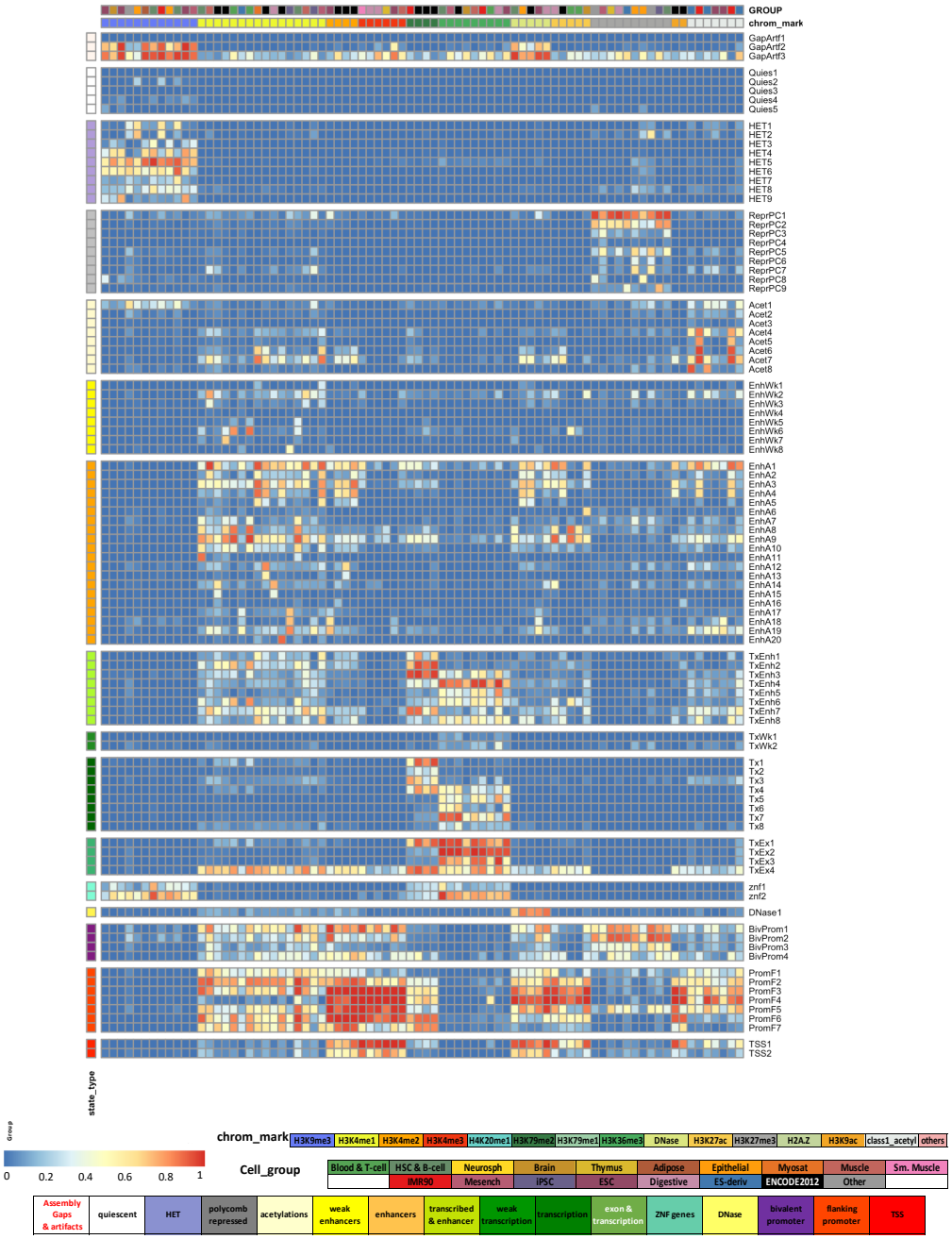
C **Maximum correlations of emission parameters of full-stack states with any state in other models**

Number of states

	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120
GapArtf1	0.50	0.52	0.50	0.51	0.56	0.53	0.54	0.56	0.57	0.61	0.59	0.60	0.60	0.98	0.99	1.00	1.00	1.00	1.00	0.87	0.90	0.97	0.95
GapArtf2	0.69	0.70	0.69	0.70	0.71	0.73	0.94	0.93	0.93	0.93	0.92	0.88	0.89	0.86	0.86	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GapArtf3	0.43	0.45	0.43	0.43	0.44	0.48	0.89	0.92	0.93	0.93	0.94	0.97	0.96	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Quies1	0.78	0.79	0.81	0.87	0.80	0.77	0.83	0.83	0.95	0.94	0.96	0.95	0.96	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.96	1.00	0.98
Quies2	0.75	0.70	0.79	0.74	0.96	0.84	0.91	0.99	0.99	0.94	0.99	0.93	0.93	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.97	0.99	0.97
Quies3	0.59	0.55	0.59	0.60	0.60	0.68	0.77	0.58	0.78	0.90	0.84	0.84	0.96	0.88	0.86	0.95	0.98	0.96	1.00	0.84	0.98	0.97	0.78
Quies4	0.80	0.82	0.81	0.85	0.96	0.91	0.93	0.99	0.99	0.98	0.98	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00	0.95	0.90	0.90	1.00
Quies5	0.30	0.32	0.34	0.35	0.35	0.35	0.37	0.93	0.93	0.37	0.93	0.67	0.94	0.92	0.79	0.93	0.99	1.00	1.00	0.98	1.00	1.00	0.99
HET1	0.65	0.72	0.74	0.67	0.95	0.97	0.97	0.94	0.94	0.98	0.94	0.99	0.98	0.97	1.00	0.99	1.00	1.00	1.00	0.95	0.96	1.00	0.97
HET2	0.43	0.47	0.57	0.69	0.81	0.79	0.79	0.88	0.88	0.81	0.90	0.99	0.78	0.86	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
HET3	0.81	0.80	0.82	0.83	0.90	0.91	0.91	0.95	0.95	0.96	0.96	0.99	0.94	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	1.00	1.00
HET4	0.89	0.88	0.91	0.91	0.94	0.97	0.96	0.99	1.00	0.99	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HET5	0.96	0.98	0.96	0.96	0.97	0.99	0.98	0.99	0.99	0.99	1.00	0.99	1.00	0.99	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HET6	0.95	0.91	0.95	0.94	0.93	0.94	0.92	0.98	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	1.00	1.00	1.00	0.98	0.97	1.00	1.00
HET7	0.86	0.84	0.85	0.84	0.82	0.83	0.87	0.95	0.95	0.93	0.94	0.92	0.92	0.91	0.90	0.90	0.92	1.00	1.00	1.00	0.99	1.00	1.00
HET8	0.87	0.83	0.88	0.87	0.86	0.87	0.85	0.88	0.87	0.85	0.89	0.87	1.00	0.99	0.99	0.99	0.99	0.99	1.00	0.95	0.94	0.95	0.97
HET9	0.72	0.75	0.73	0.77	0.87	0.83	0.85	0.88	0.87	0.88	0.88	0.98	0.87	0.94	0.96	0.96	0.99	1.00	1.00	0.99	0.98	0.97	1.00
ReprPC1	0.88	0.97	0.97	0.97	0.96	0.96	0.99	0.98	0.96	0.97	0.97	1.00	1.00	0.99	0.99	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00
ReprPC2	0.81	0.91	0.94	0.90	0.91	0.99	0.95	0.99	0.99	0.99	0.98	0.99	0.98	0.98	0.99	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00
ReprPC3	0.75	0.87	0.95	0.95	0.96	0.95	0.94	0.96	0.96	0.97	0.97	0.98	0.96	0.97	0.98	0.98	0.97	0.99	1.00	1.00	0.99	0.98	0.99
ReprPC4	0.66	0.83	0.92	0.95	0.95	0.95	0.96	0.96	0.93	0.95	0.94	0.97	0.99	0.98	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00
ReprPC5	0.89	0.96	0.89	0.96	0.94	0.93	0.91	0.95	0.94	0.96	0.96	0.95	0.96	0.94	0.96	0.99	0.99	1.00	1.00	0.98	1.00	1.00	1.00
ReprPC6	0.78	0.91	0.81	0.87	0.81	0.96	0.92	0.96	0.95	0.98	0.97	0.98	0.92	0.98	0.98	0.94	0.93	0.95	1.00	0.99	0.97	0.97	0.95
ReprPC7	0.73	0.75	0.90	0.73	0.86	0.92	0.95	0.91	0.92	0.91	0.88	0.88	0.92	0.89	0.89	0.95	0.96	0.93	1.00	0.89	0.95	0.98	0.98
ReprPC8	0.42	0.49	0.50	0.51	0.51	0.51	0.52	0.98	0.98	0.51	0.98	0.68	0.98	0.99	0.79	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ReprPC9	0.50	0.57	0.59	0.56	0.55	0.67	0.61	0.69	0.66	0.68	0.64	0.82	0.82	0.72	0.72	0.95	0.94	1.00	1.00	1.00	1.00	1.00	1.00
Acet1	0.46	0.50	0.70	0.84	0.77	0.70	0.79	0.76	0.78	0.88	0.83	0.89	0.96	0.99	0.95	0.93	0.95	0.90	1.00	0.98	0.98	0.99	0.97
Acet2	0.61	0.69	0.79	0.84	0.74	0.79	0.81	0.77	0.80	0.90	0.88	0.91	0.92	0.92	0.93	0.98	0.97	0.95	1.00	0.99	1.00	1.00	1.00
Acet3	0.60	0.58	0.72	0.77	0.71	0.81	0.81	0.79	0.84	0.97	0.92	0.96	0.98	0.97	0.95	0.97	0.96	0.97	1.00	0.99	0.99	0.98	0.98
Acet4	0.74	0.65	0.67	0.93	0.91	0.90	0.94	0.91	0.93	0.87	0.88	0.91	0.94	0.93	0.98	0.96	0.90	0.98	1.00	0.99	1.00	0.99	0.98
Acet5	0.59	0.59	0.63	0.96	0.97	0.96	0.95	0.97	0.96	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Acet6	0.85	0.90	0.91	0.90	0.90	0.92	0.90	0.89	0.90	0.96	0.93	0.95	0.92	0.93	0.95	0.89	0.96	0.88	1.00	0.98	1.00	0.98	1.00
Acet7	0.94	0.92	0.94	0.92	0.94	0.94	0.97	0.97	0.97	0.96	0.95	0.93	0.92	0.91	0.96	0.94	0.99	0.99	1.00	1.00	0.92	0.99	0.98
Acet8	0.39	0.61	0.63	0.64	0.73	0.88	0.82	0.87	0.87	0.92	0.89	0.94	0.96	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EnhWk1	0.68	0.74	0.74	0.72	0.74	0.82	0.82	0.86	0.91	0.86	0.92	0.90	0.97	0.99	0.95	0.93	0.88	0.91	1.00	0.99	0.99	0.95	0.97
EnhWk2	0.67	0.76	0.85	0.88	0.85	0.93	0.96	0.92	0.90	0.84	0.93	0.95	0.97	0.94	0.99	0.89	0.90	0.94	1.00	0.86	0.95	0.88	0.91
EnhWk3	0.55	0.57	0.58	0.65	0.77	0.75	0.76	0.72	0.70	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.93	1.00	1.00	1.00	0.98	0.98
EnhWk4	0.38	0.50	0.53	0.55	0.68	0.71	0.72	0.70	0.70	0.66	0.74	0.58	0.97	0.91	0.88	0.71	0.64	0.89	1.00	0.99	0.98	0.98	0.93
EnhWk5	0.40	0.74	0.74	0.75	0.78	0.79	0.79	0.76	0.73	0.74	0.73	0.73	0.73	0.74	0.74	0.74	1.00	0.74	1.00	0.99	1.00	1.00	0.76
EnhWk6	0.54	0.91	0.91	0.93	0.94	0.87	0.87	0.97	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
EnhWk7	0.27	0.46	0.46	0.46	0.47	0.52	0.52	0.57	0.57	0.58	0.58	0.59	0.59	0.60	0.59	0.60	1.00	0.59	1.00	1.00	1.00	1.00	1.00
EnhWk8	0.34	0.40	0.86	0.91	0.94	0.99	0.98	0.99	0.99	0.99	0.94	0.99	0.99	0.99	0.98	0.98	0.97	1.00	1.00	1.00	1.00	1.00	0.99
EnhA1	0.90	0.92	0.94	0.96	0.98	0.97	0.95	0.97	0.97	0.98	0.99	1.00	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00	0.96	1.00	0.98
EnhA2	0.82	0.79	0.87	0.81	0.82	0.80	0.82	0.81	0.81	0.85	0.87	0.85	0.86	0.86	0.84	0.99	0.99	0.98	1.00	0.98	1.00	0.98	0.99
EnhA3	0.84	0.86	0.91	0.91	0.89	0.89	0.89	0.87	0.88	0.96	0.87	0.94	0.92	0.91	0.95	0.92	0.90	0.99	1.00	0.97	0.95	0.97	0.91
EnhA4	0.85	0.93	0.91	0.95	0.94	0.94	0.97	0.94	0.96	0.94	0.94	0.97	0.97	0.97	0.93	0.98	1.00	0.98	1.00	0.96	0.95	0.95	0.99
EnhA5	0.77	0.85	0.84	0.86	0.93	0.97	0.96	0.97	0.96	0.99	0.98	0.98	0.98	0.98	1.00	0.98	0.99	0.99	1.00	0.96	0.99	0.96	0.98
EnhA6	0.27	0.35	0.37	0.37	0.81	0.82	0.81	0.85	0.87	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.99	1.00	0.99	0.99	1.00	1.00
EnhA7	0.54	0.81	0.81	0.80	0.82	0.84	0.83	0.79	0.77	0.75	0.76	0.77	0.78	0.79	0.76	0.73	0.83	0.73	1.00	0.99	0.78	0.97	0.98
EnhA8	0.45	0.94	0.94	0.94	0.90	0.97	0.98	0.95	0.93	0.97	0.97	0.98	0.98	0.99	0.96	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00
EnhA9	0.84	0.80	0.88	0.80	0.91	0.88	0.88	0.95	0.96	0.92	0.91	0.91	0.95	0.95	0.97	0.99	0.97	1.00	1.00	0.97	0.99	0.99	0.98
EnhA10	0.74	0.83	0.83	0.81	0.89	0.82	0.81	0.83	0.89	0.83	0.84	0.85	0.91	0.92	0.86	0.98	0.93	0.97	1.00	0.97	0.89	0.92	0.94
EnhA11	0.25	0.50	0.50	0.54	0.53	0.63	0.63	0.83	0.80	0.84	0.85	0.85	0.86	0.90	0.88	0.91	1.00	0.91	1.00	1.00	0.99	0.99	0.95
EnhA12	0.65	0.58	0.59	0.67	0.67	0.72	0.65	0.92	0.92	0.90	0.90	0.89	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
EnhA13	0.43	0.42	0.46	0.58	0.62	0.68	0.62	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EnhA14	0.44	0.44	0.51	0.58	0.58	0.81	0.92	0.96	0.94	0.91	0.99	0.95	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
EnhA15	0.30	0.38	0.46	0.40	0.47	0.73	0.82	0.95	0.97	0.99	0.95	0.99	0.99	1.00	0.99	0.98	1.00	1.00	1.00	0.96	0.96	0.95	0.94
EnhA16	0.25	0.39	0.43	0.44	0.45	0.42	0.44	0.33	0.30	0.36	0.49	0.86	0.88	0.94	0.88	0.87	0.87	0.89	1.00	0.86	0.86	0.85	0.86
EnhA17	0.43	0.64	0.93	0.93	0.90	0.92	0.91	0.91	0.92	0.92	0.89	0.91	0.90	0.91	0.91	0.91	0.99	0.91	1.00	1.00	1.00	1.00	0.95
EnhA18	0.39	0.54	0.88	0.88	0.94	0.94	0.90	0.94	0.96	0.98	0.94	0.98	0.97	0.97	0.97	0.96	0.97	0.97	1.00	0.97	0.98	1.00	0.98
EnhA19	0.46	0.80	0.86	0.79	0.80	0.96	0.98	0.96	0.95	0.90	1.00	0.97	0.93	0.91	0.92	0.94	0.93	0.96	1.00	1.00	0.99	0.99	0.92
EnhA20	0.31	0.34	0.35	0.36	0.37	0.39	0.38	0.40	0.40	0.41	0.45	0.41	0.44	0.48	0.46	0.40	0.71	0.53	1.00	1.00	1.00	1.00	1.00
TxEnh1	0.70	0.79	0.79	0.78	0.82	0.83	0.81	0.78	0.82	0.81	0.84	0.82	0.85	0.85	0.88	0.96	0.90	0.99	1.00	0.84	0.98	0.95	0.97
TxEnh2	0.95	0.88	0.90	0.87	0.95	0.95	0.95	0.93	0.99	0.96	0.95	0.98	0.97	0.95	0.99	0.96	0.94	1.00	1.00	0.98	0.99	0.98	0.99
TxEnh3	0.82	0.88	0.86	0.87	0.86	0.86	0.86	0.90	0.85	0.91	0.87	0.89	0.89	0.90	0.90	0.87	0.90	0.89	1.00	0.97	0.89	0.89	0.99
TxEnh4	0.96	0.97	0.97	0.97	0.96	0.96	0.96	0.97	0.96	0.97	0.99	0.98	0.97	0.99	0.98	0.98	0.98	0.98	1.00	0.98	0.97	0.99	0.98
TxEnh5	0.86	0.86	0.86	0.99	0.99	0.99	0.99	0.99	0.96	0.94	0.94	0.94	0.99	0.98	0.97	0.99	0.99	1.00	1.00	1.00	0.99	1.00	0.98
TxEnh6	0.69	0.70	0.70	0.69	0.79	0.78	0.78	0.79	0.79	0.79	0.78	0.78	0.74	0.78	0.94	0.77	0.78	1.00	1.00	0.79	1.00	1.00	1.00
TxEnh7	0.84	0.89	0.89	0.87	0.83	0.82	0.83	0.99	0.97	0.92	0.94	0.91	0.89	0.88	0.88	0.98	0.94	0.97	1.00	0.97	0.92	0.96	0.93
TxEnh8	0.66	0.67	0.67	0.71	0.84	0.84	0.84	0.79	0.84	0.91	0.93	0.92	0.97	0.92	0.93	1.00	0.93	0.99	1.00	0.99	0.93	0.97	0.97
TxWk1	0.96	0.97	0.93	0.96	0.99	0.99	0.97	0.98	0.97	1.00	0.97	0.99	1.00	0.92	0.99	1.00	0.99	1.00	1.00	0.99	0.99	1.00	0.94
TxWk2	0.71	0.71	0.71	0.88	0.88	0.88	0.87	0.88	0.95	0.98	0.98	0.97	0.99	0.91	0.99	1.00	0.99	1.00	1.00	0.99	0.98	1.00	0.96
Tx1	0.90	0.97	0.98	0.99	0.99	0.99	0.99	0.98	0.99	0.98	0.97	0.98	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
Tx2	0.84	0.95	0.94	0.96	0.98	0.98	0.98	0.99	0.98	0.99	0.99	0.99	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Tx3	0.77	0.86	0.83	0.83	0.86	0.86	0.85	0.86	0.87	0.86	0.97	0.98	0.97	0.96	0.96	1.00	0.99	0.99	1.00	0.98	0.99	0.99	0.99
Tx4	0.79	0.77	0.92	0.86	0.80	0.80	0.89	0.85	0.99	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99
Tx5	0.94	0.94	0.94	0.91	0.93	0.93	0.93	0.92	0.97	0.96	0.97	0.97	0.99	0.97	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
Tx6	0.92	0.92	0.91	0.95	0.95	0.95	0.94	0.94	0.96	0.97	0.97	0.99	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99
Tx7	0.95	0.95	0.97	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
Tx8	0.90	0.90	0.89	0.89	0.90	0.90	0.91	0.91	0.91	0.96	0.93	0.95	0.96	0.96	0.93	0.99	0.94	0.99	1.00	0.96	0.95	0.97	0.98
TxEx1	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.93	0.94	0.94	0.94	0.94	0.98	0.95	0.96	0.96	0.94	0.97	1.00	0.96	0.98	1.00	0.98
TxEx2	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
TxEx3	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98	1.00	1.00	1.00	1.00	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
TxEx4	0.73	0.74	0.82	0.93	0.83	0.80	0.82	0.80	0.80	0.84	0.90	0.89	0.93	0.96	0.94	0.99	0.97	0.99	1.00	0.98	0.99	0.87	0.94
znf1	0.63	0.75	0.86	0.86	0.86	0.91	0.90	0.89	1.00	0.88	0.87	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98
znf2	0.72	0.71	0.94	0.94	0.96	0.99	0.99	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
DNase1	0.34	0.34	0.34	0.52	0.54	0.56	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
BivProm1	0.64	0.79	0.97	0.94	0.94	0.99	0.95	0.99	0.99	0.99	0.98	0.93	0.91	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
BivProm2	0.86	0.93	0.87	0.94	0.96	0.95	0.93	0.93	0.98	0.95	0.99	0.95	0.97	0.99	0.99	0.97	0.97	1.00	1.00	1.00	1.00	0.99	1.00
BivProm3	0.96	0.92	0.91	0.90	0.90	0.90	0.86	0.90	0.90	0.91	0.96	0.96	0.90	0.92	0.93	0.89	0.87	0.87	1.00	0.97	0.96	0.92	0.97
BivProm4	0.62	0.77	0.75	0.73	0.73	0.72	0.65	0.72	0.89	0.73	0.87	0.79	0.92	0.92	0.94	0.72	0.73	0.89	1.00	0.97	0.96	0.96	0.94
PromF1	0.81	0.77	0.72	0.86	0.86	0.86	0.83	0.85	0.84	0.92	0.83	0.83	0.82	0.85	0.90	0.92	0.91	0.96	1.00	1.00	0.87	0.97	0.88
PromF2	0.85	0.88	0.88	0.87	0.89	0.89	0.90	0.88	0.89	0.96	0.97	0.96	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.95
PromF3	0.98	0.97	0.98	0.96	0.95	0.98	0.97	0.99	0.99	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
PromF4	0.91	0.98	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PromF5	0.95	0.91	0.89	0.93	0.94	0.91	0.97	0.93	0.91	0.92	0.92	0.98	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PromF6	0.85	0.88	0.81	0.97	0.97	0.92	0.95	0.95	0.93	0.95	0.97	0.96	0.96	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PromF7	0.64	0.76	0.81	0.91	0.87	0.98	0.99	0.95	0.97	0.97	0.95	0.95	0.96	0.95	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00
TSS1	0.92	0.96	0.95	0.96	0.96	0.96	0.96	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
TSS2	0.83	0.80	0.80	0.82	0.81	0.96	0.92	0.96	0.95	0.96	0.95	0.96	0.95	0.96	0.96	0.96	0.99	1.00	1.00	1.00	0.95	0.99	0.99

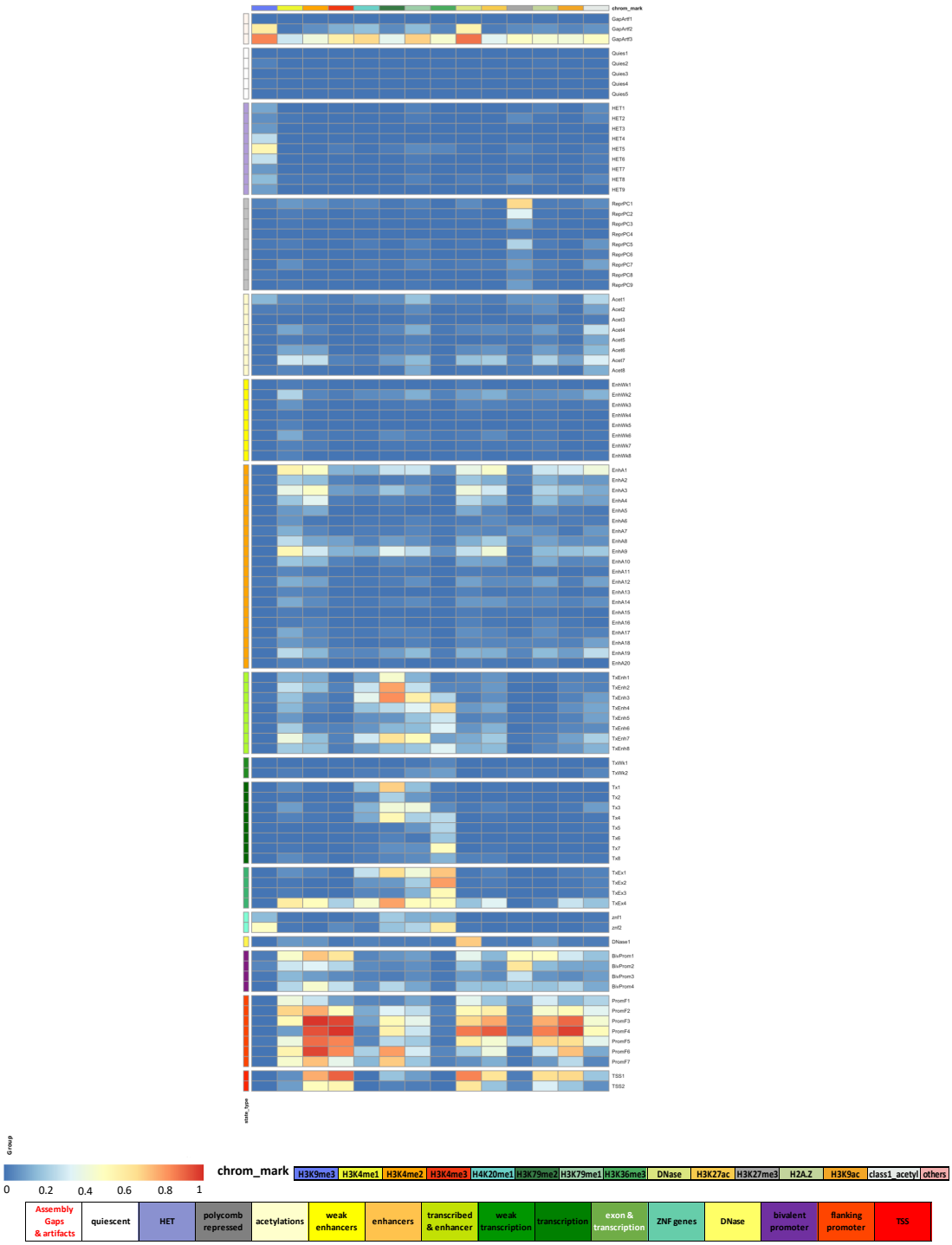
Supplementary Figure 2. 3: Evaluation of full-stack model's number of states.
(A) AUCs of full-stack models with varying number of states in recovering external genomic annotations. The Supplementary Figure 2.hows the AUC of full-stack models with 20, 40, 60, 80, 100, and 120 states at predicting the genomic locations of multiple different external genomic annotations (CpG Islands, lamina associated domains (laminB1lads), Exon, Gene body, TES, TSS, and TSS2kb regions) (**Methods**). As the number of chromatin states increases, the AUC increases, but the level of the AUC increases diminishes. **(B)** The estimated AIC-BIC curves for models with the number of states ranging from 10 to 120 (5 states apart). We calculated the AIC and BIC based on ChromHMM's output reporting the log-likelihood of observed data for 300 1-Mb regions. Neg2_log_llh: -2 * negative log likelihood of observed data. (**C**) Maximum correlations of emission parameters between each state in the 100-state model and any state for each other model. This is output from ChromHMM's CompareModels command. Rows correspond to the states of the 100-state model. Columns correspond to models with varying numbers of states. Values are the maximum correlation of any state from the model in the column (with varying number of states) with the state from the 100-state in the row. The 100-state model is boxed. This analysis can be effective at establishing some biologically motivated lower bounds on the number of states. For example, state EnhA20, a HUVEC specific enhancer state, is not captured in models with fewer than 100 states. (**D**) Maximum correlations of emission parameters associated with H3K4me1 (an enhancer mark available in all cell types) and the binary vector indicating whether the cell type associated with an emission parameter is in a tissue group (1) or not (0). The rows correspond to different tissue groups from Roadmap Epigenomics Consortium (Meuleman *et al.*, 2015). The columns correspond to different models with varying numbers of states. The values show the maximum correlations mentioned above across all states within a model. The 100-state model is boxed. The last column shows the maximum correlations observed for each tissue type (each row) across all the models.

Emission matrix of 80 representative experiments used to summarize the states



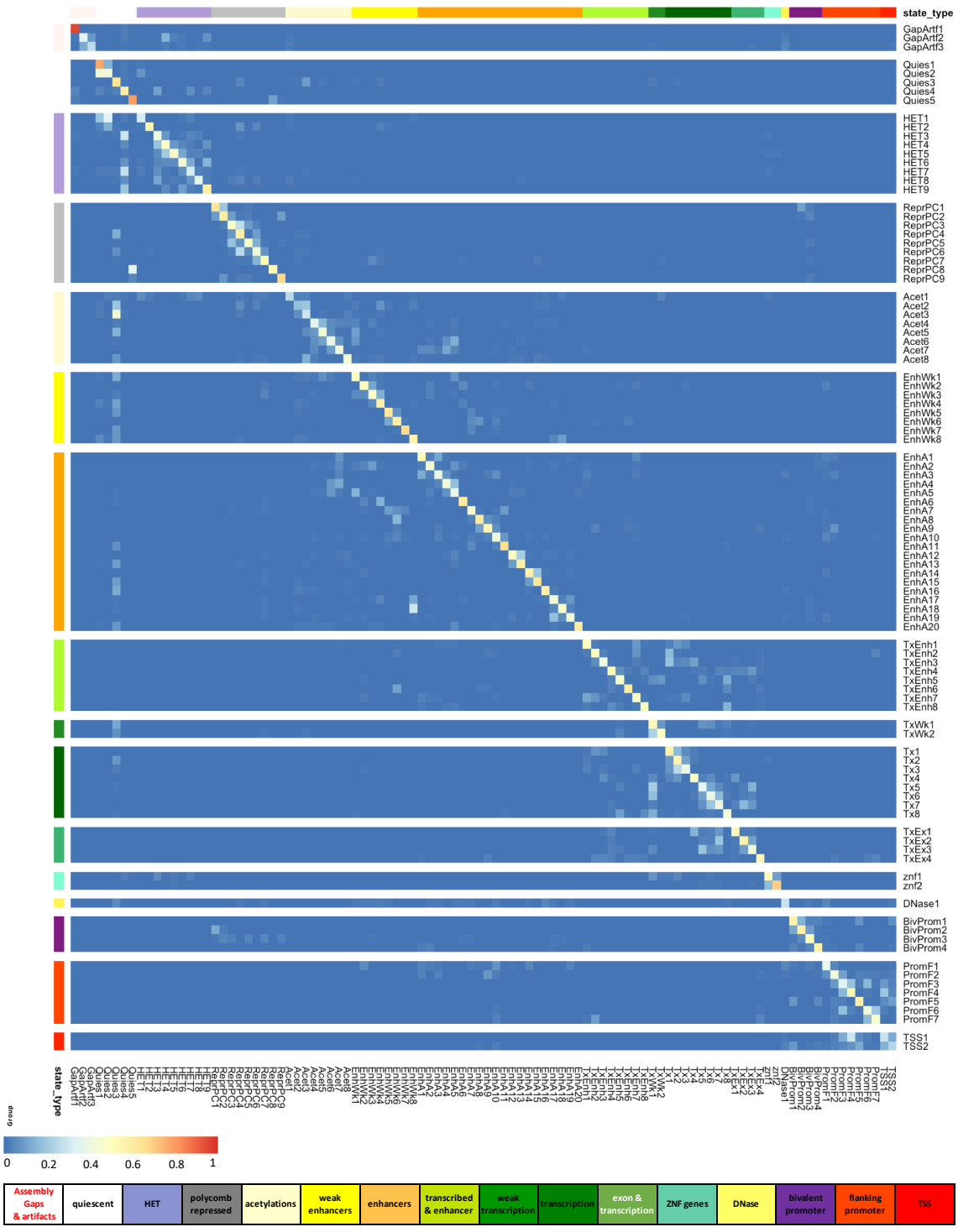
Supplementary Figure 2. 4: Emission probabilities of 80 datasets chosen to summarize the full-stack model.

Each row in the heatmap corresponds to a full-stack state. Each of the 80 columns corresponds to an dataset that has been chosen to represent the space of 1032 datasets. These datasets were chosen through a greedy search of features that optimize prediction of the full-stack annotation using Naïve Bayes with the selected features (**Methods, Supplementary Data 2. 8**). For each state and each dataset, the heatmap gives the probability within the state of observing a binary present call for the dataset's signal. States are displayed in 16 groups as in **Figure 2.2A**. Color legends for the emission values, the state groups, chromatin mark, and tissue group are shown at the bottom.



Supplementary Figure 2. 5: Full-stack states emission probabilities, averaged by chromatin marks.

Each column corresponds to an individual chromatin mark or the group of acetylation marks. The heatmap shows for each state the average emission probabilities of datasets associated with each chromatin mark or with the group of acetylations. Color legends for the emission values, the state groups, and chromatin mark are shown at the bottom.



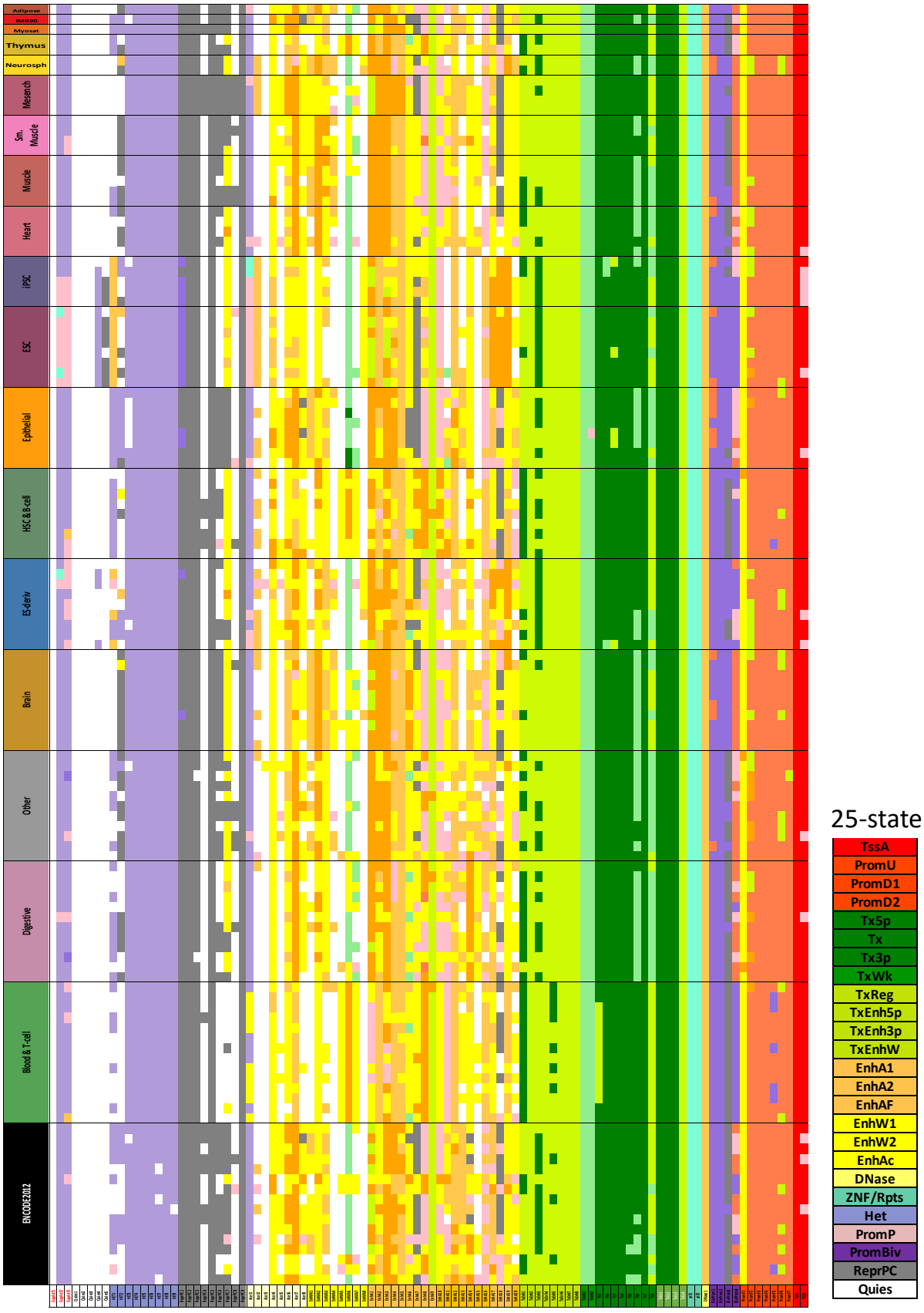
Supplementary Figure 2. 6: Full-stack states transition probabilities.
Each row and each column correspond to a full-stack state, ordered based on their associated state group. The heatmap shows for each state assigned at a current genomic position (rows) the probabilities of transitioning to another state (columns) at the subsequent genomic position. Color legends for the emission values, the state groups are shown at the bottom.

Tissue-group statistically significantly more highly emitted in full-stack states

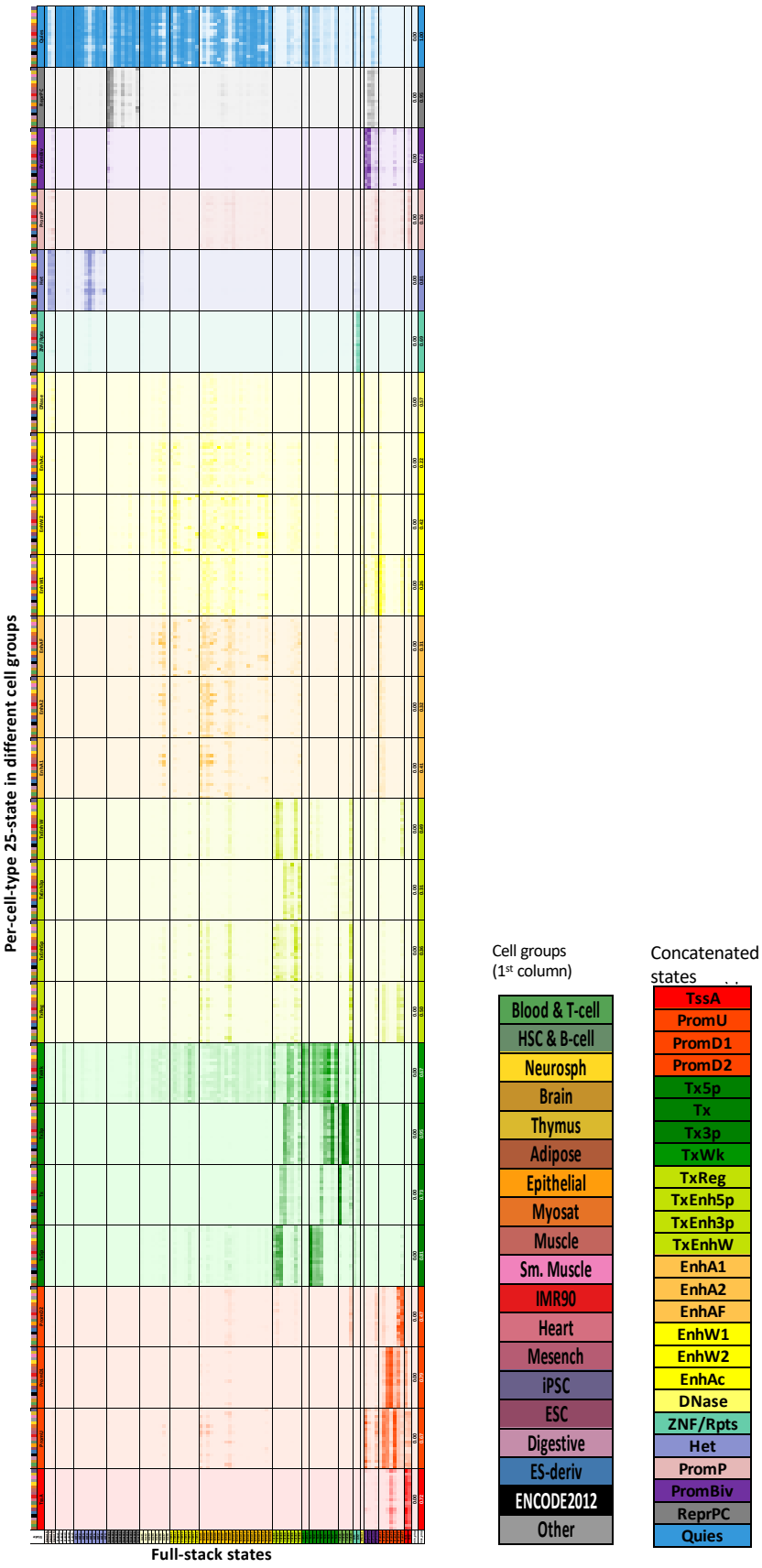
mneumonics	H3K9me3	H3K4me1	H3K4me3	H3K36me3	H3K27me3	H3K27ac	H3K9ac	DNase
HET9				ESC				
Acet5		ENCODE2012						
Acet6		ENCODE2012	ENCODE2012					
Acet7			ENCODE2012					
EnhWk3					ENCODE2012			
EnhWk4			Brain		ENCODE2012			
EnhWk5		Blood & T-cell	Blood & T-cell					
EnhWk6		Blood & T-cell	Blood & T-cell			Blood & T-cell		
EnhWk7		HSC & B-cell						
EnhA3		ENCODE2012	ENCODE2012					
EnhA4		ENCODE2012	ENCODE2012					
EnhA5		ENCODE2012	ENCODE2012			ENCODE2012		
EnhA6		Brain	Brain		ENCODE2012			
EnhA7		Blood & T-cell,HSC & B-cell		Blood & T-cell				
EnhA8		Blood & T-cell	Blood & T-cell			Blood & T-cell		
EnhA9		Blood & T-cell	Blood & T-cell					
EnhA10		HSC & B-cell						
EnhA11		HSC & B-cell	HSC & B-cell	HSC & B-cell				
EnhA14		Digestive	Digestive			Digestive		Digestive
EnhA15					ENCODE2012	Digestive		Digestive
EnhA16								
EnhA17					ENCODE2012			
EnhA18		ESC						
TxEnh1			ENCODE2012	ENCODE2012	HSC & B-cell			
TxEnh2		Blood & T-cell						
TxEnh6		Blood & T-cell	Blood & T-cell			Blood & T-cell		
TxEnh7								
TxEnh8			ENCODE2012					
Tx1		Blood & T-cell						
Tx4		Blood & T-cell						
TxEx1		Blood & T-cell						
PromF3		HSC & B-cell						
PromF4		HSC & B-cell						
PromF5		HSC & B-cell						
PromF6						ENCODE2012		
TSS1		HSC & B-cell						
TSS2		HSC & B-cell						

Supplementary Figure 2. 7 : Statistically significant tissue—group specificity in full-stack states.

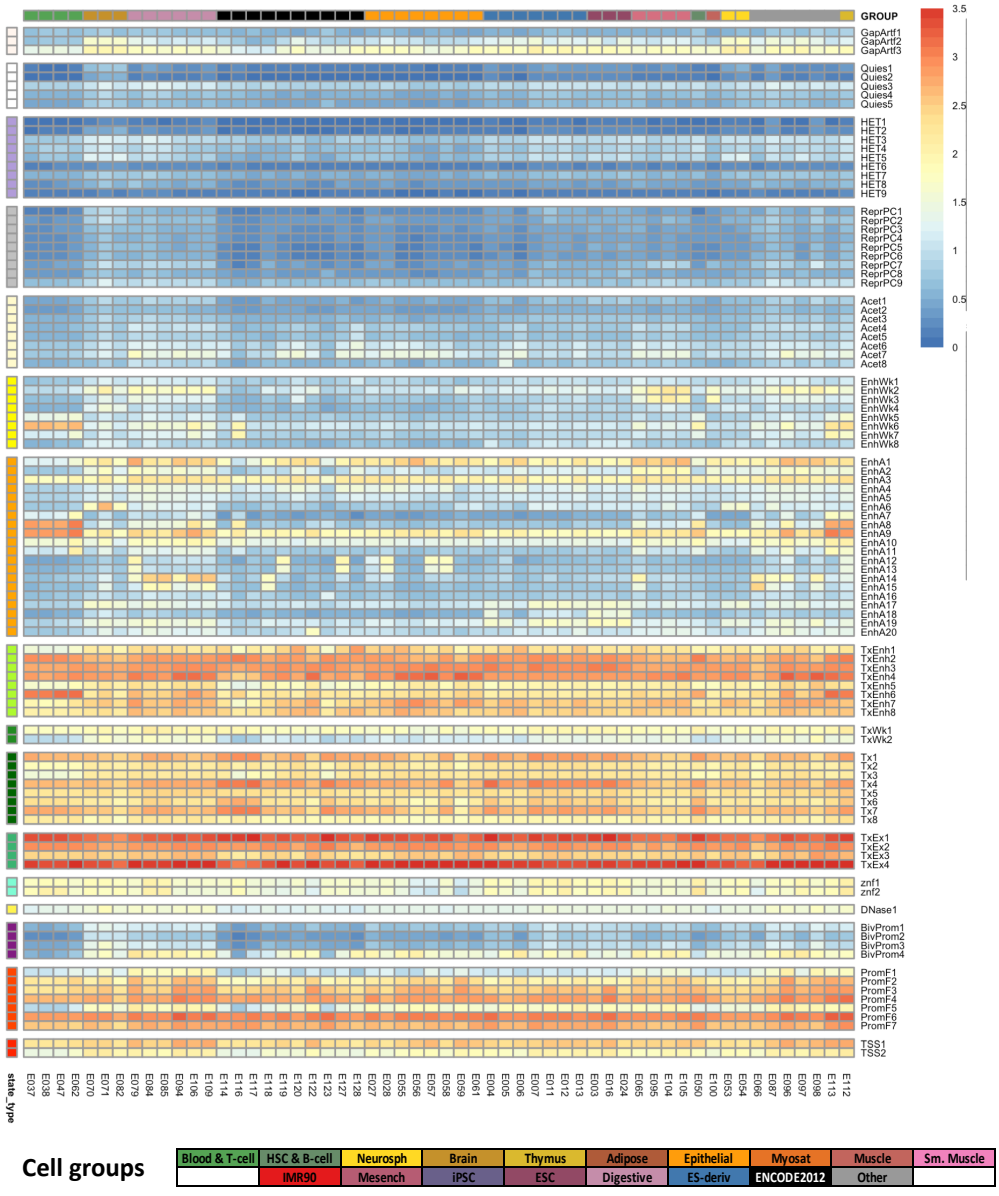
The columns correspond to the eight most frequently profiled chromatin marks (H3K9me3, H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K27ac, H3K9ac, and DNase I hypersensitivity). The rows correspond to states that for at least one chromatin mark show statistically significant higher emission probabilities for one tissue group compared to others (**Methods**). Statistical significance is based on one-sided Mann-Whitney tests at a Bonferroni-corrected p-value threshold of 3.5e-6. The entries in the grid shows the tissue groups reaching significance for each chromatin mark-full-stack state combination.


Adipose
IMR90
Myosat
Thymus
Neurosph
Mesench
Sm. Muscle
Muscle
Heart
iPSC
ESC
Epithelial
HSC & B-cell
ES-deriv
Brain
Other
Digestive
Blood & T-cell
ENCODE2012
25-state
TssA
PromU
PromD1
PromD2
Tx5p
Tx
Tx3p
TxWk
TxReg
TxEnh5p
TxEnh3p
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies

Supplementary Figure 2. 8: Full-stack states maximum-enrichments with annotated concatenated-model chromatin states in 127 reference epigenomes.
Each row corresponds to one of 127 reference epigenomes from the Roadmap Epigenomics Consortium (**Methods**). Each column corresponds to a state of the full-stack model. Each color entry corresponds to a reference epigenome- full-stack state combination. The color corresponds to the chromatin state from the 25-state model annotating the respective reference epigenome that is most enriched with the respective full-stack state. The figure highlights how some full-stack states are maximally enriched with the same concatenated-model chromatin states across all the reference epigenomes; for example, states znf1 and znf2 are maximally enriched with ZNF Gene state in all 127 reference epigenomes' 25-state concatenated annotation. At the same time, other full-stack states are enriched for distinct concatenated states, for example state EnhA8--characterized as a blood enhancer state based on emission probabilities-- is most enriched with activate/flanking enhancer in cell types of the groups Blood&Tcell, HSC&B-cell, while being most enriched with poised promoter and weak enhancer states in other cell types. Detailed description of each full-stack state enrichment patterns with concatenated states can be found in **Supplementary Data 2. 5**.

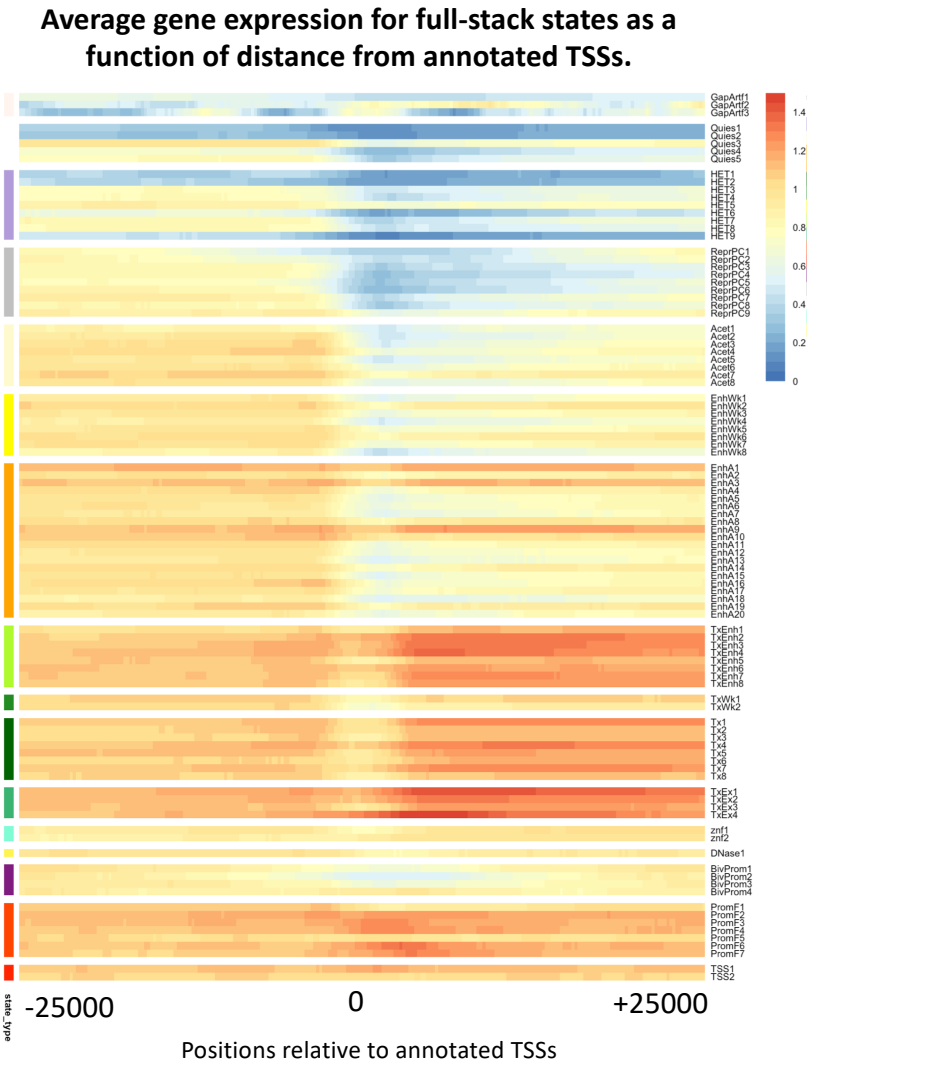

Per-cell-type 25-state in different cell groups
Full-stack states
Cell groups
(1st column)
Blood & T-cell
HSC & B-cell
Neurosph
Brain
Thymus
Adipose
Epithelial
Myosat
Muscle
Sm. Muscle
IMR90
Heart
Mesench
iPSC
ESC
Digestive
ES-deriv
ENCODE2012
Other
Concatenated
states
TssA
PromU
PromD1
PromD2
Tx5p
Tx
Tx3p
TxWk
TxReg
TxEnh5p
TxEnh3p
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies

Supplementary Figure 2. 9: Estimated probabilities of concatenated-model chromatin states overlapping with full-stack states.
The figure shows estimated probabilities of concatenated chromatin state assignments overlapping with full-stack state annotations conditioned on the cell group of the concatenated annotations. This figure is also provided as an excel file in **Supplementary Data 2. 5** where it is accompanied with detailed comments about each full-stack state. The figure is based on a 25-state per-cell type chromatin state model (Ernst and Kellis, 2015), and 19-previously defined tissue groups for the 127 reference epigenomes *(Meuleman* et al*., 2015)*. Each row corresponds to a combination of per-cell type state (among 25 states) and tissue group, as denoted in the first two columns and legend. Rows corresponding to the same concatenated-model state are grouped together. The first two columns show the colors of tissue groups and concatenated-model state, respectively, as indicated in legends on the right, and matching with the colors in **Supplementary Figure 2.8**, except we changed concatenated-model quiescent 25-state from white to blue for better visibility. The 100 following columns correspond to 100 full-stack states. Values in the heatmap correspond to the estimated probability a genomic position annotated as a full-stack state (column) is also annotated as a concatenated-model state in a cell type from the corresponding tissue group (row) (**Methods**). The last two columns show the minimum and maximum probabilities observed for each per-cell type state for any combination of tissue group and full-stack state. The heatmap colors correspond to the 25-state's colors and are scaled such that the maximum probability values in each block are colored darkest (as seen in the right most column). The figure complements **Supplementary Figure 2.8** in providing information on how each full-stack state can correspond to different 25-per-cell type states in different groups of cells, hence stratifying full-stack states' characteristics in more details. For example, full-stack state ReprPC8 shows high probabilities of overlapping ReprPC state in ESC-related cell groups (ESC, iPSC, and ES-derived), and quiescent state in other cell groups.



Supplementary Figure 2. 10: Full-stack states' average gene expression in different cell types.

Each row corresponds to one of the 100 full-stack states grouped into state groups as indicated by the legend at the bottom. Each column corresponds to one of 56 cell types whose gene expression data were available from Roadmap Epigenomics *(Meuleman* et al., *2015)*. The columns are grouped based on their associated tissue group as indicated by the legend at the bottom. Each column shows the average expression of genes in the respective cell type that overlap with each full-stack state, weighted by the extent of the overlap and the gene length (**Methods**). The figure highlights how states in the transcription and exon group show consistently high gene expression across all cell types, while cell-type-specific enhancer states tend to show higher gene expression in the cell types corresponding to those states.

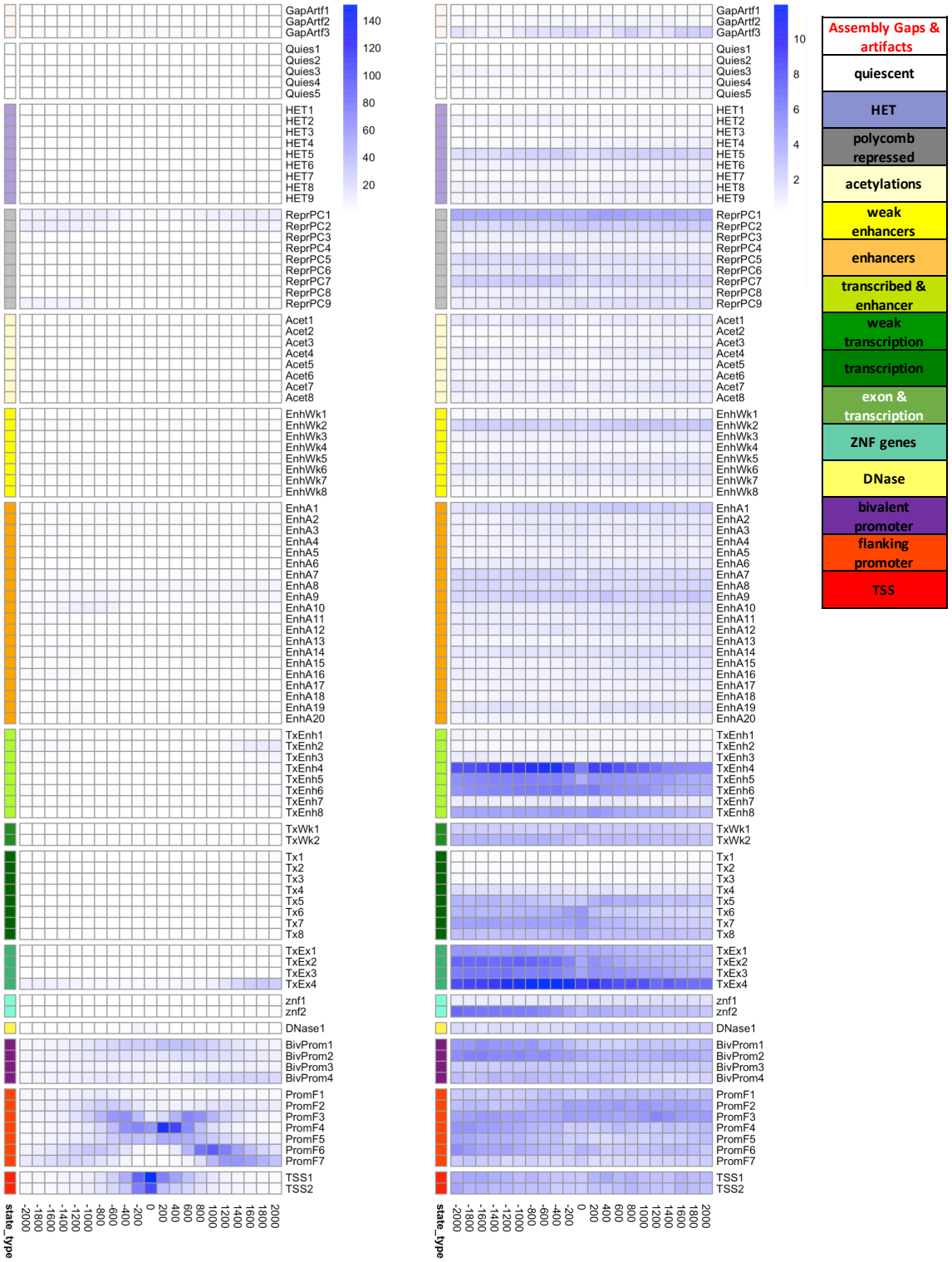


State groups

Assembly Gaps & artifacts	quiescent	HET	polycomb repressed	acetylations	weak enhancers	enhancers	transcribed & enhancer	weak transcription	transcription	exon & transcription	ZNF genes	DNase	bivalent promoter	flanking promoter	TSS

Supplementary Figure 2. 11: Full-stack states' average gene expression as a function of distance from TSS.

Each row corresponds to a full-stack state. Each column corresponds to a 200-bp bin within 25kb relative to annotated TSS, such that TSS is at position 0. Positions downstream of TSS in the direction of transcription have positive coordinate values, and those upstream have negative values. The heatmap shows for each state and position relative to the TSS, the average expression, across 56 cell types, of genes that have the state annotation at such position relative to the TSS (**Methods**). The figure highlights that states in the transcription group tend to have higher gene expression compared to other states, and the average gene expression is usually larger toward the downstream of genes. The figure also shows that for TSS and flanking promoter states, the average gene expression is relatively higher around the TSS compared to other positions.

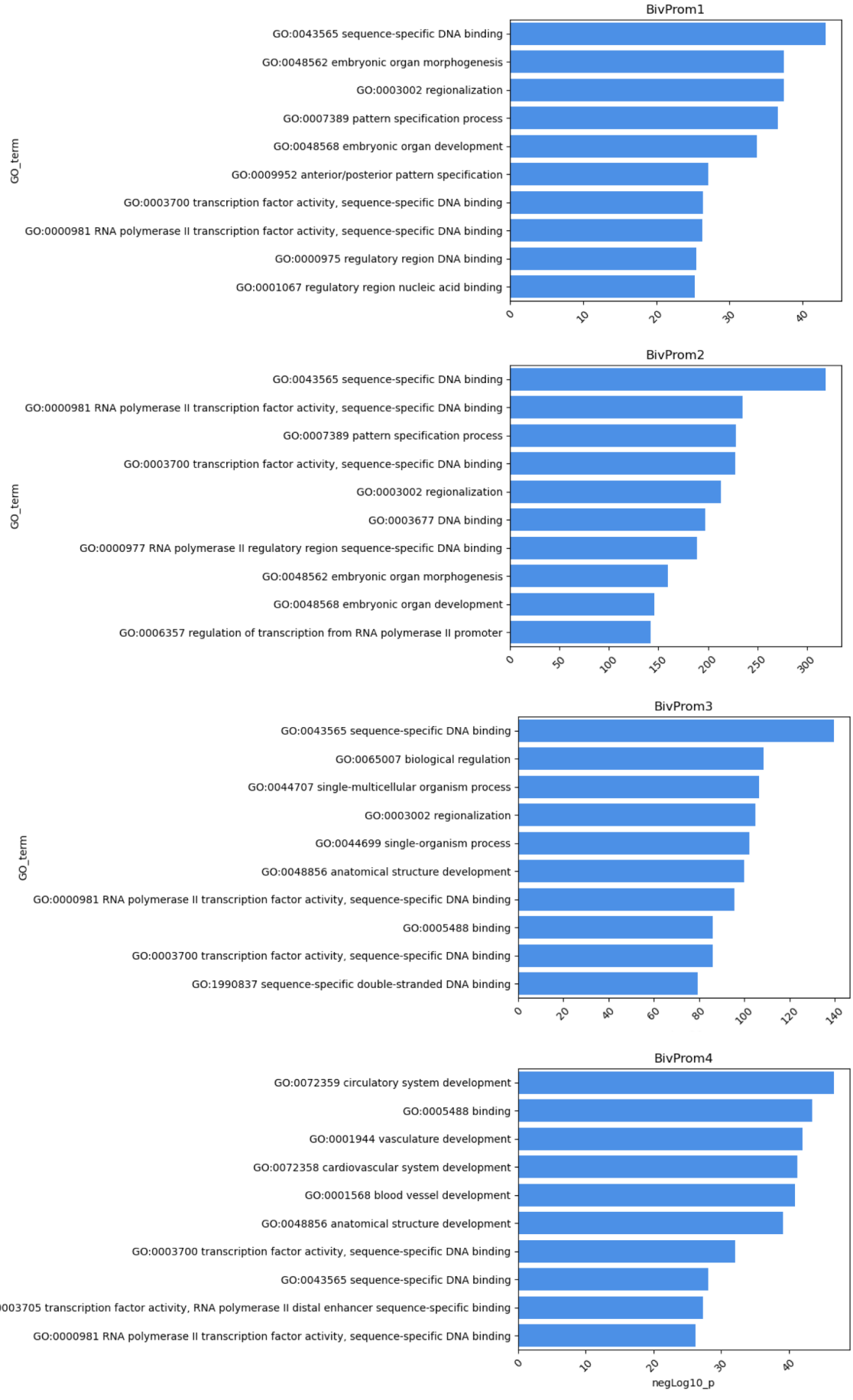


Supplementary Figure 2. 12: Positional enrichments of full-stack states around annotated transcription start sites and transcription end sites.

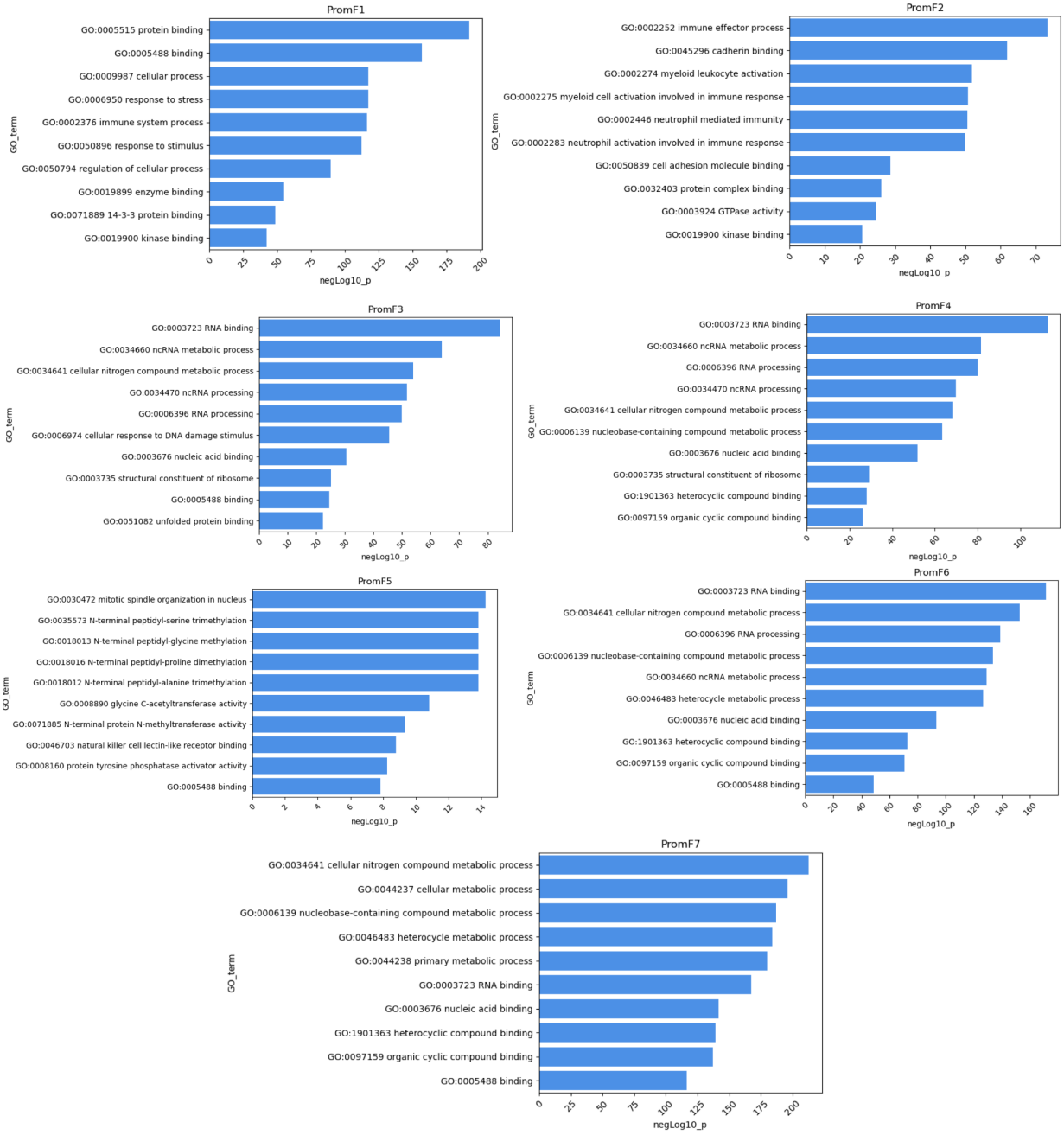
The Supplementary Figure 2.hows positional fold enrichments for positions within 2kb of annotated **(a)** transcription start sites (TSS) and **(b)** transcription end sites (TES). Each column

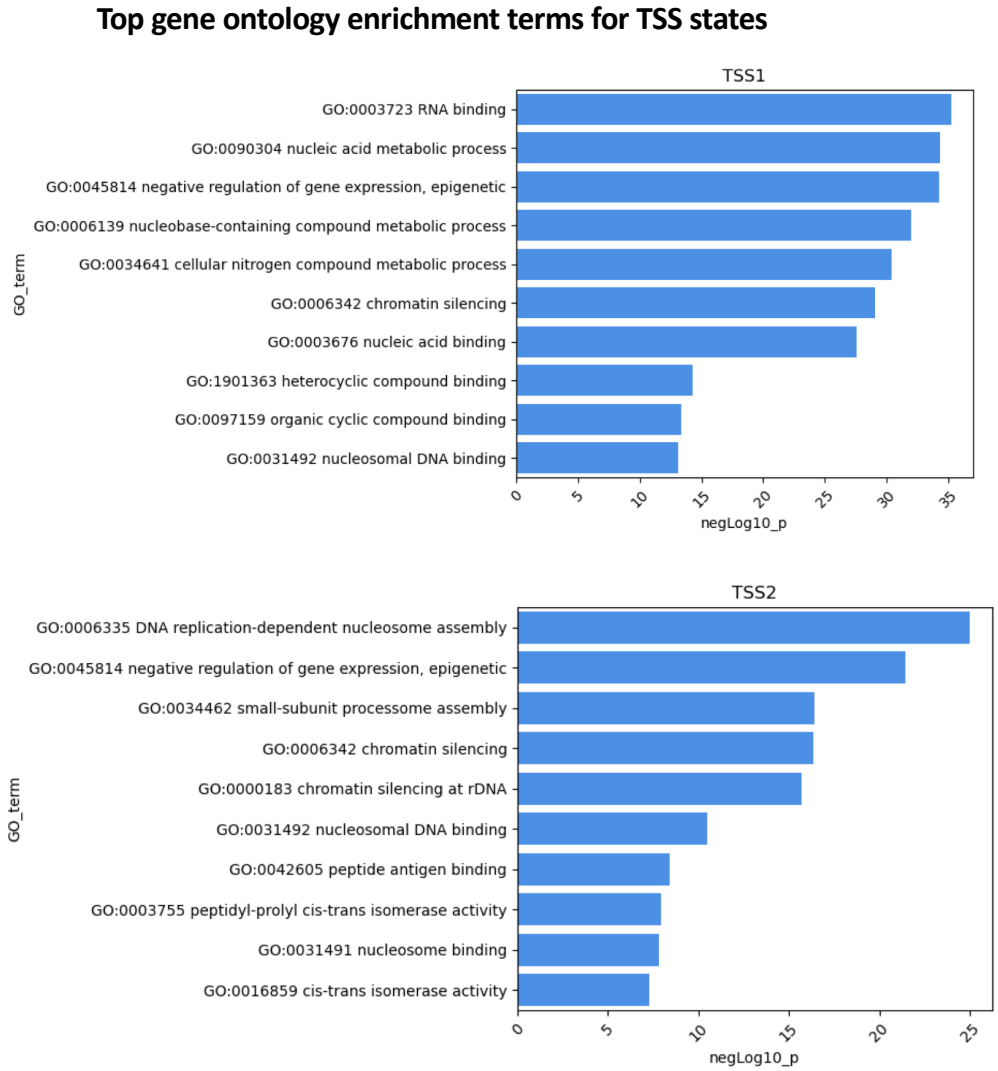
corresponds to one 200bp window as indicated at bottom. Positive coordinate values represent the number of bases downstream in the 5' to 3' direction of transcription, while negative values represent the number of bases upstream. Enrichments are calculated based on a genome-wide background. Color scale of enrichments is indicated at right for each panel. State groups' color legends are shown at right.

Top gene ontology enrichment terms for bivalent promoter states



Top gene ontology enrichment terms for flanking promoter states





Supplementary Figure 2. 13: Top GO terms for states in promoter-associated states. Each subpanel corresponds to a full-stack state (state names are shown in the plot title). In each subpanel, the top 5 most significantly enriched GO Biological Process and GO Molecular Function terms are shown on the y-axis (showing a total of 10 GO terms). The length of horizontal bars show the negative log10 (p value) of the GO enrichment, based on the Binomial Test and outputted by GREAT *(McLean* et al.*, 2010)*. The equivalent plots for all full-stack states are available in **Supplementary Data 2. 2.**

CTCF + open chromatin elements enrichments

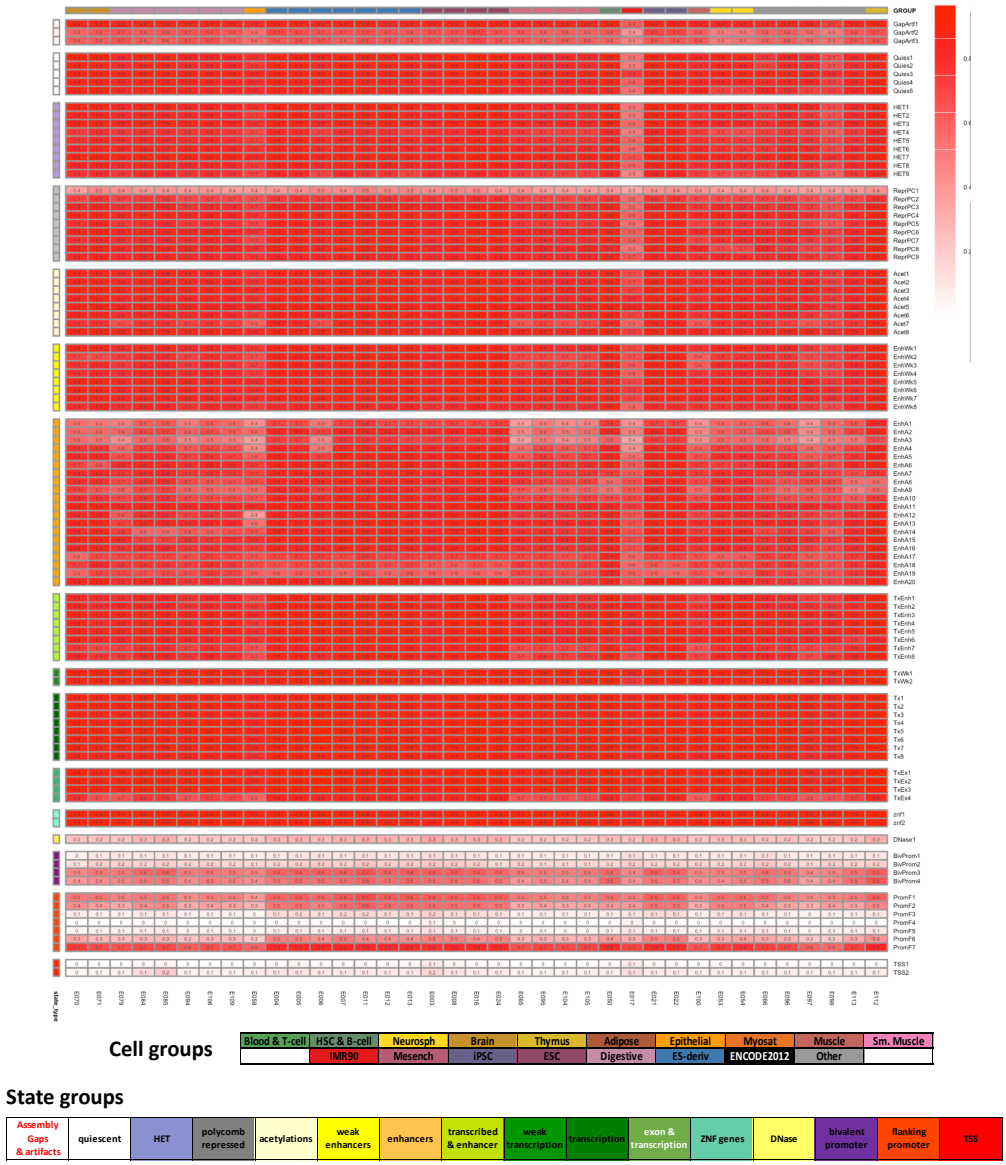
state	Genome %	Gm12878	H1hesc	Hela3	Hepg2	Huvec	K562
GapArtf1	11.9	0.0	0.0	0.0	0.0	0.0	0.0
GapArtf2	0.1	5.5	3.1	2.7	4.4	3.9	4.0
GapArtf3	0.0	2.7	1.5	1.5	3.8	0.8	1.4
Quies1	9.9	0.0	0.0	0.0	0.0	0.0	0.0
Quies2	3.1	0.0	0.1	0.1	0.0	0.0	0.1
Quies3	12.2	0.0	0.0	0.1	0.0	0.0	0.1
Quies4	4.5	0.0	0.0	0.0	0.0	0.0	0.0
Quies5	1.7	0.0	0.0	0.0	0.0	0.0	0.0
HET1	0.7	0.1	1.6	0.4	0.4	0.1	0.3
HET2	0.7	0.1	2.0	0.2	0.5	0.0	0.4
HET3	1.4	0.0	0.0	0.0	0.0	0.0	0.0
HET4	0.6	0.2	0.1	0.1	0.2	0.2	0.3
HET5	0.2	1.0	2.1	1.1	1.9	1.0	2.0
HET6	0.6	0.1	0.6	0.2	0.5	0.1	0.4
HET7	1.0	0.0	0.3	0.1	0.2	0.0	0.2
HET8	0.4	1.7	1.9	1.6	1.9	1.2	1.9
HET9	1.0	0.1	0.1	0.1	0.1	0.0	0.1
ReprPC1	0.2	0.5	1.0	2.3	1.2	0.3	1.9
ReprPC2	0.3	0.3	0.6	0.7	0.5	0.3	0.6
ReprPC3	1.1	0.1	0.4	0.2	0.2	0.1	0.3
ReprPC4	3.9	0.0	0.1	0.0	0.0	0.0	0.0
ReprPC5	0.6	0.3	3.2	1.0	0.8	0.2	0.9
ReprPC6	1.5	0.1	0.9	0.3	0.4	0.1	0.4
ReprPC7	0.6	1.0	6.5	2.9	3.4	1.1	3.1
ReprPC8	0.5	0.2	0.3	0.3	0.1	0.0	0.3
ReprPC9	0.4	0.1	0.3	0.3	0.3	0.3	0.3
Acet1	0.2	2.0	7.0	2.8	4.1	1.6	3.1
Acet2	0.9	0.4	1.9	0.7	0.9	0.3	0.9
Acet3	2.6	0.0	0.2	0.1	0.1	0.1	0.2
Acet4	0.4	0.6	2.4	2.0	1.3	1.0	1.4
Acet5	0.9	0.1	0.2	0.3	0.2	0.3	0.2
Acet6	0.4	0.2	0.5	0.8	0.3	0.7	0.3
Acet7	0.3	1.6	3.6	3.2	2.5	1.7	1.8
Acet8	0.6	0.2	1.3	0.6	0.6	0.2	0.6
EnhWk1	1.5	0.1	0.0	0.3	0.1	0.5	0.2
EnhWk2	0.4	1.9	5.6	4.3	4.2	2.7	3.4
EnhWk3	0.8	0.2	0.8	0.6	0.6	0.6	0.5
EnhWk4	2.2	0.1	0.1	0.1	0.1	0.3	0.1
EnhWk5	1.0	0.3	0.2	0.2	0.3	0.3	0.3
EnhWk6	0.6	1.0	0.5	0.6	0.5	0.7	0.8
EnhWk7	0.5	0.9	0.3	0.5	0.4	0.4	0.6
EnhWk8	1.4	0.1	1.2	0.1	0.3	0.2	0.2
EnhA1	0.2	6.8	5.3	6.0	5.9	4.4	5.0
EnhA2	0.3	1.6	3.0	3.1	2.6	3.9	2.6
EnhA3	0.2	16.5	9.9	10.7	12.3	10.0	14.0
EnhA4	0.3	1.4	1.4	2.9	1.8	1.9	1.7
EnhA5	0.7	0.6	0.4	2.0	0.8	2.3	0.8
EnhA6	0.6	0.5	1.1	0.8	0.9	0.6	0.9
EnhA7	0.4	3.0	4.8	2.7	3.3	1.5	3.6
EnhA8	0.3	1.4	3.1	2.3	2.0	1.7	2.3
EnhA9	0.2	2.5	5.7	5.7	5.3	4.5	3.1
EnhA10	0.4	3.4	1.9	3.1	2.3	3.4	3.0
EnhA11	0.7	0.4	0.3	0.4	0.5	0.3	0.5
EnhA12	0.3	1.4	3.8	3.9	2.4	1.3	2.0
EnhA13	0.8	0.2	0.3	0.6	0.2	0.2	0.2
EnhA14	0.4	1.0	2.4	2.3	1.7	1.0	1.8
EnhA15	1.0	0.1	0.2	0.3	0.4	0.1	0.3
EnhA16	0.6	5.2	5.7	5.7	5.8	6.2	6.9
EnhA17	0.5	1.5	3.4	1.7	1.9	2.2	1.8
EnhA18	0.5	0.4	2.9	0.7	1.1	0.3	0.9
EnhA19	0.3	3.1	3.9	5.2	4.4	3.4	4.1
EnhA20	0.3	0.1	0.1	0.3	0.2	0.6	0.2
TxEnh1	0.4	0.4	0.3	0.7	0.5	0.9	0.5
TxEnh2	0.4	1.0	1.0	1.3	1.3	1.6	1.4
TxEnh3	0.2	1.1	2.7	1.7	1.6	1.0	0.9
TxEnh4	0.3	3.8	7.9	7.4	6.2	5.3	2.8
TxEnh5	0.5	2.0	6.9	4.2	5.3	3.2	3.7
TxEnh6	0.2	2.9	4.5	4.2	4.1	3.6	2.5
TxEnh7	0.3	3.1	4.9	5.2	4.5	4.1	3.0
TxEnh8	0.2	4.1	4.3	5.2	4.6	4.3	3.5
TxWk1	2.8	0.1	0.1	0.1	0.1	0.1	0.1
TxWk2	0.8	0.3	2.2	0.9	1.3	0.5	1.1
Tx1	0.8	0.2	0.2	0.3	0.2	0.3	0.3
Tx2	1.6	0.1	0.1	0.1	0.1	0.2	0.1
Tx3	0.5	0.5	1.6	1.0	1.0	0.5	0.9
Tx4	0.5	0.3	0.2	0.3	0.3	0.4	0.3
Tx5	0.9	0.1	0.5	0.4	0.3	0.2	0.3
Tx6	1.1	0.1	0.0	0.1	0.1	0.1	0.1
Tx7	0.8	0.2	0.1	0.3	0.2	0.3	0.2
Tx8	0.7	1.6	1.1	1.6	1.7	2.1	1.5
TxEx1	0.3	0.3	0.6	0.4	0.2	0.3	0.2
TxEx2	0.6	0.2	1.2	0.8	0.6	0.3	0.2
TxEx3	0.7	0.2	2.3	1.3	1.3	0.6	0.9
TxEx4	0.1	2.4	3.3	3.4	2.4	2.4	1.1
znf1	0.4	0.2	0.5	0.3	0.4	0.3	0.3
znf2	0.2	0.3	1.2	0.7	0.7	0.6	0.9
DNase1	0.2	260.3	130.8	187.1	202.5	239.9	214.8
BivProm1	0.1	6.2	0.7	4.7	4.3	3.3	5.0
BivProm2	0.2	3.5	1.0	3.9	4.4	2.6	4.4
BivProm3	0.3	4.8	7.7	6.6	6.6	5.1	6.6
BivProm4	0.1	2.9	2.5	2.7	3.1	2.4	3.4
PromF1	0.2	44.4	23.9	31.9	27.5	42.3	28.1
PromF2	0.1	8.0	6.2	6.5	4.8	6.5	5.1
PromF3	0.2	0.1	0.2	0.3	0.4	0.0	0.3
PromF4	0.2	0.0	0.0	0.0	0.1	0.0	0.1
PromF5	0.1	2.8	0.3	2.9	2.2	0.3	2.9
PromF6	0.1	0.0	0.1	0.0	0.1	0.0	0.1
PromF7	0.2	0.2	0.6	0.2	0.2	0.3	0.2
TSS1	0.1	0.3	0.8	0.6	0.5	0.0	0.9
TSS2	0.1	16.3	8.8	8.9	10.9	8.2	12.2
Base	100	0.2	0.4	0.3	0.2	0.3	0.2

CTCF + *limited* open chromatin elements enrichments

state	Genome %	Gm12878_Ctcf	H1hesc_Ctcf	Hela3_Ctcf	Hepg2_Ctcf	Huvec_Ctcf	K562_Ctcf
GapArtf1	11.9	0.0	0.0	0.0	0.0	0.0	0.0
GapArtf2	0.1	0.7	0.6	0.6	0.9	0.6	0.4
GapArtf3	0.0	0.3	0.0	0.4	0.3	0.0	0.1
Quies1	9.9	0.2	0.2	0.2	0.2	0.2	0.1
Quies2	3.1	0.5	1.0	0.5	0.7	0.3	0.5
Quies3	12.2	0.5	0.4	0.5	0.5	0.8	0.6
Quies4	4.5	0.2	0.2	0.2	0.2	0.2	0.2
Quies5	1.7	0.4	0.3	0.4	0.3	0.3	0.4
HET1	0.7	2.0	4.0	1.9	2.8	1.1	1.6
HET2	0.7	1.3	2.9	1.3	2.0	0.5	1.9
HET3	1.4	0.3	0.3	0.3	0.4	0.4	0.4
HET4	0.6	0.6	0.5	0.6	0.7	0.7	0.7
HET5	0.2	2.3	2.5	2.4	2.8	2.6	2.8
HET6	0.6	1.2	1.8	1.4	2.0	1.0	1.7
HET7	1.0	0.8	1.4	1.0	1.4	0.8	1.2
HET8	0.4	2.5	2.5	2.6	2.9	2.5	2.9
HET9	1.0	0.5	0.4	0.6	0.7	0.5	0.7
ReprPC1	0.2	1.1	0.2	2.2	1.0	0.3	1.3
ReprPC2	0.3	1.4	0.7	1.6	1.2	0.8	1.7
ReprPC3	1.1	1.2	1.4	1.3	1.3	1.0	1.5
ReprPC4	3.9	0.5	0.6	0.6	0.6	0.6	0.7
ReprPC5	0.6	2.2	2.9	2.4	2.0	1.8	1.9
ReprPC6	1.5	1.4	2.4	1.6	1.7	1.3	1.8
ReprPC7	0.6	3.9	3.4	3.7	2.9	3.8	2.7
ReprPC8	0.5	1.6	1.6	1.6	1.4	1.3	1.7
ReprPC9	0.4	1.1	0.5	1.1	1.0	1.2	1.1
Acet1	0.2	3.5	4.1	3.3	3.8	3.5	2.8
Acet2	0.9	2.9	3.7	2.8	3.3	2.6	2.8
Acet3	2.6	0.9	1.2	1.0	1.1	1.0	1.1
Acet4	0.4	3.2	3.5	2.9	3.0	3.1	2.8
Acet5	0.9	1.3	1.4	1.4	1.4	1.5	1.4
Acet6	0.4	1.5	1.4	1.4	1.5	1.4	1.5
Acet7	0.3	2.8	2.5	1.7	2.0	1.7	2.4
Acet8	0.6	1.8	2.5	1.9	2.0	1.9	2.1
EnhWk1	1.5	1.3	1.0	1.2	1.2	1.7	1.2
EnhWk2	0.4	3.6	2.8	3.3	2.4	3.7	2.6
EnhWk3	0.8	1.9	2.2	1.9	1.9	2.0	2.1
EnhWk4	2.2	1.4	1.2	1.1	1.3	1.5	1.1
EnhWk5	1.0	1.5	1.2	1.2	1.3	1.7	1.5
EnhWk6	0.6	1.3	1.5	1.7	1.6	2.0	1.8
EnhWk7	0.5	0.5	1.3	1.4	1.6	1.8	1.7
EnhWk8	1.4	2.0	1.9	1.7	2.1	1.9	1.7
EnhA1	0.2	3.0	1.0	1.5	1.2	1.2	1.8
EnhA2	0.3	4.6	3.4	3.6	3.5	3.6	4.1
EnhA3	0.2	3.3	3.6	1.9	3.0	1.0	3.3
EnhA4	0.3	2.7	2.0	1.1	2.2	1.1	2.3
EnhA5	0.7	2.7	2.1	1.9	2.5	2.1	2.2
EnhA6	0.6	2.7	2.3	2.3	2.3	2.7	2.5
EnhA7	0.4	3.3	3.6	4.1	3.7	4.3	3.2
EnhA8	0.3	0.7	2.1	2.4	2.1	2.4	1.8
EnhA9	0.2	0.4	1.7	1.7	1.2	1.2	1.0
EnhA10	0.4	2.6	3.0	2.8	2.8	3.7	2.9
EnhA11	0.7	1.1	1.2	1.2	1.2	1.4	1.4
EnhA12	0.3	3.4	3.1	2.4	2.9	3.0	2.8
EnhA13	0.8	1.2	1.2	1.1	1.2	1.2	1.2
EnhA14	0.4	2.7	2.6	2.5	1.2	2.9	2.2
EnhA15	1.0	1.2	1.1	1.2	0.8	1.3	1.4
EnhA16	0.6	10.0	7.0	7.4	8.9	9.8	7.1
EnhA17	0.5	5.0	2.6	4.2	4.6	4.8	4.4
EnhA18	0.5	2.9	1.1	2.6	3.0	2.6	2.8
EnhA19	0.3	5.2	0.7	4.3	3.8	5.3	3.8
EnhA20	0.3	1.2	0.9	1.0	1.0	0.2	1.1
TxEnh1	0.4	1.4	0.9	1.2	1.0	1.0	1.1
TxEnh2	0.4	0.6	0.9	1.1	0.7	1.0	0.7
TxEnh3	0.2	0.4	1.1	0.9	0.6	0.8	0.5
TxEnh4	0.3	0.4	1.0	1.4	0.2	0.1	0.6
TxEnh5	0.5	2.1	2.2	2.9	1.9	2.1	2.2
TxEnh6	0.2	0.6	1.8	2.1	1.2	1.2	1.2
TxEnh7	0.3	1.6	1.3	1.9	1.2	1.8	1.3
TxEnh8	0.2	0.9	1.5	1.5	0.7	0.8	1.3
TxWk1	2.8	0.5	0.5	0.6	0.5	0.7	0.6
TxWk2	0.8	1.6	2.1	1.9	1.7	1.7	1.8
Tx1	0.8	0.5	0.5	0.6	0.4	0.7	0.5
Tx2	1.6	0.5	0.5	0.6	0.5	0.8	0.6
Tx3	0.5	1.3	1.6	1.5	1.3	1.5	1.3
Tx4	0.5	0.3	0.5	0.5	0.2	0.4	0.3
Tx5	0.9	0.3	0.7	0.7	0.4	0.4	0.6
Tx6	1.1	0.3	0.3	0.4	0.3	0.6	0.3
Tx7	0.8	0.2	0.3	0.3	0.1	0.2	0.2
Tx8	0.7	1.6	1.8	1.9	1.5	1.9	1.8
TxEx1	0.3	0.1	0.2	0.2	0.0	0.0	0.1
TxEx2	0.6	0.0	0.3	0.4	0.0	0.0	0.1
TxEx3	0.7	0.4	1.0	1.0	0.4	0.2	0.6
TxEx4	0.1	0.1	0.3	0.3	0.2	0.1	0.2
znf1	0.4	0.5	0.8	0.9	0.8	0.8	0.8
znf2	0.2	0.4	0.5	0.7	0.4	0.4	0.6
DNase1	0.2	31.7	32.8	26.5	37.6	18.6	28.5
BivProm1	0.1	1.4	0.0	3.3	0.9	0.4	1.2
BivProm2	0.2	1.7	0.1	2.7	1.3	0.8	1.4
BivProm3	0.3	5.9	1.5	4.9	3.3	4.3	3.5
BivProm4	0.1	3.4	1.0	2.4	2.1	1.4	2.8
PromF1	0.2	7.4	2.5	6.4	3.3	7.5	3.8
PromF2	0.1	0.7	0.6	1.4	0.6	0.3	0.9
PromF3	0.2	0.0	0.0	0.1	0.0	0.0	0.0
PromF4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
PromF5	0.1	0.6	0.0	1.7	0.4	0.1	1.0
PromF6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
PromF7	0.2	0.1	0.2	0.1	0.1	0.2	0.3
TSS1	0.1	0.0	0.0	0.3	0.1	0.0	0.1
TSS2	0.1	1.8	1.0	2.7	1.3	0.7	1.6
	100	0.6	0.7	0.8	0.6	0.6	0.7

Supplementary Figure 2. 14: Full-stack states enrichments with CTCF associated chromatin states.

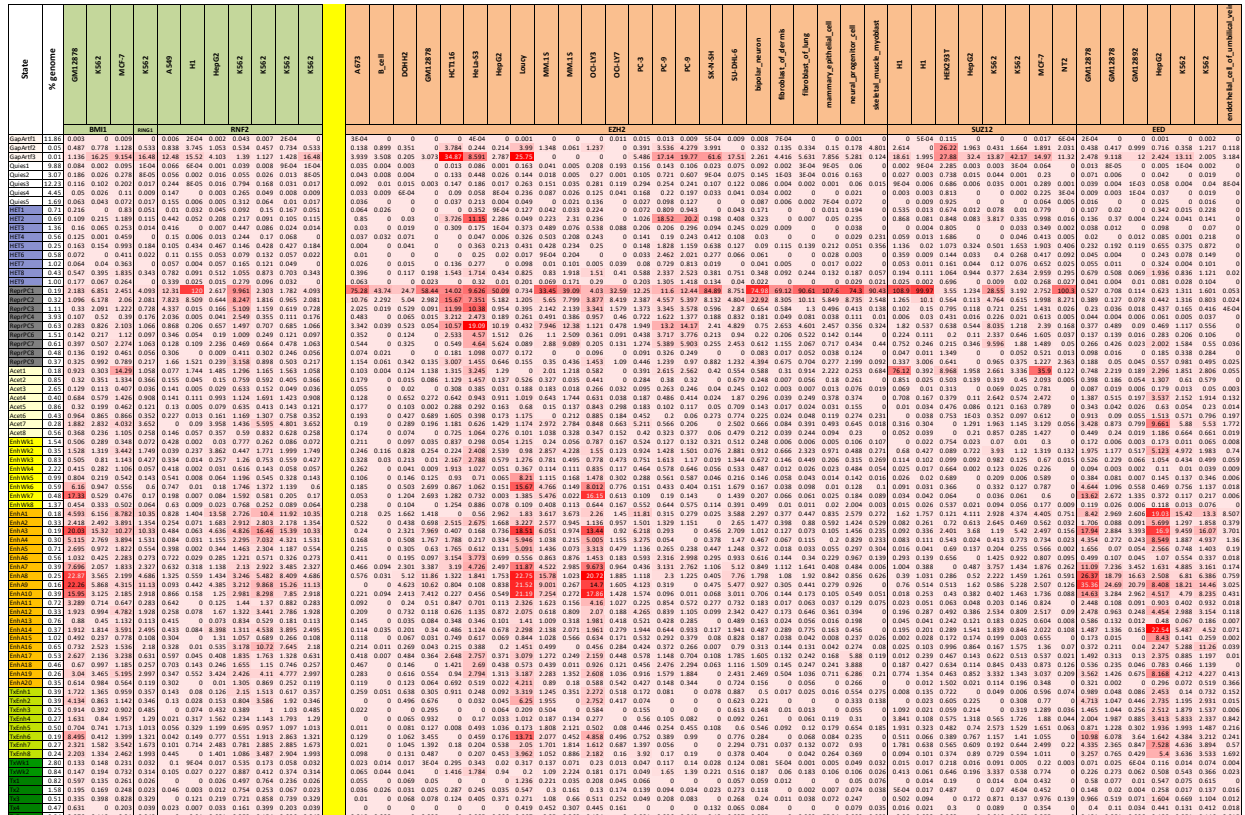
(A) The heatmap shows enrichment values for the full-stack states (rows) and a chromatin state that corresponds to CTCF with open chromatin and limited histone modification signal from concatenated annotations in six different cell types (columns). CTCF signals were included as input for training these concatenated chromatin state models (**Methods**). Coloring of enrichments is column specific. **(B)** Similar to **(A),** except showing enrichments for a state associated with CTCF datasets with limited open chromatin and limited histone modification signals.



Supplementary Figure 2. 15: Full-stack states' average DNA methylation in different cell types.

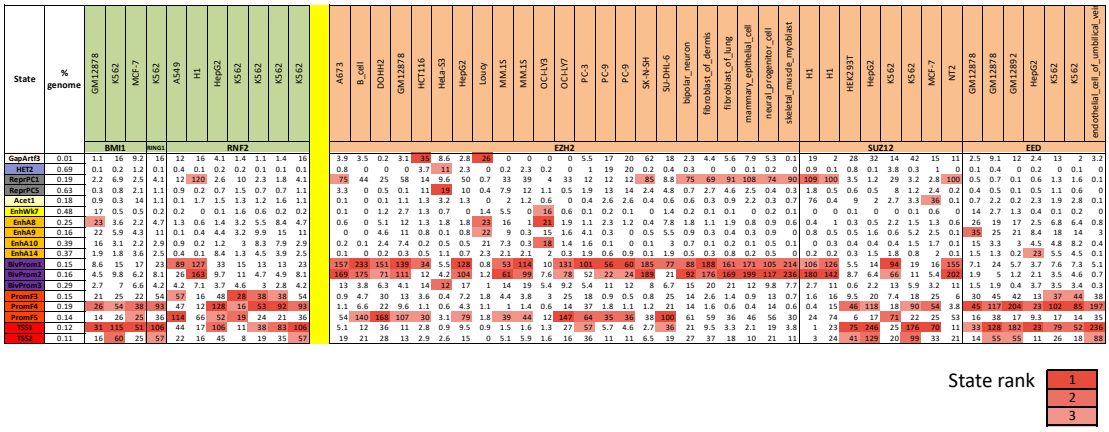
Each row corresponds to one of the 100 full-stack states grouped into state groups as indicated by the legend at the bottom. Each column corresponds to one cell type whose DNA methylation data were available from Roadmap Epigenomics. The columns are grouped based on their associated tissue group as indicated by the legend at the bottom. Each column shows the average DNA methylation level in the respective cell type that overlaps with each full-stack state (**Methods**). Among promoters-associated states, those most enriched with CpG islands also show lowest average DNA methylation levels (**Figure 2.3A**), consistent with expectation *(Jones and Takai, 2001; Weber* et al.*, 2007)*. The lower DNAme levels of IMR90 compared to other cell types might be related to a technical batch effect since it was one of two original WGBS datasets collected *(Lister* et al.*, 2009)*.

Enrichments with polycomb repressed protein groups PRC1 and PRC2

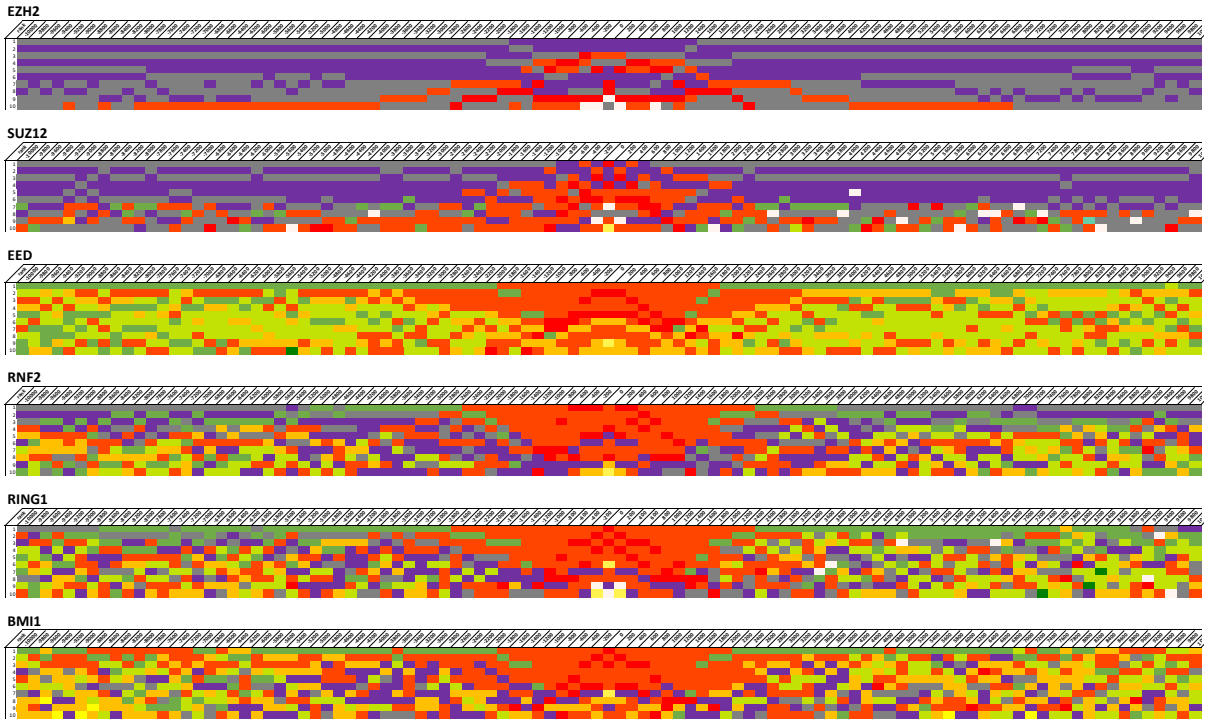


Supplementary Figure 2. 16: Full-stack states enrichments with Polycomb Repressive protein complexes PRC1 (green column headers) and PRC2 (orange column headers). The heatmap shows enrichment values for the full-stack states (rows) and the binding sites of subunits of PRC1 and PRC2 in different cell types (columns) that were available from ENCODE project (**Methods**). Coloring of enrichments is column specific with highest and lowest enrichment values in each column are colored red and white, respectively. The first two columns show state names, and their percentage of genome coverage. The last row shows the percentage of genome coverage for each type of PRCs.

Top enriched states with PRC1 and PRC2



Neighborhood enrichments with PRC1 and PRC2: ranked states by descending enrichments



Supplementary Figure 2. 17: Enrichments of selected full-stack states with PRC1 and PRC2.

(A) A subset of **Supplementary Figure 2.16** showing fold enrichment of full-stack states for binding sites of polycomb repressive protein complexes in different cell types from ENCODE (Methods). The column names highlighted in green (PRC1) and orange (PRC2) show the subunits of PRCs (BMI1, RING1 and RNF2 in PRC1, and EZH2, SUZ12, EED in PRC2) and the cell types where the PRCs were profiled in the second and first rows of column names, respectively. Each corresponding row corresponds to a state, and only states that were among the top three with greatest enrichments for at least one category of PRC complexes are shown. Top enrichment values are colored red based on the rank of the state for each score as indicated in the color legend at the bottom. Some states that show strong consensus across cell types in enrichments

with PRC1 and PRC2 include ReprPC1, BivProm1-2, PromF4-5, TSS1-2. ReprPC1 and BivProm1-2 all show strong signals of H3K27me3. **(B**) Neighborhood enrichments of full-stack states with binding sites of PRC1 and PRC2 complexes. In each subpanel, each column corresponds to a 200-bp bin across the 20,000-bp regions overlapping and surrounding annotated PRC1&2 subunit complexes. Within each column, the top 10 states most enriched at the corresponding 200-bp position (within the 20,000bp window) are shown, in descending order of enrichments, and colored based on the state groups as presented throughout the paper. **Supplementary Data 2.4** accompanies this figure to show full state names and rankings.

hg19 state enrichments with hg19 genome contexts

state	% Genome	GENCODE_coding	RefSeqExon	RefSeqGene	RefSeqTES	CpGIsland	RefSeqTSS	RefSeqTSS2kb	assembly_gap	phastCons_elements	pseudogene	repeat_elements	znf_genes
GapArtf1	11.86	0.23	0.19	0.27	0.31	0.26	0.25	0.27	8.43	0.21	1.59	0.56	0.21
GapArtf2	0.05	0.26	0.33	0.15	0.42	1.25	1.6	0.8	0.02	0.35	1.65	1.85	1.44
GapArtf3	0.012	0.69	0.91	0.3	2.03	5.01	4.87	2.23	0.01	0.71	1.74	1.7	1.26
Quies1	9.883	0.05	0.09	0.54	0.19	0.02	0.06	0.13	0	0.7	0.42	1.03	0.4
Quies2	3.07	0.15	0.17	0.56	0.22	0.04	0.19	0.17	0	0.58	0.52	1.18	0.29
Quies3	12.23	0.22	0.39	1.01	0.76	0.07	0.14	0.44	0	0.82	1.24	1.19	0.55
Quies4	4.452	0.07	0.12	0.59	0.28	0.09	0.11	0.21	0	0.35	1.46	1.73	0.8
Quies5	1.692	0.22	0.33	0.7	0.62	0.06	0.17	0.41	0	1.21	0.85	1.26	0.2
HET1	0.706	0.43	0.34	0.55	0.2	0.29	0.53	0.28	0	0.47	0.95	1.41	0.33
HET2	0.692	1.01	0.79	0.79	0.42	1.25	0.54	0.52	0	0.72	0.58	0.61	0.66
HET3	1.359	0.08	0.16	0.79	0.38	0	0.06	0.17	0	0.31	0.72	1.86	1.77
HET4	0.563	0.23	0.47	0.57	0.74	0	0.08	0.21	0	0.33	1.35	1.87	6.03
HET5	0.249	1.76	1.95	0.8	1.42	1.28	0.65	0.86	0	0.59	2.52	1.58	14
HET6	0.581	0.86	0.67	0.48	0.63	1.28	0.86	0.65	0	0.46	2.85	1.66	2.58
HET7	1.023	0.36	0.35	0.68	0.4	0.64	0.43	0.41	0	0.32	1.63	1.78	2.04
HET8	0.435	0.89	0.72	0.64	0.6	0.47	0.71	0.91	0	0.47	1.57	1.44	0.87
HET9	0.998	0.35	0.37	0.42	0.75	0.02	0.44	0.6	0	0.51	3.53	1.49	1.44
ReprPC1	0.191	2.91	3.07	1.05	3.68	8.86	3.04	8.51	0	2.51	0.89	0.37	0.3
ReprPC2	0.325	0.87	1.4	0.87	2.72	0.52	0.77	5.5	0	1.66	1.19	0.78	0.28
ReprPC3	1.107	0.55	0.68	0.71	1.06	0	0.4	1.14	0	1.16	1.39	1	0.36
ReprPC4	3.935	0.32	0.41	0.69	0.77	0.03	0.21	0.55	0	0.92	1.21	1.16	0.29
ReprPC5	0.628	2.25	1.77	0.84	1.26	1.02	1.24	1.45	0	0.98	1.34	0.68	0.38
ReprPC6	1.513	1.42	1.18	0.85	1.22	0.3	0.62	0.77	0	0.79	1.5	0.91	0.32
ReprPC7	0.614	3.99	2.85	1.21	1.54	1.56	1.64	1.54	0	1.31	0.9	0.4	0.61
ReprPC8	0.477	1.07	0.94	0.8	0.77	0.07	0.57	0.73	0	1.34	1.11	1.12	0.37
ReprPC9	0.375	0.43	0.78	0.79	1.64	0.44	0.36	3.03	0	1.12	1.26	1.02	0.53
Acet1	0.184	2.58	1.68	1.09	0.47	3.23	1.75	1.1	0	1.06	1.64	1.24	1.09
Acet2	0.855	0.77	0.59	0.83	0.25	0.11	0.52	0.55	0	0.81	1.07	1.01	0.37
Acet3	2.649	0.36	0.36	0.83	0.34	0.05	0.24	0.45	0	0.54	1.02	1.29	0.4
Acet4	0.403	0.76	0.81	0.99	0.56	0.1	0.66	0.93	0	0.78	0.82	0.83	0.38
Acet5	0.86	0.29	0.4	0.89	0.48	0.01	0.28	0.48	0	0.75	0.86	1.06	0.32
Acet6	0.428	0.36	0.56	1.04	0.76	0	0.47	0.63	0	1.13	0.68	0.87	0.46
Acet7	0.285	0.63	0.89	1.13	0.61	0.13	0.98	1.29	0	1.25	0.58	0.64	0.39
Acet8	0.562	0.53	0.58	0.76	0.53	0.21	0.58	0.81	0	0.58	0.89	1.09	0.31
EnhWk1	1.54	0.18	0.39	0.95	0.76	0.01	0.16	0.38	0	1.57	0.7	0.81	0.59
EnhWk2	0.352	1.95	2.03	1.64	1.76	0.77	1.85	2.21	0	1.39	0.36	0.42	0.66
EnhWk3	0.83	0.48	0.73	1.25	1.03	0.07	0.69	1.15	0	1.88	0.57	0.69	0.72
EnhWk4	2.216	0.19	0.38	1.12	0.68	0.02	0.22	0.45	0	2.31	0.48	0.66	0.59
EnhWk5	0.994	0.31	0.54	1.04	0.99	0.05	0.27	1.11	0	0.98	0.95	1.06	0.6
EnhWk6	0.588	0.52	0.87	1.46	1.58	0.04	0.41	1.58	0	0.96	0.65	1	0.72
EnhWk7	0.484	0.38	0.55	1.12	0.99	0.05	0.32	0.97	0	0.94	0.84	1.06	0.5
EnhWk8	1.369	0.19	0.35	0.94	0.61	0.01	0.22	0.35	0	1.66	0.57	0.85	0.42
EnhA1	0.179	1.37	1.9	1.67	2.1	0.69	3.72	3.3	0	2.32	0.33	0.32	0.73
EnhA2	0.328	0.47	0.94	1.31	1.2	0.13	1.97	1.72	0	3.36	0.54	0.42	0.8
EnhA3	0.194	0.41	0.87	1.61	1.4	0.01	1.98	2.22	0	3.05	0.46	0.43	0.82
EnhA4	0.301	0.34	0.68	1.14	1.05	0	0.91	0.91	0	2.35	0.61	0.6	0.54
EnhA5	0.714	0.29	0.58	1.09	1.16	0.01	0.43	0.62	0	2.2	0.66	0.68	0.55
EnhA6	0.563	0.74	1.01	1.61	1.19	0.14	0.93	1.31	0	1.85	0.43	0.65	0.91
EnhA7	0.393	2.15	1.83	1.07	1.32	0.58	1.97	2.48	0	0.94	0.89	0.71	0.5
EnhA8	0.255	1.57	1.78	1.45	1.76	0.57	4.94	5.54	0	1.27	0.55	0.67	0.71
EnhA9	0.162	1.24	1.78	1.86	2.12	0.39	2.78	4.24	0	1.44	0.23	0.5	0.89
EnhA10	0.395	0.42	0.89	1.19	1.39	0.21	1.41	4.58	0	1.48	0.74	0.8	0.92
EnhA11	0.715	0.4	0.67	0.98	1.15	0.03	0.58	1.31	0	0.81	1	1.14	0.73
EnhA12	0.332	1	1.08	0.92	0.9	0.22	1.76	1.61	0	1.17	0.74	0.75	0.31
EnhA13	0.764	0.29	0.42	0.88	0.69	0.01	0.43	0.72	0	1.06	0.76	0.96	0.27
EnhA14	0.366	1.09	1.22	1.14	1.28	0.18	2.84	3.02	0	1.16	0.74	0.78	0.34
EnhA15	1.017	0.43	0.6	1.02	1.06	0.03	0.72	1.42	0	1.05	0.95	1.01	0.42
EnhA16	0.645	0.42	0.59	0.91	1.01	0.37	0.89	2.09	0	1.01	1.34	1.04	0.41
EnhA17	0.526	0.3	0.65	1.36	1	0.05	0.82	1.15	0	3.15	0.45	0.53	0.94
EnhA18	0.457	0.47	0.54	0.85	0.61	0.14	0.84	0.85	0	1.23	0.7	1.03	0.49
EnhA19	0.257	1.05	1.27	1.32	0.91	0.43	2.25	1.99	0	1.95	0.47	0.61	1.12
EnhA20	0.346	0.28	0.49	1.08	0.96	0.01	0.19	0.52	0	1.53	0.72	0.76	0.77
TxEnh1	0.391	0.35	0.6	2.3	0.69	0	0.35	1.2	0	1.64	0.31	0.65	1.14
TxEnh2	0.391	0.37	0.45	2.3	0.5	0.01	0.43	2.74	0	1.08	0.27	0.85	1.95
TxEnh3	0.25	1.55	1.14	2.31	0.65	0.58	0.69	1.27	0	0.86	0.68	1.01	3.15
TxEnh4	0.267	14.2	10.4	2.26	6.15	2.58	1.3	1.62	0	3.37	0.51	0.37	1.47
TxEnh5	0.498	9.39	7.17	2.18	3.79	2.76	1.03	1.17	0	2.25	0.56	0.5	1.42
TxEnh6	0.189	6.29	5.96	2.21	5.84	0.7	0.94	1.49	0	1.81	0.46	0.61	1.14
TxEnh7	0.27	1.02	0.99	2.28	0.73	0.42	1.09	1.62	0	1.11	0.26	0.55	1.28
TxEnh8	0.243	4.84	4.96	2.22	4.27	0.23	0.82	0.81	0	2.11	0.48	0.56	1.22
TxWk1	2.8	1.39	2.28	2.15	3.09	0.08	0.18	0.35	0	1.16	0.88	1.11	1.71
TxWk2	0.842	6.61	4.8	2.11	2.51	1.58	0.71	0.72	0	1.66	0.87	0.77	1.27
Tx1	0.824	0.2	0.21	2.31	0.3	0.01	0.18	0.67	0	0.85	0.38	1.15	2.24
Tx2	1.58	0.18	0.25	2.27	0.35	0.04	0.15	0.48	0	0.84	0.61	1.17	1.74
Tx3	0.508	0.3	0.36	2.26	0.29	0.19	0.39	0.63	0	0.57	0.48	1.04	1.64
Tx4	0.47	2.66	2.15	2.31	1.38	0.04	0.39	0.65	0	1.33	0.6	1.01	2.7
Tx5	0.942	2.89	2.87	2.21	2.65	0.48	0.23	0.45	0	0.95	0.96	1.35	2.17
Tx6	1.109	3.6	4.39	2.27	4.91	0	0.14	0.25	0	2.03	0.4	0.69	1.71
Tx7	0.819	6.81	5.82	2.29	4.92	0	0.28	0.37	0	2.46	0.38	0.67	1.91
Tx8	0.681	2.51	3.51	2.2	4.03	0.02	0.34	0.57	0	1.74	0.51	0.74	1.29
TxEx1	0.265	10.4	6.47	2.32	3.88	0.6	1.27	1.3	0	2.9	0.5	0.72	2.61
TxEx2	0.556	15.8	9.97	2.32	4.1	1.19	0.73	0.76	0	3.65	0.57	0.54	2.45
TxEx3	0.663	9.74	7.04	2.25	4.26	2.73	0.6	0.78	0	2.21	0.9	0.83	2.6
TxEx4	0.098	10.7	7.83	2.26	9.28	1.99	4	8.1	0	3.09	0.37	0.41	2.17
znf1	0.406	1.06	1.22	2.13	1.39	0.3	0.23	0.4	0	0.6	1.33	1.67	20.8
znf2	0.152	10.4	7.13	2.14	3.28	0.92	0.93	0.84	0	1.16	1.88	1.21	68.6
DNase1	0.201	0.99	1.51	1.07	2.42	3.59	5.59	1.88	0	2.04	1.1	0.53	0.8
BivProm1	0.145	9.39	8.49	1.43	2.57	77	28.7	19.3	0	3.94	1.02	0.16	0.87
BivProm2	0.159	7.56	6.41	1.32	4.03	54.5	14.8	14.9	0	3.63	1.03	0.18	0.5
BivProm3	0.286	2.95	2.8	1.16	2.23	7.19	6.18	7.02	0	2.47	0.8	0.42	0.82
BivProm4	0.13	2.55	3.39	1.74	2.99	3.75	8.11	14.5	0	3.91	0.6	0.3	1.28
PromF1	0.201	2.16	2.41	1	2.35	2.6	6.51	6.46	0	1.85	0.63	0.45	0.81
PromF2	0.138	2.4	3.21	1.38	4.35	3.88	8.11	13.9	0	2.34	0.46	0.39	1.23
PromF3	0.153	4.63	5.09	1.52	4.37	25.3	13.3	25.7	0	2.19	0.53	0.31	3.39
PromF4	0.19	10.2	13.5	1.89	2.52	96.7	62.2	28	0	4.13	0.57	0.12	4.77
PromF5	0.135	7.33	8.49	1.56	2.42	71	31.9	22.4	0	3.95	0.72	0.16	2.27
PromF6	0.129	3.69	3.25	2.1	3.64	1.55	2.74	26.4	0	1.95	0.61	0.48	4.34
PromF7	0.16	1.72	2.04	2.08	2.5	0.09	1.05	20.4	0	1.53	0.71	0.68	4.41
TSS1	0.123	6.48	12.6	1.44	2.63	101	152	25.9	0	4.89	0.95	0.29	3.75
TSS2	0.114	5.53	8.12	1.15	2.44	75.4	106	20.1	0	4.46	1.74	0.57	2.13
	100	1.15	2.45	42.1	0.001	0.68	0.001	3.41	7.57	5.17	1.94	46.7	0.73

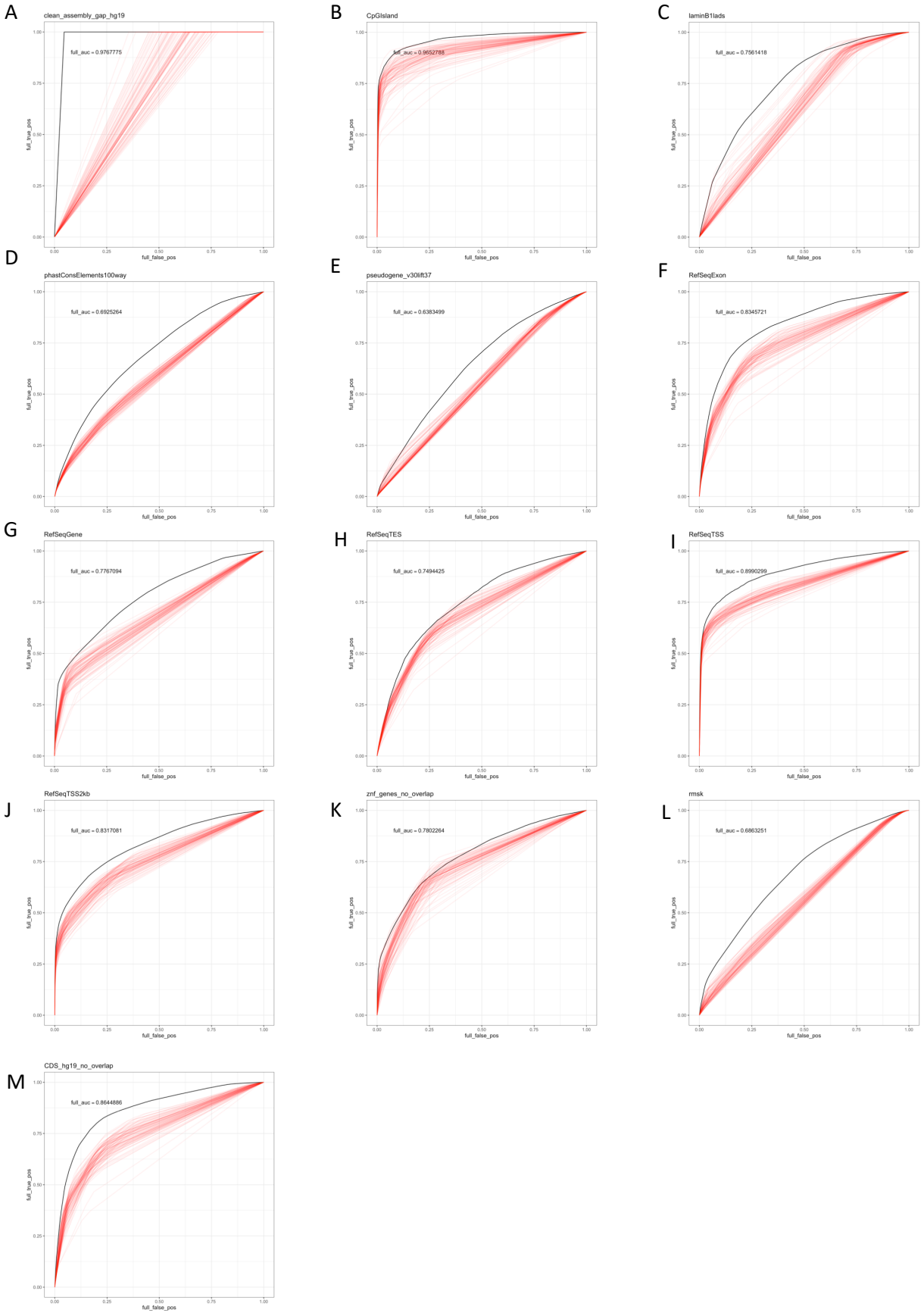
hg38 (liftedOver) state enrichments with hg38 genome contexts

state	% Genome	GENCODE_coding	RefSeqExon	RefSeqGene	RefSeqTES	CpGIsland	RefSeqTSS	RefSeqTSS2kb	assembly_gap	phastCons_elements	pseudogene	repeat_elements	znf_genes
GapArtf1	4.885	0.38	0.39	0.55	0.69	0.57	0.53	0.58	19.9	0.44	3.33	1.29	0.37
GapArtf2	0.053	0.24	0.3	0.14	0.39	1.18	1.45	0.72	0	0.8	1.59	1.4	1.37
GapArtf3	0.012	0.86	0.84	0.28	1.98	4.59	4.25	2.05	0	0.89	1.67	1.39	1.24
Quies1	10.66	0.04	0.08	0.5	0.18	0.02	0.06	0.12	0	0.66	0.4	0.95	0.37
Quies2	3.314	0.14	0.16	0.52	0.21	0.03	0.18	0.16	0	0.55	0.5	1.1	0.27
Quies3	13.17	0.21	0.36	0.94	0.71	0.08	0.12	0.41	0.02	0.76	1.17	1.11	0.52
Quies4	4.795	0.07	0.11	0.55	0.27	0.08	0.11	0.19	0	0.32	1.41	1.6	0.75
Quies5	1.82	0.21	0.3	0.66	0.59	0.05	0.16	0.38	0	1.13	0.78	1.17	0.19
HET1	0.762	0.4	0.32	0.51	0.18	0.27	0.5	0.26	0	0.43	0.91	1.31	0.31
HET2	0.746	0.94	0.73	0.75	0.38	1.17	0.5	0.48	0	0.66	0.55	0.57	0.62
HET3	1.466	0.07	0.15	0.74	0.35	0	0.07	0.16	0	0.29	0.69	1.72	1.65
HET4	0.606	0.21	0.43	0.52	0.69	0	0.08	0.2	0	0.28	1.3	1.73	5.66
HET5	0.267	1.65	1.83	0.75	1.33	1.17	0.62	0.81	0	0.49	2.39	1.46	13.1
HET6	0.627	0.8	0.62	0.45	0.58	1.13	0.83	0.61	0.99	0.37	2.75	1.52	2.41
HET7	1.103	0.34	0.33	0.63	0.38	0.59	0.41	0.38	0	0.28	1.57	1.64	1.9
HET8	0.47	0.83	0.67	0.59	0.56	0.43	0.64	0.84	0.66	0.42	1.49	1.33	0.81
HET9	1.069	0.33	0.35	0.39	0.71	0.02	0.41	0.56	0	0.46	3.41	1.38	1.35
ReprPC1	0.206	2.71	2.84	0.98	3.37	7.93	2.82	7.89	0	2.34	0.85	0.35	0.28
ReprPC2	0.353	0.81	1.29	0.8	2.5	0.74	0.71	5.08	0.88	1.54	1.18	0.72	0.26
ReprPC3	1.194	0.52	0.63	0.65	0.98	0	0.38	1.06	0	1.09	1.34	0.94	0.34
ReprPC4	4.243	0.3	0.38	0.64	0.71	0.03	0.19	0.51	0	0.87	1.16	1.08	0.27
ReprPC5	0.677	2.11	1.65	0.79	1.17	0.94	1.15	1.34	0	0.91	1.31	0.63	0.36
ReprPC6	1.631	1.33	1.09	0.79	1.13	0.28	0.58	0.72	0	0.73	1.45	0.85	0.3
ReprPC7	0.662	3.73	2.65	1.13	1.43	1.45	1.51	1.43	0	1.21	0.86	0.38	0.57
ReprPC8	0.514	1	0.88	0.74	0.71	0.06	0.53	0.68	0	1.25	1.02	1.04	0.35
ReprPC9	0.403	0.4	0.73	0.73	1.54	0.4	0.33	2.81	0	1.04	1.19	0.95	0.5
Acet1	0.204	2.37	1.55	1.05	0.46	3.1	1.57	0.97	3.04	0.81	1.54	1.04	1
Acet2	0.922	0.72	0.55	0.77	0.23	0.11	0.49	0.51	0	0.75	1.03	0.94	0.34
Acet3	2.856	0.34	0.34	0.77	0.32	0.05	0.23	0.42	0	0.5	0.97	1.2	0.37
Acet4	0.434	0.71	0.75	0.92	0.51	0.1	0.63	0.87	0	0.73	0.78	0.78	0.36
Acet5	0.927	0.27	0.37	0.82	0.44	0.01	0.26	0.45	0	0.71	0.82	0.99	0.3
Acet6	0.462	0.34	0.52	0.96	0.7	0	0.43	0.59	0	1.06	0.64	0.81	0.43
Acet7	0.307	0.59	0.82	1.06	0.58	0.11	0.91	1.2	0	1.19	0.55	0.6	0.37
Acet8	0.606	0.49	0.54	0.71	0.49	0.18	0.53	0.74	0	0.54	0.85	1.02	0.29
EnhWk1	1.661	0.17	0.36	0.88	0.7	0.01	0.15	0.35	0	1.47	0.65	0.76	0.55
EnhWk2	0.38	1.82	1.89	1.53	1.63	0.73	1.7	2.05	0	1.29	0.33	0.4	0.62
EnhWk3	0.895	0.44	0.68	1.16	0.95	0.07	0.65	1.07	0	1.78	0.55	0.65	0.68
EnhWk4	2.39	0.17	0.35	1.05	0.63	0.01	0.2	0.42	0	2.18	0.46	0.63	0.55
EnhWk5	1.069	0.29	0.5	0.97	0.92	0.05	0.27	1.04	0	0.91	0.91	0.99	0.56
EnhWk6	0.721	0.47	0.78	1.25	1.37	0.11	0.39	1.36	0	0.85	1.83	0.88	0.59
EnhWk7	0.523	0.35	0.51	1.04	0.91	0.04	0.3	0.9	0	0.89	0.8	1	0.46
EnhWk8	1.476	0.18	0.33	0.87	0.58	0.01	0.21	0.32	0	1.57	0.54	0.8	0.39
EnhA1	0.193	1.28	1.76	1.57	1.95	0.65	3.47	3.07	0	2.18	0.23	0.31	0.69
EnhA2	0.354	0.43	0.87	1.22	1.11	0.11	1.84	1.6	0	3.19	0.5	0.41	0.75
EnhA3	0.209	0.38	0.81	1.5	1.3	0.01	1.84	2.07	0	2.88	0.39	0.42	0.77
EnhA4	0.324	0.32	0.64	1.06	0.97	0	0.83	0.84	0	2.22	0.57	0.57	0.5
EnhA5	0.77	0.27	0.53	1.01	1.07	0.01	0.39	0.57	0	2.07	0.62	0.65	0.52
EnhA6	0.607	0.7	0.93	1.51	1.1	0.13	0.87	1.22	0	1.75	0.41	0.62	0.85
EnhA7	0.423	2	1.71	1	1.2	0.55	1.79	2.3	0	0.87	0.87	0.67	0.47
EnhA8	0.274	1.46	1.66	1.35	1.63	0.51	4.61	5.15	0	1.18	0.52	0.63	0.66
EnhA9	0.175	1.16	1.66	1.73	1.95	0.36	2.6	3.96	0	1.35	0.22	0.48	0.83
EnhA10	0.425	0.39	0.83	1.11	1.3	0.18	1.31	4.27	0	1.39	0.72	0.76	0.86
EnhA11	0.77	0.38	0.62	0.91	1.05	0.04	0.52	1.21	0	0.76	0.93	1.06	0.68
EnhA12	0.358	0.94	1	0.85	0.84	0.2	1.63	1.49	0	1.1	0.73	0.71	0.29
EnhA13	0.824	0.27	0.39	0.81	0.65	0.01	0.41	0.67	0	1	0.73	0.9	0.26
EnhA14	0.395	1.02	1.13	1.07	1.19	0.18	2.64	2.81	0	1.09	0.71	0.73	0.32
EnhA15	1.097	0.4	0.56	0.95	0.99	0.03	0.66	1.32	0	0.98	0.91	0.95	0.4
EnhA16	0.695	0.39	0.55	0.84	0.94	0.38	0.83	1.95	0	0.94	1.26	0.97	0.38
EnhA17	0.567	0.28	0.61	1.27	0.91	0.04	0.77	1.07	0	2.98	0.42	0.51	0.88
EnhA18	0.497	0.43	0.5	0.78	0.58	0.23	0.79	0.78	1.25	1.16	0.67	0.96	0.45
EnhA19	0.277	0.99	1.19	1.24	0.84	0.39	2.08	1.85	0	1.84	0.42	0.57	1.05
EnhA20	0.373	0.26	0.45	1.01	0.9	0.01	0.17	0.47	0	1.44	0.66	0.71	0.72
TxEnh1	0.422	0.32	0.56	2.15	0.64	0	0.34	1.11	0	1.53	0.24	0.62	1.07
TxEnh2	0.422	0.34	0.41	2.16	0.46	0.01	0.41	2.55	0	1.01	0.26	0.8	1.82
TxEnh3	0.269	1.45	1.06	2.16	0.6	0.51	0.63	1.18	0	0.78	0.65	0.94	2.94
TxEnh4	0.288	13.3	9.69	2.11	5.69	2.4	1.22	1.51	0	3.16	0.42	0.34	1.37
TxEnh5	0.537	8.78	6.66	2.04	3.52	2.57	0.96	1.09	0	2.1	0.52	0.47	1.33
TxEnh6	0.204	5.89	5.54	2.07	5.41	0.65	0.87	1.38	0	1.69	0.42	0.58	1.06
TxEnh7	0.291	0.95	0.92	2.14	0.68	0.38	1.01	1.5	0	1.03	0.19	0.52	1.2
TxEnh8	0.262	4.52	4.59	2.08	3.95	0.2	0.76	0.75	0	1.98	0.38	0.53	1.14
TxWk1	3.019	1.3	2.12	2.01	2.87	0.08	0.17	0.33	0	1.07	0.81	1.04	1.6
TxWk2	0.909	6.2	4.46	1.97	2.32	1.46	0.66	0.66	0	1.55	0.82	0.72	1.19
Tx1	0.888	0.18	0.2	2.16	0.27	0.01	0.16	0.63	0	0.78	0.36	1.08	2.09
Tx2	1.704	0.17	0.23	2.13	0.33	0.04	0.14	0.45	0	0.78	0.55	1.09	1.62
Tx3	0.548	0.28	0.33	2.12	0.27	0.2	0.36	0.59	0	0.53	0.45	0.97	1.53
Tx4	0.507	2.48	2	2.16	1.3	0.04	0.39	0.61	0	1.24	0.57	0.94	2.52
Tx5	1.016	2.7	2.66	2.07	2.46	0.45	0.21	0.41	0	0.89	0.9	1.25	2.02
Tx6	1.196	3.37	4.08	2.12	4.56	0	0.14	0.23	0	1.88	0.36	0.65	1.6
Tx7	0.884	6.37	5.41	2.14	4.58	0	0.27	0.34	0	2.31	0.33	0.63	1.78
Tx8	0.735	2.35	3.26	2.06	3.72	0.02	0.32	0.53	0	1.63	0.45	0.69	1.2
TxEx1	0.286	9.69	6.01	2.17	3.63	0.54	1.2	1.22	0	2.73	0.45	0.68	2.43
TxEx2	0.6	14.8	9.27	2.17	3.79	1.06	0.68	0.71	0	3.45	0.52	0.51	2.29
TxEx3	0.716	9.11	6.54	2.1	3.95	2.5	0.56	0.73	0	2.07	0.82	0.77	2.43
TxEx4	0.106	10	7.28	2.12	8.75	1.81	3.83	7.55	0	2.88	0.31	0.39	2.02
znf1	0.438	0.99	1.14	1.99	1.31	0.28	0.22	0.37	0	0.53	1.26	1.54	19.4
znf2	0.164	9.69	6.63	2	3.04	0.92	0.85	0.78	0	0.97	1.77	1.12	64
DNase1	0.216	0.93	1.41	1	2.28	3.89	5.19	1.75	0	1.87	1.04	0.48	0.75
BivProm1	0.156	8.78	7.89	1.33	2.36	70.4	26.7	17.9	0	3.45	0.96	0.16	0.81
BivProm2	0.172	7.04	5.92	1.23	3.72	49.6	13.7	13.8	0	3.24	0.97	0.17	0.46
BivProm3	0.308	2.76	2.6	1.08	2.06	6.76	5.73	6.52	0	2.26	0.75	0.39	0.77
BivProm4	0.14	2.38	3.15	1.62	2.78	3.42	7.52	13.5	0	3.66	0.53	0.29	1.2
PromF1	0.217	2.02	2.24	0.94	2.17	2.43	6.07	6.02	0	1.69	0.6	0.42	0.76
PromF2	0.149	2.24	2.99	1.29	4.09	3.49	7.58	12.9	0	2.15	0.44	0.37	1.15
PromF3	0.165	4.33	4.73	1.42	4.15	23.4	12.4	23.9	0	1.98	0.48	0.3	3.17
PromF4	0.205	9.54	12.5	1.76	2.34	88.3	57.9	26.1	0	3.64	0.52	0.12	4.46
PromF5	0.146	6.85	7.89	1.45	2.2	65.1	29.6	20.8	0	3.48	0.65	0.15	2.12
PromF6	0.139	3.45	3.02	1.96	3.41	1.45	2.56	24.6	0	1.81	0.55	0.45	4.05
PromF7	0.173	1.61	1.89	1.95	2.3	0.08	0.98	18.9	0	1.41	0.66	0.64	4.12
TSS1	0.132	6.08	11.7	1.34	2.82	93.9	141	24.2	0	4.01	0.88	0.24	3.51
TSS2	0.122	5.19	7.58	1.07	2.54	74.2	98.8	18.7	0	3.46	1.57	0.43	2
	100	1.23	2.64	45	0	0.75	0	3.67	0.56	5.42	2.03	50.7	0.78

													Avg. acros
Corr. Enr. Hg19 &hg38		1	1	1	1	1	1	1	0.98	0.99	0.93	0.97	1	0.99

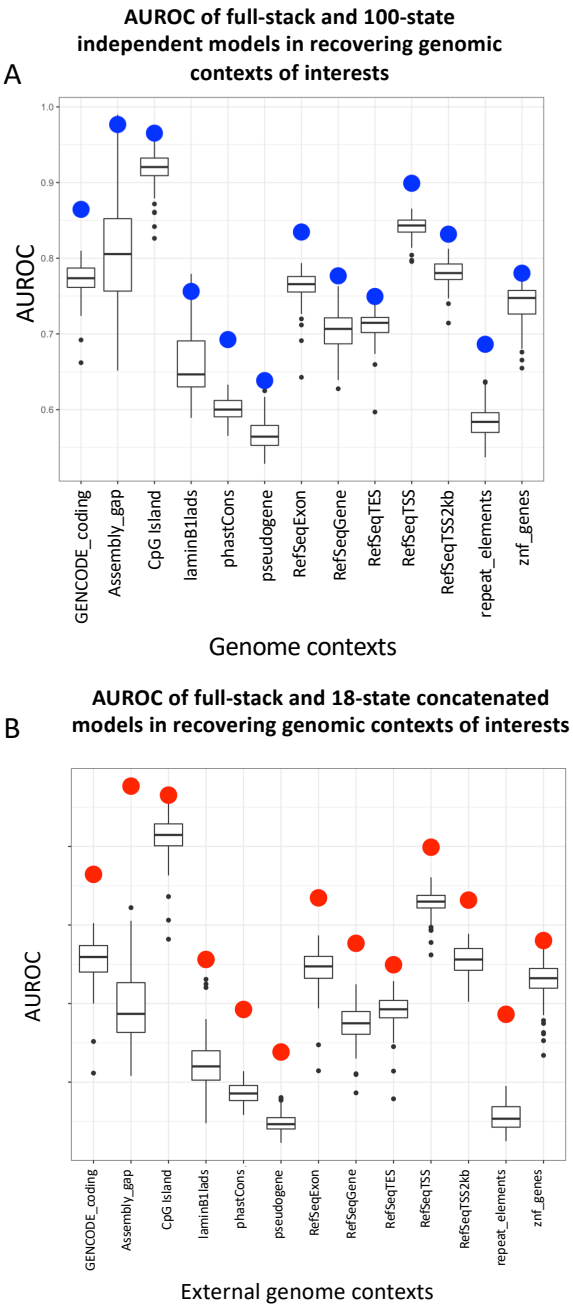
Supplementary Figure 2. 18: Comparison of hg19 and hg38 full-stack states enrichments with annotated genomic contexts.

The heatmaps show enrichment values for the full-stack states (rows) and different external genome annotations from **Figure 2.3A** (columns) in hg19 (A) and hg38 (B) (**Methods**). Panel **(A)** is similar to **Figure 2.3A**, but we present it here for better comparison with the hg38 enrichment heatmap. Results in **(B)** are based on (1) lifting over the full-stack annotation from hg19 to hg38 (**Methods**), and (2) doing enrichment analysis with annotated genome contexts derived from various databases in hg38 (**Methods**). In each heatmap, coloring of enrichments is column specific with highest and lowest enrichment values in each column are colored red and white, respectively. The first two columns of each heatmap show state labels and their percentage of genome coverage. The last row of each heatmap shows the percentage of genome coverage for each type of genome contexts. Below the heatmap in **(B)** is the correlation of the enrichments across states based on hg19 and hg38 for each corresponding annotation column as well as the average of them.


A
clean_assembly_gap_hg19
full_auc = 0.9767775
B
CpGIsland
full_auc = 0.9652788
C
laminB1lads
full_auc = 0.7561418
D
phastConsElements100way
full_auc = 0.6925264
E
pseudogene_v30lift37
full_auc = 0.6383499
F
RefSeqExon
full_auc = 0.8345721
G
RefSeqGene
full_auc = 0.7767094
H
RefSeqTES
full_auc = 0.7494425
I
RefSeqTSS
full_auc = 0.8990299
J
RefSeqTSS2kb
full_auc = 0.8317081
K
znf_genes_no_overlap
full_auc = 0.7802264
L
rmsk
full_auc = 0.6863251
M
CDS_hg19_no_overlap
full_auc = 0.8644886
full_true_pos
full_false_pos

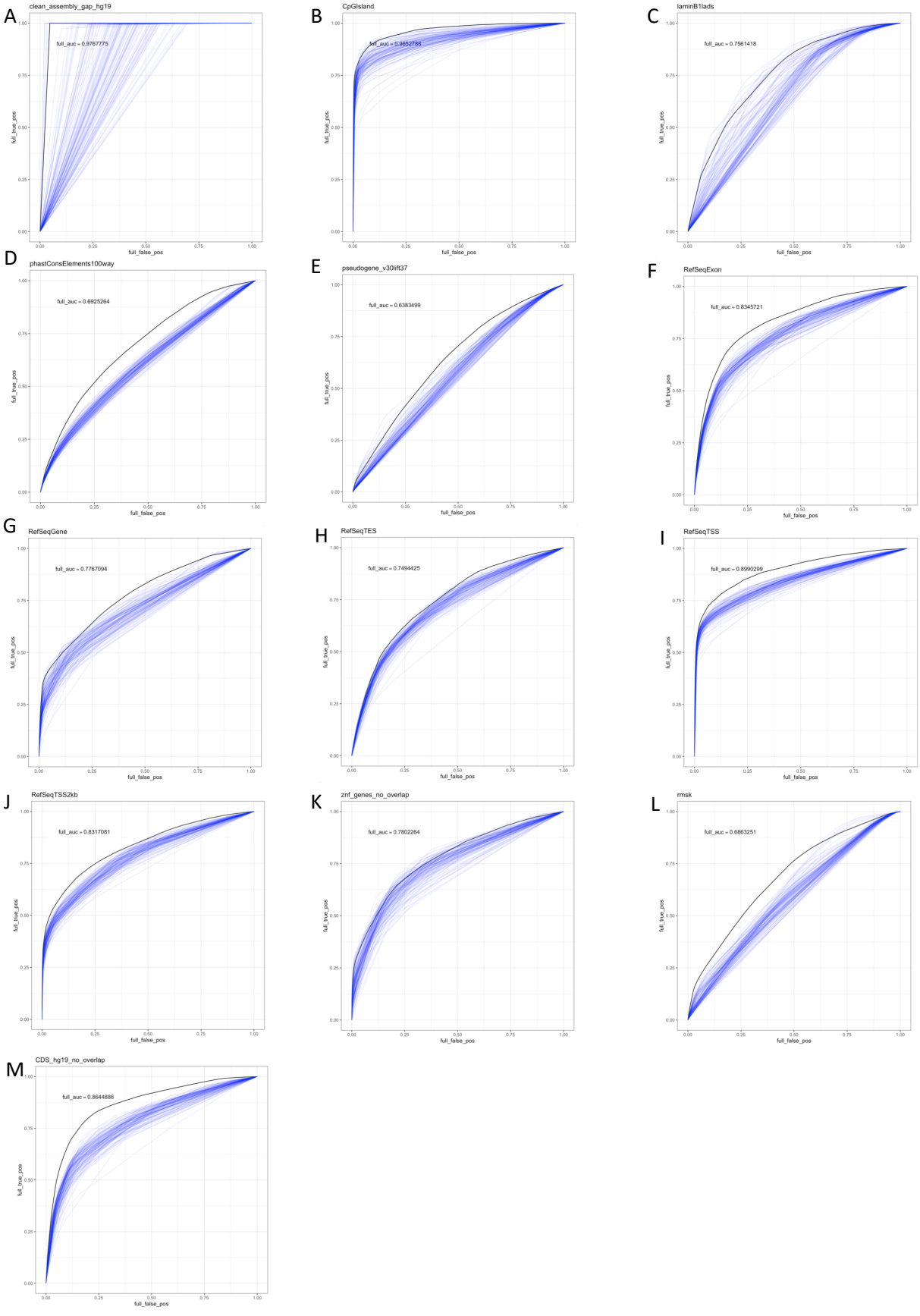
Supplementary Figure 2. 19: ROC comparison of full-stack model annotations and the 18-state concatenated model annotations for predicting various external genomic annotations.

Each panel shows the ROC curves from using the full-stack model annotations and 98 chromatin state annotations from a concatenated model to predict different external genomic annotations (**Methods**). The concatenated annotations are from a previously learned 18-state concatenated model *(Meuleman* et al*., 2015)*. The full-stack annotations' ROC curves are in black, and 98 concatenated annotations' ROCs are in red. The respective genomic contexts for panels A-M are assembly gaps, CpG Islands, lamina associated domains (laminB1lads), phastCons elements, pseudogenes, exons, gene bodies, transcription end sites (TES), transcription start sites (TSS), 2kb regions surrounding transcription start sites (TSS2kb), ZNF genes, repeat elements in UCSC Genome Browser's repeatMasker track and coding sequences.

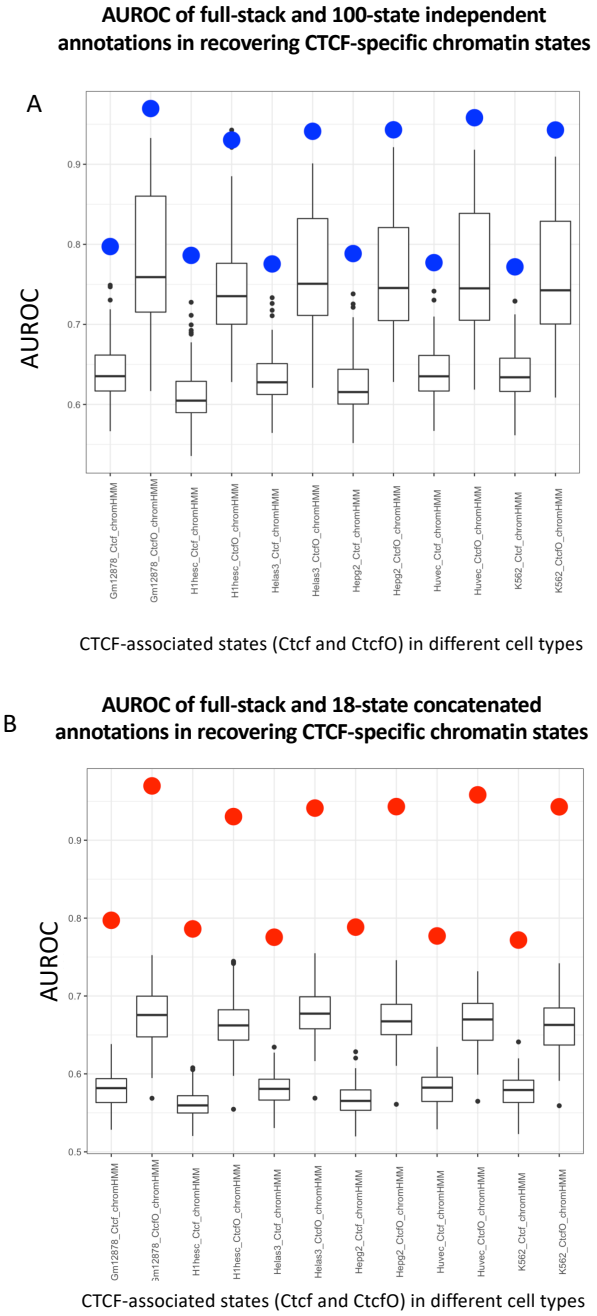


Supplementary Figure 2. 20: AUROC comparison of the full- stacked and the concatenated and independent chromatin state annotations at predicting various external genomics annotations.

(A) AUROC values for ROC curves in **Supplementary Figure 2.19**. The x-axis represents different genomic contexts. The box-plots show AUROC of the 127 100-state annotations based on models learned independently in 127 cell types at predicting locations of the external annotations. The blue dots show the AUROC for the full-stack chromatin state annotations. **(B)** Similar to **(A),** but showing AUROC values for ROC curves in **Supplementary Figure 2.21**. but the boxplots show the AUROC values for 98 18-state annotations based on concatenated models in 98 cell types. The red dots show the AUROC for the full-stack chromatin state annotations.


A
clean_assembly_gap_hg19
full_auc = 0.9767775
B
CpGIsland
full_auc = 0.9652788
C
laminB1lads
full_auc = 0.7561418
D
phastConsElements100way
full_auc = 0.6925264
E
pseudogene_v30lift37
full_auc = 0.6383499
F
RefSeqExon
full_auc = 0.8345721
G
RefSeqGene
full_auc = 0.7767094
H
RefSeqTES
full_auc = 0.7494425
I
RefSeqTSS
full_auc = 0.8990299
J
RefSeqTSS2kb
full_auc = 0.8317081
K
znf_genes_no_overlap
full_auc = 0.7802264
L
rmsk
full_auc = 0.6863251
M
CDS_hg19_no_overlap
full_auc = 0.8644886
full_true_pos
full_false_pos

Supplementary Figure 2. 21: ROC comparison of full-stack model annotations and the 100-state independent model annotations for predicting various external genomic annotations. Each panel shows the ROC curves from using the full-stack model annotations and the 127 independent model chromatin state annotations to predict different external genomic annotations (**Methods**). The independent models were 100 state models learned separately using all available data from each cell type. The full-stack annotations' ROC curves are in black, and independent annotations' ROCs are in blue. The respective genomic contexts for panels A-M are assembly gaps, CpG Islands, lamina associated domains (laminB1lads), PhastCons elements, pseudogenes, exons, gene bodies, transcription end sites (TES), transcription start sites (TSS), 2kb regions surrounding transcription start sites (TSS2kb), ZNF genes, repeat elements from all classes and families in UCSC Genome Browser's repeatMasker track and coding sequences.

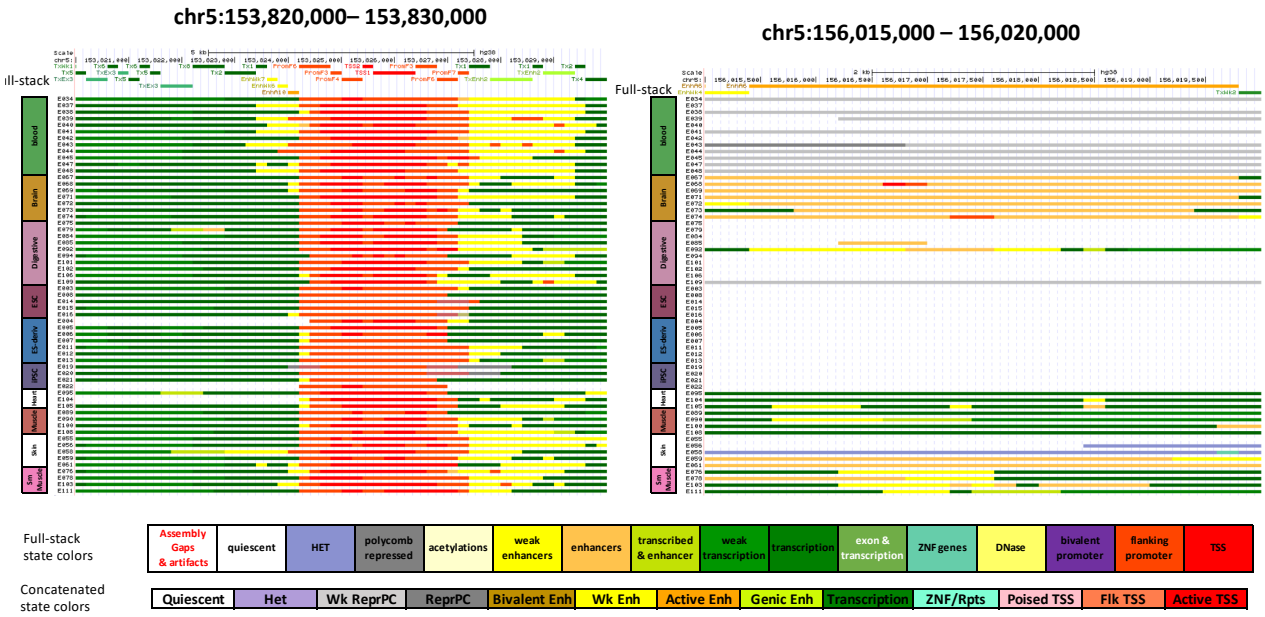


Supplementary Figure 2. 22: AUROC comparison of the full-stack and concatenated and independent chromatin state annotations at predicting CTCF-specific chromatin states. Box-plots show AUROC of **(A)** 127 100-state independent and **(B)** 98 18-state concatenated model annotations, which did not include CTCF, at predicting bases in sets of CTCF-associated chromatin states. In both panels, the x-axis represents sets of chromatin states associated with CTCF signal and limited histone modification signal in one of six cell types from a previously published chromatin state model that included CTCF *(Hoffman* et al.*, 2013)* (**Methods**). CtcfO corresponds to a state that also had open chromatin signals, while state Ctcf lacked those signals. The dots colored **(A)** blue and **(B)** red show the AUROC for the full-stack chromatin state annotations, which were not trained using CTCF signals data.

Group	DNase	H3K27ac	H3K4me1	H3K4me2	H3K4me3	H3K9ac
weak enhancers	1.2	1.4	1.3	1.0	1.6	1.2
enhancers	1.0	1.1	1.0	0.9	1.4	1.0
TSS	0.1	0.3	0.7	0.2	0.1	0.4
promoters	0.3	0.3	0.4	0.2	0.2	0.3
bivalent promoters	0.5	0.6	0.4	0.4	0.5	0.5

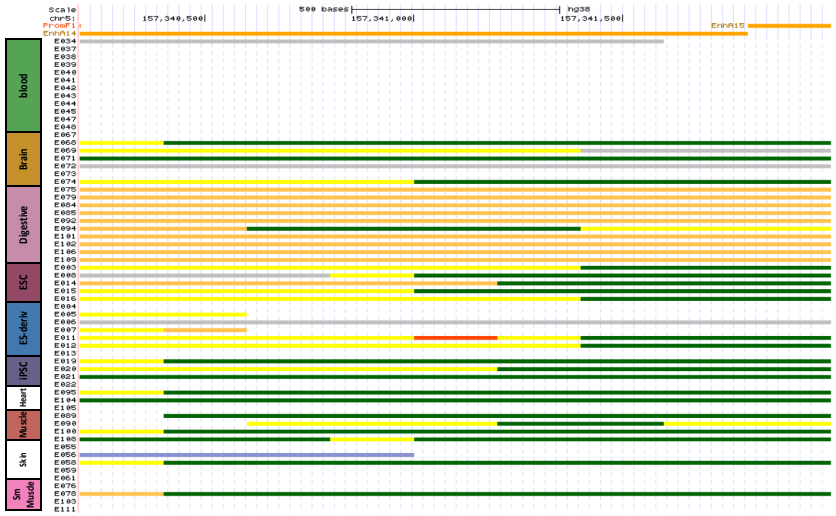
Supplementary Figure 2. 23: Coefficient of variations of emission probabilities across different cell groups.

Average coefficient of variations for the five enhancer and promoter state groups of full-stack states (rows) and six chromatin marks that are associated with enhancer and promoter activities. For a mark and state group combination, the coefficient of variation for the mark emission was computed separately for each state and then averaged among states in the group. The enhancer and weak enhancer group showed greater than two-fold higher coefficient of variations compared to the promoter group.

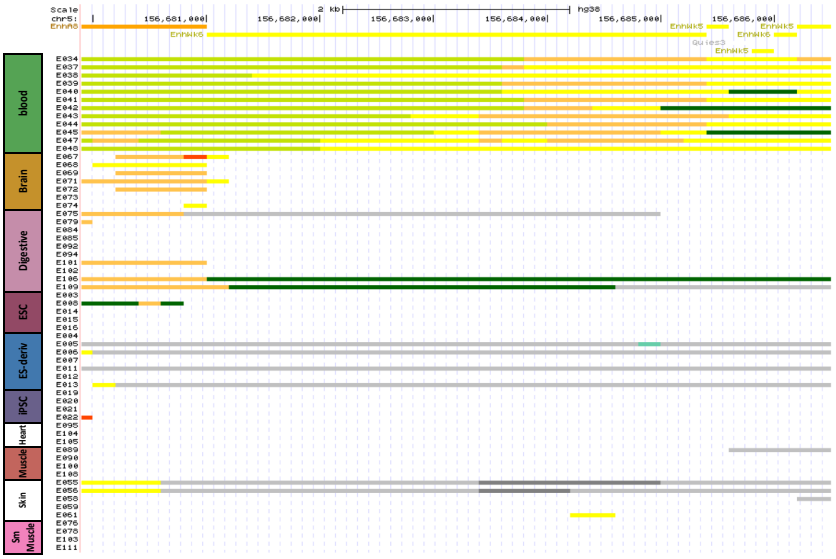


Supplementary Figure 2. 24: Illustration of the full-stack annotations at two distinct loci. Two loci representing regions that are in transcribed and active promoter states across cell types (left), and in an enhancer state specifically in brain (right). The loci correspond to those presented in **Figure 2.1**. The top track shows the full-stack state annotations. The following tracks show concatenated annotations from 18-state models based on observed data *(Meuleman* et al.*, 2015)*. The cell types are ordered based on their associated cell groups. A color legend for the states is shown along the bottom.

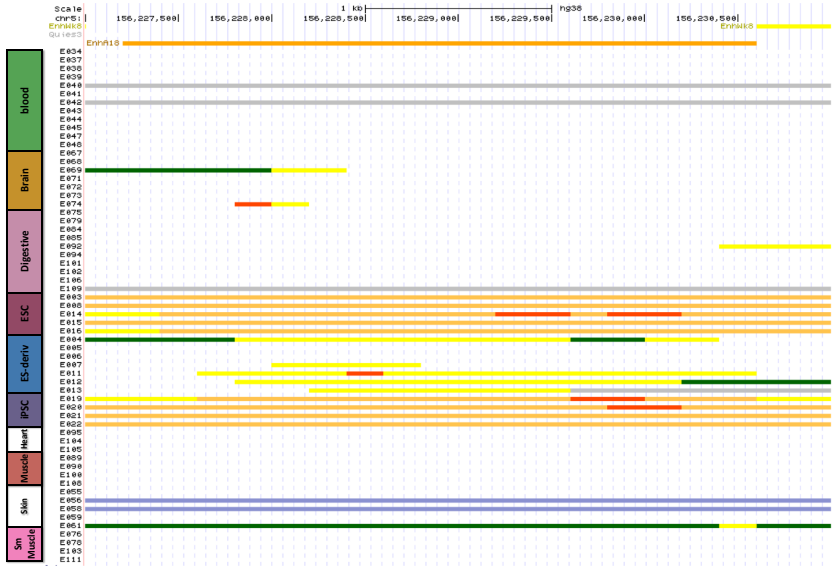
A



B



C



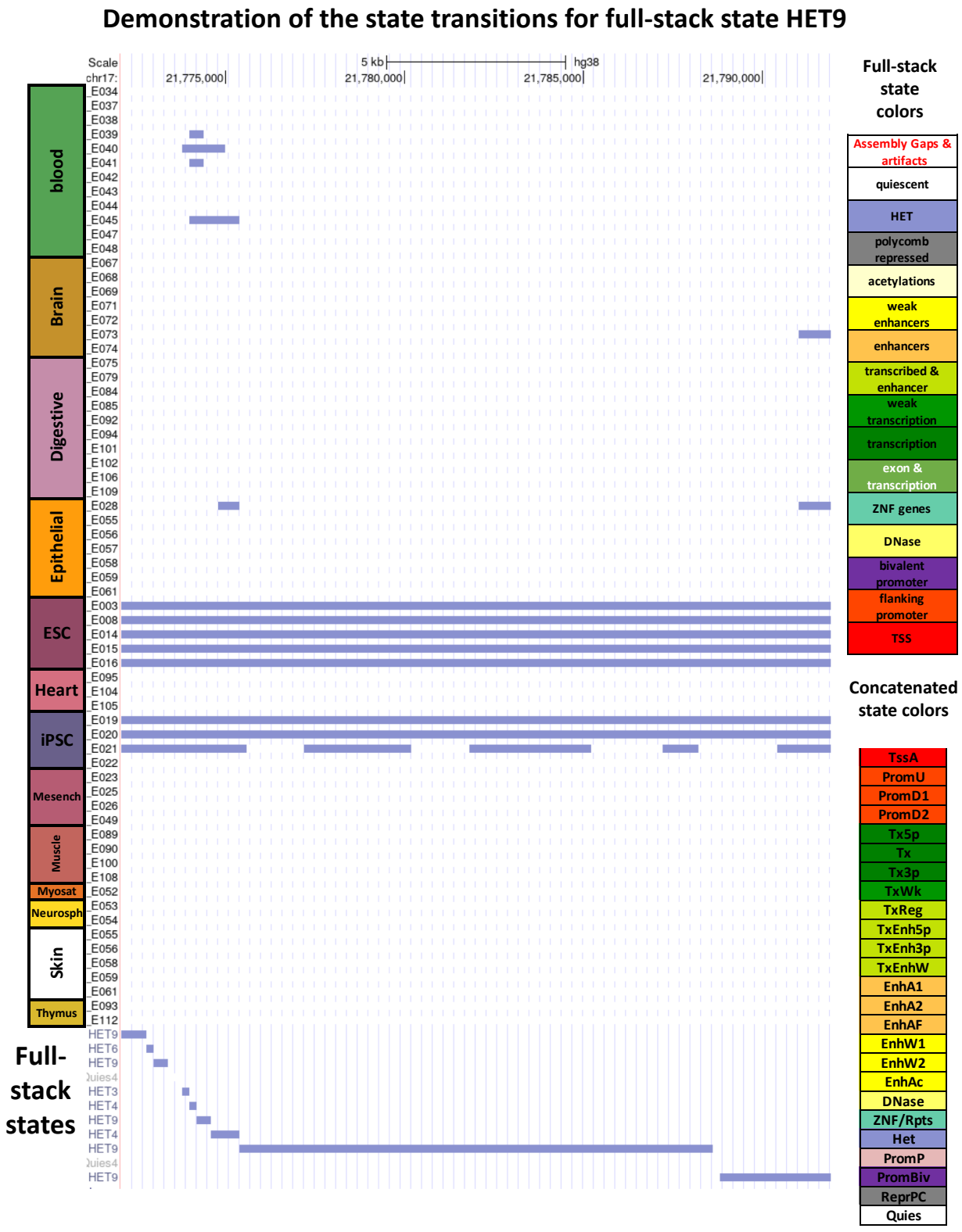
Full-stack state colors

Assembly Gaps & artifacts
quiescent
HET
polycomb repressed
acetylations
weak enhancers
enhancers
transcribed & enhancer
weak transcription
transcription
exon & transcription
ZNF genes
DNase
bivalent promoter
flanking promoter
TSS

Concatenated state colors

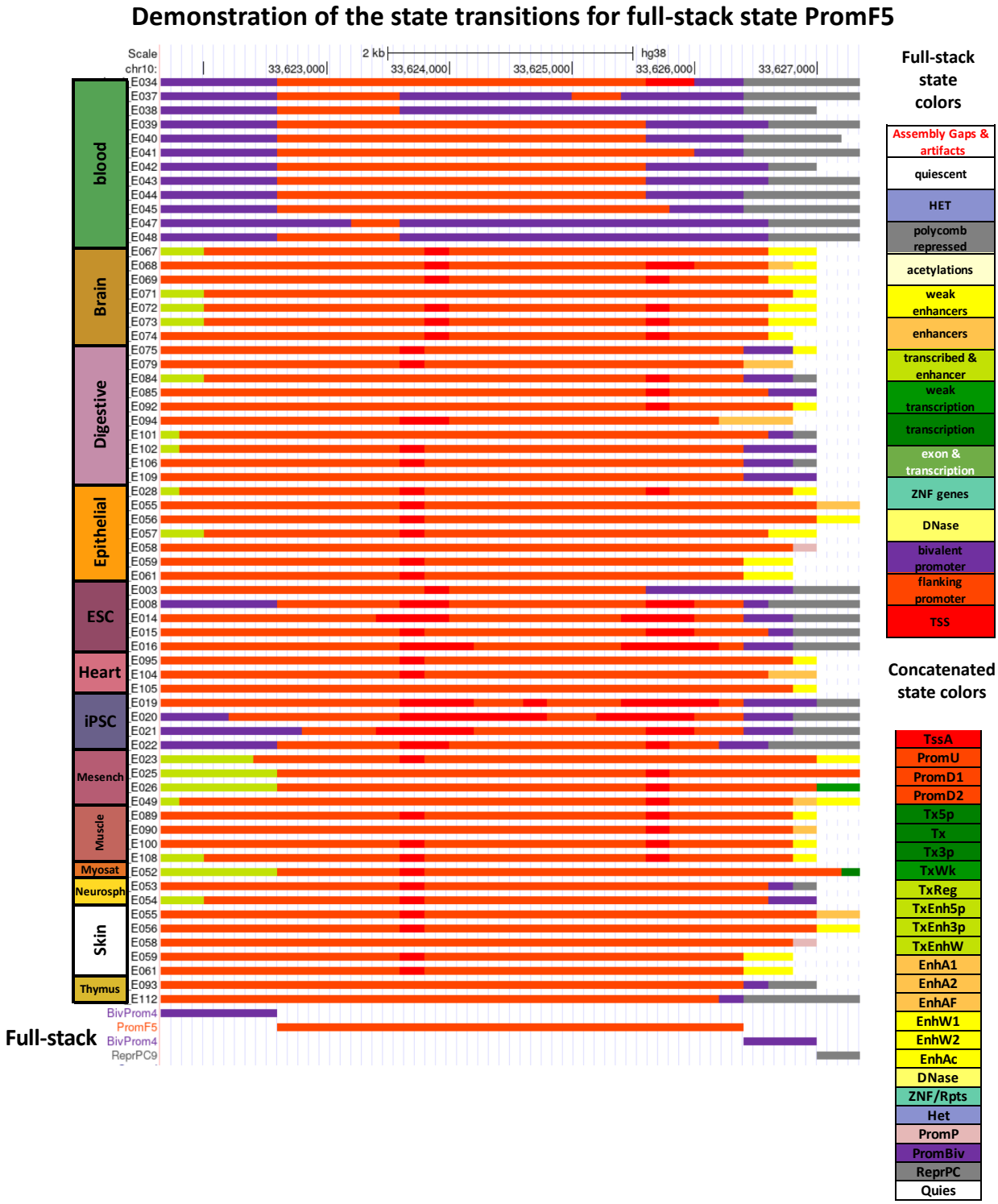
Quiescent
Het
Wk ReprPC
ReprPC
Bivalent Enh
Wk Enh
Active Enh
Genic Enh
Transcription
ZNF/Rpts
Poised TSS
Flk TSS
Active TSS

Supplementary Figure 2. 25: Illustration of full-stack cell-type-specific enhancer states. **(A-C)** The first track in each panel demonstrates the full-stack state annotation. Each of the following tracks show chromatin state annotations from a 18-state concatenated model *(Meuleman* et al.*, 2015)*. The individual reference epigenomes IDs and their tissue groups are labeled on left. The chromatin state coloring is labeled on right. **(A)** A genomic region (chr5:157340200-157342000) annotated to an active enhancer state in digestive cells in the full-stack model (EnhA14). **(B)** A genomic region (chr5:156679900-156686500) annotated to blood enhancer states in the full-stack model (EnhWk6 and EnhA8). **(C)** A genomic region (chr5:156227000-156231000) annotated as an ESC/iPSC-specific enhancer state in the full-stack model (state EnhA18).



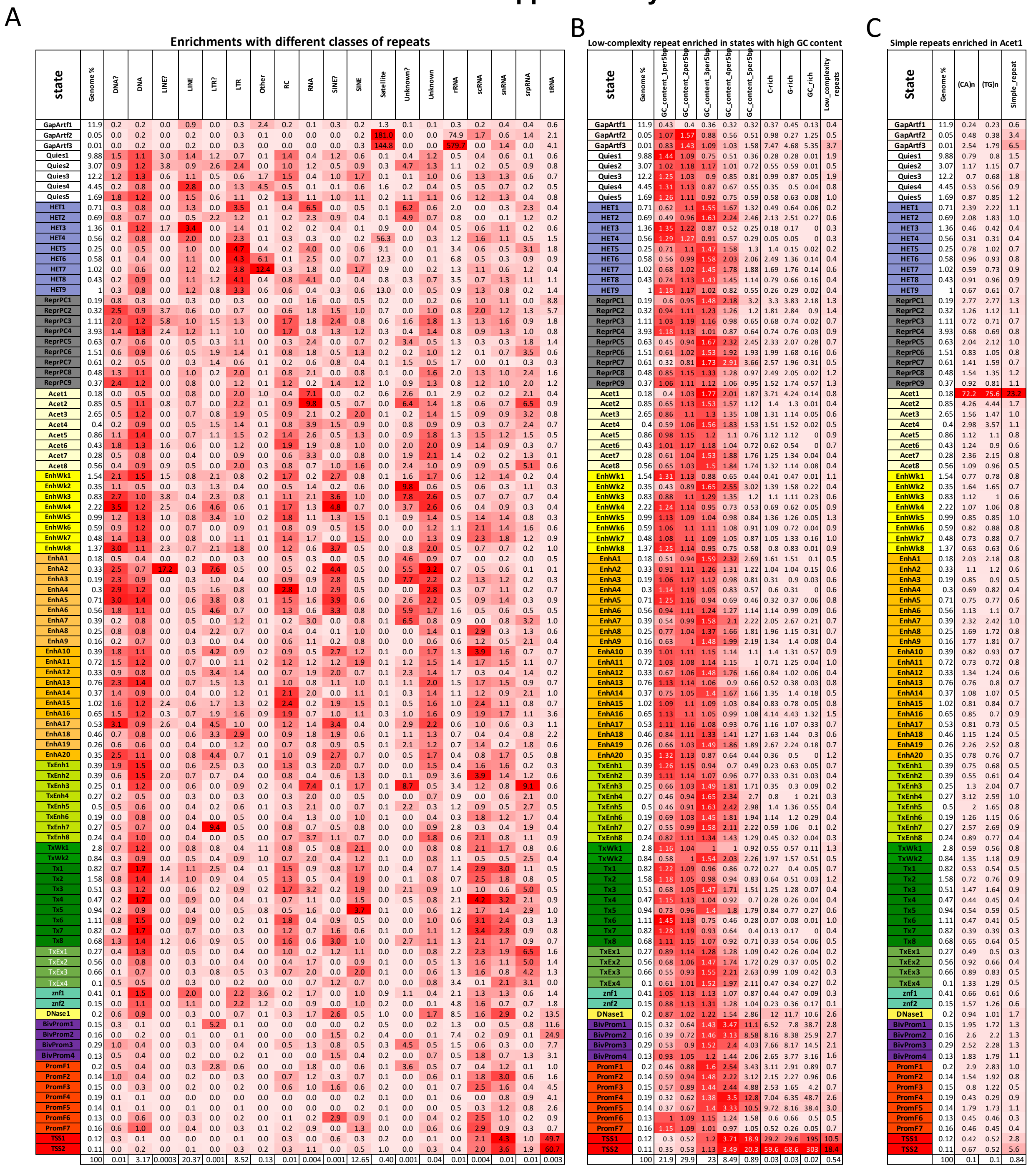
Supplementary Figure 2. 26: Illustration of full-stack heterochromatin state HET9. The figure captures the concatenated chromatin state maps for various reference epigenomes, and the corresponding full-stack chromatin state maps at region chr17:21772099-21791900. The first 66 tracks show chromatin state annotations from a 25-state concatenated model for 66

reference epigenomes (equivalently, in this paper, cell types) (Ernst and Kellis, 2015). The individual reference epigenomes IDs and their tissue groups are labeled on the left. The chromatin state colors are explained on the right. The last track, shown in full mode to display all state labels on the right, corresponds to the full-stack chromatin state map at this region. State HET9 is characterized, based on our analysis, as an ESC-group-related heterochromatin state (**Figure 2.2C, Supplementary Figure 2.8-9, Supplementary Data 2. 3, 5**). Detailed characterizations of all full-stack states are in **Supplementary Data 2. 3**.



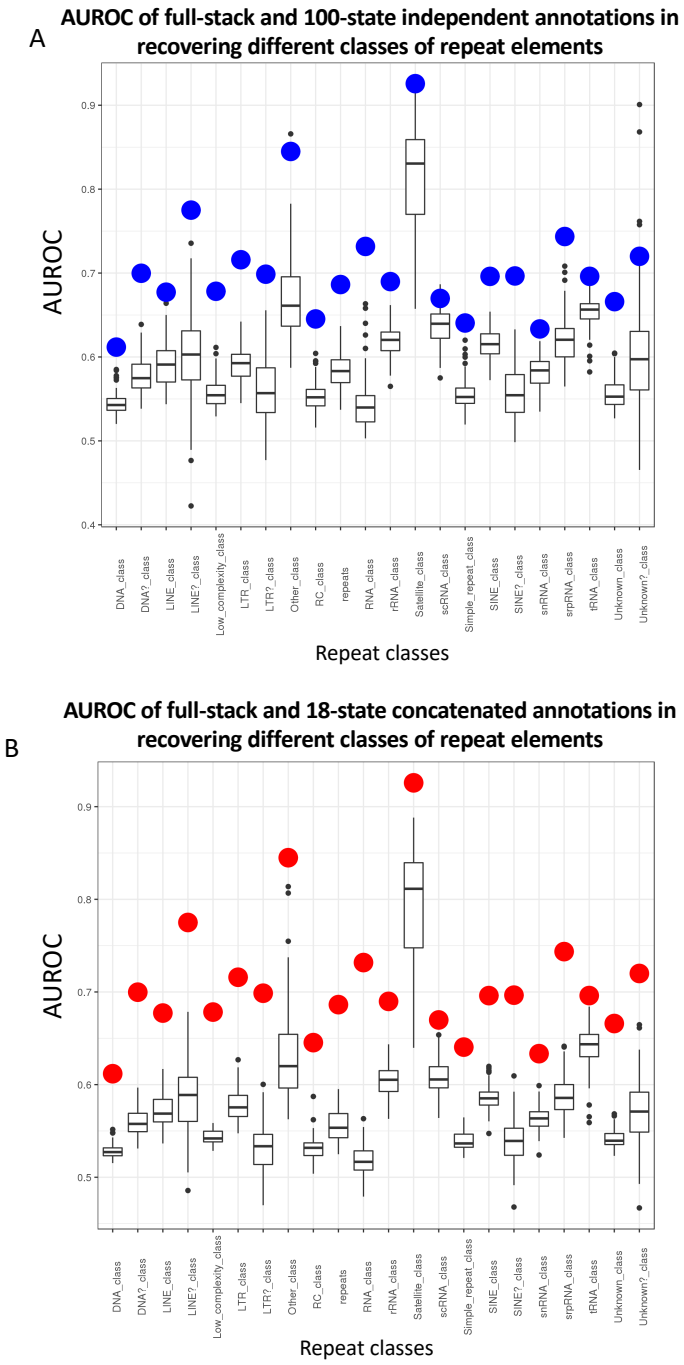
Supplementary Figure 2. 27: Illustration of full-stack flanking promoter state PromF5. The figure captures the concatenated chromatin state maps for various reference epigenomes, and the corresponding full-stack chromatin state maps at region chr10:33621649-33627350. The first 66 tracks show chromatin state annotations from a 25-state concatenated model for 66 reference epigenomes (equivalently, in this paper, cell types) (Ernst and Kellis, 2015). The individual reference epigenomes IDs and their tissue groups are labeled on the left. The chromatin

state coloring is labeled on the right. The last track, shown in full mode to display all state labels on the left, corresponds to the full-stack chromatin state map at this region. State PromF5 is characterized, based on our various analyses as state frequently found at flanking promoter regions with some upstream bias, and sometimes, this state overlaps with regions of bivalent promoters in Blood-related and ESC-related groups (Blood & T cells, HSC & B cells, ESC, iPSC and ES-deriv) (**Supplementary Figure 2. 8-9, Supplementary Data 2. 3, 5**). Detailed characterization of all full-stack states are in **Supplementary Data 2. 3-5**.



Supplementary Figure 2. 28: Full stack states enrichments with RepeatMasker classes of repeats (A), low-complexity repeats and GC content (B), simple repeats (C).

This is an extended version of **Figure 2.4C**. In each panel, the rows correspond to full-stack states. The second column reports the percentage of the genome that each full-stack state occupies. In **(A)**, columns 3-21 correspond to different repeat classes. In **(B)**, columns 3-7 correspond to 5bp windows in the genome are stratified by the number of G/C bases in them, columns 8-10 correspond to regions enriched with C-rich, G-rich, and GC-rich low complexity sequences, respectively, and column 11 shows enrichments for all low complexity sequences from RepeatMasker. States TSS1-2 are most enriched with Low complexity repeat class, which is consistent with these states having a high enrichment (19-20 fold) for windows in which all bases are a G or C. In **(C)**, columns 3-4 correspond to simple repeats of repeated (CA) and (TG) sequences, and column 5 shows enrichments for all simple repeats. State Acet1 is most enriched with simple repeats and this enrichment is mostly driven by enrichments with repeated CA and TG dinucleotides. In each panel, the values in all columns except the first and second columns correspond to fold enrichment for different repeat contexts in the full-stack states. Values are colored on a column-specific color scale. The last row gives the percentage of the genome that each repeat class occupies.

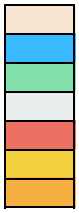


Supplementary Figure 2. 29: AUROC comparison of the full-stack, concatenated and independent chromatin state annotations at predicting different classes of repeat elements.

Box-plots showing AUROC of the **(A)** 127 100-state annotations from independent models and **(B)** 98 18-state annotations from a concatenated model at predicting bases in different repeat classes labeled on the x-axis. The dots colored **(A)** blue and **(B)** red show the AUROC for the full-stack chromatin state annotations.

A

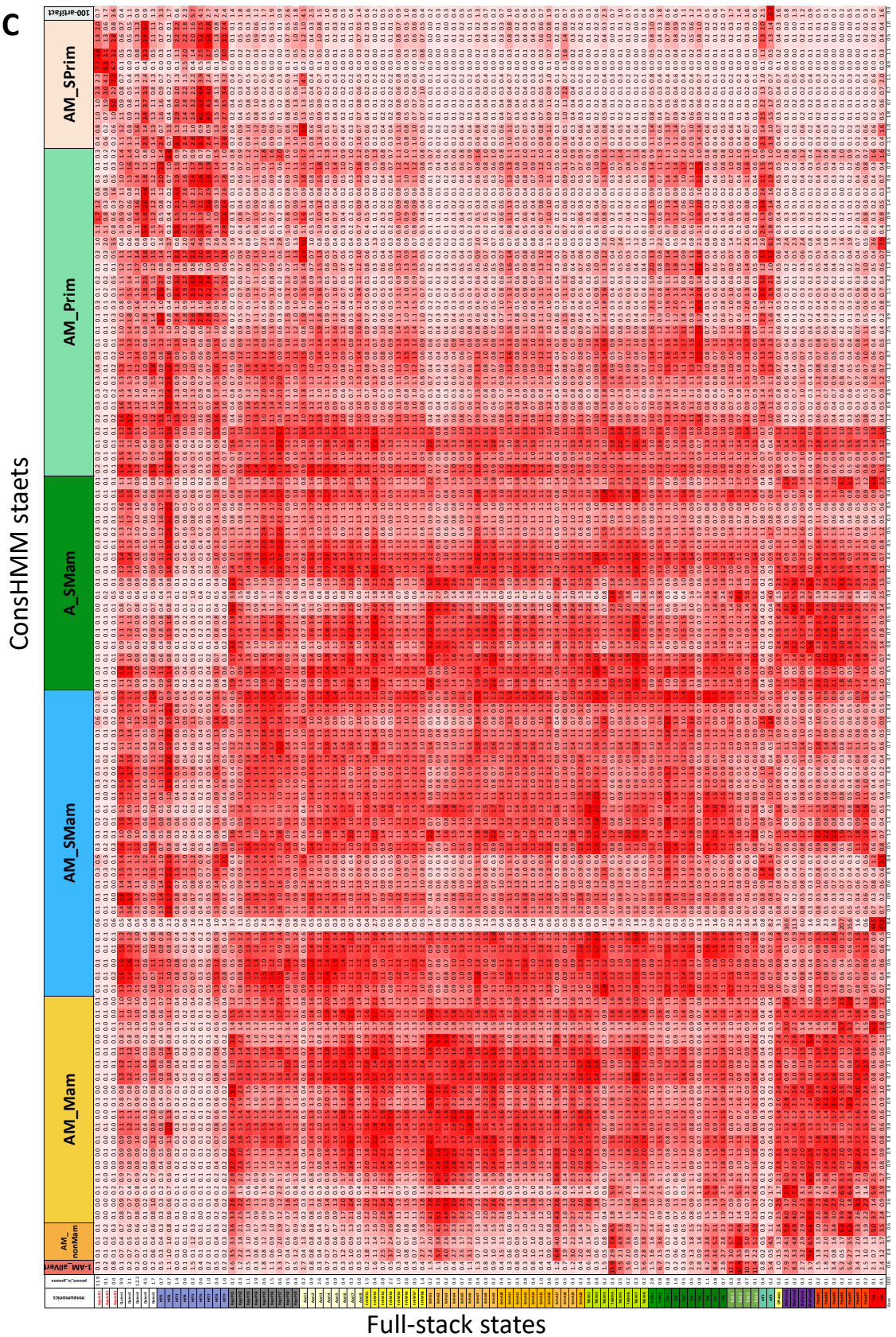
state	max_enrich	consHMM_state
GapArtf1	7.7	96-AM_SPrim
GapArtf2	8.1	96-AM_SPrim
GapArtf3	7.2	95-AM_SPrim
Quies1	1.8	30-AM_SMam
Quies2	1.8	76-AM_Prim
Quies3	1.7	80-AM_Prim
Quies4	4.6	86-AM_Prim
Quies5	1.8	72-AM_Prim
HET1	3.4	76-AM_Prim
HET2	3.7	82-AM_Prim
HET3	4.4	86-AM_Prim
HET4	2.8	93-AM_SPrim
HET5	5.7	100-artifact
HET6	5.3	93-AM_SPrim
HET7	4.3	93-AM_SPrim
HET8	2.6	82-AM_Prim
HET9	3.5	93-AM_SPrim
ReprPC1	4.6	1-AM_allVert
ReprPC2	2.5	5-AM_Mam
ReprPC3	1.6	100-artifact
ReprPC4	1.5	69-AM_Prim
ReprPC5	2.7	82-AM_Prim
ReprPC6	1.9	100-artifact
ReprPC7	4.6	28-AM_SMam
ReprPC8	2.6	100-artifact
ReprPC9	1.8	100-artifact
Acet1	9.0	82-AM_Prim
Acet2	2.5	100-artifact
Acet3	1.9	75-AM_Prim
Acet4	2.0	14-AM_Mam
Acet5	1.6	75-AM_Prim
Acet6	2.1	9-AM_Mam
Acet7	2.6	14-AM_Mam
Acet8	1.6	75-AM_Prim
EnhWk1	2.2	5-AM_Mam
EnhWk2	2.7	28-AM_SMam
EnhWk3	2.8	5-AM_Mam
EnhWk4	3.7	2-AM_nonMam
EnhWk5	1.4	75-AM_Prim
EnhWk6	1.6	13-AM_Mam
EnhWk7	1.5	77-AM_Prim
EnhWk8	2.7	2-AM_nonMam
EnhA1	4.1	5-AM_Mam
EnhA2	5.7	2-AM_nonMam
EnhA3	4.8	5-AM_Mam
EnhA4	3.7	5-AM_Mam
EnhA5	3.2	5-AM_Mam
EnhA6	3.1	2-AM_nonMam
EnhA7	2.0	28-AM_SMam
EnhA8	2.3	15-AM_Mam
EnhA9	2.8	14-AM_Mam
EnhA10	2.2	6-AM_Mam
EnhA11	1.7	77-AM_Prim
EnhA12	2.1	14-AM_Mam
EnhA13	1.6	9-AM_Mam
EnhA14	2.2	14-AM_Mam
EnhA15	1.5	6-AM_Mam
EnhA16	1.9	100-artifact
EnhA17	6.4	2-AM_nonMam
EnhA18	2.2	94-AM_SPrim
EnhA19	3.0	2-AM_nonMam
EnhA20	2.1	5-AM_Mam
TxEnh1	2.6	6-AM_Mam
TxEnh2	2.2	37-AM_SMam
TxEnh3	2.3	82-AM_Prim
TxEnh4	15.9	1-AM_allVert
TxEnh5	9.5	1-AM_allVert
TxEnh6	5.6	1-AM_allVert
TxEnh7	2.4	47-A_SMam
TxEnh8	5.6	1-AM_allVert
TxWk1	2.4	7-AM_Mam
TxWk2	6.7	1-AM_allVert
Tx1	1.9	44-AM_SMam
Tx2	2.0	77-AM_Prim
Tx3	2.1	75-AM_Prim
Tx4	2.9	1-AM_allVert
Tx5	4.0	77-AM_Prim
Tx6	5.9	7-AM_Mam
Tx7	8.4	1-AM_allVert
Tx8	2.8	3-AM_nonMam
TxEx1	12.1	1-AM_allVert
TxEx2	18.4	1-AM_allVert
TxEx3	10.1	1-AM_allVert
TxEx4	11.4	1-AM_allVert
znf1	3.0	79-AM_Prim
znf2	10.9	100-artifact
DNase1	3.1	28-AM_SMam
BivProm1	16.6	28-AM_SMam
BivProm2	11.3	28-AM_SMam
BivProm3	6.0	28-AM_SMam
BivProm4	7.8	2-AM_nonMam
PromF1	3.0	28-AM_SMam
PromF2	3.8	5-AM_Mam
PromF3	3.9	28-AM_SMam
PromF4	20.7	28-AM_SMam
PromF5	15.6	28-AM_SMam
PromF6	3.4	1-AM_allVert
PromF7	2.7	6-AM_Mam
TSS1	44.5	28-AM_SMam
TSS2	47.8	28-AM_SMam



High align and match frequencies for a few primates
High align and march for mammals, but missing notable subsets
High align and match frequencies for primates
Putative artifact
High align and match frequency for all vertebrates
High align and match for mammals
High align and match for mammals and some non-mammals

B

ConsHMM_state	Characterization based on Supp. Data. File 1 from Arneson & Ernst, 2019
1-AM_allVert	**Most conserved**. High align and match frequency with all vertebrates' genomes. Most enriched for CDS, UTRs, exons of protein coding genes with preference for 1st and 2nd codon position, canonical splice-site.
2-AM_nonMam	**Conserved enhancers**. Most enriched for TES of protein coding genes, enhancer-group chromatin states in ct-spec annotations
3-AM_nonMam	Enicheed for CDS and exons of protein coding genes with 3rd codon position preference; DHS in non-exons
5-AM_Mam	**Conserved enhancer and DNase** (top2 most enriched with enhancer-related and DNase states in ct-spec annotations)
14-AM_Mam	Most enriched state for TxEnh5' and TxEnhW chromatin states in ct-spec annotations
28-AM_SMam	Strongest signals of overlapping **TSS and promoter** states in ct-spec annotations, CpG islands, Low Complexity class and family Repeats, DHS in non-exons
82-AM_Prim	**Moderate conservation** (align and match with primates); **ReprPC** state in ct-spec annotations
75-AM_Prim	Most enriched state for **RNA, scRNA, snRNA and srpRNA class and family repeats**
76-AM_Prim	Most enriched state for **LTR class and ERVL-MaLR, PiggyBac, and TcMar-Mariner family repeats**
77-AM_Prim	Most enriched state for **SINE class, and Alu family repeats** and Tx5' and TxWk chromatin states
86-AM_Prim	Most enriched state for **LINE class, L1 family repeats**, and **Quies** chromatin state in ct-spec annotations
100-artifact	**Alignment artifacts**. Most enriched state for TSS, Exons, TES of pseudogenes, ZNF/Rpts chromatin state in ct-spec annotations. tRNA class and family repeats, protein-DNA complex assembly genes
93-AM_SPrim	Most enriched state for **HET** chromatin state in ct-spec annotations, **ERV1 and ERVK family repeats**
95-AM_SPrim	Most enriched state for **simple and rRNA class and family and acro and telo family repeats**
96-AM_SPrim	Most enriched state for assembly gaps, centr family repeats


C
ConsHMM staets
Full-stack states
AM_Mam
AM_SMam
A_SMam
AM_Prim
AM_SPrim
AM_nonMam
1-AM_allVer
100-artifact
mnemonics
percent_in_genome

Supplementary Figure 2. 30: Full-stack states enrichments with conservation states.
(A) The first column gives the label of the full-stack states. The second column shows the maximum fold enrichment for each full-stack state for any ConsHMM state defined to annotate nucleotides based on sequence conservation patterns (Arneson and Ernst, 2019) (**Methods**). The third column shows the ConsHMM state that had the highest fold-enrichment in each full-stack state. One notable ConsHMM state is state 1 (1-AM_allVert), representing regions with high probabilities of aligning and matching the human reference genome for all vertebrates and the most enriched for exons. Full-stack states in the transcription-exon group (TxEx) are all maximally enriched with ConsHMM state 1. Another notable ConsHMM state, state 28 (28-AM_SMam), was the ConsHMM most strongly enriched for overlapping annotated TSS. Consistent with this, this state is also the maximum-enriched ConsHMM state in many full-stack states in TSS and Promoter flanking groups. **(B)** Characterizations of notable ConsHMM states. **(C)** Enrichments of full-stack states for each ConsHMM state from a 100-state model based on a 100-way vertebrate alignment (Arneson and Ernst, 2019). Rows (vertical) correspond to different full-stack states. The header row gives the ConsHMM state labels, where ConsHMM states are placed in groups previously defined based on their patterns of sequence alignment with other vertebrates (Arneson and Ernst, 2019), colored as in **(A)**. The second column (horizontal) shows the percentage of the genome that each full-stack state falls into. Each of the remaining columns (horizontal) corresponds to one ConsHMM state. Values in the columns are colored on a column specific coloring scale. The last row (vertical) in the heatmap gives the percentage of the genome that is covered by each ConsHMM state. The corresponding excel file for this figure is provided in **Supplementary Data 2. 7.**

Enrichment of full-stack states major ZNF gene states

State	Genome %	ZNF* gene	ZNF* C2H2 gene	ZNF* not C2H2
GapArtf1	11.86	0.21	0.14	0.42
GapArtf2	0.05	1.44	1.68	0.79
GapArtf3	0.01	1.26	1.54	0.60
Quies1	9.88	0.40	0.03	1.51
Quies2	3.07	0.29	0.11	0.83
Quies3	12.23	0.55	0.41	0.99
Quies4	4.45	0.80	0.66	1.20
Quies5	1.69	0.20	0.22	0.13
HET1	0.71	0.33	0.11	1.00
HET2	0.69	0.66	0.80	0.17
HET3	1.36	1.77	1.83	1.58
HET4	0.56	6.03	7.02	3.68
HET5	0.25	13.97	16.70	7.34
HET6	0.58	2.58	2.93	1.43
HET7	1.02	2.04	2.15	1.70
HET8	0.43	0.87	0.81	0.98
HET9	1.00	1.44	1.56	1.00
ReprPC1	0.19	0.30	0.12	0.86
ReprPC2	0.32	0.28	0.17	0.59
ReprPC3	1.11	0.36	0.28	0.59
ReprPC4	3.93	0.29	0.23	0.46
ReprPC5	0.63	0.38	0.27	0.70
ReprPC6	1.51	0.32	0.25	0.52
ReprPC7	0.61	0.61	0.61	0.58
ReprPC8	0.48	0.37	0.31	0.56
ReprPC9	0.37	0.53	0.47	0.70
Acet1	0.18	1.09	1.00	1.42
Acet2	0.85	0.37	0.24	0.73
Acet3	2.65	0.40	0.28	0.74
Acet4	0.40	0.38	0.41	0.25
Acet5	0.86	0.32	0.30	0.36
Acet6	0.43	0.46	0.48	0.38
Acet7	0.28	0.39	0.45	0.19
Acet8	0.56	0.31	0.26	0.45
EnhWk1	1.54	0.59	0.35	1.30
EnhWk2	0.35	0.66	0.64	0.75
EnhWk3	0.83	0.72	0.62	1.02
EnhWk4	2.22	0.59	0.42	1.08
EnhWk5	0.99	0.60	0.49	0.91
EnhWk6	0.59	0.72	0.62	0.99
EnhWk7	0.48	0.50	0.58	0.21
EnhWk8	1.37	0.42	0.28	0.84
EnhA1	0.18	0.73	0.88	0.28
EnhA2	0.33	0.80	0.85	0.61
EnhA3	0.19	0.82	1.00	0.20
EnhA4	0.30	0.54	0.58	0.37
EnhA5	0.71	0.55	0.47	0.77
EnhA6	0.56	0.91	0.98	0.65
EnhA7	0.39	0.50	0.39	0.80
EnhA8	0.25	0.71	0.56	1.09
EnhA9	0.16	0.89	1.14	0.07
EnhA10	0.39	0.92	1.03	0.61
EnhA11	0.72	0.73	0.78	0.55
EnhA12	0.33	0.31	0.24	0.50
EnhA13	0.76	0.27	0.19	0.52
EnhA14	0.37	0.34	0.29	0.52
EnhA15	1.02	0.42	0.32	0.71
EnhA16	0.65	0.41	0.35	0.57
EnhA17	0.53	0.94	0.99	0.73
EnhA18	0.46	0.49	0.49	0.47
EnhA19	0.26	1.12	1.25	0.64
EnhA20	0.35	0.77	0.77	0.69
TxEnh1	0.39	1.14	1.40	0.27
TxEnh2	0.39	1.95	2.14	1.24
TxEnh3	0.25	3.15	3.76	1.05
TxEnh4	0.27	1.47	1.76	0.58
TxEnh5	0.50	1.42	1.69	0.61
TxEnh6	0.19	1.14	1.44	0.11
TxEnh7	0.27	1.28	1.59	0.24
TxEnh8	0.24	1.22	1.41	0.57
TxWk1	2.80	1.71	1.84	1.37
TxWk2	0.84	1.27	1.28	1.21
Tx1	0.82	2.24	2.25	2.11
Tx2	1.58	1.74	1.88	1.28
Tx3	0.51	1.64	2.03	0.32
Tx4	0.47	2.70	3.01	1.55
Tx5	0.94	2.17	2.51	1.03
Tx6	1.11	1.71	1.79	1.58
Tx7	0.82	1.91	2.13	1.23
Tx8	0.68	1.29	1.40	0.92
TxEx1	0.27	2.61	2.97	1.27
TxEx2	0.56	2.45	2.85	1.10
TxEx3	0.66	2.60	3.14	0.94
TxEx4	0.10	2.17	2.55	0.91
znf1	0.41	20.83	25.12	8.91
znf2	0.15	68.59	86.77	17.81
DNase1	0.20	0.80	0.77	0.86
BivProm1	0.15	0.87	0.66	1.44
BivProm2	0.16	0.50	0.34	0.95
BivProm3	0.29	0.82	0.80	0.83
BivProm4	0.13	1.28	1.13	1.78
PromF1	0.20	0.81	0.76	0.91
PromF2	0.14	1.23	1.34	0.85
PromF3	0.15	3.39	3.95	1.60
PromF4	0.19	4.78	5.61	2.20
PromF5	0.14	2.27	2.34	2.19
PromF6	0.13	4.34	5.14	2.05
PromF7	0.16	4.41	5.24	1.76
TSS1	0.12	3.75	4.31	1.98
TSS2	0.11	2.13	2.27	1.76
Base %	100	0.73	0.57	0.18

Supplementary Figure 2. 31: Full-stack states enrichments with different subsets of ZNF genes.

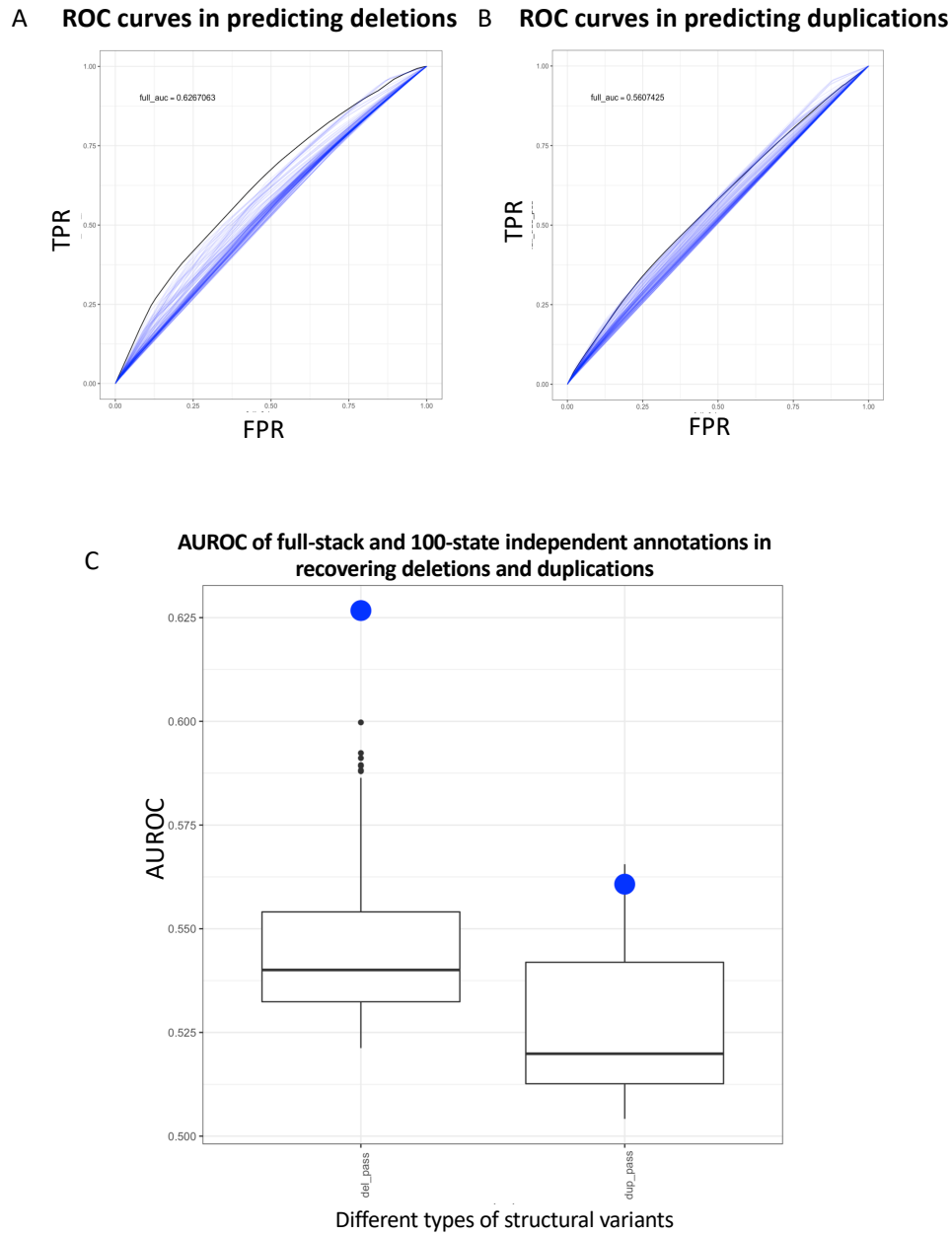
The rows correspond to full-stack states. The first column presents the state labels, the second presents the percentage of the genome that each state occupies, and the remaining three columns enrichments for different subsets of zinc finger genes. The first of these is all genes with a ZNF symbol. The second is the subset of ZNF genes also annotated as C2H2 genes and the third those that are not C2H2 genes. The values correspond to the full-stack states' fold enrichment for the ZNF gene families. Values are colored on a column-specific color scale. The last row gives the percentage of the genome that each type of ZNF gene family occupies

Enrichment of full-stack states with structural variants

state	% Genome	deletion	duplication
GapArtf1	4.88	0.71	0.76
GapArtf2	0.05	1.29	1.58
GapArtf3	0.01	0.89	1.22
Quies1	10.7	1.63	1.21
Quies2	3.31	1.57	1.26
Quies3	13.2	1.04	0.94
Quies4	4.8	1.23	1.14
Quies5	1.82	0.32	1.19
HET1	0.76	1.58	1.3
HET2	0.75	1	1.45
HET3	1.47	1.16	1.03
HET4	0.61	1.54	1.47
HET5	0.27	1.23	1.22
HET6	0.63	1.36	1.37
HET7	1.1	1.18	1.13
HET8	0.47	0.96	1.08
HET9	1.07	1.35	1.46
ReprPC1	0.21	0.5	0.81
ReprPC2	0.35	0.63	0.86
ReprPC3	1.19	0.77	0.88
ReprPC4	4.24	0.86	0.92
ReprPC5	0.68	0.81	1.08
ReprPC6	1.63	0.81	0.97
ReprPC7	0.66	0.73	1.03
ReprPC8	0.51	0.16	1.22
ReprPC9	0.4	0.86	1.01
Acet1	0.2	1	1.19
Acet2	0.92	1	0.94
Acet3	2.86	1.06	0.96
Acet4	0.43	0.87	0.89
Acet5	0.93	0.99	0.92
Acet6	0.46	0.94	0.88
Acet7	0.31	0.82	0.88
Acet8	0.61	0.9	0.98
EnhWk1	1.66	1.04	0.93
EnhWk2	0.38	0.73	0.91
EnhWk3	0.89	0.85	0.88
EnhWk4	2.39	1.04	0.91
EnhWk5	1.07	1.01	0.94
EnhWk6	0.72	0.7	0.72
EnhWk7	0.52	0.95	0.96
EnhWk8	1.48	1.17	1.02
EnhA1	0.19	0.66	0.86
EnhA2	0.35	0.82	0.86
EnhA3	0.21	0.75	0.84
EnhA4	0.32	0.87	0.87
EnhA5	0.77	0.96	0.9
EnhA6	0.61	0.79	0.84
EnhA7	0.42	0.81	0.99
EnhA8	0.27	0.73	0.81
EnhA9	0.17	0.67	0.75
EnhA10	0.43	0.82	0.91
EnhA11	0.77	0.93	1.01
EnhA12	0.36	0.82	0.93
EnhA13	0.82	0.97	0.95
EnhA14	0.39	0.82	0.99
EnhA15	1.1	0.93	0.98
EnhA16	0.69	1.03	1
EnhA17	0.57	0.9	0.88
EnhA18	0.5	1.04	1.06
EnhA19	0.28	0.8	0.9
EnhA20	0.37	1.04	0.92
TxEnh1	0.42	0.7	0.85
TxEnh2	0.42	0.7	0.83
TxEnh3	0.27	0.75	0.91
TxEnh4	0.29	0.51	0.86
TxEnh5	0.54	0.62	0.94
TxEnh6	0.2	0.59	0.8
TxEnh7	0.29	0.66	0.86
TxEnh8	0.26	0.57	0.78
TxWk1	3.02	0.74	0.88
TxWk2	0.91	0.7	0.96
Tx1	0.89	0.77	0.87
Tx2	1.7	0.85	0.92
Tx3	0.55	0.82	0.96
Tx4	0.51	0.63	0.87
Tx5	1.02	0.66	0.94
Tx6	1.2	0.56	0.85
Tx7	0.88	0.52	0.84
Tx8	0.73	0.63	0.8
TxEx1	0.29	0.53	0.82
TxEx2	0.6	0.49	0.93
TxEx3	0.72	0.57	0.97
TxEx4	0.11	0.45	0.8
znf1	0.44	0.94	1.02
znf2	0.16	1.09	1.26
DNase1	0.22	0.81	0.98
BivProm1	0.16	0.53	0.9
BivProm2	0.17	0.52	0.88
BivProm3	0.31	0.66	1
BivProm4	0.14	0.54	0.84
PromF1	0.22	0.75	0.99
PromF2	0.15	0.6	0.86
PromF3	0.16	0.56	0.87
PromF4	0.21	0.51	0.88
PromF5	0.15	0.6	0.9
PromF6	0.14	0.52	0.88
PromF7	0.17	0.58	0.87
TSS1	0.13	0.53	0.9
TSS2	0.12	0.62	0.99
	100	29.1	31.6

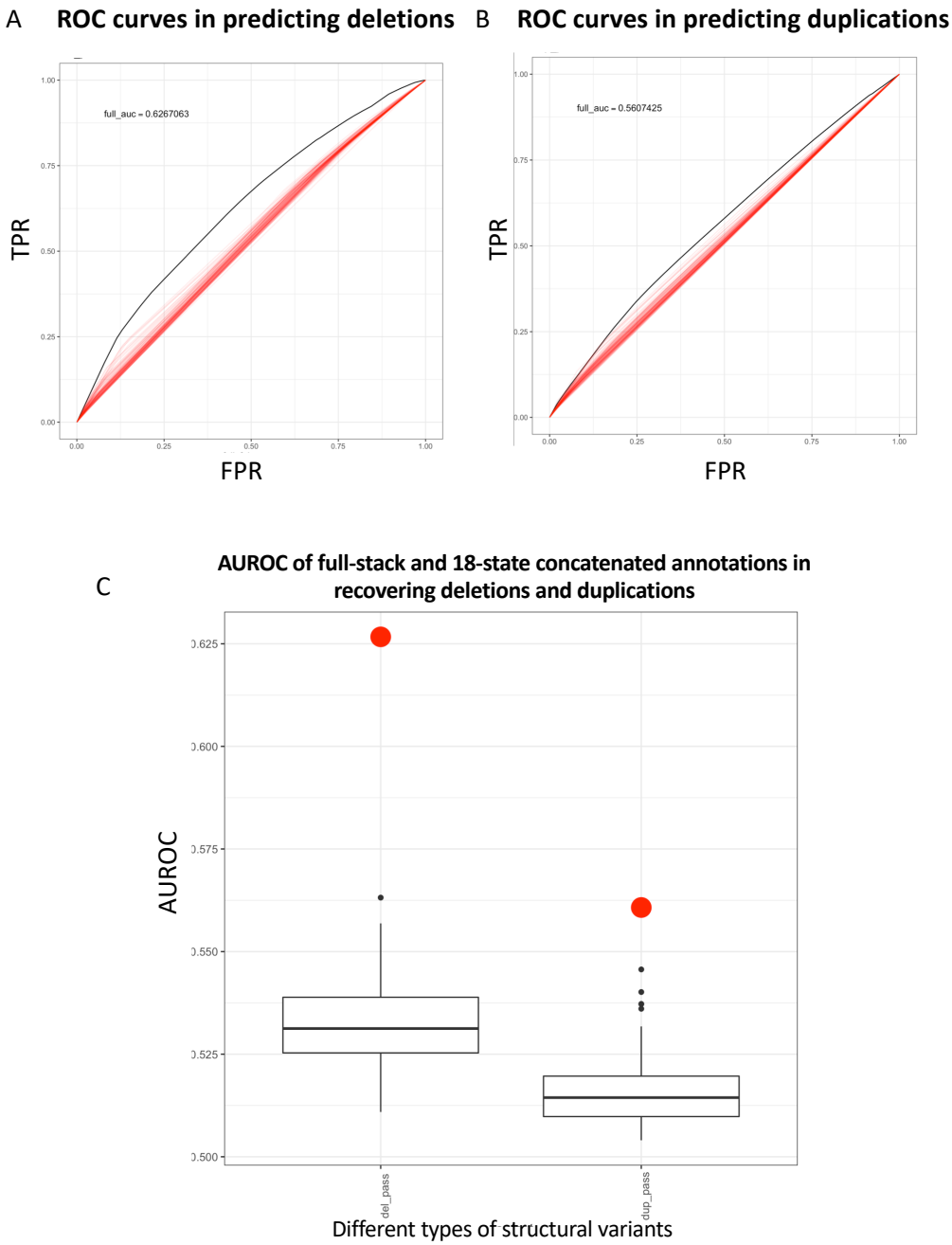
Supplementary Figure 2. 32: Full-stack states enrichments with structural variants.
The rows correspond to full-stack states. The first column presents the state labels, the second presents the percentage of the genome in hg38 that each state occupies, and the last two columns

correspond to two different types of structural variants: deletions and duplications. The values correspond to the full-stack states' fold enrichment for the structural variant type. Values are colored on a column-specific color scale. The last row gives the percentage of the genome that each type of structural variants occupies.



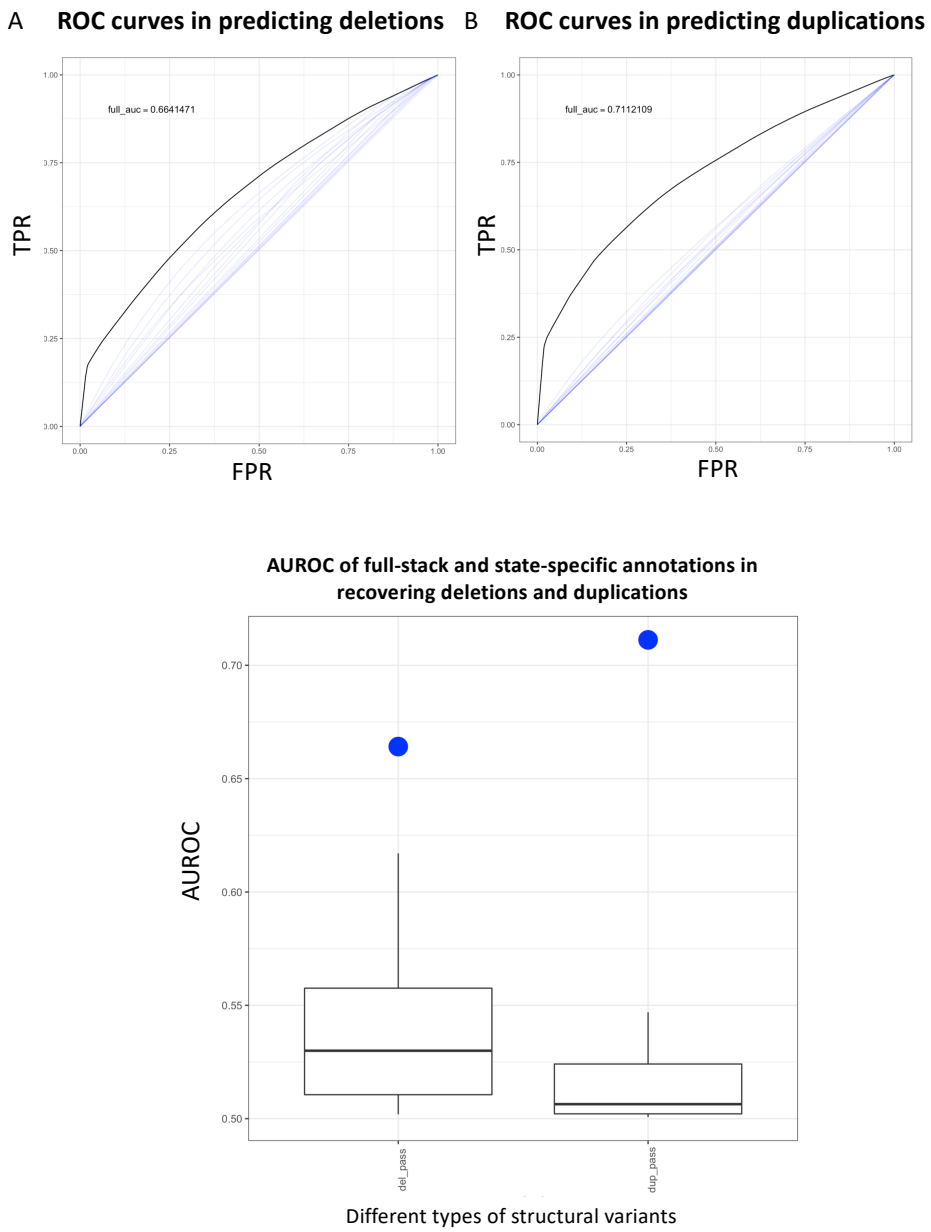
Supplementary Figure 2. 33: Comparison of full-stack model annotations and the 100-state independent model annotations in predicting structural variants of type deletions and duplications.

(A) ROC curves for the full-stack model and the 127 100-state independent models' chromatin state annotations at predicting bases covered by deletions (**Methods**). The full-stack model's annotation ROC curve is in black and the 127 100-state independent models' annotation ROCs are shown in blue. **(B)** Similar plot as **(A),** but for duplications. **(C)** Comparison of the AUROC in predicting structural variants. The x-axis represents different types of structural variants. The box-plots show AUROC for 127 100-state independent models' in predicting deletions and duplications. The blue dots show the AUROC of the full-stack chromatin state annotation.



Supplementary Figure 2. 34: Comparison of full-stack model annotations and 18-state concatenated model annotations in predicting structural variants of type deletions and duplications.

(A) ROC curves for the full-stack model and 98 concatenated models' chromatin state annotations at predicting bases covered by deletion (**Methods)**. The full-stack model's annotation ROC curve is in black and the 98 18-state annotations from concatenated models ROCs are shown in red. (**B)** Similar plot as (A), but for duplications. **(C)** Comparison of the AUROC in predicting structural variants. The x-axis represents different types of structural variants. The box-plots show AUROC of 98 18-state concatenated models' in predicting deletions and duplications. The red dots show the AUROC of the full-stack chromatin state annotations in predicting bases in each type of structural variant.



Supplementary Figure 2. 35:Comparison of full-stack states vs. state-specific annotations in predicting structural variants of types deletions and duplications.
We followed the procedure outlined in *(Abel* et al.*, 2020)* to compute the enrichments between annotations associated with one chromatin state and structural variants. In particular, we utilized 15-state chromatin state annotation for 127 reference epigenomes from Roadmap Epigenomics Consortium. Then, for each of the 15 states, we stratified genomic positions based on the number of cell types in which the state is present (ranging from 0 to 127), resulting in 15 state-specific models' annotations (**Methods**). **(A)** ROC curves for the full-stack model and 15 state-specific models' annotations at predicting bases covered by deletions (**Methods**) The full-stack model's ROC curve is in black, and state-specific models' ROCs are shown in blue. (**B)** Similar plot as (A), but for duplications. **(C)** Comparison of the AUROC in predicting structural variants. The x-axis represents different types of structural variants. The box-plots show AUROC of 15 state-specific models' annotation in predicting deletions and duplications. The blue dots show the AUROC of the full-stack chromatin state annotations in predicting respective types of structural variants.

A Top 10% non-coding prioritized variants

state	% genome	FIRE	fitCons	FATHMM	GERP	LINSIGHT	PhastCons	phyloP	DANN	CDTS	CADD	REMM	Eigen	Eigen_PC	funSeq2
GapArtf2	0.05	0.2	0.3	3.1	0.1	0.2	0.3	0.1	0.2	0.1	0.2	0.3	0.4	2.2	0.7
GapArtf3	0.01	0.1	0.5	3.7	0.1	0.4	0.4	0.3	0.4	0.2	0.9	0.8	2.2	6.9	1.5
EnhA2	0.36	0.6	2.2	2.5	2.7	3.8	2	2	1	1.5	2.7	4.2	3.7	4.2	2.8
EnhA3	0.21	0.9	5.9	1.9	2.6	4.1	1.9	1.9	1.1	1.7	2.8	4.8	5.5	8.6	5.3
EnhA17	0.58	0.5	1.7	2.4	2.5	3.2	2	2	0.9	1.3	2.5	3	3	2.5	2.2
TxEnh4	0.25	5.8	6	1.3	1.4	2.4	0.8	1.1	1	2.4	1.3	3.4	2.3	4.6	4.9
Tx7	0.83	5.9	6.7	1.3	1.7	1.9	1.1	1.5	1	1.8	1.4	1.1	1.5	0.7	3.1
TxEx2	0.5	6.6	6.9	1.4	1.3	1.8	0.8	1.1	0.9	2.5	1.2	2	1.4	1.5	3.6
TxEx4	0.1	4.8	4.6	1.2	1.6	2.9	1	1.3	1.1	2.3	1.7	5.2	3.5	6.7	6.6
BivProm1	0.14	1.3	8.1	2.4	2.2	4.7	1.8	1.9	3.9	7.2	4.3	6.9	7	9.3	5
BivProm2	0.16	0.9	7	2.5	2.2	4	1.6	1.8	3	6.6	3.1	5.7	5.9	8.6	3.4
BivProm4	0.14	1.2	3.6	2.5	2.8	4.6	2.2	2.2	1.6	3.4	3.2	5.4	5.1	6.6	4.1
PromF2	0.15	3.7	6.1	1.5	1.9	3.9	1.3	1.5	1.7	4.1	2.6	5.9	6.8	9.2	6.5
PromF3	0.16	6.2	3.4	1	1.7	4.3	1.1	1.4	2.3	6.7	3	6.3	8.1	9.8	8.4
PromF4	0.19	6.7	0.9	1.6	2.3	5.7	1.8	1.8	4	7.3	4.4	7.5	8.9	9.7	8.8
PromF5	0.14	2.3	7.3	2	2.3	5.2	1.9	2	4	6.9	4.4	7.3	7.9	9.5	6.5
TSS1	0.13	5.1	3.1	2.1	2.6	5.9	2.5	2.3	4.9	5.3	4.9	7.7	7.9	9.4	7.4
TSS2	0.12	2.4	5.6	2.2	2.1	5	2.4	2.1	4.7	3.5	4.1	6.4	6.2	8.5	5

B Top 5% non-coding prioritized variants

state	% genome	FIRE	fitCons	FATHMM	GERP	LINSIGHT	PhastCons	phyloP	DANN	CDTS	CADD	REMM	Eigen	Eigen_PC	funSeq2
GapArtf2	0.05	0	0.5	4.2	0.1	0.1	0.3	0.1	0.2	0.1	0.1	0.2	0.5	1.9	1
GapArtf3	0.01	0	1	5	0.1	0.2	0.5	0.2	0.4	0.3	0.4	0.7	2.8	9	2.5
EnhA2	0.36	0.5	3.1	3.4	3.5	4.9	3.3	2.5	0.7	1.4	3.9	5.1	4.4	3.5	3.7
EnhA3	0.21	0.6	6.3	2.6	3.2	4.9	3	2.3	0.8	1.7	3.7	5.3	6.4	12	7.8
EnhA17	0.58	0.4	2.8	3.3	3.4	4.1	3.1	2.5	0.6	1.2	3.5	3.7	3.7	1.5	2.6
TxEnh4	0.25	8.1	7.4	1.8	1.5	2.4	1.1	1.2	0.8	2.8	1.5	3.4	2.3	5.4	5.9
TxEx2	0.5	8.4	6.6	2	1.4	1.8	1	1.2	0.8	2.9	1.2	1.8	1.2	1.3	3.3
TxEx3	0.65	7.6	4.5	1.1	0.9	1.1	0.8	0.9	0.9	2.3	0.8	1.3	0.8	1.8	1.9
BivProm1	0.14	0.8	17	3.2	2.7	5.4	2.9	2.3	5.5	13	4.9	10	9.6	17	7.5
BivProm2	0.16	0.5	15	3.3	2.7	4.7	2.7	2.3	3.7	12	3.9	7.8	7.1	15	4.9
BivProm4	0.14	0.8	7	3.7	3.9	5.8	3.6	2.9	1.4	4.6	4.5	7.2	6.3	7.8	5.9
PromF2	0.15	4.2	9.7	2	2.2	4.1	2.1	1.8	1.5	5.9	3.2	7.8	9.8	16	10
PromF3	0.16	8	6.8	1.2	1.8	3.9	1.6	1.5	2.5	12	3	9.2	13	18	13
PromF4	0.19	9.2	1.9	2.2	2.8	6.1	2.9	2.3	5.6	14	4.9	12	16	19	14
PromF5	0.14	2	15	2.8	2.9	5.9	3.1	2.5	5.6	13	5.1	12	12	18	9.9
TSS1	0.13	5.7	6.4	2.8	3.2	7.2	4.3	3	7.7	10	6.4	13	13	18	12
TSS2	0.12	2.1	12	3.1	2.4	6.2	3.9	2.6	7.5	6.2	5.7	9.8	9.6	15	7.9

C Top 1% non-coding prioritized variants

state	% genome	FIRE	fitCons	FATHMM	GERP	LINSIGHT	PhastCons	phyloP	DANN	CDTS	CADD	REMM	Eigen	Eigen_PC	funSeq2
ReprPC1	0.2	0	1.1	4.4	3.9	3.9	3.6	3.8	0.7	11	4.9	5	5.2	4.3	1.5
EnhA2	0.36	0.3	1.9	4.5	5.6	5.2	4.8	4.2	0.2	0.8	6.3	7.4	6.2	0.8	4.6
EnhA3	0.21	0.3	9.2	3.2	4.5	4.5	3.9	3.4	0.3	0.8	5.6	5.6	8	7.9	12
EnhA17	0.58	0.2	1	4.8	5.7	4.8	4.7	4.3	0.3	0.6	5.9	5.8	5.5	0.1	2.3
TxEnh4	0.25	15	17	2.9	2	2.2	1.5	1.8	0.7	2.8	2.1	2.8	2	4.1	7.8
TxEnh8	0.25	4.1	9	2	2.2	2.3	1.6	1.8	0.4	0.8	1.9	2.6	2.3	2.5	7.7
TxEx1	0.26	9.9	10	2.5	2.1	1.9	1.7	2	0.6	1.2	2	2.2	1.8	0.1	3.2
TxEx2	0.5	13	14	3.2	1.9	1.8	1.4	1.8	0.8	2.5	1.6	1.9	1.1	0.3	2.1
TxEx3	0.65	13	8.7	1.8	1	1.1	0.9	1.2	0.7	2.1	0.9	1.1	0.7	0.8	1.1
DNase1	0.22	0.3	3	10	2.2	3	2.4	2.5	2.2	2.3	2.8	3.2	4.3	10	7.1
BivProm1	0.14	0.2	1.1	4.5	2.8	4.1	4.1	3.8	8.8	52	8.7	13	11	32	14
BivProm2	0.16	0	1.3	4.8	3.5	4.2	4.2	4.1	4.8	41	7.3	8.9	8.6	16	6.9
BivProm4	0.14	0.5	1.7	5.4	6.6	5.8	5.7	5.4	0.7	7.8	8.4	9.7	9.8	6.5	8.6
PromF2	0.15	4.1	5.5	2.6	2.6	3.5	2.7	2.5	0.9	11	5	5.3	10	40	21
PromF3	0.16	11	0.8	1.2	1.4	2.6	2	1.8	2.3	35	4.2	6.7	16	64	33
PromF4	0.19	13	0.2	2.4	2.6	4.5	4	3.4	9.8	56	8.7	19	32	73	38
PromF5	0.14	0.8	1.1	3.7	3	4.2	4.4	3.9	8.9	48	9.1	17	16	52	21
TSS1	0.13	6	0.8	3	3.4	5.3	6.1	4.8	17	41	12	24	26	61	30
TSS2	0.12	1.5	2	4.2	2.5	3.7	5.3	4	19	23	9.6	14	16	40	16

State rank: 1 2 3 4 5

Fold Enrichment highlights: FE >= 1, FE < 1

Supplementary Figure 2. 36: Enrichment of selected full-stack states with prioritized variants, non-coding genome.
Extended version of figure **Figure 2.5C** showing fold enrichment of full-stack states for genomic bases prioritized in the **(A)** top 10% **(B)** top 5%, and **(C)** top 1% among non-coding bases by 14-different variant prioritization scores previously curated in (Arneson and Ernst, 2019) (Methods). Only states that were among the top five with greatest enrichments for at least one score are shown. Top enrichment values are colored red based on the rank of the state for each score as indicated in the color legend at the bottom. Depletions are shown in yellow.

Top 1% prioritized variants, non-coding

state	% NC genome	CADD	CDTS	DANN	Eigen_PC	Eigen	FATHMM	FIRE	GERP	LINSIGHT	PhastCons	REMM	fitCons	funSeq2	phyloP
GapArtf1	4.19	0.07	0.09	2.31	0.11	0.04	0.34	0.06	0.06	0.05	0.36	0.03	0.79	0.01	0.36
GapArtf2	0.05	0.05	0.19	0.22	1.6	0.55	2.62	0	0.04	0.05	0.29	0.09	0.36	4.06	0.17
GapArtf3	0.01	0.16	0.76	0.36	12.5	5.85	3.36	0	0.15	0.14	0.62	0.23	3.1	9.64	0.43
Quies1	10.8	0.6	0.06	0.74	0	0.39	0.65	0	0.69	0.6	0.67	0.28	0.02	0.01	0.65
Quies2	3.38	0.61	0.1	0.46	0	0.4	0.67	0	0.6	0.59	0.67	0.52	0.05	0.01	0.62
Quies3	13.4	0.51	0.22	1.03	0.03	0.39	0.6	0.56	0.66	0.62	0.67	0.27	0.24	0.02	0.7
Quies4	4.79	0.08	0.13	2.7	0	0.04	0.14	0.08	0.09	0.09	0.19	0.04	0.05	0	0.19
Quies5	1.86	0.4	2.77	1.77	0.01	0.03	0.06	0.04	0.55	0.18	0.84	0.31	0.02	0.01	0.8
HET1	0.78	0.28	0.3	0.63	0.01	0.19	0.49	0.01	0.28	0.3	0.56	0.33	0.17	0.02	0.6
HET2	0.76	0.61	0.63	0.28	0.07	0.44	0.92	0	0.62	0.69	0.76	0.73	0.13	0.02	0.8
HET3	1.5	0.06	0.12	2.72	0	0.03	0.09	0.05	0.06	0.08	0.12	0.05	0.07	0.02	0.13
HET4	0.62	0.03	0.07	1.9	0.01	0.02	0.58	0.05	0.03	0.05	0.15	0.03	0.15	0.12	0.21
HET5	0.27	0.07	0.43	1.2	0.14	0.04	0.42	0.02	0.08	0.11	0.38	0.11	0.69	0.43	0.52
HET6	0.63	0.02	0.3	1.2	0.06	0.01	0.33	0.12	0.03	0.04	0.41	0.04	0.24	0.03	0.48
HET7	1.11	0.05	0.31	1.69	0.01	0.03	0.21	0.15	0.05	0.07	0.3	0.06	0.22	0.01	0.34
HET8	0.48	0.14	0.35	0.88	0.11	0.09	0.32	0.12	0.13	0.23	0.35	0.19	0.32	0.27	0.41
HET9	1.08	0.11	0.11	1.52	0	0.06	0.4	0.21	0.11	0.11	0.41	0.09	0.09	0.04	0.41
ReprPC1	0.2	4.91	11.5	0.65	4.3	5.23	4.44	0.02	3.9	3.95	3.59	5	1.15	1.45	3.85
ReprPC2	0.36	2.33	1.41	0.33	0.48	2.06	2.28	0.14	2.31	2.25	2.13	2.07	0.6	0.4	2.13
ReprPC3	1.22	1.28	0.31	0.47	0.02	1.02	1.37	0.27	1.28	1.33	1.31	1.14	0.21	0.13	1.29
ReprPC4	4.33	0.79	0.17	0.76	0.01	0.59	0.9	0.25	0.84	0.86	0.9	0.55	0.1	0.02	0.87
ReprPC5	0.68	0.74	1.05	0.33	0.64	0.55	1.07	0.28	0.63	0.96	0.88	0.98	0.54	0.22	0.99
ReprPC6	1.65	0.56	0.48	0.38	0.18	0.4	0.83	0.43	0.53	0.74	0.75	0.64	0.3	0.06	0.82
ReprPC7	0.65	1.01	1.23	0.34	1.94	0.82	1.27	0.64	0.8	1.27	1.03	1.53	0.62	0.94	1.13
ReprPC8	0.52	0.52	4.8	1.12	0	0.01	0.02	0	0.55	0.33	1.03	0.8	0.01	0.05	1.05
ReprPC9	0.41	1.07	0.81	0.6	0.22	0.84	1.14	0.74	1.14	1.12	1.17	0.87	0.39	0.23	1.17
Acet1	0.2	0.2	1.26	0.61	1.37	0.15	0.62	0.13	0.2	0.3	0.59	0.36	1.02	0.82	0.7
Acet2	0.94	0.76	0.37	0.34	0.03	0.57	0.92	0.31	0.76	0.89	0.9	0.88	0.24	0.15	0.96
Acet3	2.91	0.37	0.28	0.59	0.01	0.25	0.48	0.5	0.4	0.48	0.51	0.35	0.16	0.03	0.54
Acet4	0.44	0.71	0.49	0.23	0.41	0.58	0.71	0.41	0.69	1.04	0.73	1.06	0.6	1.53	0.79
Acet5	0.95	0.72	0.21	0.36	0.02	0.53	0.68	0.3	0.76	0.87	0.75	0.72	0.17	0.18	0.77
Acet6	0.47	1.26	0.23	0.21	0.18	1.1	0.98	0.25	1.31	1.47	1.16	1.48	0.71	1.41	1.11
Acet7	0.31	1.66	0.61	0.2	5.21	2.12	0.95	0.33	1.34	2.01	1.3	2.42	2.52	7.82	1.24
Acet8	0.62	0.48	0.51	0.52	0.1	0.34	0.57	0.38	0.46	0.66	0.58	0.59	0.28	0.45	0.62
EnhWk1	1.7	1.7	0.22	0.41	0.01	1.29	1.45	0.19	1.87	1.73	1.7	1.27	0.19	0.19	1.54
EnhWk2	0.38	1.62	1.16	0.26	4.04	1.63	1.06	1.42	1.26	2.11	1.35	2.71	1.54	4.62	1.35
EnhWk3	0.91	2.73	0.48	0.22	0.18	2.18	2.12	0.51	2.6	2.66	2.38	2.93	0.38	1.05	2.14
EnhWk4	2.44	3.49	0.2	0.26	0.02	2.77	3.04	0.16	3.58	3.03	3.1	2.46	0.11	0.15	2.72
EnhWk5	1.09	0.89	0.47	0.62	0.05	0.71	0.84	0.6	1.03	0.97	0.97	0.69	0.27	0.16	0.95
EnhWk6	0.64	0.8	0.62	0.48	0.15	0.74	0.67	1	0.91	1.06	0.87	0.87	1.14	1.9	0.92
EnhWk7	0.53	0.79	0.36	0.56	0.14	0.76	0.69	0.68	0.84	0.94	0.84	0.93	1.55	0.61	0.83
EnhWk8	1.51	2.29	0.34	0.61	0	1.81	2.11	0.12	2.44	2.02	2.1	1.83	0.3	0.18	1.91
EnhA1	0.19	4.39	1.91	0.42	21	7.39	1.9	1.29	2.95	3.89	2.82	5.37	6.9	12.5	2.58
EnhA2	0.36	6.3	0.78	0.2	0.85	6.17	4.46	0.31	5.57	5.24	4.78	7.39	1.86	4.55	4.23
EnhA3	0.21	5.64	0.81	0.33	7.88	8	3.21	0.35	4.53	4.55	3.9	5.6	9.25	12.3	3.44
EnhA4	0.33	3.59	0.36	0.21	1.96	4.42	2.16	0.09	3.1	3.18	2.74	3.57	5.43	7.05	2.37
EnhA5	0.79	2.94	0.28	0.3	0.17	2.69	2.07	0.14	2.82	2.7	2.56	2.74	1.78	1.8	2.21
EnhA6	0.62	2.66	0.54	0.22	0.24	2.34	2.25	0.55	2.79	2.67	2.36	2.7	0.42	1.13	2.26
EnhA7	0.42	0.81	1.16	0.33	1.76	0.75	0.98	0.87	0.68	1.03	0.8	1.19	1.92	2.47	0.87
EnhA8	0.28	1.52	1.58	0.36	4.89	1.97	1.05	1.31	1.21	1.74	1.19	2.15	2.92	8.02	1.21
EnhA9	0.18	2.11	1.62	0.37	13	3.42	0.93	2.47	1.38	2.25	1.44	2.82	5.94	13.1	1.4
EnhA10	0.43	1.85	1.4	0.41	1.62	2.09	1.29	1.24	1.7	1.92	1.56	2.18	2.81	4.74	1.47
EnhA11	0.79	0.62	0.5	0.6	0.07	0.49	0.61	1	0.67	0.77	0.71	0.58	0.52	0.39	0.72
EnhA12	0.36	1.33	0.69	0.31	2.06	1.4	0.97	0.48	1.14	1.62	1.14	2.03	0.87	5.3	1.11
EnhA13	0.84	1.05	0.31	0.41	0.06	0.84	0.91	0.28	1.1	1.17	1.03	1.04	0.2	0.8	1
EnhA14	0.4	1.29	0.77	0.35	2.15	1.37	1.03	1.04	1.16	1.6	1.15	1.64	0.75	4.67	1.16
EnhA15	1.12	1.03	0.37	0.53	0.11	0.8	0.96	0.73	1.09	1.15	1.05	0.83	0.22	0.57	1.02
EnhA16	0.71	0.81	0.97	0.8	0.44	0.7	0.92	0.78	0.85	0.89	0.95	0.8	0.34	0.73	0.98
EnhA17	0.58	5.89	0.57	0.27	0.14	5.47	4.85	0.21	5.66	4.83	4.74	5.83	0.96	2.32	4.32
EnhA18	0.5	1.87	0.81	0.76	0.09	1.53	1.73	0.21	1.75	1.63	1.62	2.15	1.43	1.77	1.53
EnhA19	0.28	3.35	1.64	0.35	2.23	3.52	2.46	0.89	2.74	2.98	2.51	4.42	3.24	8.15	2.36
EnhA20	0.38	1.74	0.28	0.3	0.07	1.55	1.54	0.13	1.91	1.71	1.62	1.48	2.5	0.42	1.56
TxEnh1	0.43	1.59	0.34	0.24	0.08	1.65	0.92	0.34	1.94	2.03	1.58	1.89	1.12	2.4	1.59
TxEnh2	0.43	0.72	0.61	0.32	0.21	0.68	0.36	1.87	0.75	1.04	0.74	1.11	1.42	3.1	0.85
TxEnh3	0.27	0.44	0.83	0.53	0.72	0.37	0.55	3.18	0.43	0.61	0.63	0.71	2.14	2.59	0.72
TxEnh4	0.25	2.06	2.81	0.67	4.08	2	2.9	14.5	1.96	2.22	1.54	2.84	17.2	7.85	1.81
TxEnh5	0.49	1.25	2.04	0.57	2.35	1.21	1.91	5.89	1.33	1.6	1.15	1.82	7.23	2.82	1.37
TxEnh6	0.19	1.25	1.54	0.44	3.02	1.41	1.66	7.08	1.34	1.62	1.1	1.86	8.85	6.75	1.28
TxEnh7	0.3	1.27	0.96	0.23	3.97	1.49	0.65	1.38	1.03	1.71	1.06	2.18	2.3	5.29	1.07
TxEnh8	0.25	1.92	0.85	0.36	2.53	2.31	1.99	4.1	2.17	2.3	1.63	2.56	8.98	7.73	1.75
TxWk1	3.03	0.59	0.58	0.87	0.02	0.71	1.07	2.95	1.11	0.9	0.9	0.53	1.92	0.16	1.07
TxWk2	0.86	0.69	1.32	0.57	0.53	0.55	1.32	2.12	0.84	1.02	0.86	0.93	2.7	0.41	1.08
Tx1	0.91	0.28	0.46	0.81	0.01	0.23	0.23	0.92	0.34	0.45	0.42	0.34	0.42	0.66	0.5
Tx2	1.74	0.4	0.43	0.88	0.01	0.36	0.32	0.86	0.57	0.63	0.56	0.38	0.3	0.27	0.63
Tx3	0.56	0.42	0.53	0.4	0.24	0.38	0.33	0.94	0.47	0.73	0.53	0.65	0.48	1.1	0.59
Tx4	0.5	0.69	0.51	0.69	0.03	0.62	1	4.09	0.93	0.87	0.79	0.75	3.07	0.94	1
Tx5	1	0.37	1	0.76	0.12	0.33	0.8	9.56	0.5	0.52	0.53	0.46	3.32	0.29	0.64
Tx6	1.17	1.11	0.55	0.62	0	1.33	2.22	3.31	2.1	1.57	1.48	0.96	5.37	0.45	1.86
Tx7	0.83	1.4	0.64	0.57	0.01	1.43	2.68	7.44	2.17	1.76	1.51	1.41	7.92	1.07	1.94
Tx8	0.73	1.5	0.54	0.38	0.14	1.62	1.77	3	2.07	1.91	1.61	1.71	4.66	1.72	1.72
TxEx1	0.26	1.98	1.22	0.62	0.15	1.76	2.49	9.93	2.12	1.93	1.69	2.22	10.2	3.21	2.03
TxEx2	0.5	1.64	2.49	0.78	0.31	1.14	3.21	13.3	1.92	1.84	1.4	1.9	13.7	2.1	1.81
TxEx3	0.65	0.9	2.13	0.71	0.77	0.67	1.8	12.7	0.98	1.09	0.93	1.13	8.72	1.07	1.17
TxEx4	0.1	3.01	2.69	0.53	10.4	3.9	2.43	8.09	2.28	2.78	2.03	3.84	8.28	12.6	2.2
znf1	0.44	0.08	0.29	2.09	0.03	0.05	0.24	0.13	0.09	0.13	0.26	0.11	1.12	0.25	0.32
znf2	0.15	0.2	0.31	1.35	0.17	0.09	0.73	0.11	0.16	0.22	0.39	0.26	5.07	0.61	0.59
DNase1	0.22	2.78	2.32	2.17	10.4	4.3	10.1	0.34	2.23	2.95	2.41	3.17	3.03	7.12	2.45
BivProm1	0.14	8.74	52	8.77	31.7	11.3	4.48	0.16	2.8	4.06	4.14	12.7	1.06	13.9	3.81
BivProm2	0.16	7.25	41.1	4.76	16.1	8.61	4.81	0.04	3.45	4.2	4.22	8.94	1.25	6.89	4.14
BivProm3	0.3	4.01	5.6	1.24	6.14	4.06	3.44	0.47	2.83	3.24	3.14	5.01	1.26	3.33	3
BivProm4	0.14	8.44	7.77	0.72	6.48	9.81	5.41	0.49	6.65	5.8	5.66	9.73	1.66	8.63	5.39
PromF1	0.22	2.94	5.38	0.82	20.5	4.58	3.41	1.66	1.69	2.68	1.98	4.22	4.59	12.1	1.85
PromF2	0.15	5.02	10.8	0.92	39.6	10.5	2.57	4.13	2.6	3.48	2.75	5.25	5.45	20.9	2.52
PromF3	0.16	4.2	35.4	2.26	63.7	15.6	1.24	10.8	1.38	2.61	1.96	6.69	0.76	32.8	1.84
PromF4	0.19	8.7	55.7	9.76	72.8	31.5	2.4	13.1	2.61	4.48	4	19	0.21	37.9	3.39
PromF5	0.14	9.06	48.3	8.88	51.6	16.5	3.7	0.85	3.02	4.25	4.38	17.2	1.13	21.4	3.88
PromF6	0.14	2.68	9.81	0.52	19.8	4.8	0.91	7.92	1.53	2	1.57	2.92	0.56	16.2	1.67
PromF7	0.17	1.47	3.75	0.41	3.19	1.54	0.72	6.51	1.27	1.51	1.16	1.89	1.35	5.38	1.34
TSS1	0.13	12.1	41.2	17.4	61.4	25.7	3.01	6	3.43	5.3	6.09	23.5	0.83	30.3	4.8
TSS2	0.12	9.59	23.1	18.7	39.7	16	4.24	1.49	2.48	3.71	5.27	13.6	1.98	16	4.04
	100	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.99	1.71	1.01	1.01	0.99	0.99

Top 1% prioritized variants, whole-genome

state	% genome	CADD	CDTS	DANN	Eigen_PC	Eigen	FATHMM	FIRE	GERP	PhastCons	REMM	fitCons	phyloP
GapArtf1	12	0.07	0.03	0.31	0.04	0.01	0.1	0.02	0.01	0.15	0.01	0.17	0.13
GapArtf2	0.1	0.18	0.27	0.21	1.6	0.5	1.28	0	0.08	0.31	0.12	0.26	0.19
GapArtf3	0	0.62	0.96	0.6	12	5.25	2.22	0	0.43	0.8	0.28	0.62	0.71
Quies1	9.9	0.19	0.05	0.23	0	0.39	0.48	0	0.48	0.54	0.15	0.05	0.36
Quies2	3.1	0.28	0.09	0.19	0	0.41	0.55	0	0.47	0.57	0.42	0.16	0.45
Quies3	12	0.27	0.2	0.41	0.03	0.4	0.49	0.6	0.51	0.59	0.16	0.24	0.47
Quies4	4.5	0.05	0.11	0.73	0	0.04	0.11	0.08	0.06	0.18	0.03	0.07	0.14
Quies5	1.7	0.19	2.57	0.63	0.01	0.03	0.04	0.04	0.4	0.72	0.2	0.02	0.47
HET1	0.7	0.39	0.32	0.39	0.01	0.21	0.51	0.01	0.37	0.55	0.41	0.48	0.63
HET2	0.7	0.82	0.72	0.77	0.07	0.49	0.99	0	0.7	0.82	0.87	1.04	0.92
HET3	1.4	0.05	0.11	0.66	0	0.04	0.07	0.06	0.05	0.11	0.04	0.08	0.09
HET4	0.6	0.09	0.06	0.47	0.01	0.02	0.31	0.06	0.03	0.14	0.03	0.23	0.19
HET5	0.2	1.01	0.81	1.31	0.18	0.07	0.55	0.02	0.28	0.47	0.37	1.83	0.78
HET6	0.6	0.32	0.33	0.56	0.08	0.02	0.33	0.13	0.07	0.4	0.09	0.87	0.51
HET7	1	0.17	0.32	0.5	0.01	0.03	0.21	0.17	0.08	0.29	0.1	0.32	0.36
HET8	0.4	0.44	0.38	0.6	0.19	0.1	0.36	0.14	0.2	0.38	0.27	0.79	0.46
HET9	1	0.16	0.1	0.48	0.01	0.06	0.3	0.23	0.09	0.36	0.07	0.35	0.34
ReprPC1	0.2	4.61	12	2.45	4.95	5.45	4.49	0.02	3.69	3.58	5.07	2.86	4.12
ReprPC2	0.3	1.4	1.39	0.7	0.56	2.11	2.02	0.16	1.93	1.9	1.72	0.85	1.79
ReprPC3	1.1	0.72	0.28	0.45	0.02	1.03	1.17	0.3	1.08	1.15	0.93	0.54	1
ReprPC4	3.9	0.44	0.15	0.38	0.01	0.6	0.75	0.27	0.69	0.79	0.4	0.34	0.64
ReprPC5	0.6	1.62	1.25	1.67	0.77	0.66	1.4	0.3	0.95	1.15	1.49	2.21	1.42
ReprPC6	1.5	1.02	0.51	1.06	0.21	0.46	1	0.47	0.72	0.87	0.9	1.46	1.01
ReprPC7	0.6	2.82	1.52	2.98	2.21	1.07	2.13	0.71	1.49	1.65	2.6	3.96	2.14
ReprPC8	0.5	0.7	4.67	0.88	0	0.01	0.02	0	0.6	1.04	0.86	0.02	0.94
ReprPC9	0.4	0.61	0.78	0.43	0.26	0.87	1.03	0.8	0.93	1.05	0.7	0.39	0.91
Acet1	0.2	1.82	1.9	2.04	1.61	0.35	1.37	0.14	0.97	1.11	1.39	2.68	1.63
Acet2	0.9	0.77	0.37	0.63	0.04	0.61	0.95	0.33	0.83	0.9	0.97	0.82	0.96
Acet3	2.6	0.35	0.26	0.36	0.01	0.27	0.46	0.54	0.39	0.49	0.37	0.39	0.48
Acet4	0.4	0.75	0.45	0.62	0.47	0.62	0.77	0.45	0.75	0.77	1.1	0.8	0.78
Acet5	0.9	0.41	0.19	0.26	0.02	0.55	0.59	0.33	0.63	0.67	0.61	0.3	0.55
Acet6	0.4	0.6	0.2	0.3	0.22	1.13	0.83	0.27	1.03	1.02	1.13	0.39	0.77
Acet7	0.3	0.96	0.57	0.5	5.82	2.2	0.85	0.37	1.1	1.19	1.8	0.65	0.95
Acet8	0.6	0.48	0.49	0.47	0.12	0.36	0.57	0.41	0.48	0.58	0.62	0.52	0.59
EnhWk1	1.5	0.57	0.19	0.22	0.01	1.31	1.1	0.21	1.36	1.41	0.81	0.19	0.96
EnhWk2	0.4	1.7	1.26	1.49	4.43	1.8	1.35	1.56	1.4	1.51	2.83	1.99	1.53
EnhWk3	0.8	1.12	0.46	0.39	0.2	2.22	1.73	0.55	2.02	2.03	2.34	0.49	1.56
EnhWk4	2.2	1.14	0.18	0.19	0.02	2.8	2.34	0.18	2.67	2.54	1.65	0.2	1.85
EnhWk5	1	0.42	0.44	0.34	0.06	0.72	0.68	0.65	0.81	0.84	0.49	0.28	0.66
EnhWk6	0.6	0.54	0.58	0.45	0.18	0.77	0.6	1.07	0.8	0.81	0.71	0.47	0.7
EnhWk7	0.5	0.46	0.34	0.37	0.27	0.79	0.59	0.73	0.68	0.75	0.73	0.35	0.6
EnhWk8	1.4	0.83	0.3	0.28	0.01	1.83	1.65	0.13	1.85	1.74	1.26	0.2	1.31
EnhA1	0.2	2.67	1.99	1.13	22.3	7.57	1.75	1.46	2.49	2.57	3.68	1.42	2.17
EnhA2	0.3	2.49	0.72	0.42	1	6.22	3.52	0.35	4.29	3.96	5.58	0.52	3.11
EnhA3	0.2	2.47	0.74	0.44	8.92	8.12	2.47	0.38	3.46	3.26	3.59	0.43	2.44
EnhA4	0.3	1.45	0.33	0.33	2.33	4.48	1.68	0.1	2.32	2.3	2.32	0.36	1.6
EnhA5	0.7	1.06	0.25	0.28	0.2	2.72	1.59	0.16	2.09	2.13	1.85	0.29	1.46
EnhA6	0.6	1.3	0.55	0.63	0.27	2.41	1.95	0.6	2.36	2.11	2.32	0.77	1.85
EnhA7	0.4	1.52	1.29	1.58	2.03	0.86	1.24	0.95	0.95	1	1.51	2.03	1.15
EnhA8	0.3	1.45	1.61	1.16	5.57	2.14	1.07	1.42	1.17	1.22	2.02	1.39	1.12
EnhA9	0.2	1.62	1.67	1.04	14.1	3.61	0.97	2.73	1.27	1.43	2.19	1.3	1.25
EnhA10	0.4	0.86	1.31	0.41	1.86	2.13	1.05	1.34	1.32	1.35	1.57	0.41	1.02
EnhA11	0.7	0.4	0.47	0.39	0.08	0.51	0.53	1.07	0.55	0.64	0.47	0.41	0.52
EnhA12	0.3	1.07	0.68	0.79	2.37	1.48	0.95	0.53	1.06	1.11	1.79	1.02	0.98
EnhA13	0.8	0.49	0.28	0.28	0.08	0.86	0.75	0.3	0.85	0.89	0.76	0.3	0.66
EnhA14	0.4	1.06	0.74	0.84	2.38	1.43	1.07	1.14	1.1	1.12	1.52	1.08	1.04
EnhA15	1	0.56	0.34	0.4	0.13	0.82	0.85	0.79	0.88	0.93	0.65	0.45	0.74
EnhA16	0.6	0.5	0.92	0.49	0.5	0.73	0.79	0.84	0.69	0.85	0.62	0.43	0.74
EnhA17	0.5	2.15	0.52	0.29	0.18	5.51	3.8	0.23	4.4	3.89	4.37	0.31	3.19
EnhA18	0.5	0.98	0.78	0.48	0.11	1.56	1.48	0.24	1.45	1.41	1.8	0.45	1.24
EnhA19	0.3	1.89	1.65	0.86	2.58	3.63	2.15	1	2.35	2.26	3.7	1.05	2.01
EnhA20	0.3	0.72	0.24	0.27	0.09	1.57	1.21	0.15	1.44	1.37	0.92	0.29	1.04
TxEnh1	0.4	0.63	0.32	0.33	0.09	1.71	0.76	0.38	1.53	1.4	1.38	0.36	1.11
TxEnh2	0.4	0.44	0.55	0.33	0.24	0.73	0.37	2.01	0.66	0.72	0.88	0.37	0.59
TxEnh3	0.2	1.22	0.92	1.25	0.83	0.54	0.94	3.43	0.93	0.96	1.31	1.59	1.21
TxEnh4	0.3	11.6	4.65	11.8	4.16	3.37	7.86	14.1	7.68	5.51	9.96	15.5	8.5
TxEnh5	0.5	7.11	3.25	7.54	2.63	1.97	4.76	5.99	4.41	3.41	5.84	10.1	5.17
TxEnh6	0.2	4.83	2.02	5	3.31	1.95	3.16	7.3	3.19	2.48	4.12	6.77	3.49
TxEnh7	0.3	1.04	1.02	0.8	4.43	1.64	0.79	1.53	1.04	1.11	2	1.08	1.06
TxEnh8	0.2	4.42	1	4.13	2.76	2.81	3.5	4.28	3.94	2.8	4.13	5.27	3.66
TxWk1	2.8	1.2	0.55	1.3	0.03	0.79	1.28	3.14	1.4	1.14	0.65	1.43	1.34
TxWk2	0.8	4.88	2.02	5.3	0.65	0.97	3.35	2.19	3.02	2.43	3.53	6.85	3.7
Tx1	0.8	0.2	0.41	0.28	0.02	0.25	0.24	1.02	0.32	0.41	0.27	0.2	0.34
Tx2	1.6	0.21	0.39	0.3	0.01	0.38	0.29	0.95	0.47	0.51	0.27	0.18	0.42
Tx3	0.5	0.3	0.5	0.26	0.29	0.41	0.34	1.02	0.44	0.51	0.62	0.32	0.49
Tx4	0.5	2.27	0.54	2.33	0.03	0.86	1.73	4.31	2.01	1.49	1.66	2.84	2
Tx5	0.9	2.16	1.11	2.34	0.15	0.5	1.56	9.87	1.51	1.22	1.53	2.93	1.77
Tx6	1.1	3.33	0.54	3.29	0	1.58	2.98	3.48	3.38	2.41	1.47	3.8	3.12
Tx7	0.8	6.12	0.67	6.1	0.01	1.93	4.77	7.43	5.32	3.57	3.82	7.43	5.09
Tx8	0.7	2.36	0.52	2.15	0.17	1.82	2.24	3.16	2.71	2.04	2.22	2.7	2.36
TxEx1	0.3	9.04	2.18	8.81	0.19	2.68	6.21	9.54	7.07	4.74	7.48	11.3	7.09
TxEx2	0.6	13.1	3.66	13.5	0.36	2.32	8.73	12.4	9.42	6.2	9.82	17.4	9.87
TxEx3	0.7	7.38	3.3	7.88	0.89	1.36	4.67	12.5	4.51	3.48	5.29	10.3	5.46
TxEx4	0.1	9.49	4.54	8.66	10.4	5.16	6.3	7.94	5.69	4.73	7.72	11.6	6.7
znf1	0.4	0.65	0.36	0.98	0.05	0.09	0.42	0.16	0.35	0.42	0.34	1.05	0.57
znf2	0.2	5.7	1.02	7.38	0.18	0.2	1.26	0.12	0.84	1.1	1.69	11	2.16
DNase1	0.2	1.76	2.26	1.89	11.5	4.42	8.48	0.38	1.88	2.18	2.61	1	1.94
BivProm1	0.1	11.7	52.9	10.5	32.5	12.1	5.48	0.19	3.08	5.23	12.7	7.8	5.61
BivProm2	0.2	9.42	42.3	7.56	17.3	9.29	5.76	0.05	3.63	5.07	9.92	7.35	5.8
BivProm3	0.3	3.38	6.04	2.64	6.88	4.29	3.38	0.53	2.57	3.09	4.82	2.85	3.04
BivProm4	0.1	5.74	7.96	2.13	7.11	9.97	4.84	0.56	5.77	5.13	8.18	2.37	4.92
PromF1	0.2	2.58	5.63	1.8	21.8	4.84	3.19	1.91	1.61	1.99	3.81	2.17	1.81
PromF2	0.1	4.09	11.1	2.05	40.5	10.7	2.52	4.58	2.34	2.66	3.94	2.31	2.37
PromF3	0.2	5.61	35.9	4.12	62.6	16.1	2.2	11.5	1.91	2.63	6.38	3.15	2.78
PromF4	0.2	11.9	56.6	11.5	68.2	30.9	3.61	13.5	3.41	5.42	16.6	2.29	5.3
PromF5	0.1	10.3	49.1	9.24	51.5	16.9	4.36	1.01	2.98	5.05	14.7	5.6	4.91
PromF6	0.1	4.01	10.2	3.02	20.2	5.37	1.97	8.23	2.21	2.25	3.49	3.67	2.63
PromF7	0.2	1.77	3.73	1.44	3.38	1.72	1.05	6.9	1.45	1.41	1.97	1.69	1.49
TSS1	0.1	10.8	41	13.6	59.6	25.3	2.67	6.49	2.58	5.88	17.5	2.76	4.28
TSS2	0.1	7.96	22.6	14.1	39.7	15.9	3.71	1.59	1.8	5	9.9	3.88	3.61
	100	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.00	2.13	1.02	1.02	1.00

Supplementary Figure 2. 37: Enrichment of all full-stack states for top 1% bases prioritized by variant prioritization scores.
Extended version of figure **Figure 2.5C** showing the enrichment values of all full-stack states for genomic bases prioritized in the top 1% prioritized bases **(A)** in non-coding genome, and **(B)** genome-wide, by various variant prioritization scores. Coloring of enrichments is column specific. The second column in each heatmap, to the right of the state labels, shows the percentage of the background region (non-coding genome in **(A)** and whole genome in **(B)**) that each full-stack state covers. The last line in both heatmaps gives the actual percentage of the background region that is covered by each set of prioritized variants, which can differ from 1% exactly because of how ties of prioritization scores among bases were handled.

Top 5% prioritized variants, non-coding

state	% NC genome	CADD	CDTS	DANN	Eigen_PC	Eigen	FATHMM	FIRE	GERP	LINSIGHT	PhastCons	REMM	fitCons	funSeq2	phyloP
GapArtf1	4.19	0.32	0.13	1.57	0.11	0.12	0.48	0.18	0.12	0.13	0.57	0.08	0.48	0.03	0.4
GapArtf2	0.05	0.13	0.12	0.21	1.91	0.46	4.19	0.03	0.05	0.09	0.29	0.25	0.51	1.02	0.14
GapArtf3	0.01	0.39	0.34	0.42	9.02	2.83	5.02	0.03	0.09	0.19	0.51	0.72	1.04	2.53	0.23
Quies1	10.8	0.68	0.32	0.78	0.03	0.58	1	0.01	0.88	0.58	0.75	0.29	0.02	0.1	0.97
Quies2	3.38	0.59	0.36	0.54	0.04	0.49	0.82	0.01	0.68	0.54	0.6	0.41	0.08	0.12	0.75
Quies3	13.4	0.77	0.63	1.08	0.09	0.58	0.81	0.72	0.85	0.61	0.88	0.3	0.19	0.26	0.94
Quies4	4.79	0.5	0.33	1.8	0.02	0.1	0.3	0.2	0.18	0.13	0.55	0.06	0.06	0.04	0.35
Quies5	1.86	0.73	2.53	1.44	0.01	0.04	0.07	0.04	0.9	0.48	1.37	0.38	0.02	0.2	1.1
HET1	0.78	0.38	0.5	0.7	0.15	0.23	0.52	0.01	0.32	0.31	0.47	0.36	0.35	0.11	0.49
HET2	0.76	0.51	0.74	0.48	0.66	0.49	0.86	0.01	0.6	0.64	0.61	0.68	0.28	0.15	0.75
HET3	1.5	0.67	0.42	1.91	0.04	0.08	0.21	0.29	0.15	0.12	0.61	0.07	0.13	0.09	0.32
HET4	0.62	0.54	0.28	1.41	0.08	0.06	2.25	0.2	0.09	0.1	0.55	0.06	0.24	0.14	0.28
HET5	0.27	0.53	0.5	1.21	0.45	0.13	0.61	0.24	0.13	0.22	0.61	0.29	0.87	0.37	0.42
HET6	0.63	0.31	0.29	1.13	0.16	0.06	0.39	0.17	0.05	0.13	0.49	0.09	0.21	0.05	0.29
HET7	1.11	0.34	0.43	1.38	0.08	0.07	0.24	0.3	0.1	0.13	0.4	0.11	0.22	0.06	0.27
HET8	0.48	0.53	0.59	1	0.68	0.2	0.42	0.27	0.28	0.27	0.51	0.56	0.63	0.33	0.54
HET9	1.08	0.44	0.27	1.21	0.04	0.14	0.73	0.27	0.18	0.17	0.62	0.1	0.09	0.05	0.44
ReprPC1	0.2	2.55	5.34	1.13	7.87	3.79	2.76	0.34	2.37	3.33	2.15	3.69	5.48	1.91	1.96
ReprPC2	0.36	1.68	1.69	0.68	1.29	1.69	1.83	0.61	1.74	2.05	1.57	1.69	0.92	0.81	1.47
ReprPC3	1.22	1.17	0.92	0.71	0.22	0.97	1.26	0.55	1.17	1.15	1.16	0.97	0.32	0.36	1.16
ReprPC4	4.33	0.93	0.66	0.9	0.07	0.69	1.01	0.43	0.92	0.76	0.97	0.46	0.1	0.18	0.98
ReprPC5	0.68	0.7	1.31	0.64	1.98	0.71	0.96	0.42	0.75	0.88	0.73	1.27	0.7	0.46	0.9
ReprPC6	1.65	0.6	0.91	0.63	0.59	0.5	0.81	0.5	0.66	0.65	0.65	0.75	0.3	0.27	0.81
ReprPC7	0.65	0.86	1.5	0.71	3.99	1.12	1.06	0.68	0.92	1.26	0.83	1.85	1.01	1.08	1
ReprPC8	0.52	0.82	3.68	1.15	0.01	0.01	0.02	0.01	1.03	0.68	1.36	0.93	0.01	0.31	1.17
ReprPC9	0.41	1.07	1.14	0.8	0.44	0.89	1.12	0.9	1.13	1.09	1.09	0.85	0.55	0.48	1.08
Acet1	0.2	0.36	0.99	0.74	3.65	0.32	0.52	0.26	0.25	0.38	0.75	0.85	1.45	0.9	0.59
Acet2	0.94	0.73	0.85	0.54	0.36	0.59	0.89	0.39	0.82	0.75	0.74	1.11	0.57	0.42	0.88
Acet3	2.91	0.55	0.71	0.75	0.09	0.35	0.6	0.61	0.58	0.43	0.56	0.41	0.2	0.2	0.68
Acet4	0.44	0.75	1.03	0.5	1.56	0.74	0.78	0.53	0.91	0.92	0.69	1.95	1.03	1.51	0.87
Acet5	0.95	0.78	0.68	0.57	0.27	0.65	0.83	0.41	0.95	0.76	0.76	0.84	0.33	0.47	0.93
Acet6	0.47	1.23	0.79	0.49	1.44	1.26	1.19	0.38	1.46	1.41	1.1	1.93	0.98	1.49	1.19
Acet7	0.31	1.48	1.25	0.59	7.51	2.77	1.16	0.54	1.53	2.09	1.16	4.08	3.1	4.93	1.19
Acet8	0.62	0.58	0.92	0.7	0.53	0.45	0.65	0.49	0.64	0.59	0.56	0.91	0.51	0.53	0.69
EnhWk1	1.7	1.6	0.68	0.69	0.22	1.5	1.64	0.33	1.81	1.66	1.57	1.11	0.35	0.64	1.51
EnhWk2	0.38	1.42	1.68	0.75	5.21	2.11	1.11	1.34	1.45	2.15	1.14	3.35	2.12	3.63	1.2
EnhWk3	0.91	2.04	1.03	0.54	0.7	2.02	1.89	0.57	2.1	2.42	1.81	2.31	0.76	1.44	1.6
EnhWk4	2.44	2.46	0.71	0.55	0.14	2.44	2.57	0.25	2.6	2.6	2.29	1.66	0.21	0.69	1.97
EnhWk5	1.09	0.97	0.85	0.82	0.23	0.8	0.95	0.72	1.07	0.91	1	0.65	0.44	0.5	1.03
EnhWk6	0.64	0.99	1.14	0.77	1.45	0.89	0.77	1.16	1.09	1.09	0.95	1.23	1.4	1.99	1.02
EnhWk7	0.53	1.01	0.92	0.81	2.82	0.96	0.86	0.87	1.02	0.96	0.96	1.51	2.57	1.23	1
EnhWk8	1.51	1.74	0.76	0.72	0.17	1.61	1.85	0.22	1.89	1.75	1.68	1.32	0.81	0.55	1.53
EnhA1	0.19	2.94	2.32	1.09	13.8	7.27	1.88	1.53	2.48	4.09	2.13	6.79	6.4	8.17	1.83
EnhA2	0.36	3.88	1.42	0.68	3.52	4.41	3.38	0.47	3.52	4.88	3.26	5.13	3.11	3.74	2.54
EnhA3	0.21	3.74	1.66	0.8	12	6.45	2.58	0.63	3.25	4.88	2.95	5.34	6.33	7.79	2.34
EnhA4	0.33	2.73	1	0.62	8.45	3.96	2.19	0.27	2.6	3.48	2.29	3.57	4.43	4.57	1.92
EnhA5	0.79	2.39	0.87	0.68	2.52	2.54	2.12	0.34	2.37	2.77	2.16	2.36	2.47	1.91	1.85
EnhA6	0.62	1.95	1.15	0.56	0.83	1.97	1.77	0.68	2.11	2.35	1.75	2.03	0.67	1.62	1.62
EnhA7	0.42	0.8	1.49	0.64	4.15	1.01	0.9	0.96	0.82	1.02	0.73	2.28	2.53	2.11	0.89
EnhA8	0.28	1.38	1.83	0.73	8.16	2.51	1.15	1.37	1.37	1.99	1.11	3.84	3.31	5.3	1.14
EnhA9	0.18	1.8	2.12	0.86	11.6	4.47	1.04	2.34	1.6	2.57	1.28	5.55	4.78	8.59	1.28
EnhA10	0.43	1.7	1.7	0.76	5.19	2.2	1.35	1.36	1.64	2.11	1.44	2.97	3.63	3.68	1.36
EnhA11	0.79	0.84	1.02	0.85	0.59	0.65	0.78	1.13	0.85	0.78	0.84	0.77	0.88	0.68	0.9
EnhA12	0.36	1.29	1.19	0.67	5.56	1.79	1.14	0.63	1.29	1.69	1.07	3.19	2.33	3.33	1.1
EnhA13	0.84	1.12	0.79	0.67	0.97	1.02	1.1	0.47	1.26	1.13	1.07	1.25	0.56	0.88	1.14
EnhA14	0.4	1.23	1.34	0.68	3.79	1.6	1.12	1.19	1.29	1.63	1.06	2.17	1.52	3.35	1.1
EnhA15	1.12	1.08	0.88	0.77	0.48	0.92	1.06	0.87	1.19	1.09	1.05	0.76	0.37	0.86	1.09
EnhA16	0.71	1.01	1.22	0.94	1.63	0.89	0.97	0.9	0.99	0.95	1.03	1.21	0.7	1.04	1.05
EnhA17	0.58	3.53	1.19	0.62	1.5	3.67	3.32	0.38	3.36	4.09	3.12	3.66	2.83	2.57	2.51
EnhA18	0.5	1.35	1.1	0.85	0.84	1.18	1.36	0.35	1.29	1.39	1.23	1.85	3.61	1.39	1.11
EnhA19	0.28	2.24	1.85	0.81	5.47	2.84	1.94	0.94	2.02	2.85	1.82	4.23	5.68	5.19	1.59
EnhA20	0.38	1.56	0.77	0.63	0.92	1.58	1.61	0.29	1.82	1.7	1.51	1.49	3.7	1.02	1.51
TxEnh1	0.43	1.75	1.16	0.69	1.29	1.82	1.01	0.67	1.99	2.18	1.58	2.12	1.53	3.58	1.58
TxEnh2	0.43	1.16	1.4	0.78	1.76	1.08	0.48	1.85	1.19	1.34	1.06	2.66	2.19	4.36	1.22
TxEnh3	0.27	0.65	1.56	0.84	1.6	0.51	0.43	2.78	0.54	0.67	0.68	2.02	2.2	3.61	0.78
TxEnh4	0.25	1.48	2.85	0.83	5.35	2.26	1.83	8.07	1.54	2.38	1.14	3.44	7.4	5.95	1.25
TxEnh5	0.49	1.02	2.18	0.79	3.88	1.45	1.27	4.1	1.18	1.61	0.9	2.28	3.92	3.07	1.05
TxEnh6	0.19	1.18	2.08	0.75	4.8	1.77	1.21	4.8	1.3	1.8	0.95	3.02	4.72	5.27	1.1
TxEnh7	0.3	1.24	1.71	0.7	5.58	1.93	0.65	1.37	1.24	1.76	0.96	4.04	3.02	5.67	1.07
TxEnh8	0.25	1.68	1.76	0.71	4.96	2.45	1.53	3.87	1.82	2.54	1.34	3.36	4.56	5.94	1.39
TxWk1	3.03	0.91	1.19	1.08	0.13	0.74	0.87	3.29	1.06	0.91	1.02	0.49	1.25	0.94	1.09
TxWk2	0.86	0.7	1.64	0.75	1.3	0.69	0.87	1.99	0.87	0.96	0.72	1.14	1.72	1.12	0.89
Tx1	0.91	0.81	1.17	1.12	0.23	0.47	0.28	1.38	0.67	0.56	0.92	0.7	0.7	1.81	1.01
Tx2	1.74	0.81	1.02	1.09	0.14	0.57	0.38	1.32	0.84	0.66	0.9	0.47	0.39	1.09	0.96
Tx3	0.56	0.6	1.23	0.7	0.89	0.48	0.32	1.07	0.68	0.64	0.57	1.38	0.88	1.93	0.73
Tx4	0.5	0.96	1.33	0.99	0.29	0.7	0.75	4.16	0.97	0.97	0.98	0.9	2.21	2.23	1.14
Tx5	1	0.55	1.59	0.96	0.38	0.37	0.57	6.86	0.5	0.52	0.59	0.51	2.22	0.88	0.61
Tx6	1.17	1.42	1.42	1	0.17	1.3	1.66	4.61	1.81	1.67	1.49	0.79	2.65	2	1.71
Tx7	0.83	1.42	1.68	0.84	0.33	1.31	1.83	6.97	1.77	1.77	1.34	1.06	3.94	2.68	1.61
Tx8	0.73	1.51	1.43	0.75	1.11	1.52	1.41	3.25	1.74	1.94	1.38	1.67	2.7	2.81	1.45
TxEx1	0.26	1.33	1.87	0.86	0.98	1.21	1.45	6.9	1.37	1.68	1.21	2.02	5.95	4.35	1.36
TxEx2	0.5	1.24	2.87	0.76	1.26	1.22	1.96	8.42	1.44	1.81	1.04	1.83	6.57	3.32	1.22
TxEx3	0.65	0.81	2.29	0.85	1.76	0.81	1.13	7.64	0.86	1.11	0.76	1.27	4.54	1.9	0.86
TxEx4	0.1	2	2.67	0.96	7.73	3.9	1.68	5.77	1.78	2.93	1.47	6.15	7.51	9.25	1.48
znf1	0.44	0.67	0.71	1.71	0.22	0.12	0.24	1.23	0.18	0.21	0.76	0.26	1.17	0.69	0.45
znf2	0.15	0.62	0.83	1.33	0.55	0.23	0.69	1.18	0.26	0.4	0.75	0.5	3.38	1.19	0.62
DNase1	0.22	2.16	1.99	1.43	11.7	4.68	3.31	0.65	1.98	2.95	1.9	2.66	4.25	5.5	1.68
BivProm1	0.14	4.9	13.4	5.46	16.6	9.6	3.24	0.78	2.69	5.36	2.88	10.5	17.3	7.45	2.33
BivProm2	0.16	3.88	11.9	3.72	14.8	7.12	3.29	0.48	2.67	4.69	2.72	7.79	15	4.87	2.28
BivProm3	0.3	2.5	3.32	1.32	7.32	3.57	2.43	0.87	2.12	3.3	2.13	4.16	5.24	2.77	1.83
BivProm4	0.14	4.51	4.58	1.42	7.8	6.33	3.71	0.84	3.88	5.76	3.61	7.17	7	5.87	2.89
PromF1	0.22	2.13	3.47	1.26	12.9	5.55	1.98	1.94	1.67	3.01	1.59	5.37	8.38	6.76	1.44
PromF2	0.15	3.16	5.86	1.54	15.9	9.78	1.96	4.16	2.22	4.1	2.07	7.79	9.74	10.3	1.77
PromF3	0.16	2.96	11.6	2.51	18.5	13.5	1.24	8.02	1.75	3.88	1.59	9.18	6.83	13.5	1.49
PromF4	0.19	4.87	13.8	5.63	19.4	15.9	2.15	9.21	2.77	6.11	2.91	12.5	1.89	14.4	2.27
PromF5	0.14	5.13	12.8	5.58	17.6	12.4	2.77	1.95	2.85	5.86	3.1	11.7	15.5	9.86	2.45
PromF6	0.14	2.28	5.91	1.06	10.6	6.12	0.93	6.59	1.78	3.32	1.42	5.69	5.97	9.39	1.45
PromF7	0.17	1.69	3.33	0.88	3.18	2.22	0.82	5.37	1.63	2.44	1.28	4.11	4.18	5.01	1.41
TSS1	0.13	6.36	10.1	7.68	18.1	13.3	2.84	5.74	3.21	7.19	4.26	12.7	6.44	12	2.98
TSS2	0.12	5.67	6.24	7.48	14.9	9.56	3.11	2.07	2.44	6.21	3.94	9.83	11.8	7.89	2.64
	100	4.96	4.97	4.95	4.96	4.97	4.96	4.98	4.98	4.97	4.96	5.00	5.02	4.96	4.97

Top 5% prioritized variants, whole-genome

state	% genome	CADD	CDTS	DANN	Eigen_PC	Eigen	FATHMM	FIRE	GERP	PhastCons	REMM	fitCons	phyloP
GapArtf1	12	0.12	0.04	0.51	0.04	0.04	0.16	0.07	0.04	0.22	0.03	0.19	0.15
GapArtf2	0.1	0.15	0.14	0.23	1.79	0.42	3.13	0.04	0.06	0.35	0.25	0.46	0.15
GapArtf3	0	0.44	0.4	0.48	8.38	2.57	4	0.04	0.18	0.7	0.75	0.95	0.32
Quies1	9.9	0.62	0.33	0.71	0.03	0.57	0.95	0.01	0.84	0.7	0.28	0.03	0.93
Quies2	3.1	0.57	0.38	0.51	0.05	0.5	0.8	0.01	0.67	0.58	0.4	0.12	0.74
Quies3	12	0.72	0.65	1.02	0.11	0.59	0.8	0.77	0.83	0.83	0.3	0.24	0.92
Quies4	4.5	0.42	0.34	1.67	0.03	0.1	0.29	0.22	0.17	0.47	0.06	0.07	0.34
Quies5	1.7	0.67	2.58	1.36	0.01	0.04	0.07	0.04	0.88	1.25	0.38	0.02	1.07
HET1	0.7	0.41	0.53	0.71	0.18	0.27	0.55	0.01	0.36	0.49	0.39	0.44	0.52
HET2	0.7	0.63	0.81	0.58	0.74	0.57	0.94	0.01	0.69	0.7	0.75	0.5	0.82
HET3	1.4	0.56	0.44	1.81	0.04	0.09	0.2	0.32	0.14	0.49	0.07	0.15	0.31
HET4	0.6	0.47	0.3	1.36	0.09	0.07	2	0.23	0.09	0.46	0.06	0.28	0.28
HET5	0.2	0.73	0.65	1.39	0.53	0.23	0.71	0.3	0.26	0.65	0.43	1.22	0.52
HET6	0.6	0.37	0.31	1.15	0.21	0.09	0.41	0.19	0.09	0.49	0.14	0.4	0.33
HET7	1	0.33	0.45	1.33	0.1	0.08	0.25	0.33	0.11	0.38	0.13	0.28	0.29
HET8	0.4	0.57	0.64	1.03	0.75	0.25	0.46	0.3	0.32	0.52	0.57	0.79	0.57
HET9	1	0.42	0.28	1.18	0.05	0.14	0.69	0.29	0.19	0.57	0.11	0.16	0.45
ReprPC1	0.2	2.86	5.62	1.48	7.94	3.93	2.98	0.39	2.61	2.46	3.85	4.11	2.21
ReprPC2	0.3	1.7	1.79	0.74	1.36	1.74	1.87	0.67	1.77	1.62	1.7	0.91	1.51
ReprPC3	1.1	1.14	0.97	0.72	0.25	0.99	1.26	0.59	1.17	1.14	0.96	0.42	1.16
ReprPC4	3.9	0.88	0.69	0.86	0.09	0.71	1	0.47	0.91	0.94	0.47	0.17	0.98
ReprPC5	0.6	0.99	1.44	0.9	2.11	0.87	1.15	0.47	0.95	0.96	1.43	1.11	1.08
ReprPC6	1.5	0.76	0.99	0.77	0.67	0.61	0.92	0.54	0.77	0.78	0.85	0.6	0.91
ReprPC7	0.6	1.42	1.7	1.24	4.17	1.41	1.47	0.77	1.32	1.28	2.19	1.72	1.36
ReprPC8	0.5	0.88	3.81	1.21	0.02	0.01	0.02	0.01	1.07	1.35	0.97	0.02	1.22
ReprPC9	0.4	1.05	1.2	0.81	0.49	0.92	1.15	0.96	1.13	1.1	0.85	0.57	1.09
Acet1	0.2	0.75	1.22	1.1	3.84	0.57	0.83	0.29	0.56	1.04	1.15	1.84	0.87
Acet2	0.9	0.78	0.9	0.6	0.41	0.66	0.95	0.42	0.88	0.8	1.12	0.72	0.93
Acet3	2.6	0.54	0.75	0.73	0.12	0.38	0.61	0.65	0.59	0.56	0.42	0.27	0.69
Acet4	0.4	0.81	1.09	0.56	1.64	0.81	0.85	0.57	0.96	0.75	1.88	1.14	0.92
Acet5	0.9	0.75	0.72	0.55	0.3	0.67	0.82	0.45	0.94	0.74	0.81	0.38	0.92
Acet6	0.4	1.18	0.82	0.48	1.49	1.28	1.18	0.42	1.44	1.09	1.83	1.02	1.18
Acet7	0.3	1.47	1.31	0.61	7.61	2.79	1.19	0.59	1.54	1.19	3.83	3.04	1.21
Acet8	0.6	0.6	0.98	0.71	0.57	0.48	0.67	0.53	0.66	0.59	0.89	0.59	0.72
EnhWk1	1.5	1.5	0.71	0.64	0.24	1.5	1.6	0.36	1.76	1.51	1.08	0.38	1.47
EnhWk2	0.4	1.62	1.83	0.95	5.31	2.23	1.31	1.48	1.61	1.33	3.36	2.16	1.36
EnhWk3	0.8	1.98	1.09	0.54	0.75	2.04	1.89	0.61	2.08	1.8	2.24	0.82	1.6
EnhWk4	2.2	2.33	0.74	0.51	0.16	2.44	2.51	0.27	2.53	2.22	1.62	0.25	1.93
EnhWk5	1	0.92	0.89	0.79	0.25	0.81	0.94	0.76	1.05	0.97	0.64	0.47	1.02
EnhWk6	0.6	0.97	1.2	0.76	1.52	0.92	0.79	1.24	1.11	0.95	1.2	1.34	1.04
EnhWk7	0.5	0.96	0.96	0.78	2.94	0.99	0.87	0.94	1.02	0.94	1.44	2.39	1.01
EnhWk8	1.4	1.65	0.79	0.68	0.19	1.61	1.81	0.24	1.84	1.62	1.28	0.81	1.5
EnhA1	0.2	2.97	2.48	1.16	13.8	7.25	1.98	1.68	2.54	2.22	6.49	5.74	1.91
EnhA2	0.3	3.75	1.48	0.66	3.61	4.42	3.36	0.51	3.48	3.22	4.95	3.05	2.53
EnhA3	0.2	3.59	1.73	0.75	12.1	6.44	2.57	0.69	3.21	2.91	5.07	6.01	2.33
EnhA4	0.3	2.61	1.05	0.58	8.57	3.95	2.17	0.31	2.56	2.25	3.38	4.38	1.9
EnhA5	0.7	2.27	0.91	0.64	2.59	2.54	2.08	0.38	2.33	2.1	2.27	2.45	1.82
EnhA6	0.6	1.94	1.23	0.61	0.89	2.02	1.82	0.73	2.13	1.79	2.03	0.78	1.66
EnhA7	0.4	1.06	1.62	0.9	4.27	1.14	1.07	1.05	1	0.92	2.33	2.62	1.04
EnhA8	0.3	1.52	1.96	0.88	8.31	2.61	1.27	1.48	1.49	1.22	3.77	3.17	1.24
EnhA9	0.2	1.87	2.26	0.95	11.7	4.5	1.15	2.53	1.69	1.38	5.29	4.29	1.36
EnhA10	0.4	1.63	1.77	0.73	5.28	2.2	1.35	1.45	1.63	1.42	2.81	3.29	1.35
EnhA11	0.7	0.81	1.06	0.82	0.62	0.66	0.79	1.21	0.85	0.81	0.75	0.89	0.9
EnhA12	0.3	1.34	1.26	0.74	5.64	1.84	1.2	0.69	1.35	1.13	3.06	2.41	1.15
EnhA13	0.8	1.06	0.82	0.64	1.01	1.03	1.09	0.51	1.24	1.04	1.19	0.6	1.12
EnhA14	0.4	1.28	1.42	0.76	3.87	1.65	1.2	1.29	1.34	1.12	2.13	1.6	1.16
EnhA15	1	1.04	0.92	0.75	0.51	0.94	1.07	0.94	1.18	1.03	0.76	0.45	1.09
EnhA16	0.6	0.97	1.27	0.92	1.68	0.91	0.98	0.96	0.99	1.01	1.16	0.73	1.05
EnhA17	0.5	3.38	1.24	0.58	1.56	3.66	3.28	0.42	3.3	3.05	3.54	2.67	2.47
EnhA18	0.5	1.32	1.15	0.84	0.89	1.21	1.36	0.38	1.3	1.23	1.79	3.32	1.13
EnhA19	0.3	2.25	1.96	0.88	5.57	2.89	2.01	1.03	2.07	1.89	4.09	5.02	1.66
EnhA20	0.3	1.48	0.81	0.59	0.96	1.59	1.58	0.32	1.78	1.47	1.43	3.62	1.49
TxEnh1	0.4	1.67	1.24	0.66	1.36	1.86	1.04	0.74	1.98	1.57	2.04	1.44	1.59
TxEnh2	0.4	1.1	1.47	0.74	1.83	1.12	0.52	1.97	1.2	1.05	2.5	1.92	1.23
TxEnh3	0.2	0.84	1.72	1.01	1.74	0.67	0.61	3.01	0.73	0.87	2.09	2.04	0.95
TxEnh4	0.3	3.84	3.74	3.15	6.01	3.59	3.66	8.18	3.42	3.35	5.19	8.04	3.03
TxEnh5	0.5	2.52	2.75	2.25	4.38	2.29	2.43	4.29	2.34	2.23	3.42	5	2.13
TxEnh6	0.2	2.12	2.45	1.69	5.08	2.27	1.91	5.02	2	1.78	3.59	5.15	1.77
TxEnh7	0.3	1.31	1.84	0.77	5.69	2	0.76	1.51	1.32	1.06	3.88	2.5	1.15
TxEnh8	0.2	2.39	2.01	1.45	5.29	2.9	2.15	4.02	2.4	2.06	3.79	4.88	1.97
TxWk1	2.8	1.05	1.26	1.2	0.24	0.87	1.02	3.45	1.19	1.17	0.64	1.46	1.22
TxWk2	0.8	1.75	2.01	1.75	1.76	1.31	1.7	2.07	1.69	1.65	1.98	2.84	1.65
Tx1	0.8	0.73	1.23	1.05	0.26	0.49	0.3	1.52	0.67	0.87	0.67	0.67	1
Tx2	1.6	0.74	1.07	1.02	0.16	0.58	0.39	1.44	0.83	0.85	0.47	0.4	0.95
Tx3	0.5	0.58	1.29	0.67	0.93	0.5	0.35	1.17	0.69	0.58	1.3	0.81	0.74
Tx4	0.5	1.32	1.47	1.35	0.52	1.01	1.08	4.39	1.31	1.35	1.24	2.44	1.44
Tx5	0.9	0.97	1.75	1.33	0.59	0.65	0.9	7	0.86	0.98	0.88	2.42	0.94
Tx6	1.1	1.92	1.54	1.52	0.49	1.69	2.08	4.81	2.24	2.01	1.27	3.24	2.12
Tx7	0.8	2.51	1.89	1.95	0.97	2.12	2.7	7.07	2.67	2.45	2.05	4.87	2.46
Tx8	0.7	1.81	1.54	1.09	1.32	1.77	1.7	3.42	2.01	1.71	1.92	2.94	1.71
TxEx1	0.3	3.1	2.51	2.56	1.91	2.41	2.83	6.9	2.83	2.9	3.52	6.38	2.7
TxEx2	0.6	3.94	3.68	3.4	2.57	2.97	3.98	8.4	3.62	3.65	4.13	7.81	3.29
TxEx3	0.7	2.4	2.88	2.37	2.44	1.74	2.31	7.69	2.13	2.22	2.6	5.3	2.05
TxEx4	0.1	3.73	3.45	2.66	8.09	4.71	3.03	5.97	3.1	3.05	7.01	8.31	2.74
znf1	0.4	0.72	0.8	1.76	0.28	0.2	0.33	1.37	0.27	0.75	0.34	1.29	0.53
znf2	0.2	2.14	1.3	2.81	0.75	0.65	1.17	1.28	0.79	1.2	1.28	4.93	1.15
DNase1	0.2	2.17	2.07	1.48	11.8	4.66	3.36	0.72	2.02	1.95	2.63	4.11	1.73
BivProm1	0.1	5.92	13.5	6.41	15.9	9.44	3.92	0.86	3.45	3.83	10.8	13.3	3.1
BivProm2	0.2	4.8	12.1	4.61	14.4	7.21	3.88	0.54	3.33	3.54	8.16	10.8	2.96
BivProm3	0.3	2.79	3.55	1.65	7.39	3.68	2.63	0.95	2.32	2.39	4.27	4.26	2.05
BivProm4	0.1	4.63	4.82	1.67	7.9	6.37	3.86	0.93	4.03	3.8	7.11	4.97	3.06
PromF1	0.2	2.32	3.67	1.45	12.8	5.56	2.15	2.14	1.82	1.76	5.23	7.15	1.59
PromF2	0.1	3.31	6.12	1.74	15.7	9.7	2.15	4.47	2.38	2.25	7.53	7.44	1.93
PromF3	0.2	3.47	11.9	3.01	18.1	13.2	1.71	8.41	2.17	2.11	9.11	5.48	1.91
PromF4	0.2	5.99	14.1	6.63	18.3	15.1	2.91	9.4	3.59	4	12.7	3.05	3.12
PromF5	0.1	5.86	12.9	6.26	17	12	3.32	2.14	3.4	3.83	11.9	12.1	3.04
PromF6	0.1	2.72	6.28	1.55	10.7	6.27	1.36	6.96	2.14	1.89	5.75	4.4	1.82
PromF7	0.2	1.82	3.51	1.06	3.32	2.35	1.01	5.64	1.78	1.47	4.03	3.36	1.56
TSS1	0.1	6.81	10.1	8.06	17.3	12.9	3.03	6.09	3.47	4.71	12.8	5.57	3.32
TSS2	0.1	6.02	6.16	7.76	14.4	9.26	3.23	2.2	2.62	4.27	9.88	9.16	2.91
	100	5.00	5.00	5.00	5.00	4.98	5.00	5.01	5.02	5.00	5.03	5.03	5.00

Supplementary Figure 2. 38: Enrichment of all full-stack states for top 5% bases prioritized by variant prioritization scores.
Analogous to **Supplementary Figure 2.37**, but for top 5% prioritized bases.

Top 10% prioritized variants, non-coding

state	% NC genome	CADD	CDTS	DANN	Eigen_PC	Eigen	FATHMM	FIRE	GERP	LINSIGHT	PhastCons	REMM	fitCons	funSeq2	phyloP
GapArtf1	4.19	0.53	0.14	1.21	0.14	0.14	0.48	0.3	0.17	0.17	1.26	0.1	0.24	0.12	0.45
GapArtf2	0.05	0.21	0.09	0.22	2.17	0.41	3.12	0.17	0.06	0.18	0.3	0.31	0.26	0.68	0.15
GapArtf3	0.01	0.91	0.23	0.35	6.86	2.23	3.72	0.13	0.08	0.37	0.43	0.75	0.5	1.47	0.25
Quies1	10.8	0.7	0.46	0.82	0.09	0.61	1.29	0.02	0.95	0.57	0.78	0.25	0.01	0.16	1.03
Quies2	3.38	0.59	0.48	0.61	0.14	0.48	1.06	0.02	0.76	0.51	0.59	0.44	0.04	0.16	0.8
Quies3	13.4	0.87	0.76	1.06	0.18	0.69	0.93	0.8	0.92	0.68	0.96	0.3	0.23	0.45	0.99
Quies4	4.79	0.78	0.42	1.37	0.06	0.17	0.4	0.28	0.25	0.16	1.34	0.07	0.06	0.11	0.44
Quies5	1.86	0.86	2.01	1.24	0.01	0.04	0.09	0.04	0.98	0.59	1.56	0.4	0.01	0.24	1.13
HET1	0.78	0.48	0.56	0.71	0.32	0.24	0.66	0.02	0.39	0.31	0.61	0.5	0.17	0.15	0.52
HET2	0.76	0.47	0.75	0.64	1.07	0.48	1.07	0.02	0.66	0.65	0.48	0.86	0.15	0.23	0.81
HET3	1.5	1.01	0.55	1.45	0.14	0.19	0.3	0.42	0.23	0.17	1.5	0.11	0.16	0.21	0.44
HET4	0.62	0.82	0.39	1.08	0.4	0.16	1.77	0.39	0.15	0.15	1.25	0.1	0.2	0.23	0.36
HET5	0.27	0.8	0.56	1.05	1.02	0.23	0.54	0.58	0.21	0.3	1.08	0.47	0.63	0.46	0.48
HET6	0.63	0.53	0.31	0.94	0.28	0.1	0.49	0.23	0.08	0.17	0.95	0.16	0.13	0.09	0.3
HET7	1.11	0.55	0.48	1.14	0.14	0.11	0.29	0.41	0.15	0.15	0.92	0.18	0.18	0.16	0.31
HET8	0.48	0.74	0.69	0.98	0.94	0.33	0.57	0.41	0.42	0.38	0.82	0.87	0.37	0.39	0.65
HET9	1.08	0.67	0.33	0.98	0.13	0.18	0.73	0.31	0.25	0.22	1.17	0.13	0.07	0.08	0.5
ReprPC1	0.2	1.94	3.64	1.33	5.65	3.58	2.31	0.64	1.92	2.91	1.33	3.29	2.58	1.5	1.58
ReprPC2	0.36	1.43	1.63	0.89	1.46	1.66	1.69	0.84	1.54	1.96	1.13	1.67	0.47	0.81	1.34
ReprPC3	1.22	1.12	1.07	0.84	0.5	0.99	1.35	0.7	1.18	1.14	1	1.07	0.19	0.43	1.16
ReprPC4	4.33	0.98	0.82	0.93	0.19	0.76	1.16	0.54	0.98	0.8	0.99	0.49	0.06	0.27	1.03
ReprPC5	0.68	0.7	1.28	0.84	2.15	0.82	1.13	0.51	0.87	0.94	0.58	1.58	0.36	0.51	0.94
ReprPC6	1.65	0.65	0.99	0.8	0.82	0.57	0.98	0.57	0.79	0.71	0.59	0.99	0.17	0.38	0.86
ReprPC7	0.65	0.78	1.41	0.95	3.72	1.28	1.06	0.73	0.99	1.31	0.58	2.13	0.54	1.08	0.99
ReprPC8	0.52	0.94	2.76	1.13	0.02	0.01	0.03	0.03	1.14	0.84	1.39	1.16	0.01	0.36	1.18
ReprPC9	0.41	1.04	1.16	0.89	0.57	0.96	1.18	0.93	1.12	1.16	0.96	0.92	0.32	0.57	1.07
Acet1	0.2	0.5	0.88	0.77	3.15	0.48	0.54	0.4	0.33	0.61	0.76	1.19	0.88	0.9	0.67
Acet2	0.94	0.75	0.96	0.69	0.68	0.65	1.04	0.46	0.93	0.78	0.66	1.4	0.31	0.51	0.89
Acet3	2.91	0.65	0.83	0.81	0.21	0.44	0.77	0.66	0.7	0.49	0.66	0.56	0.14	0.32	0.75
Acet4	0.44	0.76	1.12	0.71	1.91	0.93	0.9	0.62	1.03	1.02	0.57	2.4	0.75	1.33	0.91
Acet5	0.95	0.81	0.84	0.71	0.51	0.75	0.99	0.49	1.05	0.81	0.69	1.07	0.27	0.6	0.96
Acet6	0.47	1.13	0.96	0.69	2.19	1.47	1.23	0.53	1.44	1.42	0.86	2.35	0.98	1.36	1.18
Acet7	0.31	1.3	1.31	0.87	6.43	3	1.13	0.7	1.49	2.1	0.83	4.19	2.72	3.41	1.15
Acet8	0.62	0.65	0.98	0.82	0.76	0.54	0.8	0.55	0.76	0.65	0.59	1.2	0.34	0.55	0.74
EnhWk1	1.7	1.38	0.85	0.85	0.6	1.49	1.64	0.45	1.66	1.57	1.21	1.07	0.27	0.77	1.43
EnhWk2	0.38	1.21	1.62	1.03	4.41	2.22	1	1.33	1.43	2.11	0.77	3.35	1.61	2.87	1.14
EnhWk3	0.91	1.59	1.13	0.79	1.23	1.83	1.65	0.64	1.81	2.1	1.22	2.19	0.56	1.35	1.43
EnhWk4	2.44	1.84	0.88	0.78	0.53	2.1	2.18	0.31	2.12	2.17	1.52	1.34	0.16	0.83	1.71
EnhWk5	1.09	0.97	0.93	0.9	0.46	0.86	1.01	0.78	1.08	0.95	0.92	0.7	0.4	0.69	1.04
EnhWk6	0.64	1.03	1.22	0.89	1.71	1.1	0.78	1.31	1.14	1.23	0.86	1.51	1.46	1.91	1.05
EnhWk7	0.53	1.04	1.04	0.9	2.96	1.24	0.94	1.02	1.07	1.07	0.92	1.84	2.09	1.32	1.03
EnhWk8	1.51	1.42	0.89	0.84	0.52	1.47	1.68	0.3	1.63	1.52	1.32	1.19	0.45	0.66	1.39
EnhA1	0.19	2.24	2.11	1.37	8.95	5.85	1.45	1.59	2.09	3.55	1.35	5.5	5.53	5.56	1.56
EnhA2	0.36	2.71	1.46	1.03	4.25	3.7	2.53	0.63	2.66	3.78	2	4.22	2.16	2.84	2.04
EnhA3	0.21	2.77	1.68	1.15	8.57	5.5	1.89	0.94	2.55	4.06	1.89	4.84	5.91	5.27	1.95
EnhA4	0.33	2.14	1.17	0.91	7.47	3.95	1.86	0.5	2.18	3.04	1.55	3.64	4.18	3.24	1.7
EnhA5	0.79	1.9	1.03	0.91	3.47	2.48	1.87	0.53	2.02	2.43	1.52	2.32	1.85	1.67	1.66
EnhA6	0.62	1.54	1.26	0.81	1.25	1.82	1.43	0.77	1.82	2.07	1.18	1.92	0.59	1.62	1.44
EnhA7	0.42	0.81	1.43	0.83	3.64	1.28	0.94	1.04	0.94	1.21	0.61	2.58	1.76	1.83	0.92
EnhA8	0.28	1.27	1.71	0.93	5.88	2.82	1.04	1.53	1.35	2.17	0.84	3.79	2.86	3.9	1.11
EnhA9	0.18	1.59	2	1.11	7.83	4.28	0.87	2.23	1.55	2.68	0.89	4.99	4.4	6	1.21
EnhA10	0.43	1.51	1.62	0.96	4.76	2.42	1.22	1.49	1.54	2.13	1.09	3.18	2.92	2.82	1.29
EnhA11	0.79	0.91	1.09	0.93	0.92	0.78	0.9	1.2	0.93	0.9	0.84	0.96	0.7	0.8	0.95
EnhA12	0.36	1.19	1.25	0.89	4.91	2.12	1.15	0.78	1.3	1.74	0.85	3.43	1.38	2.42	1.08
EnhA13	0.84	1.08	0.92	0.81	1.45	1.16	1.21	0.61	1.28	1.18	0.94	1.5	0.37	0.83	1.14
EnhA14	0.4	1.14	1.38	0.88	3.57	1.78	1.07	1.29	1.3	1.72	0.85	2.38	1.05	2.5	1.08
EnhA15	1.12	1.05	1	0.87	0.84	1	1.1	0.97	1.19	1.12	0.94	0.83	0.31	0.9	1.09
EnhA16	0.71	1.06	1.22	0.97	2.33	1.09	1	1.01	1.04	1.08	1.01	1.48	0.42	1.04	1.07
EnhA17	0.58	2.5	1.3	0.91	2.51	3.03	2.43	0.53	2.54	3.18	1.98	3.03	1.71	2.2	2.01
EnhA18	0.5	1.15	1.1	0.92	1.29	1.12	1.28	0.44	1.16	1.23	1.09	1.95	2.05	1.23	1.02
EnhA19	0.28	1.73	1.71	1.06	4.89	2.71	1.56	1.03	1.72	2.47	1.25	3.89	3.88	3.8	1.38
EnhA20	0.38	1.37	0.92	0.81	1.95	1.71	1.56	0.45	1.67	1.63	1.15	1.64	3.15	1.03	1.44
TxEnh1	0.43	1.59	1.34	0.92	2.25	2.04	0.77	0.99	1.85	2.21	1.18	2.39	1.89	3.24	1.5
TxEnh2	0.43	1.26	1.5	0.95	2.58	1.44	0.4	1.95	1.33	1.63	0.93	3.13	2.13	4.12	1.24
TxEnh3	0.27	0.78	1.56	0.95	1.88	0.69	0.32	2.66	0.68	0.9	0.68	2.43	3.4	3.71	0.83
TxEnh4	0.25	1.27	2.43	1.01	4.61	2.32	1.31	5.85	1.39	2.4	0.78	3.42	6.02	4.88	1.14
TxEnh5	0.49	0.95	1.94	0.97	3.35	1.59	0.93	3.36	1.16	1.68	0.66	2.49	3.99	2.9	1.01
TxEnh6	0.19	1.14	1.94	0.95	4.09	2.02	0.9	3.94	1.28	1.96	0.72	3.15	4.68	4.26	1.07
TxEnh7	0.3	1.13	1.71	0.97	4.93	2.19	0.5	1.45	1.3	1.87	0.7	4.04	2.87	4.64	1.06
TxEnh8	0.25	1.49	1.78	0.95	4.87	2.66	1.1	3.6	1.65	2.5	0.95	3.57	5.22	4.62	1.3
TxWk1	3.03	1.02	1.27	1.1	0.29	0.87	0.66	3.12	1.06	1	1.01	0.52	2.66	1.48	1.09
TxWk2	0.86	0.75	1.55	0.89	1.42	0.84	0.67	1.9	0.95	1.07	0.63	1.41	1.74	1.4	0.9
Tx1	0.91	1.04	1.3	1.12	0.64	0.71	0.25	1.71	0.88	0.82	1.03	1.03	2.18	2.57	1.09
Tx2	1.74	0.96	1.14	1.09	0.35	0.74	0.32	1.49	0.95	0.82	0.97	0.6	1.35	1.68	1.01
Tx3	0.56	0.7	1.31	0.87	1.19	0.64	0.25	1.16	0.83	0.77	0.57	1.8	1.51	2.25	0.8
Tx4	0.5	1.1	1.44	1.05	0.67	0.9	0.55	3.97	1.05	1.15	0.98	1.16	5.18	2.83	1.14
Tx5	1	0.67	1.54	1	0.48	0.46	0.43	5.24	0.55	0.6	0.68	0.65	4.82	1.71	0.68
Tx6	1.17	1.47	1.56	1.11	0.47	1.51	1.19	4.55	1.7	1.8	1.23	0.77	5.31	2.48	1.59
Tx7	0.83	1.43	1.77	0.98	0.71	1.51	1.28	5.94	1.66	1.87	1.06	1.14	6.66	3.07	1.49
Tx8	0.73	1.4	1.54	0.94	1.66	1.67	1.02	3.19	1.61	1.93	1.05	1.84	4.05	2.82	1.36
TxEx1	0.26	1.23	1.8	0.98	1.48	1.25	0.99	5.76	1.26	1.66	0.93	2.16	6.09	4.42	1.23
TxEx2	0.5	1.17	2.48	0.9	1.53	1.37	1.37	6.57	1.32	1.82	0.77	1.96	6.91	3.61	1.13
TxEx3	0.65	0.83	2.01	0.96	1.76	0.93	0.82	5.65	0.85	1.18	0.66	1.46	5.79	2.49	0.86
TxEx4	0.1	1.68	2.29	1.13	6.74	3.53	1.21	4.83	1.59	2.89	0.98	5.19	4.58	6.58	1.31
znf1	0.44	1.01	0.85	1.39	0.49	0.27	0.2	1.97	0.28	0.31	1.39	0.41	2.89	1.18	0.56
znf2	0.15	0.92	0.97	1.2	0.89	0.36	0.55	2.97	0.36	0.57	1.09	0.7	5.32	1.5	0.72
DNase1	0.22	1.73	1.72	1.31	9.02	4.49	2.09	0.98	1.71	2.83	1.29	2.58	2.22	4.04	1.45
BivProm1	0.14	4.33	7.17	3.91	9.27	6.97	2.38	1.34	2.24	4.71	1.76	6.88	8.09	5.05	1.87
BivProm2	0.16	3.1	6.63	2.98	8.6	5.88	2.53	0.89	2.16	4.05	1.64	5.69	7.01	3.38	1.79
BivProm3	0.3	1.89	2.54	1.41	5.4	3.28	1.92	0.97	1.78	2.88	1.34	3.68	2.58	2.17	1.53
BivProm4	0.14	3.22	3.37	1.61	6.59	5.09	2.53	1.17	2.82	4.62	2.18	5.37	3.6	4.07	2.23
PromF1	0.22	1.75	2.66	1.39	8.17	4.79	1.52	1.95	1.54	2.98	1.07	4.73	4.77	4.55	1.3
PromF2	0.15	2.6	4.06	1.66	9.24	6.81	1.48	3.72	1.91	3.85	1.34	5.89	6.07	6.55	1.53
PromF3	0.16	2.97	6.67	2.29	9.85	8.13	1.03	6.18	1.67	4.32	1.05	6.33	3.35	8.35	1.35
PromF4	0.19	4.42	7.32	3.99	9.74	8.93	1.63	6.73	2.29	5.69	1.76	7.52	0.89	8.75	1.82
PromF5	0.14	4.39	6.9	4	9.53	7.9	1.99	2.35	2.34	5.15	1.89	7.31	7.3	6.45	1.96
PromF6	0.14	2.21	4.13	1.25	8.46	4.98	0.74	5.64	1.7	3.74	1.01	5.21	2.91	6.32	1.37
PromF7	0.17	1.71	2.73	1.07	4.24	2.48	0.66	4.61	1.62	2.84	1.01	4.42	2.32	4.08	1.37
TSS1	0.13	4.91	5.35	4.9	9.36	7.85	2.14	5.06	2.56	5.89	2.54	7.7	3.06	7.37	2.27
TSS2	0.12	4.14	3.47	4.65	8.47	6.17	2.23	2.42	2.09	5.03	2.43	6.42	5.64	4.97	2.07
	100	9.92	9.93	9.91	9.91	9.93	9.93	9.92	10.0	9.93	9.93	9.96	10.7	9.91	9.93

Top 10% prioritized variants, whole-genome

state	% genome	CADD	CDTS	DANN	Eigen_PC	Eigen	FATHMM	FIRE	GERP	PhastCons	REMM	fitCons	phyloP
GapArtf1	12	0.21	0.05	0.41	0.05	0.05	0.17	0.11	0.06	0.48	0.04	0.09	0.18
GapArtf2	0.1	0.23	0.1	0.24	2.1	0.4	2.8	0.17	0.06	0.33	0.32	0.26	0.16
GapArtf3	0	0.9	0.27	0.39	6.46	2.08	3.31	0.16	0.13	0.54	0.77	0.52	0.32
Quies1	9.9	0.7	0.49	0.82	0.1	0.63	1.33	0.03	0.97	0.79	0.26	0.02	1.05
Quies2	3.1	0.6	0.52	0.63	0.16	0.51	1.11	0.02	0.79	0.61	0.45	0.06	0.83
Quies3	12	0.88	0.81	1.08	0.2	0.73	0.97	0.86	0.95	0.98	0.32	0.25	1.02
Quies4	4.5	0.77	0.44	1.34	0.07	0.18	0.41	0.3	0.25	1.35	0.07	0.07	0.45
Quies5	1.7	0.87	2.08	1.26	0.01	0.05	0.09	0.05	1.01	1.57	0.42	0.01	1.16
HET1	0.7	0.52	0.61	0.75	0.36	0.28	0.72	0.03	0.43	0.64	0.54	0.22	0.56
HET2	0.7	0.55	0.82	0.72	1.15	0.55	1.16	0.03	0.74	0.54	0.94	0.26	0.88
HET3	1.4	1	0.6	1.46	0.15	0.2	0.31	0.45	0.24	1.51	0.12	0.16	0.46
HET4	0.6	0.82	0.42	1.1	0.44	0.17	1.77	0.44	0.16	1.26	0.11	0.22	0.38
HET5	0.2	0.94	0.67	1.19	1.13	0.32	0.65	0.7	0.3	1.14	0.58	0.79	0.56
HET6	0.6	0.59	0.34	1	0.33	0.13	0.53	0.26	0.11	0.99	0.2	0.22	0.35
HET7	1	0.57	0.51	1.15	0.16	0.13	0.31	0.45	0.17	0.95	0.2	0.22	0.34
HET8	0.4	0.8	0.75	1.04	1.02	0.39	0.63	0.46	0.47	0.87	0.93	0.46	0.7
HET9	1	0.69	0.35	1.01	0.15	0.2	0.75	0.33	0.26	1.2	0.14	0.1	0.54
ReprPC1	0.2	2.15	3.82	1.56	5.74	3.74	2.48	0.72	2.11	1.52	3.46	2.76	1.75
ReprPC2	0.3	1.5	1.74	0.97	1.56	1.75	1.77	0.91	1.62	1.19	1.75	0.55	1.41
ReprPC3	1.1	1.15	1.15	0.88	0.55	1.05	1.41	0.75	1.23	1.03	1.11	0.24	1.2
ReprPC4	3.9	0.99	0.88	0.96	0.22	0.8	1.22	0.59	1.02	1.01	0.51	0.1	1.07
ReprPC5	0.6	0.88	1.4	1.01	2.27	0.97	1.28	0.57	1.02	0.72	1.73	0.58	1.07
ReprPC6	1.5	0.76	1.07	0.9	0.91	0.66	1.09	0.62	0.89	0.68	1.09	0.32	0.95
ReprPC7	0.6	1.1	1.57	1.25	3.88	1.51	1.32	0.84	1.25	0.84	2.38	0.92	1.22
ReprPC8	0.5	1.01	2.87	1.2	0.02	0.01	0.03	0.03	1.21	1.45	1.23	0.01	1.25
ReprPC9	0.4	1.09	1.24	0.95	0.64	1.02	1.25	0.99	1.17	1.02	0.97	0.36	1.12
Acet1	0.2	0.73	1.04	0.99	3.31	0.67	0.72	0.45	0.52	0.93	1.4	1.12	0.85
Acet2	0.9	0.8	1.03	0.75	0.75	0.72	1.12	0.5	1	0.71	1.47	0.39	0.96
Acet3	2.6	0.68	0.89	0.84	0.24	0.47	0.82	0.7	0.74	0.68	0.59	0.18	0.79
Acet4	0.4	0.82	1.2	0.78	2	1.01	0.97	0.69	1.11	0.62	2.48	0.82	0.98
Acet5	0.9	0.82	0.9	0.73	0.55	0.79	1.04	0.53	1.09	0.71	1.11	0.3	1
Acet6	0.4	1.15	1.02	0.72	2.29	1.54	1.29	0.59	1.49	0.89	2.4	1.01	1.22
Acet7	0.3	1.34	1.4	0.92	6.56	3.11	1.19	0.78	1.56	0.87	4.26	2.75	1.21
Acet8	0.6	0.69	1.05	0.86	0.81	0.58	0.86	0.6	0.81	0.63	1.25	0.39	0.79
EnhWk1	1.5	1.38	0.9	0.87	0.65	1.54	1.68	0.49	1.7	1.22	1.09	0.29	1.47
EnhWk2	0.4	1.35	1.75	1.18	4.53	2.35	1.14	1.47	1.57	0.9	3.48	1.77	1.26
EnhWk3	0.8	1.61	1.21	0.83	1.31	1.9	1.71	0.7	1.87	1.25	2.24	0.6	1.48
EnhWk4	2.2	1.84	0.94	0.79	0.59	2.16	2.23	0.35	2.16	1.53	1.36	0.18	1.74
EnhWk5	1	0.98	1	0.92	0.5	0.9	1.05	0.84	1.12	0.95	0.73	0.43	1.08
EnhWk6	0.6	1.06	1.3	0.94	1.78	1.17	0.83	1.41	1.2	0.9	1.57	1.49	1.1
EnhWk7	0.5	1.07	1.12	0.94	3.06	1.32	0.99	1.09	1.13	0.94	1.9	2.11	1.08
EnhWk8	1.4	1.42	0.95	0.86	0.57	1.51	1.72	0.33	1.67	1.33	1.21	0.47	1.42
EnhA1	0.2	2.33	2.25	1.48	9	5.95	1.56	1.75	2.19	1.44	5.56	5.55	1.66
EnhA2	0.3	2.73	1.55	1.08	4.39	3.8	2.59	0.69	2.73	2.03	4.27	2.19	2.09
EnhA3	0.2	2.79	1.79	1.19	8.68	5.63	1.94	1.06	2.63	1.92	4.91	5.91	2.01
EnhA4	0.3	2.16	1.25	0.95	7.59	4.06	1.92	0.57	2.24	1.58	3.7	4.18	1.75
EnhA5	0.7	1.91	1.1	0.94	3.59	2.56	1.92	0.59	2.07	1.53	2.37	1.86	1.7
EnhA6	0.6	1.6	1.36	0.88	1.34	1.9	1.5	0.83	1.9	1.23	1.99	0.66	1.51
EnhA7	0.4	0.98	1.55	1	3.76	1.43	1.09	1.15	1.08	0.73	2.72	1.92	1.05
EnhA8	0.3	1.4	1.84	1.06	6.02	2.98	1.18	1.67	1.48	0.93	3.93	2.96	1.22
EnhA9	0.2	1.68	2.14	1.21	7.91	4.41	0.96	2.43	1.66	0.98	5.08	4.45	1.3
EnhA10	0.4	1.54	1.7	0.99	4.87	2.51	1.27	1.61	1.6	1.12	3.24	2.93	1.34
EnhA11	0.7	0.93	1.16	0.96	0.97	0.82	0.94	1.29	0.98	0.86	1	0.73	0.99
EnhA12	0.3	1.26	1.34	0.97	5.02	2.22	1.23	0.87	1.39	0.91	3.51	1.46	1.15
EnhA13	0.8	1.09	0.98	0.83	1.52	1.22	1.26	0.66	1.32	0.95	1.54	0.4	1.18
EnhA14	0.4	1.21	1.48	0.97	3.67	1.88	1.16	1.4	1.38	0.91	2.46	1.14	1.16
EnhA15	1	1.07	1.07	0.91	0.89	1.05	1.15	1.04	1.24	0.97	0.87	0.35	1.13
EnhA16	0.6	1.08	1.29	1.01	2.41	1.14	1.04	1.09	1.08	1.04	1.53	0.46	1.11
EnhA17	0.5	2.51	1.39	0.94	2.62	3.1	2.47	0.59	2.59	1.99	3.07	1.72	2.05
EnhA18	0.5	1.18	1.17	0.96	1.37	1.17	1.33	0.48	1.21	1.12	2	2.06	1.07
EnhA19	0.3	1.8	1.82	1.14	5	2.81	1.64	1.14	1.81	1.32	3.97	3.91	1.46
EnhA20	0.3	1.38	0.98	0.84	2.04	1.77	1.61	0.5	1.72	1.16	1.68	3.14	1.48
TxEnh1	0.4	1.63	1.45	0.96	2.37	2.15	0.81	1.11	1.93	1.22	2.46	1.92	1.57
TxEnh2	0.4	1.29	1.6	0.99	2.69	1.53	0.44	2.12	1.4	0.97	3.21	2.17	1.31
TxEnh3	0.2	0.92	1.71	1.09	2.02	0.82	0.43	2.89	0.82	0.81	2.58	3.51	0.96
TxEnh4	0.3	2.51	2.97	2.18	5.11	3.12	2.28	5.99	2.38	1.93	4.2	6.47	2.09
TxEnh5	0.5	1.76	2.31	1.72	3.74	2.14	1.58	3.54	1.8	1.37	3.08	4.5	1.61
TxEnh6	0.2	1.66	2.2	1.45	4.32	2.37	1.3	4.13	1.69	1.17	3.5	4.95	1.46
TxEnh7	0.3	1.21	1.85	1.06	5.07	2.31	0.58	1.6	1.39	0.78	4.14	2.94	1.15
TxEnh8	0.2	1.9	2	1.36	5.12	2.97	1.45	3.76	2	1.35	3.86	5.4	1.63
TxWk1	2.8	1.13	1.36	1.21	0.39	0.99	0.76	3.26	1.17	1.12	0.63	2.73	1.2
TxWk2	0.8	1.33	1.81	1.42	1.79	1.26	1.13	2.01	1.41	1.13	1.88	2.26	1.33
Tx1	0.8	1.06	1.39	1.15	0.69	0.76	0.28	1.89	0.92	1.06	1.07	2.21	1.14
Tx2	1.6	0.98	1.22	1.12	0.38	0.79	0.34	1.62	0.99	1	0.63	1.36	1.06
Tx3	0.5	0.73	1.41	0.9	1.24	0.69	0.27	1.26	0.87	0.6	1.86	1.54	0.84
Tx4	0.5	1.34	1.59	1.27	0.87	1.12	0.75	4.19	1.27	1.21	1.39	5.31	1.35
Tx5	0.9	0.92	1.68	1.22	0.65	0.66	0.63	5.33	0.77	0.9	0.88	4.9	0.88
Tx6	1.1	1.77	1.69	1.4	0.74	1.78	1.44	4.72	1.97	1.53	1.06	5.41	1.85
Tx7	0.8	2.02	1.98	1.55	1.21	2.01	1.76	6.04	2.15	1.65	1.68	6.79	1.96
Tx8	0.7	1.61	1.67	1.14	1.86	1.87	1.2	3.35	1.8	1.25	2.04	4.17	1.54
TxEx1	0.3	2.17	2.24	1.84	2.19	2	1.73	5.84	2.04	1.81	2.9	6.41	1.95
TxEx2	0.6	2.59	3.01	2.2	2.51	2.44	2.44	6.63	2.45	2.12	3.08	7.25	2.22
TxEx3	0.7	1.69	2.39	1.73	2.29	1.54	1.47	5.73	1.55	1.43	2.16	6.07	1.51
TxEx4	0.1	2.58	2.77	2.01	6.97	4.02	1.94	5.08	2.31	1.81	5.62	5.11	2
znf1	0.4	1.09	0.95	1.47	0.56	0.34	0.26	2.18	0.35	1.47	0.49	2.97	0.64
znf2	0.2	1.75	1.32	1.97	1.23	0.71	0.88	3.32	0.74	1.34	1.21	5.79	1.04
DNase1	0.2	1.8	1.8	1.39	9.04	4.59	2.16	1.09	1.8	1.36	2.67	2.28	1.53
BivProm1	0.1	4.8	7.2	4.44	9.06	6.93	2.78	1.48	2.71	2.29	7.11	8.09	2.32
BivProm2	0.2	3.58	6.71	3.48	8.5	5.96	2.87	1.01	2.58	2.1	5.94	7.1	2.19
BivProm3	0.3	2.1	2.7	1.63	5.49	3.41	2.08	1.09	1.95	1.51	3.85	2.75	1.7
BivProm4	0.1	3.36	3.54	1.81	6.69	5.2	2.66	1.31	2.98	2.33	5.48	3.72	2.37
PromF1	0.2	1.9	2.81	1.55	8.18	4.88	1.66	2.15	1.68	1.19	4.83	4.82	1.42
PromF2	0.1	2.74	4.23	1.84	9.21	6.86	1.62	3.98	2.06	1.47	5.96	6.08	1.66
PromF3	0.2	3.26	6.81	2.62	9.73	8.1	1.31	6.4	1.95	1.35	6.46	3.56	1.61
PromF4	0.2	4.91	7.4	4.55	9.42	8.62	2.08	6.84	2.8	2.36	7.72	1.62	2.32
PromF5	0.1	4.76	6.95	4.42	9.33	7.79	2.33	2.58	2.72	2.31	7.48	7.34	2.32
PromF6	0.1	2.48	4.36	1.56	8.54	5.14	0.99	5.89	1.96	1.28	5.39	3.13	1.61
PromF7	0.2	1.84	2.9	1.22	4.43	2.63	0.79	4.84	1.76	1.15	4.55	2.43	1.5
TSS1	0.1	5.17	5.33	5.19	9.1	7.67	2.31	5.32	2.81	2.83	7.82	3.28	2.51
TSS2	0.1	4.38	3.42	4.88	8.24	6.06	2.35	2.6	2.3	2.67	6.56	5.67	2.28
	100	10.00	10.00	10.00	9.95	9.95	10.00	10.03	10.02	10.00	10.03	10.88	10.00

Supplementary Figure 2. 39: Enrichment of all full-stack states for top 10% bases prioritized by variant prioritization scores.
Analogous to **Supplementary Figure 2.37** and **Supplementary Figure 2.38**, but for top 10% prioritized bases.

A

Top 10% whole-genome prioritized variants

state	genome %	FIRE	fitCons	FATHMM	GERP	phyloP	PhastCons	DANN	CDTS	CADD	REMM	Eigen	Eigen_PC
GapArtf2	0.05	0.17	0.26	2.8	0.06	0.16	0.33	0.24	0.1	0.23	0.32	0.4	2.1
GapArtf3	0.01	0.16	0.52	3.31	0.13	0.32	0.54	0.39	0.27	0.9	0.77	2.08	6.46
EnhA2	0.33	0.69	2.19	2.59	2.73	2.09	2.03	1.08	1.55	2.73	4.27	3.8	4.39
TxEnh4	0.27	5.99	6.47	2.28	2.38	2.09	1.93	2.18	2.97	2.51	4.2	3.12	5.11
Tx7	0.82	6.04	6.79	1.76	2.15	1.96	1.65	1.55	1.98	2.02	1.68	2.01	1.21
TxEx2	0.56	6.63	7.25	2.44	2.45	2.22	2.12	2.2	3.01	2.59	3.08	2.44	2.51
BivProm1	0.15	1.48	8.09	2.78	2.71	2.32	2.29	4.44	7.2	4.8	7.11	6.93	9.06
BivProm2	0.16	1.01	7.1	2.87	2.58	2.19	2.1	3.48	6.71	3.58	5.94	5.96	8.5
BivProm4	0.13	1.31	3.72	2.66	2.98	2.37	2.33	1.81	3.54	3.36	5.48	5.2	6.69
PromF2	0.14	3.98	6.08	1.62	2.06	1.66	1.47	1.84	4.23	2.74	5.96	6.86	9.21
PromF3	0.15	6.4	3.56	1.31	1.95	1.61	1.35	2.62	6.81	3.26	6.46	8.1	9.73
PromF4	0.19	6.84	1.62	2.08	2.8	2.32	2.36	4.55	7.4	4.91	7.72	8.62	9.42
PromF5	0.14	2.58	7.34	2.33	2.72	2.32	2.31	4.42	6.95	4.76	7.48	7.79	9.33
TSS1	0.12	5.32	3.28	2.31	2.81	2.51	2.83	5.19	5.33	5.17	7.82	7.67	9.1
TSS2	0.11	2.6	5.67	2.35	2.3	2.28	2.67	4.88	3.42	4.38	6.56	6.06	8.24

B

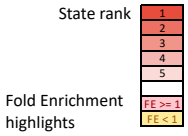
Top 5% whole-genome prioritized variants

state	genome %	FIRE	fitCons	FATHMM	GERP	phyloP	PhastCons	DANN	CDTS	CADD	REMM	Eigen	Eigen_PC
GapArtf3	0.01	0.04	0.95	4	0.18	0.32	0.7	0.48	0.4	0.44	0.75	2.57	8.38
EnhA2	0.33	0.51	3.05	3.36	3.48	2.53	3.22	0.66	1.48	3.75	4.95	4.42	3.61
TxEnh4	0.27	8.18	8.04	3.66	3.42	3.03	3.35	3.15	3.74	3.84	5.19	3.59	6.01
TxEx2	0.56	8.4	7.81	3.98	3.62	3.29	3.65	3.4	3.68	3.94	4.13	2.97	2.57
TxEx3	0.66	7.69	5.3	2.31	2.13	2.05	2.22	2.37	2.88	2.4	2.6	1.74	2.44
TxEx4	0.1	5.97	8.31	3.03	3.1	2.74	3.05	2.66	3.45	3.73	7.01	4.71	8.09
BivProm1	0.15	0.86	13.3	3.92	3.45	3.1	3.83	6.41	13.5	5.92	10.8	9.44	15.9
BivProm2	0.16	0.54	10.8	3.88	3.33	2.96	3.54	4.61	12.1	4.8	8.16	7.21	14.4
BivProm4	0.13	0.93	4.97	3.86	4.03	3.06	3.8	1.67	4.82	4.63	7.11	6.37	7.9
PromF2	0.14	4.47	7.44	2.15	2.38	1.93	2.25	1.74	6.12	3.31	7.53	9.7	15.7
PromF3	0.15	8.41	5.48	1.71	2.17	1.91	2.11	3.01	11.9	3.47	9.11	13.2	18.1
PromF4	0.19	9.4	3.05	2.91	3.59	3.12	4	6.63	14.1	5.99	12.7	15.1	18.3
PromF5	0.14	2.14	12.1	3.32	3.4	3.04	3.83	6.26	12.9	5.86	11.9	12	17
TSS1	0.12	6.09	5.57	3.03	3.47	3.32	4.71	8.06	10.1	6.81	12.8	12.9	17.3
TSS2	0.11	2.2	9.16	3.23	2.62	2.91	4.27	7.76	6.16	6.02	9.88	9.26	14.4

C

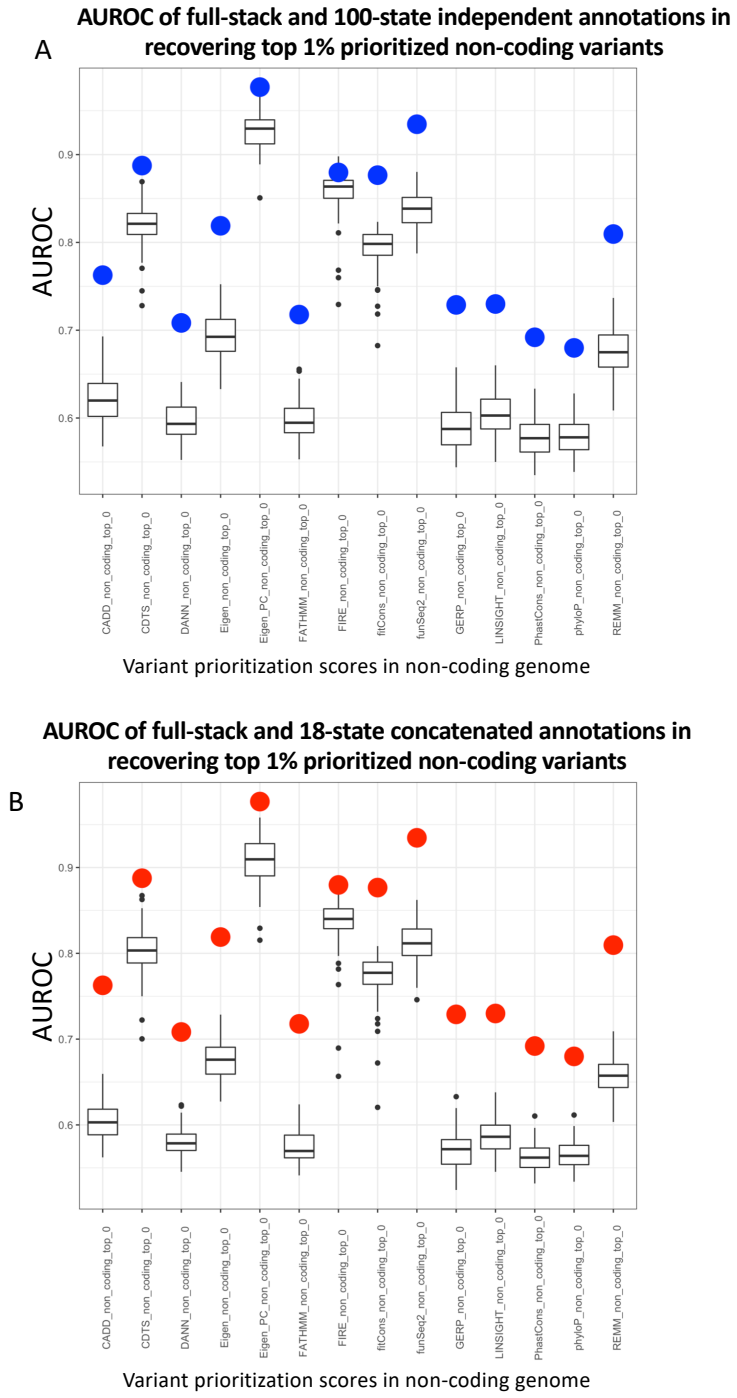
Top 1% whole-genome prioritized variants

state	genome %	FIRE	fitCons	FATHMM	GERP	phyloP	PhastCons	DANN	CDTS	CADD	REMM	Eigen	Eigen_PC
TxEnh4	0.27	14.1	15.5	7.86	7.68	8.5	5.51	11.8	4.65	11.6	9.96	3.37	4.16
TxEx1	0.27	9.54	11.3	6.21	7.07	7.09	4.74	8.81	2.18	9.04	7.48	2.68	0.19
TxEx2	0.56	12.4	17.4	8.73	9.42	9.87	6.2	13.5	3.66	13.1	9.82	2.32	0.36
TxEx3	0.66	12.5	10.3	4.67	4.51	5.46	3.48	7.88	3.3	7.38	5.29	1.36	0.89
TxEx4	0.1	7.94	11.6	6.3	5.69	6.7	4.73	8.66	4.54	9.49	7.72	5.16	10.4
znf2	0.15	0.12	11	1.26	0.84	2.16	1.1	7.38	1.02	5.7	1.69	0.2	0.18
DNase1	0.2	0.38	1	8.48	1.88	1.94	2.18	1.89	2.26	1.76	2.61	4.42	11.5
BivProm1	0.15	0.19	7.8	5.48	3.08	5.61	5.23	10.5	52.9	11.7	12.7	12.1	32.5
BivProm2	0.16	0.05	7.35	5.76	3.63	5.8	5.07	7.56	42.3	9.42	9.92	9.29	17.3
BivProm4	0.13	0.56	2.37	4.84	5.77	4.92	5.13	2.13	7.96	5.74	8.18	9.97	7.11
PromF2	0.14	4.58	2.31	2.52	2.34	2.37	2.66	2.05	11.1	4.09	3.94	10.7	40.5
PromF3	0.15	11.5	3.15	2.2	1.91	2.78	2.63	4.12	35.9	5.61	6.38	16.1	62.6
PromF4	0.19	13.5	2.29	3.61	3.41	5.3	5.42	11.5	56.6	11.9	16.6	30.9	68.2
PromF5	0.14	1.01	5.6	4.36	2.98	4.91	5.05	9.24	49.1	10.3	14.7	16.9	51.5
TSS1	0.12	6.49	2.76	2.67	2.58	4.28	5.88	13.6	41	10.8	17.5	25.3	59.6
TSS2	0.11	1.59	3.88	3.71	1.8	3.61	5	14.1	22.6	7.96	9.9	15.9	39.7



Supplementary Figure 2. 40: Enrichment of selected full-stack states with prioritized variants, whole genome.

A similar figure to **Supplementary Figure 2.36**, except showing the top enriched states for prioritized variants from the whole genome not restricted to non-coding regions. Fold enrichment of full-stack states for genomic bases prioritized in the **(A)** top 10% **(B)** top 5%, and **(C)** top 1% bases by 12-different variant prioritization scores (**Methods**). Only states that were among the five with greatest enrichment for at least one score are shown. Top enrichment values are colored red based on the rank of the state for the score as indicated in the color legend at the bottom. Depletions are shown in yellow.



Supplementary Figure 2. 41: Comparison of the full-stack model annotations against the concatenated and independent model annotations at predicting top 1% non-coding bases prioritized by various variant prioritization scores.

The box-plots show AUROC of the **(A)** 127 100-state annotations from independent models and **(B)** 98 18-state annotations from concatenated models at predicting locations of the top 1% non-coding prioritized variants. In both panels, the x-axis represents different groups of top 1% non-coding bases prioritized variants previously curated in (Arneson and Ernst, 2019), based on 14-different variant-prioritization scores (**Methods**). The **(A)** blue and **(B)** red dots show the AUROC for the full-stack chromatin state annotations.

A

Enrichment of full-stack states with GNOMAD variants

state	% Genome	maf_0_0.0001	maf_0.0001_0.0005	maf_0.0005_0.001	maf_0.001_0.005	maf_0.005_0.01	maf_0.01_0.05	maf_0.05_0.1	maf_0.1_0.2	maf_0.2_0.3	maf_0.3_0.4	maf_0.4_0.5	ucsc_snp151_common
GapArtf1	11.9	0.25	0.35	0.34	0.35	0.34	0.34	0.33	0.32	0.32	0.31	0.3	0.29
GapArtf2	0.05	0.74	1.3	1.22	1.21	1.17	1.3	1.41	1.61	1.43	1.11	1.27	2.49
GapArtf3	0.01	0.59	1.27	1.39	1.29	1.28	1.38	1.73	1.81	1.27	1	0.92	2.64
Quies1	9.88	1.12	1.1	1.12	1.14	1.15	1.17	1.19	1.19	1.22	1.23	1.24	1.18
Quies2	3.07	1.19	1.11	1.11	1.12	1.1	1.11	1.11	1.13	1.16	1.17	1.2	1.17
Quies3	12.2	1.08	1.17	1.22	1.23	1.28	1.31	1.37	1.34	1.32	1.3	1.3	1.21
Quies4	4.45	1.09	1.15	1.16	1.16	1.16	1.16	1.17	1.18	1.16	1.21	1.19	1.18
Quies5	1.69	0.69	0.93	0.9	0.87	0.87	0.85	0.82	0.76	0.74	0.73	0.7	0.78
HET1	0.71	1.26	1.19	1.18	1.18	1.15	1.16	1.12	1.16	1.16	1.21	1.19	1.24
HET2	0.69	1.29	1.29	1.25	1.26	1.25	1.28	1.26	1.3	1.32	1.35	1.31	1.39
HET3	1.36	1.07	1	1	0.99	0.97	0.96	0.94	0.96	0.98	1.01	1	1.03
HET4	0.56	1.17	1.13	1.11	1.08	1.12	1.07	0.94	0.92	0.91	0.94	1.1	1.21
HET5	0.25	1.09	1.07	1.05	1.03	1.03	1.02	1.01	1	1.08	1.01	1.02	1.3
HET6	0.58	1.28	1.35	1.31	1.29	1.28	1.28	1.21	1.22	1.22	1.24	1.21	1.39
HET7	1.02	1.23	1.36	1.33	1.34	1.34	1.33	1.34	1.36	1.34	1.38	1.35	1.37
HET8	0.43	1.11	1.07	1.06	1.06	1.04	1.05	1.04	1.13	1.14	1.13	1.08	1.17
HET9	1	1.11	1.12	1.1	1.07	1.09	1.06	1.05	1.04	0.99	1.03	1.1	1.2
ReprPC1	0.19	1.15	1.03	1.01	1	0.96	0.93	0.97	0.93	0.95	0.94	0.87	0.98
ReprPC2	0.32	1.06	0.99	0.99	0.99	0.99	0.99	1.04	1.08	1.05	0.97	1	1.03
ReprPC3	1.11	1.05	0.95	0.94	0.94	0.93	0.95	0.96	1	1.02	1	1	1.01
ReprPC4	3.93	1.07	1.02	1.03	1.04	1.06	1.08	1.12	1.14	1.16	1.17	1.18	1.11
ReprPC5	0.63	1.19	1.15	1.11	1.11	1.08	1.09	1.07	1.12	1.16	1.14	1.11	1.19
ReprPC6	1.51	1.17	1.13	1.11	1.11	1.09	1.12	1.13	1.17	1.19	1.2	1.19	1.2
ReprPC7	0.61	1.26	1.24	1.19	1.17	1.11	1.12	1.07	1.12	1.09	1.12	1.12	1.21
ReprPC8	0.48	0.68	0.86	0.83	0.8	0.75	0.72	0.65	0.58	0.57	0.58	0.55	0.66
ReprPC9	0.37	1.05	1.03	1.04	1.05	1.09	1.08	1.11	1.13	1.17	1.12	1.1	1.1
Acet1	0.18	1.38	1.86	1.97	1.97	1.96	1.99	1.99	1.98	1.92	1.88	1.85	2.09
Acet2	0.85	1.15	1.07	1.05	1.06	1.03	1.04	1.03	1.05	1.07	1.04	1.07	1.1
Acet3	2.65	1.13	1.1	1.1	1.11	1.11	1.13	1.16	1.17	1.16	1.15	1.16	1.15
Acet4	0.4	1.15	1.08	1.06	1.06	1.03	1.04	1.02	1.05	1.08	1.05	1.08	1.1
Acet5	0.86	1.1	1.04	1.04	1.05	1.03	1.06	1.04	1.1	1.09	1.11	1.09	1.09
Acet6	0.43	1.07	0.99	0.99	0.98	0.97	1	0.97	0.99	0.98	1.01	1	1.02
Acet7	0.28	1.14	1.05	1.03	1.02	0.99	0.98	0.94	0.97	0.99	1.06	1.02	1.05
Acet8	0.56	1.17	1.14	1.11	1.13	1.11	1.13	1.12	1.15	1.17	1.18	1.2	1.18
EnhWk1	1.54	1.04	0.98	0.98	1	1	1	1	1.03	1.02	1.03	1.04	1.01
EnhWk2	0.35	1.2	1.15	1.12	1.08	1.02	1.03	0.93	0.96	1	0.96	0.99	1.07
EnhWk3	0.83	1.1	1.03	1.01	1.01	0.99	1	0.99	1	1	0.99	1.06	1.02
EnhWk4	2.22	1.07	1	1	1	0.99	1	1	1.01	1	0.99	1	1
EnhWk5	0.99	1.07	1.07	1.09	1.1	1.13	1.14	1.16	1.16	1.17	1.19	1.17	1.12
EnhWk6	0.59	1.05	1.03	1.03	1.03	1.04	1.05	1.03	1.06	1.07	1.04	1.03	1.04
EnhWk7	0.48	1.07	1.04	1.06	1.04	1.08	1.07	1.12	1.13	1.2	1.16	1.16	1.11
EnhWk8	1.37	1.07	1	1	1	1	0.99	0.97	1	1.03	1.04	1.04	1.04
EnhA1	0.18	1.18	1.09	1.05	1.02	0.97	0.92	0.84	0.86	0.87	0.83	0.84	0.95
EnhA2	0.33	1.1	1.01	0.98	0.98	0.94	0.92	0.89	0.88	0.94	0.86	0.86	0.93
EnhA3	0.19	1.05	0.94	0.91	0.91	0.89	0.87	0.83	0.81	0.84	0.86	0.82	0.86
EnhA4	0.3	1.04	0.94	0.92	0.93	0.91	0.9	0.87	0.89	0.89	0.9	0.85	0.92
EnhA5	0.71	1.03	0.94	0.95	0.94	0.94	0.92	0.93	0.94	0.92	0.94	0.94	0.94
EnhA6	0.56	1.09	1	0.99	0.98	0.95	0.93	0.9	0.93	0.96	0.94	0.9	0.95
EnhA7	0.39	1.16	1.11	1.07	1.07	1.06	1.06	1.04	1.1	1.09	1.1	1.11	1.16
EnhA8	0.25	1.12	1.06	1.04	1.02	1.03	1.02	1	1.01	1.06	1.02	0.99	1.07
EnhA9	0.16	1.14	1.05	1.02	0.98	0.94	0.91	0.87	0.89	0.87	0.78	0.79	0.93
EnhA10	0.39	1.05	1	1.01	0.99	0.99	0.99	0.98	1	0.99	0.99	0.96	0.98
EnhA11	0.72	1.07	1.07	1.09	1.1	1.13	1.14	1.19	1.2	1.18	1.22	1.17	1.13
EnhA12	0.33	1.12	1.05	1.03	1.01	1	1	0.97	0.98	1.04	1	1.03	1.05
EnhA13	0.76	1.06	1	1.01	1	1	1.02	1.02	1.05	1.07	1.06	1.07	1.04
EnhA14	0.37	1.11	1.04	1.02	1.01	1	1	0.97	1.01	1.02	1.04	1.07	1.05
EnhA15	1.02	1.07	1.03	1.04	1.05	1.06	1.07	1.08	1.11	1.11	1.15	1.16	1.08
EnhA16	0.65	1.07	1.06	1.08	1.08	1.1	1.1	1.12	1.13	1.15	1.14	1.13	1.11
EnhA17	0.53	1.06	0.95	0.94	0.93	0.91	0.9	0.86	0.88	0.89	0.86	0.87	0.91
EnhA18	0.46	1.16	1.11	1.09	1.09	1.08	1.08	1.07	1.09	1.1	1.09	1.11	1.15
EnhA19	0.26	1.14	1.05	1.03	1.02	0.96	0.97	0.92	0.94	0.94	0.92	0.93	1.01
EnhA20	0.35	1.04	0.96	0.97	0.97	0.97	0.97	0.95	0.97	1	1.02	1.04	1
TxEnh1	0.39	1.03	0.93	0.92	0.91	0.89	0.86	0.81	0.81	0.77	0.8	0.81	0.86
TxEnh2	0.39	1.09	0.98	0.97	0.95	0.93	0.86	0.82	0.79	0.76	0.71	0.73	0.83
TxEnh3	0.25	1.2	1.11	1.08	1.04	1.01	0.93	0.83	0.84	0.8	0.77	0.76	0.92
TxEnh4	0.27	1.22	1.13	1.05	0.99	0.9	0.85	0.75	0.72	0.66	0.71	0.72	0.86
TxEnh5	0.5	1.22	1.17	1.11	1.06	1	0.97	0.89	0.88	0.87	0.9	0.94	1.01
TxEnh6	0.19	1.14	1.05	1	0.96	0.92	0.89	0.9	0.85	0.8	0.8	0.8	0.92
TxEnh7	0.27	1.19	1.11	1.06	1.03	0.98	0.93	0.85	0.81	0.87	0.87	0.86	0.95
TxEnh8	0.24	1.1	0.98	0.93	0.92	0.86	0.81	0.75	0.78	0.75	0.77	0.75	0.86
TxWk1	2.8	1.06	1.08	1.09	1.09	1.11	1.09	1.08	1.04	1.03	1.02	1.01	1.04
TxWk2	0.84	1.16	1.12	1.09	1.07	1.04	1.02	0.98	0.99	0.98	0.95	0.96	1.06
Tx1	0.82	1.09	1.04	1.03	1.02	1.01	0.96	0.91	0.88	0.84	0.84	0.8	0.88
Tx2	1.58	1.08	1.09	1.1	1.1	1.12	1.1	1.09	1.04	1.03	1.02	1.01	1.01
Tx3	0.51	1.14	1.09	1.06	1.06	1.01	1.01	0.96	0.93	0.96	0.96	0.95	1.02
Tx4	0.47	1.08	1	0.98	0.95	0.94	0.87	0.85	0.8	0.78	0.75	0.72	0.86
Tx5	0.94	1.15	1.19	1.17	1.15	1.16	1.11	1.09	1.06	1.03	1.04	1.02	1.05
Tx6	1.11	1	0.9	0.88	0.88	0.86	0.8	0.77	0.73	0.73	0.71	0.74	0.81
Tx7	0.82	1.03	0.9	0.87	0.84	0.81	0.75	0.68	0.67	0.68	0.63	0.64	0.75
Tx8	0.68	1.04	0.95	0.94	0.91	0.91	0.88	0.82	0.82	0.83	0.82	0.82	0.88
TxEx1	0.27	1.14	1	0.96	0.9	0.85	0.79	0.72	0.67	0.65	0.61	0.66	0.8
TxEx2	0.56	1.13	1	0.94	0.88	0.82	0.76	0.67	0.66	0.63	0.6	0.62	0.76
TxEx3	0.66	1.21	1.17	1.1	1.05	1.01	0.95	0.89	0.85	0.87	0.84	0.83	0.95
TxEx4	0.1	1.22	1.07	1.02	0.93	0.88	0.8	0.73	0.72	0.69	0.64	0.68	0.84
znf1	0.41	1.1	1.08	1.05	1.05	1.01	0.98	0.94	0.91	0.92	0.96	0.93	1
znf2	0.15	1.1	1.06	0.99	1.01	0.97	0.9	0.84	0.94	0.82	0.95	0.93	0.97
DNase1	0.2	1.16	1.15	1.15	1.14	1.13	1.15	1.14	1.14	1.14	1.12	1.11	1.16
BivProm1	0.15	1.29	1.11	1.05	1.03	0.97	0.95	0.9	0.9	0.82	0.81	0.87	0.99
BivProm2	0.16	1.25	1.09	1.06	1.03	0.97	0.94	0.95	0.89	0.87	0.85	0.83	1
BivProm3	0.29	1.2	1.15	1.1	1.09	1.05	1.05	1.02	1.04	1	0.98	0.97	1.1
BivProm4	0.13	1.09	0.99	0.97	0.96	0.94	0.93	0.9	0.9	0.88	0.84	0.84	0.9
PromF1	0.2	1.21	1.13	1.08	1.06	1.04	1	0.97	0.96	0.96	0.94	0.97	1.05
PromF2	0.14	1.17	1.04	1.01	0.95	0.91	0.89	0.81	0.8	0.81	0.78	0.8	0.89
PromF3	0.15	1.22	1.02	0.97	0.94	0.89	0.84	0.77	0.77	0.68	0.71	0.69	0.86
PromF4	0.19	1.53	1.29	1.23	1.16	1.05	0.95	0.86	0.8	0.79	0.74	0.8	0.98
PromF5	0.14	1.3	1.12	1.08	1.06	0.96	0.95	0.92	0.85	0.83	0.85	0.82	0.98
PromF6	0.13	1.13	0.97	0.95	0.92	0.87	0.84	0.83	0.78	0.77	0.71	0.76	0.86
PromF7	0.16	1.08	0.96	0.96	0.94	0.93	0.89	0.88	0.83	0.84	0.84	0.82	0.89
TSS1	0.12	1.68	1.65	1.52	1.43	1.33	1.22	1.16	1.07	1.02	1.02	1.1	1.28
TSS2	0.11	1.5	1.73	1.63	1.55	1.46	1.38	1.35	1.29	1.18	1.15	1.2	1.35
	100	4.95	0.96	0.24	0.41	0.14	0.22	0.07	0.07	0.04	0.03	0.02	0.52

B

CG Sites enrichments

State	% Genome	CGSites
GapArtf1	12	0.4
GapArtf2	0.1	1.9
GapArtf3	0	2.5
Quies1	9.9	0.6
Quies2	3.1	0.9
Quies3	12	1
Quies4	4.5	0.7
Quies5	1.7	0.8
HET1	0.7	1.2
HET2	0.7	2
HET3	1.4	0.5
HET4	0.6	0.5
HET5	0.2	1.1
HET6	0.6	1.5
HET7	1	1.5
HET8	0.4	1
HET9	1	0.7
ReprPC1	0.2	2.4
ReprPC2	0.3	1.1
ReprPC3	1.1	0.8
ReprPC4	3.9	0.8
ReprPC5	0.6	1.8
ReprPC6	1.5	1.5
ReprPC7	0.6	2.3
ReprPC8	0.5	0.9
ReprPC9	0.4	1
Acet1	0.2	2.1
Acet2	0.9	1.2
Acet3	2.6	1.2
Acet4	0.4	1.4
Acet5	0.9	0.9
Acet6	0.4	0.8
Acet7	0.3	1.4
Acet8	0.6	1.5
EnhWk1	1.5	0.6
EnhWk2	0.4	2
EnhWk3	0.8	1.1
EnhWk4	2.2	0.7
EnhWk5	1	0.9
EnhWk6	0.6	1
EnhWk7	0.5	1
EnhWk8	1.4	0.7
EnhA1	0.2	1.9
EnhA2	0.3	1.2
EnhA3	0.2	1
EnhA4	0.3	0.8
EnhA5	0.7	0.6
EnhA6	0.6	1.1
EnhA7	0.4	1.7
EnhA8	0.3	1.5
EnhA9	0.2	1.6
EnhA10	0.4	1.1
EnhA11	0.7	1.1
EnhA12	0.3	1.4
EnhA13	0.8	0.8
EnhA14	0.4	1.4
EnhA15	1	0.9
EnhA16	0.6	1.1
EnhA17	0.5	0.9
EnhA18	0.5	1.3
EnhA19	0.3	1.6
EnhA20	0.3	0.6
TxEnh1	0.4	0.7
TxEnh2	0.4	0.9
TxEnh3	0.2	1.7
TxEnh4	0.3	2.2
TxEnh5	0.5	2.2
TxEnh6	0.2	1.6
TxEnh7	0.3	1.7
TxEnh8	0.2	1.3
TxWk1	2.8	1.1
TxWk2	0.8	1.8
Tx1	0.8	0.9
Tx2	1.6	1
Tx3	0.5	1.5
Tx4	0.5	1
Tx5	0.9	1.8
Tx6	1.1	0.6
Tx7	0.8	0.7
Tx8	0.7	0.9
TxEx1	0.3	1.3
TxEx2	0.6	1.7
TxEx3	0.7	2.2
TxEx4	0.1	1.8
znf1	0.4	1
znf2	0.2	1.2
DNase1	0.2	1.9
BivProm1	0.1	7.3
BivProm2	0.2	5.8
BivProm3	0.3	2.6
BivProm4	0.1	1.9
PromF1	0.2	2.4
PromF2	0.1	2.4
PromF3	0.2	4.2
PromF4	0.2	8.6
PromF5	0.1	7
PromF6	0.1	1.9
PromF7	0.2	1.3
TSS1	0.1	10
TSS2	0.1	9.7
	100	1.8

Supplementary Figure 2. 42: Full-stack states enrichments with variants from GNOMAD stratified by minor allele frequencies, common variants (A) and CpG dinucleotides (B).
In each subpanel, each row corresponds to a full-stack state. The first column gives the state labels, the second gives the percent of the genome that each state covers. The heatmap colors are on a column specific coloring scale. The last row shows the percentage of the genome that each group of variants occupy. **(A)** The last column shows the enrichment of full-stack states with common variants from UCSC Genome Browser' snp151 track (Methods). Other columns show fold enrichments of full-stack states for GNOMAD variants with the specified ranges of MAF, which are ordered in increasing MAF *(Karczewski* et al*., 2020)*. **(B)** The last column shows fold enrichment of full-stack states with CpG dinucleotides**.** The three states showing highest enrichment if variants of lowest MAF (0 < MAF <= 0.0001) (TSS1-2, PromF4) are also the states most enriched states with CpG dinucleotide sites, likely reflecting the higher mutation rates associated with CpG dinucleotide sites *(Karczewski* et al*., 2020)*. We note the distinction between this panel, which shows the enrichments of states with CpG dinucleotide sites, and Supplementary Figure 2.28B, which highlights the relative higher enrichments of TSS-associated states with regions that are rich with G, C or both nucleotides and hence with low-complexity repeat class.

	background: genome		background: common snps		background: genome		
state	% background	gwas_catalog	% background	gwas_catalog bg: ucsc_snp151_common	autosomal	chrX	chrY
GapArtf1	11.861	0.254	3.449	0.873	0.871	1.142	6.888
GapArtf2	0.050	0.267	0.125	0.107	0.911	0.716	6.047
GapArtf3	0.012	0.279	0.031	0.106	0.923	0.473	6.106
Quies1	9.883	0.737	11.671	0.624	1.048	0.349	0.355
Quies2	3.070	0.793	3.586	0.679	1.054	0.323	0.131
Quies3	12.226	1.098	14.755	0.910	1.057	0.240	0.235
Quies4	4.452	0.803	5.259	0.679	0.930	2.184	1.317
Quies5	1.692	0.180	1.323	0.230	0.087	18.310	0.044
HET1	0.706	0.738	0.874	0.597	1.053	0.389	0.028
HET2	0.692	1.021	0.964	0.733	1.058	0.301	0.004
HET3	1.359	0.950	1.397	0.925	1.005	1.273	0.035
HET4	0.563	0.949	0.683	0.784	1.024	0.808	0.319
HET5	0.249	0.940	0.325	0.721	1.053	0.251	0.384
HET6	0.581	0.723	0.808	0.520	1.006	1.196	0.174
HET7	1.023	0.952	1.405	0.693	0.980	1.533	0.581
HET8	0.435	1.228	0.508	1.050	1.044	0.541	0.047
HET9	0.998	0.637	1.195	0.531	0.929	2.391	0.831
ReprPC1	0.191	1.540	0.188	1.565	1.056	0.339	0.000
ReprPC2	0.325	1.602	0.334	1.560	1.051	0.429	0.000
ReprPC3	1.107	1.536	1.124	1.513	1.062	0.239	0.003
ReprPC4	3.935	1.262	4.385	1.132	1.069	0.108	0.008
ReprPC5	0.628	1.547	0.746	1.302	1.065	0.176	0.001
ReprPC6	1.513	1.356	1.818	1.129	1.073	0.032	0.003
ReprPC7	0.614	1.573	0.739	1.305	1.065	0.178	0.000
ReprPC8	0.477	0.173	0.314	0.263	0.046	19.086	0.011
ReprPC9	0.375	1.025	0.412	0.933	1.018	1.016	0.094
Acet1	0.184	1.154	0.386	0.552	1.047	0.494	0.024
Acet2	0.855	1.209	0.936	1.104	1.069	0.096	0.018
Acet3	2.649	1.113	3.035	0.971	1.064	0.123	0.201
Acet4	0.403	1.348	0.443	1.224	1.064	0.186	0.001
Acet5	0.860	1.214	0.934	1.117	1.059	0.293	0.000
Acet6	0.428	1.210	0.435	1.190	1.057	0.322	0.002
Acet7	0.285	1.479	0.298	1.415	1.064	0.190	0.000
Acet8	0.562	1.386	0.661	1.179	1.051	0.388	0.113
EnhWk1	1.540	1.046	1.561	1.033	1.044	0.529	0.115
EnhWk2	0.352	1.535	0.377	1.435	1.066	0.165	0.000
EnhWk3	0.830	1.289	0.844	1.267	1.051	0.426	0.012
EnhWk4	2.216	1.038	2.209	1.042	1.057	0.313	0.033
EnhWk5	0.994	1.171	1.114	1.045	1.024	0.815	0.329
EnhWk6	0.588	1.472	0.612	1.414	1.030	0.766	0.140
EnhWk7	0.484	1.509	0.540	1.354	1.047	0.496	0.034
EnhWk8	1.369	0.945	1.419	0.912	1.010	1.129	0.165
EnhA1	0.179	1.605	0.170	1.689	1.071	0.068	0.000
EnhA2	0.328	1.315	0.307	1.407	1.061	0.258	0.000
EnhA3	0.194	1.478	0.166	1.725	1.059	0.293	0.000
EnhA4	0.301	1.245	0.275	1.361	1.052	0.424	0.002
EnhA5	0.714	1.158	0.671	1.233	1.042	0.578	0.056
EnhA6	0.563	1.111	0.537	1.164	1.051	0.429	0.001
EnhA7	0.393	1.647	0.454	1.425	1.045	0.541	0.021
EnhA8	0.255	1.747	0.272	1.633	1.041	0.611	0.040
EnhA9	0.162	1.568	0.151	1.678	1.057	0.333	0.002
EnhA10	0.395	1.486	0.389	1.509	1.044	0.542	0.044
EnhA11	0.715	1.217	0.807	1.078	1.041	0.604	0.033
EnhA12	0.332	1.376	0.347	1.316	1.051	0.444	0.004
EnhA13	0.764	1.149	0.795	1.105	1.037	0.689	0.041
EnhA14	0.366	1.474	0.385	1.401	1.056	0.344	0.019
EnhA15	1.017	1.214	1.097	1.126	1.040	0.599	0.105
EnhA16	0.645	1.137	0.715	1.026	1.024	0.905	0.075
EnhA17	0.526	1.215	0.477	1.338	1.045	0.533	0.017
EnhA18	0.457	1.114	0.526	0.969	1.013	1.120	0.060
EnhA19	0.257	1.372	0.260	1.354	1.054	0.383	0.001
EnhA20	0.346	1.052	0.346	1.054	1.026	0.807	0.243
TxEnh1	0.391	1.265	0.335	1.476	1.060	0.268	0.012
TxEnh2	0.391	1.384	0.325	1.666	1.056	0.339	0.023
TxEnh3	0.250	1.603	0.230	1.741	1.063	0.214	0.005
TxEnh4	0.267	1.858	0.229	2.164	1.069	0.102	0.000
TxEnh5	0.498	1.484	0.502	1.472	1.061	0.245	0.012
TxEnh6	0.189	1.531	0.174	1.661	1.058	0.293	0.012
TxEnh7	0.270	1.526	0.257	1.606	1.070	0.077	0.000
TxEnh8	0.243	1.603	0.210	1.856	1.066	0.151	0.000
TxWk1	2.800	1.152	2.899	1.113	1.029	0.766	0.215
TxWk2	0.842	1.336	0.893	1.260	1.039	0.649	0.028
Tx1	0.824	1.279	0.728	1.447	1.049	0.448	0.060
Tx2	1.580	1.208	1.602	1.192	1.035	0.673	0.144
Tx3	0.508	1.297	0.517	1.275	1.058	0.300	0.019
Tx4	0.470	1.326	0.402	1.548	1.054	0.349	0.098
Tx5	0.942	1.393	0.989	1.326	1.044	0.538	0.097
Tx6	1.109	1.183	0.899	1.459	1.038	0.647	0.067
Tx7	0.819	1.209	0.617	1.606	1.061	0.255	0.000
Tx8	0.681	1.302	0.599	1.480	1.055	0.343	0.041
TxEx1	0.265	1.600	0.212	2.005	1.064	0.190	0.000
TxEx2	0.556	1.630	0.424	2.137	1.066	0.152	0.000
TxEx3	0.663	1.606	0.633	1.683	1.052	0.410	0.002
TxEx4	0.098	1.952	0.083	2.314	1.065	0.181	0.000
znf1	0.406	1.247	0.407	1.242	1.039	0.623	0.072
znf2	0.152	1.087	0.147	1.119	1.063	0.212	0.000
DNase1	0.201	1.371	0.232	1.187	1.032	0.768	0.037
BivProm1	0.145	1.760	0.144	1.776	1.054	0.388	0.002
BivProm2	0.159	1.536	0.159	1.539	1.051	0.434	0.002
BivProm3	0.286	1.483	0.315	1.343	1.045	0.525	0.040
BivProm4	0.130	1.550	0.117	1.718	1.032	0.777	0.021
PromF1	0.201	1.775	0.212	1.684	1.057	0.333	0.002
PromF2	0.138	2.049	0.123	2.301	1.058	0.313	0.002
PromF3	0.153	1.847	0.132	2.137	1.063	0.220	0.000
PromF4	0.190	1.453	0.187	1.477	1.046	0.536	0.000
PromF5	0.135	1.667	0.133	1.694	1.043	0.589	0.002
PromF6	0.129	1.331	0.111	1.551	1.043	0.581	0.000
PromF7	0.160	1.545	0.143	1.732	1.038	0.665	0.029
TSS1	0.123	1.686	0.158	1.315	1.011	1.162	0.033
TSS2	0.114	1.567	0.154	1.160	0.991	1.375	0.450
	100	0.0029	100	0.5556	93.07	5.02	1.92

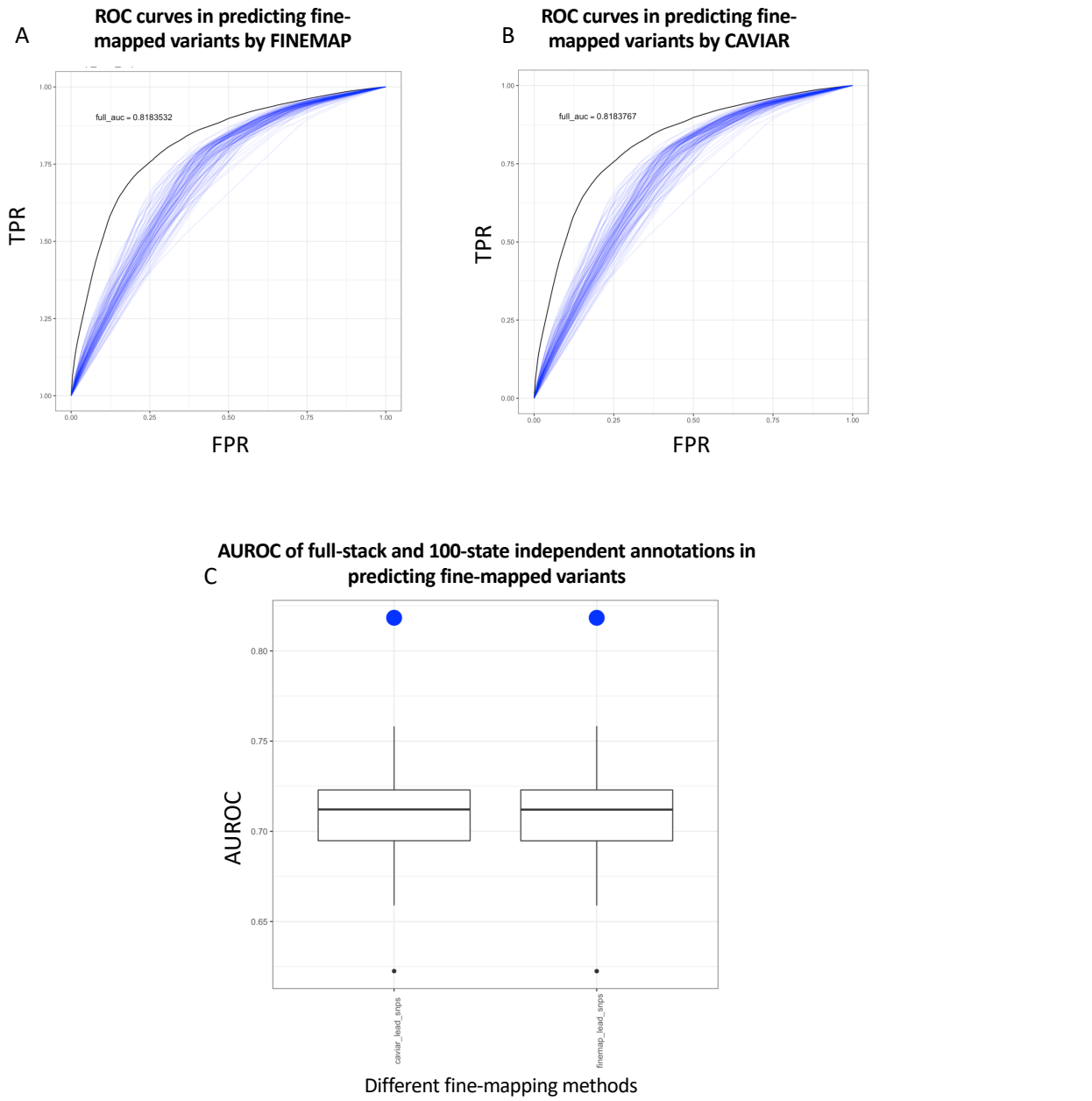
Supplementary Figure 2. 43: Full-stack states enrichments with GWAS catalog variants *(Welter* et al.*, 2014)* **and sex chromosomes.**
Each row corresponds to a full-stack state. The first column gives the state labels and the second column shows the percentage in the genome that each full-stack state occupies. The third column shows the fold enrichments of full-stack states with GWAS catalog variants against the whole-genome. The fourth column shows the percentage of the background context (UCSC snp151 common variants) that each full-stack state occupies. The fifth column shows fold enrichments with GWAS catalog variants against the background of common variants (**Methods**). The sixth, seventh and eight columns report the autosomal, chrX, and chrY fold enrichments, respectively. Columns are colored on a column specific coloring scale. The last row reports the percent of the background context (whole genome and set of common variants) that each annotation category covers.

Enrichment of full-stack states with fine-mapped variants

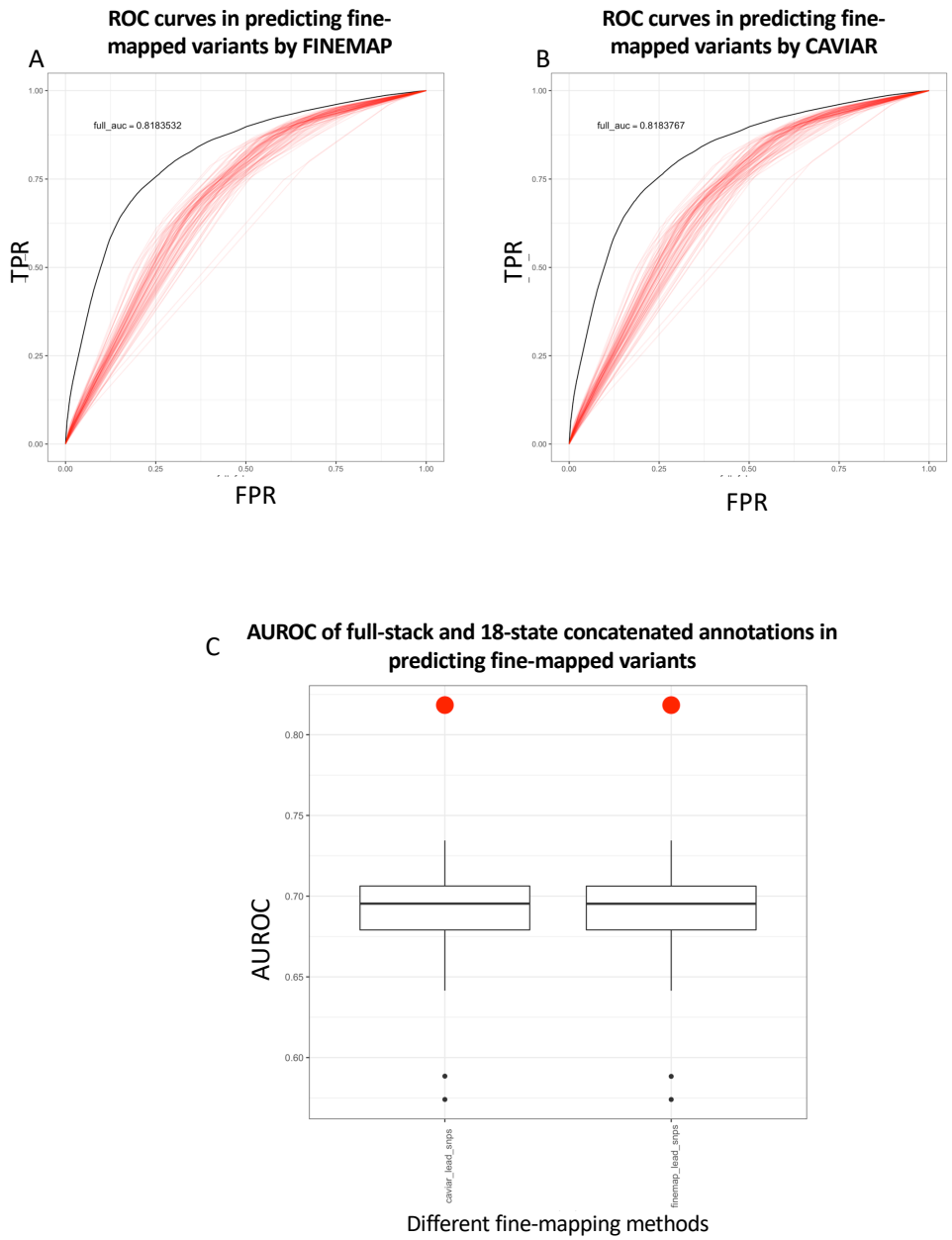
state	% background	caviar	finemap
GapArtf1	3.4	0.3	0.3
GapArtf2	0.1	0.1	0.1
GapArtf3	0	0.2	0.2
Quies1	12	0.5	0.5
Quies2	3.6	0.6	0.6
Quies3	15	0.8	0.8
Quies4	5.3	0.5	0.5
Quies5	1.3	0.1	0.1
HET1	0.9	0.5	0.5
HET2	1	0.5	0.5
HET3	1.4	0.7	0.7
HET4	0.7	0.7	0.7
HET5	0.3	0.6	0.6
HET6	0.8	0.5	0.5
HET7	1.4	0.6	0.6
HET8	0.5	0.8	0.8
HET9	1.2	0.4	0.4
ReprPC1	0.2	2.3	2.3
ReprPC2	0.3	1.7	1.7
ReprPC3	1.1	1.3	1.3
ReprPC4	4.4	0.9	0.9
ReprPC5	0.7	1.2	1.2
ReprPC6	1.8	1	1
ReprPC7	0.7	1.4	1.4
ReprPC8	0.3	0.1	0.1
ReprPC9	0.4	1.1	1.1
Acet1	0.4	0.5	0.5
Acet2	0.9	1.1	1.1
Acet3	3	1	1
Acet4	0.4	1.5	1.5
Acet5	0.9	1.1	1.1
Acet6	0.4	1.2	1.2
Acet7	0.3	1.9	1.9
Acet8	0.7	1.1	1.1
EnhWk1	1.6	1.1	1.1
EnhWk2	0.4	2.2	2.2
EnhWk3	0.8	1.5	1.5
EnhWk4	2.2	1.1	1.1
EnhWk5	1.1	1.2	1.2
EnhWk6	0.6	1.6	1.6
EnhWk7	0.5	1.3	1.3
EnhWk8	1.4	0.9	0.9
EnhA1	0.2	2.9	2.9
EnhA2	0.3	2	2
EnhA3	0.2	2.6	2.6
EnhA4	0.3	1.6	1.6
EnhA5	0.7	1.3	1.3
EnhA6	0.5	1.4	1.4
EnhA7	0.5	1.6	1.6
EnhA8	0.3	2.3	2.3
EnhA9	0.2	2.7	2.7
EnhA10	0.4	1.8	1.8
EnhA11	0.8	1.3	1.3
EnhA12	0.3	1.3	1.3
EnhA13	0.8	0.9	0.9
EnhA14	0.4	1.5	1.5
EnhA15	1.1	1.1	1.1
EnhA16	0.7	1.2	1.2
EnhA17	0.5	1.5	1.5
EnhA18	0.5	0.9	0.9
EnhA19	0.3	1.6	1.6
EnhA20	0.3	1.3	1.3
TxEnh1	0.3	2	2
TxEnh2	0.3	2.2	2.2
TxEnh3	0.2	2.1	2.1
TxEnh4	0.2	2.9	2.9
TxEnh5	0.5	2	2
TxEnh6	0.2	2.6	2.6
TxEnh7	0.3	2.5	2.5
TxEnh8	0.2	2.4	2.4
TxWk1	2.9	1.2	1.2
TxWk2	0.9	1.5	1.5
Tx1	0.7	1.8	1.8
Tx2	1.6	1.3	1.3
Tx3	0.5	1.6	1.6
Tx4	0.4	1.8	1.8
Tx5	1	1.5	1.5
Tx6	0.9	1.5	1.5
Tx7	0.6	1.9	1.9
Tx8	0.6	1.8	1.8
TxEx1	0.2	2	2
TxEx2	0.4	2.5	2.5
TxEx3	0.6	2	2
TxEx4	0.1	3.4	3.4
znf1	0.4	1	1
znf2	0.1	1.1	1.1
DNase1	0.2	1.2	1.2
BivProm1	0.1	2.6	2.6
BivProm2	0.2	2.5	2.5
BivProm3	0.3	1.6	1.6
BivProm4	0.1	2.9	2.9
PromF1	0.2	2.5	2.5
PromF2	0.1	3.1	3
PromF3	0.1	3	3
PromF4	0.2	3.1	3.1
PromF5	0.1	2.9	2.9
PromF6	0.1	2.3	2.3
PromF7	0.1	2.1	2.1
TSS1	0.2	2.4	2.4
TSS2	0.2	1.5	1.6
	100	0.3	0.3

Supplementary Figure 2. 44: Full-stack states enrichment values for fine-mapped variants at phenotype associated loci.

At phenotype associated loci, causal variants were fine-mapped by two methods, CAVIAR and Finemap *(Benner* et al*., 2016; Tate* et al*., 2019)*. A set of lead fine-mapped variants in 1MB loci across the genome were identified (**Methods**). The Supplementary Figure 2.hows the full-stack states' enrichment values for these fine-mapped variants calculated against a background of common variants. The rows correspond to full-stack states. The first column gives the state labels, the second column the percent of the genome that each state covers, followed by columns with the fold enrichment for fine-mapped variants by CAVIAR and Finemap. The heatmap colors are on a column specific coloring scale. The last row shows the percentage of the background set of variants that the sets of lead fine-mapped variants occupy.



Supplementary Figure 2. 45: Comparison of full-stack model annotations and the 100-state annotations from independent models in predicting fine-mapped variants.
(A) ROC curves for the full-stack model and the 127 100-state independent models' chromatin state annotations at predicting variants that show highest probabilities of being causal according to fine-mapping method FINEMAP *(Benner* et al.*, 2016)* against the background of common variants (**Methods**). The full-stack annotation model's ROC curve is in black and the 127 100-state annotations from independent models' ROCs are shown in blue. **(B)** Similar plot as (A), but for variants evaluated by fine-mapping method CAVIAR *(Chen* et al.*, 2015)*. **(C)** Comparison of the AUROC in predicting fine-mapped variants from a background of common variants. The x-axis represents two different fine-mapping methods used to evaluate variants' potential for causing diseases. The box-plots show AUROC of 127 100 annotations from independent models in predicting these variants. The blue dots show the AUROC of the full-stack chromatin state annotations.

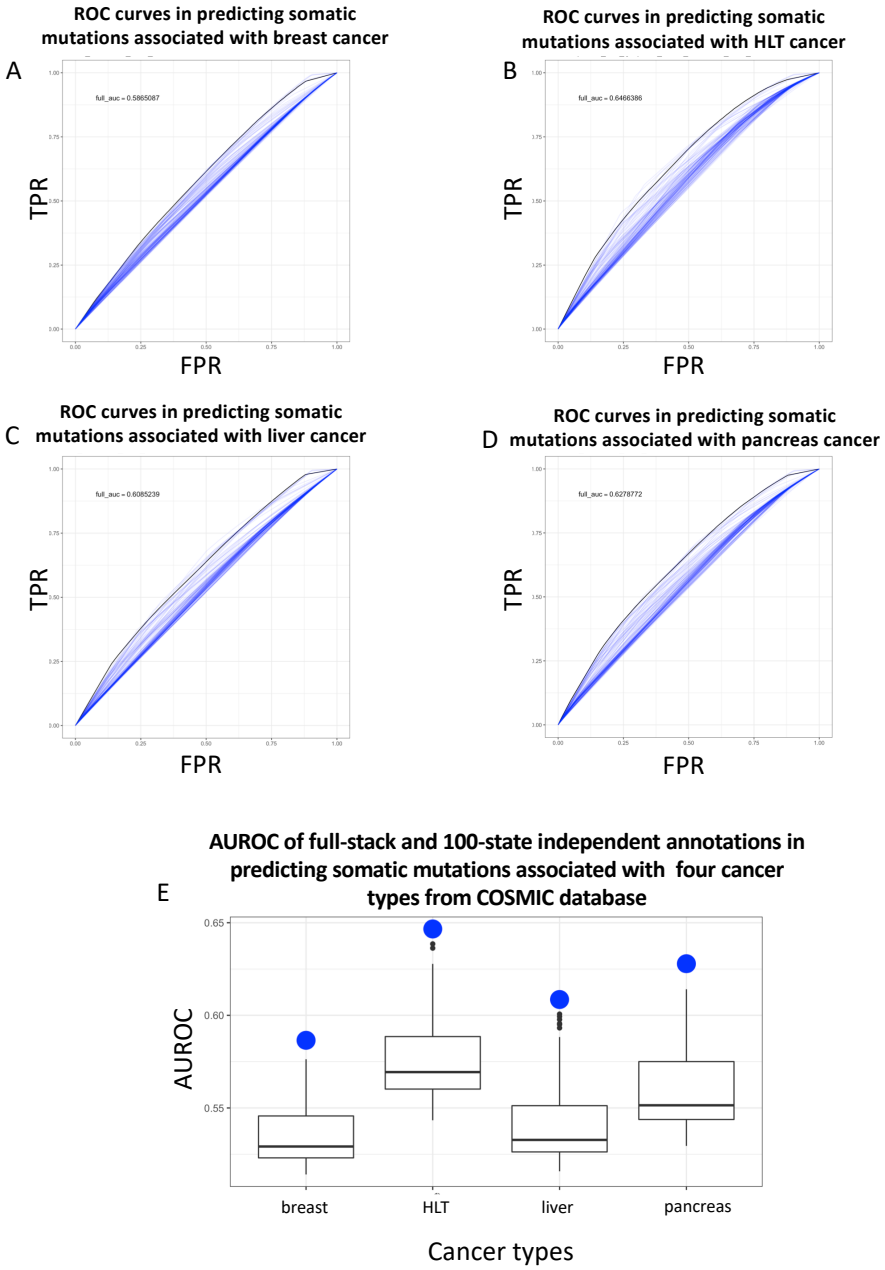


Supplementary Figure 2. 46: Comparison of full-stack model annotations and the 18-state annotations from a concatenated model in predicting fine-mapped variants.
(A) ROC curves for the full-stack model and the 98 18-state annotations from concatenated models' chromatin state annotations at predicting variants that show highest probabilities of being causal according to fine-mapping method FINEMAP *(Benner* et al.*, 2016)* against the background of common variants (**Methods**). The full-stack model annotation's ROC curve is in black and the 98 18-state annotations from a concatenated model ROCs are shown in red. **(B)** Similar plot as (A), but for variants evaluated by fine-mapping method CAVIAR *(Chen* et al.*, 2015)*. **(C)** Comparison of the AUROC in predicting fine-mapped variants from a background of common variants. The x-axis represents two different fine-mapping methods used to evaluate variants' potential for causing diseases. The box-plots show AUROC of 98 18-state annotations from concatenated models in predicting these variants. The red dots show the AUROC of the full-stack chromatin state annotations.

Enrichment of full-stack states with COSMIC database's cancer associated variants

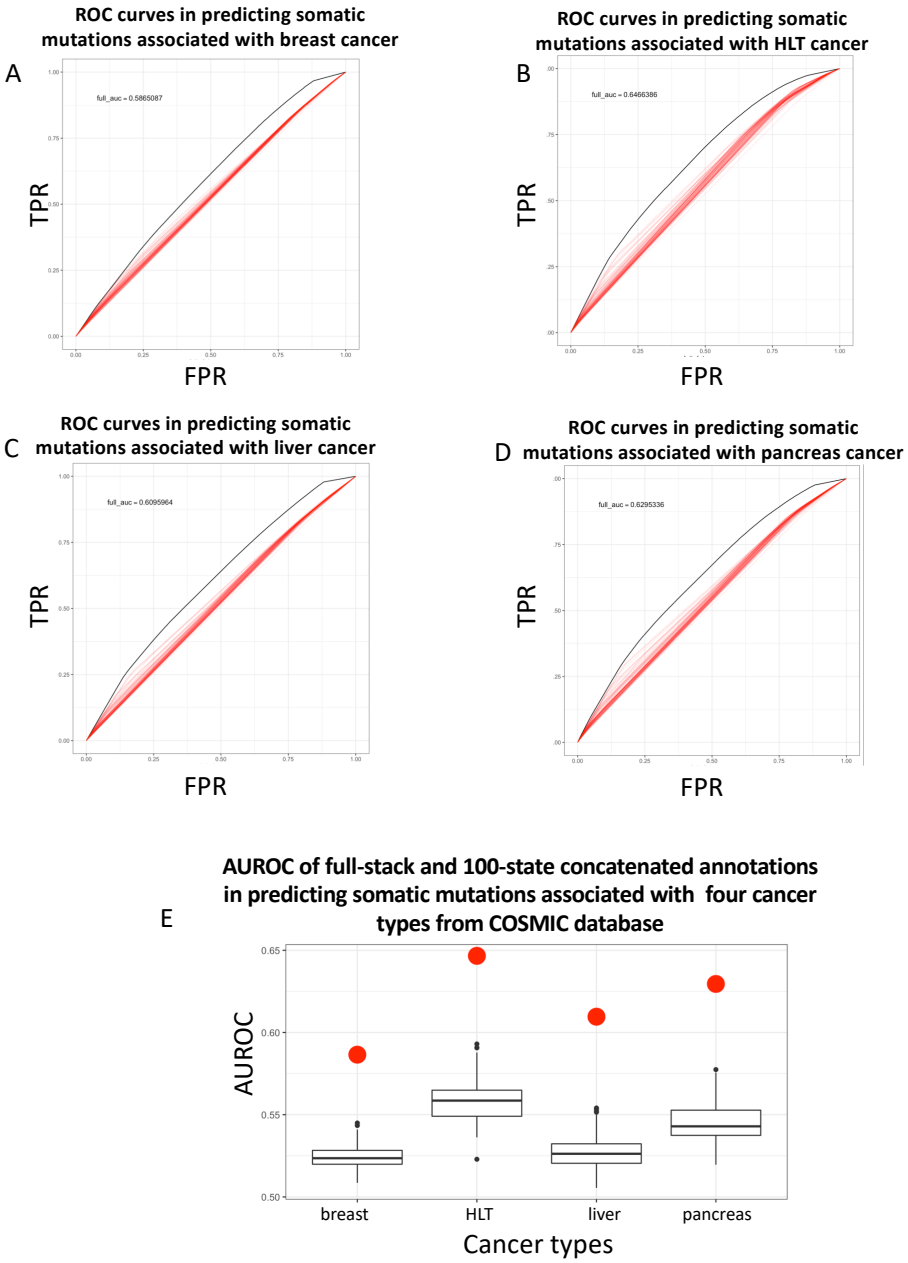
state	% background	breast	haematopoietic lymphoid_tissu	liver	pancreas
GapArtf1	4.51	0.7	0.6	0.54	0.55
GapArtf2	0.05	0.76	4.88	2.07	4.09
GapArtf3	0.01	1.35	5.58	5.38	4.22
Quies1	10.8	1.21	1.69	1.54	1.57
Quies2	3.36	1.26	1.41	1.49	1.75
Quies3	13.4	0.97	1.23	0.95	0.94
Quies4	4.86	1.18	1.25	1.09	1.23
Quies5	1.85	1.34	0.96	0.79	0.98
HET1	0.77	1.25	1.38	1.53	1.95
HET2	0.75	1.29	0.92	1.2	1.68
HET3	1.49	1.11	0.89	1.01	1.06
HET4	0.61	0.87	1.14	0.89	1.47
HET5	0.27	1.04	0.97	1.23	1.32
HET6	0.63	1.37	1.29	1.31	1.89
HET7	1.12	1.19	1.17	1.03	1.37
HET8	0.47	1.15	0.85	0.95	0.98
HET9	1.08	1.44	1.34	1.18	1.56
ReprPC1	0.2	0.97	0.72	0.95	0.92
ReprPC2	0.35	1.06	0.76	0.89	0.84
ReprPC3	1.21	0.96	0.65	0.84	0.81
ReprPC4	4.3	1.02	0.88	0.92	0.9
ReprPC5	0.67	0.91	0.63	0.9	0.94
ReprPC6	1.63	0.92	0.77	0.88	0.94
ReprPC7	0.64	0.8	0.6	0.88	0.93
ReprPC8	0.52	1.24	0.75	0.64	0.77
ReprPC9	0.41	0.93	1.05	0.96	0.97
Acet1	0.2	0.83	2.67	1.23	1.37
Acet2	0.93	1	0.88	0.96	1.03
Acet3	2.9	1.04	0.98	0.93	1.01
Acet4	0.44	0.85	0.65	0.81	0.85
Acet5	0.94	0.95	0.92	0.93	0.93
Acet6	0.47	0.9	0.84	0.86	0.83
Acet7	0.31	0.77	0.65	0.77	0.77
Acet8	0.61	0.93	0.75	0.87	1
EnhWk1	1.69	1.03	0.99	1	0.93
EnhWk2	0.38	0.77	0.56	0.83	0.71
EnhWk3	0.91	0.86	0.77	0.86	0.83
EnhWk4	2.43	1.07	1.07	1.1	1.04
EnhWk5	1.09	1	0.93	0.96	0.9
EnhWk6	0.64	0.98	0.72	0.82	0.7
EnhWk7	0.53	1	0.88	0.83	0.79
EnhWk8	1.5	1.06	1.1	1.16	1.12
EnhA1	0.19	0.84	0.51	0.82	0.6
EnhA2	0.36	1.07	0.72	0.87	0.77
EnhA3	0.21	1.18	0.52	0.82	0.61
EnhA4	0.33	1	0.76	0.87	0.69
EnhA5	0.78	1.02	0.85	0.92	0.82
EnhA6	0.61	1	0.75	0.9	0.91
EnhA7	0.42	0.84	0.78	0.83	0.8
EnhA8	0.27	0.85	0.84	0.79	0.69
EnhA9	0.18	1.02	0.49	0.75	0.63
EnhA10	0.43	0.86	0.64	0.76	0.64
EnhA11	0.78	0.93	0.79	0.82	0.79
EnhA12	0.36	0.89	0.65	0.9	0.77
EnhA13	0.84	0.98	0.95	0.96	0.85
EnhA14	0.4	0.91	0.65	0.8	0.61
EnhA15	1.11	0.91	0.91	0.87	0.74
EnhA16	0.7	1.06	0.94	0.92	0.88
EnhA17	0.58	0.95	0.81	0.95	0.83
EnhA18	0.5	1.04	0.88	1.02	1.1
EnhA19	0.28	0.93	0.65	0.8	0.77
EnhA20	0.38	0.93	1.03	1.03	0.88
TxEnh1	0.43	0.83	0.62	0.78	0.59
TxEnh2	0.43	0.73	0.43	0.73	0.55
TxEnh3	0.27	0.76	0.46	0.73	0.6
TxEnh4	0.25	0.68	0.37	0.96	0.63
TxEnh5	0.49	0.82	0.48	0.83	0.67
TxEnh6	0.19	0.74	0.38	0.8	0.6
TxEnh7	0.29	0.69	0.48	0.76	0.59
TxEnh8	0.25	0.83	0.41	0.84	0.59
TxWk1	3.03	0.84	0.76	0.82	0.68
TxWk2	0.85	0.85	0.6	0.85	0.8
Tx1	0.9	0.76	0.62	0.7	0.55
Tx2	1.73	0.82	0.85	0.77	0.68
Tx3	0.56	0.75	0.62	0.73	0.67
Tx4	0.5	0.77	0.44	0.78	0.54
Tx5	1	0.76	0.58	0.72	0.66
Tx6	1.17	0.75	0.46	0.91	0.54
Tx7	0.83	0.71	0.37	1	0.49
Tx8	0.73	0.81	0.48	0.85	0.58
TxEx1	0.26	0.7	0.33	0.87	0.54
TxEx2	0.5	0.66	0.35	0.95	0.57
TxEx3	0.65	0.75	0.44	0.82	0.62
TxEx4	0.09	0.94	0.33	0.95	0.56
znf1	0.44	0.85	0.53	0.73	0.7
znf2	0.15	0.77	0.42	0.83	0.72
DNase1	0.22	1	0.72	1.09	0.78
BivProm1	0.14	0.78	0.63	1.01	0.8
BivProm2	0.16	0.79	0.68	1.06	1.05
BivProm3	0.3	0.87	0.72	0.91	0.83
BivProm4	0.14	0.92	0.8	0.89	0.78
PromF1	0.22	0.91	0.48	0.82	0.66
PromF2	0.15	0.89	0.53	0.81	0.55
PromF3	0.16	0.77	0.51	0.83	0.58
PromF4	0.18	0.79	0.63	0.97	0.58
PromF5	0.14	0.83	0.62	0.94	0.67
PromF6	0.14	0.75	0.46	0.83	0.51
PromF7	0.17	0.69	0.54	0.78	0.57
TSS1	0.12	0.95	0.57	0.94	0.66
TSS2	0.12	0.81	0.81	0.88	0.68
	100	0.01	0.02	0.05	0.02

Supplementary Figure 2. 47: Full-stack states enrichments with cancer-associated somatic mutations in the non-coding genome.
Each row corresponds to a full-stack state. The first column gives the state labels and the second column shows the percentage in the background genome context that each full-stack state occupies. For this analysis, the background context is the non-coding genome (**Methods**). The following columns correspond to one of four cancer types with the most number of mutations in the COSMIC database *(Tate* et al.*, 2019)*. These columns give the enrichments of full-stack states for mutations that appear at least once in the database for the cancer types. The heatmap colors are on a column specific coloring scale. The last row shows the percentage of the genome that mutations associated with each cancer type occupy.



Supplementary Figure 2. 48: Comparison of full-stack model annotation and the 100-state independent annotations in predicting somatic mutations associated with four cancer types from COSMIC database *(Tate* et al.*, 2019)*.

(A) ROC curves for the full-stack model's annotations and the 127 100-state annotations from independent models at predicting somatic mutations associated with breast cancer against the background of non-coding genome (**Methods**). The full-stack model annotation's ROC curve is in black and the 127 100-state independent model annotations' ROCs are shown in blue. **(B-D)** Similar plot as (A), but for mutations associated with **(B)** haematopoietic and lymphoid tissue (HLT) cancer, **(C)** liver cancer and **(D)** pancreas cancer, respectively. **(E)** Comparison of the AUROC in predicting cancer-associated somatic mutations from a background of non-coding genome. The x-axis represents four different cancer types that we considered in this analysis. The box-plots show AUROC of 127 100-state independent models' in predicting these mutations. The blue dots show the AUROC of the full-stack chromatin state annotations.



Supplementary Figure 2. 49: Comparison of full-stack model annotation and the 18-state concatenated annotations in predicting somatic mutations associated with four cancer types from COSMIC database *(Tate* et al.*, 2019)*.

(A) ROC curves for the full-stack model and the 98 18-state concatenatedmodels' chromatin state annotations at predicting somatic mutations associated with breast cancer against the background of non-coding genome (**Methods**). The full-stack model's ROC curve is in black and the 98 18-state annotations from a concatenated model ROCs are shown in blue. **(B-D)** Similar plot as (A), but for mutations associated with haematopoietic and lymphoid tissue (HLT) cancer, liver cancer and pancreas cancer, respectively. **(C)** Comparison of the AUROC in predicting cancer-associated somatic mutations from a background of non-coding genome. The x-axis represents four different cancer types that we considered in this analysis. The box-plots show AUROC of 98 18-state annotations from a concatenated model in predicting these mutations. The blue dots show the AUROC of the full-stack chromatin state annotations.

state	percent_in_genome	unmapped_hg19_to_hg38		percent_unmapped_in_state
GapArtf1	11.86	8.40		99.59
GapArtf2	0.05	0.14		0.01
GapArtf3	0.01	0.35		0.00
Quies1	9.88	0.00		0.02
Quies2	3.07	0.00		0.00
Quies3	12.23	0.01		0.09
Quies4	4.45	0.01		0.05
Quies5	1.69	0.01		0.02
HET1	0.71	0.00		0.00
HET2	0.69	0.01		0.00
HET3	1.36	0.00		0.00
HET4	0.56	0.01		0.01
HET5	0.25	0.01		0.00
HET6	0.58	0.01		0.01
HET7	1.02	0.01		0.01
HET8	0.43	0.01		0.00
HET9	1.00	0.02		0.02
ReprPC1	0.19	0.01		0.00
ReprPC2	0.32	0.01		0.00
ReprPC3	1.11	0.00		0.00
ReprPC4	3.93	0.00		0.01
ReprPC5	0.63	0.00		0.00
ReprPC6	1.51	0.00		0.01
ReprPC7	0.61	0.00		0.00
ReprPC8	0.48	0.01		0.00
ReprPC9	0.37	0.01		0.00
Acet1	0.18	0.02		0.00
Acet2	0.85	0.00		0.00
Acet3	2.65	0.00		0.01
Acet4	0.40	0.00		0.00
Acet5	0.86	0.00		0.00
Acet6	0.43	0.00		0.00
Acet7	0.28	0.00		0.00
Acet8	0.56	0.00		0.00
EnhWk1	1.54	0.00		0.00
EnhWk2	0.35	0.00		0.00
EnhWk3	0.83	0.00		0.00
EnhWk4	2.22	0.00		0.00
EnhWk5	0.99	0.01		0.01
EnhWk6	0.59	0.01		0.01
EnhWk7	0.48	0.00		0.00
EnhWk8	1.37	0.00		0.00
EnhA1	0.18	0.00		0.00
EnhA2	0.33	0.00		0.00
EnhA3	0.19	0.00		0.00
EnhA4	0.30	0.00		0.00
EnhA5	0.71	0.00		0.00
EnhA6	0.56	0.00		0.00
EnhA7	0.39	0.00		0.00
EnhA8	0.25	0.01		0.00
EnhA9	0.16	0.00		0.00
EnhA10	0.39	0.00		0.00
EnhA11	0.72	0.01		0.00
EnhA12	0.33	0.00		0.00
EnhA13	0.76	0.00		0.00
EnhA14	0.37	0.00		0.00
EnhA15	1.02	0.00		0.00
EnhA16	0.65	0.01		0.01
EnhA17	0.53	0.00		0.00
EnhA18	0.46	0.00		0.00
EnhA19	0.26	0.00		0.00
EnhA20	0.35	0.00		0.00
TxEnh1	0.39	0.00		0.00
TxEnh2	0.39	0.00		0.00
TxEnh3	0.25	0.00		0.00
TxEnh4	0.27	0.00		0.00
TxEnh5	0.50	0.00		0.00
TxEnh6	0.19	0.00		0.00
TxEnh7	0.27	0.00		0.00
TxEnh8	0.24	0.00		0.00
TxWk1	2.80	0.00		0.01
TxWk2	0.84	0.00		0.00
Tx1	0.82	0.00		0.00
Tx2	1.58	0.00		0.00
Tx3	0.51	0.00		0.00
Tx4	0.47	0.00		0.00
Tx5	0.94	0.00		0.00
Tx6	1.11	0.00		0.00
Tx7	0.82	0.00		0.00
Tx8	0.68	0.00		0.00
TxEx1	0.27	0.00		0.00
TxEx2	0.56	0.00		0.00
TxEx3	0.66	0.00		0.00
TxEx4	0.10	0.00		0.00
znf1	0.41	0.00		0.00
znf2	0.15	0.00		0.00
DNase1	0.20	0.01		0.00
BivProm1	0.15	0.01		0.00
BivProm2	0.16	0.01		0.00
BivProm3	0.29	0.01		0.00
BivProm4	0.13	0.01		0.00
PromF1	0.20	0.00		0.00
PromF2	0.14	0.00		0.00
PromF3	0.15	0.00		0.00
PromF4	0.19	0.00		0.00
PromF5	0.14	0.00		0.00
PromF6	0.13	0.00		0.00
PromF7	0.16	0.00		0.00
TSS1	0.12	0.01		0.00
TSS2	0.11	0.03		0.00
percent in genome	100	7.665		100

B

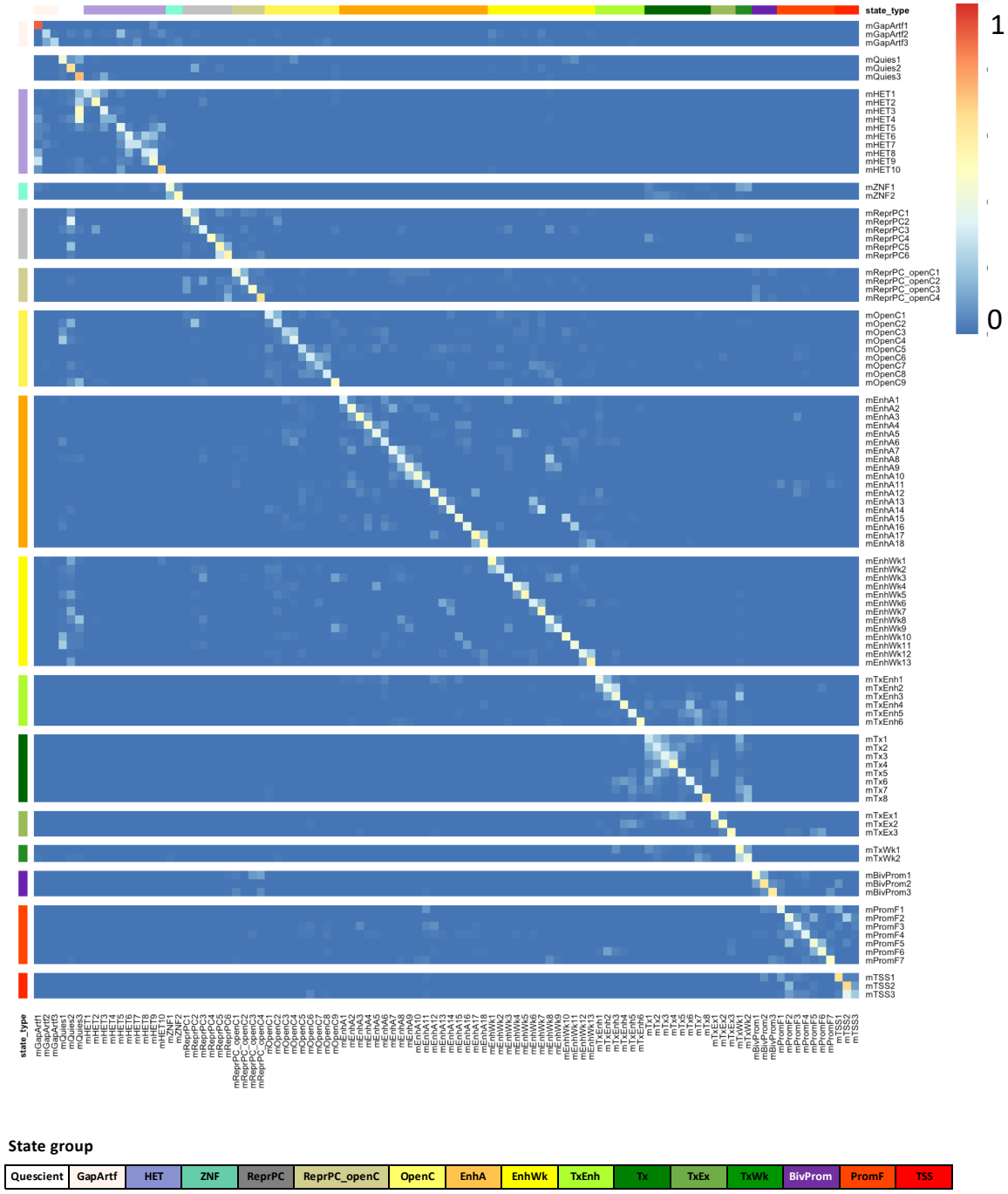
Enrichment of full-stack states with bases in hg19 that were unmapped to hg38, and analysis of unmapped bases with assembly gaps (all analyses in chr1-22,X,Y)

# bp in hg38 not mapped to a state	218,892,096
# bp in hg38 not mapped to a state AND overlapping hg38 assembly gaps	134,483,364
Fraction of hg38 bases that are unannotated to a state after liftOver from hg19 AND overlap hg38 assembly gaps	0.61
# bp in hg19 unmapped to hg38	237,275,800
# bp in hg19 unmapped to hg38 AND overlapping assembly gaps	234,342,292
Fraction of unmapped_to_hg38 bp in hg19 that also overlap hg19 assembly gaps	0.99
# bp in hg38 that are not assembly gaps	2,937,659,104
# bp in hg38 that are not assembly gaps AND annotated as a state after liftOver	2,853,200,372
Fraction of hg38 bases that are non-assembly gaps AND annotated to a state after liftOver	0.97

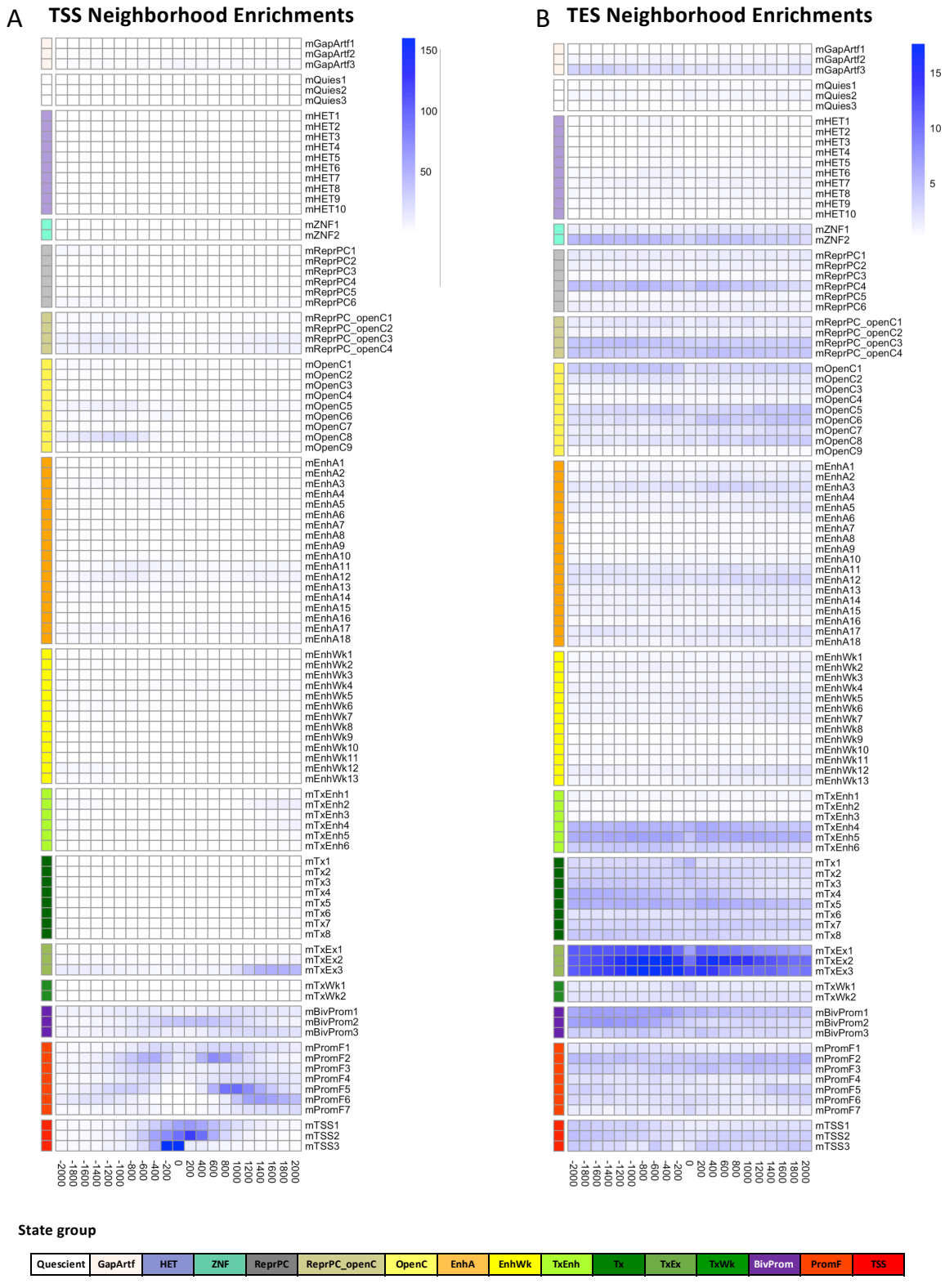
Supplementary Figure 2. 50: Full-stack states enrichments of bases that were not lifted over from hg19 to hg38.
(A) The heatmap shows enrichment values for the full-stack states (rows) of genomic bases that were unmapped when lifting the state annotation from hg19 to hg38 (Methods). The first column shows the state label and the second column shows the percentage of the genome that each state covers. The third column shows enrichment values, colored such that highest enrichment values are colored red and lowest ones are colored white. The fourth column shows the percentage of the unmapped regions (from hg19 to hg38) in each state. **(B)** Table showing details of numbers of bases involved in liftOver procedure, highlighting the overlap between the unmapped and unannotated regions with assembly gaps. As part of the liftOver procedure, bases in hg38 that are mapped to from multiple bases in hg19 are left unannotated to any state in hg38 (Methods). All results are reported in chromosomes 1-22, X, Y.

Universal chromatin state annotation of the mouse genome

Mouse full-stack model transition probabilities



Supplementary Figure 3. 1: Mouse full-stack states transition probabilities.
Each row and each column correspond to a full-stack state, ordered based on their associated state group. The heatmap shows for each state assigned at a current genomic position (rows) the probabilities of transitioning to another state (columns) at the subsequent genomic position. The state groups are shown at the bottom.



Supplementary Figure 3. 2: Positional enrichments of full-stack states around annotated transcription start sites and transcription end sites.

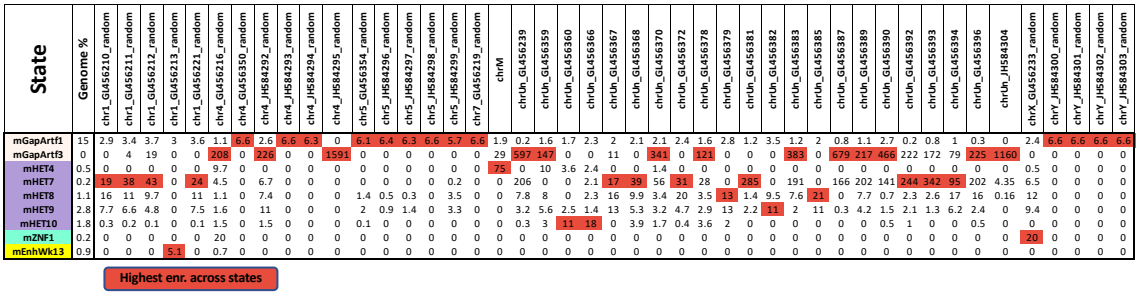
The figure shows positional fold enrichments for positions within 2kb of annotated (**a**) transcription start sites (TSS) and (**b**) transcription end sites (TES). Each column corresponds to one 200bp

window as indicated at bottom. Positive coordinate values represent the number of bases downstream in the 5' to 3' direction of transcription, while negative values represent the number of bases upstream. Enrichments are calculated based on a genome-wide background. Color scale of enrichments is indicated at right for each panel. State groups' color legends are shown at the bottom.

A **Enrichment of mouse full-stack states with chromosomes**

state	Genome %	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chrM	chrX	chrY
mGapArtf1	15.10	0.63	0.63	0.74	0.80	0.72	0.71	1.01	0.65	0.60	0.66	0.47	0.73	0.76	0.97	0.63	0.65	0.71	0.69	0.75	1.88	2.16	6.38
mGapArtf2	0.25	0.89	1.01	0.87	1.17	1.56	1.01	1.12	1.16	1.11	1.25	1.25	1.04	1.17	0.75	0.97	0.89	1.59	0.86	1.08	0.00	0.28	0.02
mGapArtf3	0.04	1.17	0.44	0.44	2.17	1.23	0.57	0.84	0.80	0.79	0.73	1.38	0.88	1.02	0.51	0.44	0.69	2.01	1.48	0.51	29.46	0.56	0.30
mQuies1	2.79	1.14	1.15	1.40	0.96	0.94	1.05	0.62	1.03	1.06	1.38	0.84	1.10	1.35	1.07	1.24	1.20	0.81	1.43	1.12	0.00	0.22	0.00
mQuies2	8.01	1.04	1.04	0.99	0.78	0.82	1.10	0.99	0.87	1.15	0.99	1.25	0.94	1.26	1.08	1.03	0.97	0.89	1.12	1.32	0.31	1.28	0.01
mQuies3	16.39	1.28	1.00	1.40	1.09	1.03	1.14	0.58	1.17	0.78	1.23	0.65	1.18	0.89	1.26	1.23	1.37	0.86	1.21	0.67	0.38	0.59	0.00
mHET1	0.81	1.12	1.10	1.12	1.14	1.57	0.96	0.82	1.46	0.95	1.41	0.62	1.08	1.06	1.13	1.20	1.11	1.30	0.98	0.49	0.00	0.06	0.00
mHET2	2.41	1.08	1.20	0.90	0.95	1.34	0.88	1.32	1.69	1.22	1.15	0.84	1.15	1.06	1.08	1.17	0.94	1.17	0.91	0.58	0.00	0.07	0.00
mHET3	3.56	1.48	0.93	1.32	1.08	1.03	1.15	0.63	1.30	0.62	1.09	0.63	1.05	0.64	1.32	1.34	1.44	1.01	1.02	0.56	2.77	0.78	0.00
mHET4	0.52	1.08	0.92	1.31	0.96	1.10	1.02	0.66	1.11	0.78	1.11	0.61	1.02	0.88	1.29	1.13	1.41	0.86	1.02	0.70	75.36	1.47	0.00
mHET5	1.23	1.27	0.88	1.17	0.93	1.20	1.03	0.96	1.07	0.93	1.12	0.71	0.89	0.96	0.92	1.09	0.99	1.19	0.90	0.61	4.01	1.39	0.06
mHET6	0.74	0.57	0.55	0.76	1.33	0.98	0.92	3.50	0.65	0.95	0.60	0.51	1.84	1.25	0.57	0.48	0.55	1.91	0.80	1.11	0.00	0.94	0.03
mHET7	0.20	0.62	0.55	0.75	1.60	1.34	0.73	2.43	0.63	0.85	0.87	0.54	1.99	1.55	0.44	0.51	0.58	1.93	1.34	0.75	0.00	0.38	0.01
mHET8	1.10	0.46	0.50	0.72	1.23	0.74	0.98	3.43	0.64	0.86	0.51	0.44	1.72	1.18	0.61	0.35	0.46	1.74	0.62	1.22	0.00	1.87	0.32
mHET9	2.80	0.66	0.58	0.88	0.91	0.65	1.03	2.02	0.66	0.77	0.61	0.45	1.11	0.93	0.77	0.49	0.56	1.29	0.64	0.91	0.00	3.35	0.85
mHET10	1.77	1.36	0.84	1.35	1.08	0.96	1.14	0.68	1.03	1.11	1.23	0.82	0.83	1.24	1.13	1.03	1.03	0.96	1.14	0.63	0.00	0.75	0.18
mZNF1	0.24	0.81	1.02	0.84	0.89	1.16	0.85	1.24	1.05	1.14	1.35	0.88	1.56	1.78	0.66	0.76	1.02	1.67	1.21	0.88	0.00	0.42	0.01
mZNF2	0.13	0.59	0.89	0.53	1.00	1.44	0.71	1.73	1.43	1.23	1.36	1.00	1.08	1.71	0.52	0.63	1.20	2.39	1.08	0.87	0.00	0.19	0.01
mReprPC1	0.76	0.69	1.39	0.51	1.45	1.28	1.01	1.38	1.14	1.42	0.58	1.99	0.72	0.85	0.69	1.08	0.79	1.32	0.59	1.69	0.00	0.56	0.00
mReprPC2	2.65	0.95	1.26	0.76	0.93	1.00	1.18	1.35	0.94	1.39	0.82	1.74	0.88	1.15	1.02	1.13	0.94	1.07	1.00	1.34	0.00	0.32	0.00
mReprPC3	0.94	0.96	1.08	0.69	1.43	1.42	0.93	1.35	1.47	1.61	0.66	1.36	1.13	1.15	1.04	0.97	0.62	0.99	0.99	0.86	0.00	0.09	0.00
mReprPC4	0.10	0.27	0.39	0.10	0.34	0.46	0.27	0.56	0.30	0.22	0.25	0.43	0.12	0.19	0.11	0.09	0.24	0.19	0.23	0.60	0.00	11.90	0.00
mReprPC5	1.21	0.43	0.56	0.37	0.48	0.54	0.37	0.48	0.56	0.58	0.33	0.47	0.43	0.71	0.53	0.22	0.33	0.33	0.38	0.91	0.00	9.22	0.00
mReprPC6	0.49	0.36	0.55	0.29	0.59	0.69	0.37	0.58	0.51	0.54	0.33	0.60	0.48	0.57	0.33	0.37	0.37	0.56	0.38	1.12	0.00	8.94	0.00
mReprPC_openC1	0.15	0.67	1.44	0.43	2.26	1.29	0.67	1.20	1.25	1.35	0.48	2.19	0.69	0.61	0.54	1.37	0.57	0.98	0.68	1.91	0.00	0.52	0.00
mReprPC_openC2	0.38	0.63	1.35	0.44	1.87	1.49	0.82	1.15	1.44	1.53	0.49	1.78	0.85	0.92	0.75	0.91	0.67	0.82	0.69	1.54	0.00	0.72	0.00
mReprPC_openC3	0.22	0.54	1.12	0.52	1.59	0.95	0.75	1.35	1.30	1.16	0.90	2.30	0.63	0.63	0.62	1.26	0.73	1.44	0.67	1.66	0.00	1.13	0.00
mReprPC_openC4	0.22	0.87	1.56	0.58	1.41	1.62	0.87	1.34	0.88	1.07	0.68	1.74	0.91	1.10	0.54	1.25	0.50	1.06	0.55	2.43	0.00	0.35	0.00
mOpenC1	0.46	0.72	1.44	0.44	1.66	1.26	0.86	1.35	1.34	1.17	0.86	2.24	0.89	0.51	0.56	1.39	0.76	1.73	0.66	1.46	0.00	0.02	0.00
mOpenC2	1.15	0.94	1.32	0.65	1.11	1.23	1.05	1.36	1.05	1.39	0.94	1.76	0.94	0.88	0.89	1.22	0.96	1.40	0.91	1.19	0.00	0.02	0.00
mOpenC3	0.43	1.06	1.11	0.90	0.92	1.06	1.10	1.06	1.05	1.26	1.17	1.26	1.04	1.27	0.98	1.10	1.04	1.29	1.30	1.27	0.00	0.12	0.00
mOpenC4	0.90	1.15	1.19	1.20	0.86	0.99	1.07	0.84	0.97	1.07	1.32	1.06	1.01	1.37	0.99	1.18	1.14	1.02	1.31	1.18	0.00	0.22	0.00
mOpenC5	0.24	0.84	1.35	0.71	1.31	1.13	0.95	1.26	1.33	1.12	1.12	1.95	1.08	0.67	0.58	1.21	0.90	1.58	0.74	1.35	0.00	0.08	0.00
mOpenC6	0.14	0.97	1.23	0.85	1.15	1.07	1.07	1.33	1.09	1.26	1.14	1.57	0.89	0.91	0.78	1.14	0.86	1.33	0.92	1.38	0.00	0.20	0.01
mOpenC7	0.20	0.97	1.13	0.84	1.06	1.27	1.10	1.13	1.01	1.30	1.05	1.63	0.88	0.89	0.91	1.08	0.92	1.22	0.92	1.26	0.00	0.46	0.01
mOpenC8	0.22	1.13	1.04	1.26	1.00	0.87	1.04	1.35	1.00	0.98	1.35	0.81	0.96	1.29	0.93	0.81	0.91	1.90	0.96	0.74	0.00	0.33	0.01
mOpenC9	1.38	1.17	1.16	1.13	0.91	1.04	1.14	0.84	1.10	1.27	1.24	0.93	1.10	1.08	1.14	1.24	1.34	0.88	1.20	0.83	0.00	0.19	0.00
mEnhA1	0.26	1.05	1.15	0.95	1.14	1.12	1.14	1.02	1.04	1.30	1.17	1.10	1.13	1.17	0.93	1.01	1.04	1.19	1.33	1.07	0.00	0.08	0.00
mEnhA2	0.24	0.97	1.41	0.77	1.38	1.18	0.99	1.00	1.29	1.27	0.89	1.75	1.20	0.81	0.77	0.98	0.79	1.29	1.18	0.95	0.00	0.06	0.00
mEnhA3	0.15	0.77	1.37	0.61	1.42	1.13	0.77	1.16	1.50	1.12	1.11	2.22	0.98	0.63	0.79	1.16	0.72	1.32	1.13	1.40	0.00	0.03	0.00
mEnhA4	0.28	0.94	1.23	0.66	1.15	1.02	1.09	1.20	1.30	1.34	1.06	1.61	0.99	1.05	0.81	1.06	1.10	1.26	1.19	1.22	0.00	0.06	0.00
mEnhA5	0.28	0.90	1.24	0.78	1.20	1.15	1.06	1.29	1.20	1.17	1.08	1.67	0.92	0.94	0.76	1.17	0.95	1.44	0.86	1.54	0.00	0.04	0.00
mEnhA6	0.40	1.10	1.14	1.01	0.99	0.95	1.28	0.99	1.15	1.20	1.22	1.08	0.99	1.22	0.95	1.09	1.08	1.01	1.41	1.19	0.00	0.12	0.00
mEnhA7	0.38	1.08	1.28	0.81	1.27	1.15	0.97	0.98	1.24	1.41	0.86	1.65	1.15	1.04	0.88	0.90	0.77	1.20	1.31	1.00	0.00	0.05	0.00
mEnhA8	0.67	1.12	1.20	1.10	1.09	1.06	1.06	0.95	1.16	1.28	0.97	1.23	1.19	1.20	1.07	0.87	1.03	0.91	1.32	1.01	0.00	0.12	0.00
mEnhA9	0.36	1.11	1.18	1.17	1.20	1.10	1.08	0.96	1.04	1.32	0.98	0.99	1.41	1.19	1.02	0.94	1.02	0.80	1.27	0.85	0.00	0.14	0.00
mEnhA10	0.20	1.07	1.29	1.03	1.29	1.10	1.02	0.95	1.08	1.38	0.88	1.26	1.42	1.20	0.91	0.89	0.88	0.86	1.29	1.00	0.00	0.10	0.00
mEnhA11	0.13	0.98	1.19	0.85	1.29	1.18	0.81	1.08	1.30	1.17	0.99	1.73	1.03	1.02	0.79	1.13	0.91	1.15	1.24	1.14	0.00	0.15	0.00
mEnhA12	0.18	0.76	1.16	0.78	1.23	1.32	0.85	1.29	1.31	1.02	1.16	2.13	0.83	0.83	0.66	1.46	0.76	1.42	0.69	1.78	0.00	0.08	0.00
mEnhA13	0.28	1.03	1.16	0.77	0.94	1.16	1.12	1.17	0.90	1.21	1.20	1.73	0.99	1.04	0.82	1.26	1.07	1.31	0.78	1.63	0.00	0.12	0.00
mEnhA14	0.46	1.11	1.12	0.83	0.90	0.99	1.09	1.11	0.81	1.34	1.18	1.84	1.06	1.13	0.86	1.24	1.12	1.41	0.73	1.35	0.00	0.12	0.00
mEnhA15	0.34	1.12	1.17	0.95	0.87	1.07	1.14	1.07	1.32	1.16	1.03	1.30	0.77	1.09	1.03	0.94	1.07	1.12	1.43	1.55	0.00	0.15	0.00
mEnhA16	0.40	1.04	1.21	0.85	1.17	1.00	1.21	1.10	1.20	1.32	1.01	1.40	0.99	1.33	1.00	0.84	0.91	1.18	1.04	1.11	0.00	0.09	0.00
mEnhA17	0.23	0.83	1.14	0.90	1.15	1.22	1.16	1.27	1.06	1.11	1.00	1.92	0.87	1.07	0.63	1.18	0.93	1.28	0.92	1.56	0.00	0.16	0.00
mEnhA18	0.31	0.99	1.13	1.00	1.07	1.03	1.19	1.10	0.77	1.12	1.03	1.55	0.95	1.41	0.69	0.92	0.99	1.09	1.15	1.55	0.00	0.47	0.00
mEnhWk1	0.80	1.16	0.93	0.95	0.95	1.12	1.14	0.97	1.22	1.16	1.29	1.34	0.84	1.14	0.95	1.05	1.16	1.19	1.09	1.07	0.00	0.33	0.00
mEnhWk2	0.45	1.03	1.16	0.82	1.16	1.27	1.07	0.96	1.19	1.27	1.23	1.64	0.94	1.05	0.86	1.01	0.93	1.35	0.93	1.20	0.00	0.06	0.00
mEnhWk3	0.47	0.91	1.11	0.71	1.43	1.41	1.06	1.22	1.05	1.71	0.81	1.44	1.15	0.95	0.90	1.25	0.84	1.11	1.03	0.84	0.00	0.04	0.00
mEnhWk4	0.41	1.02	1.17	1.02	0.96	1.15	1.14	1.11	1.04	1.25	1.12	1.42	0.80	1.08	0.87	1.08	1.04	1.16	1.12	1.59	0.00	0.15	0.00
mEnhWk5	0.87	1.14	1.02	1.19	0.91	1.13	1.18	0.95	1.13	1.13	1.17	0.99	0.88	1.32	0.97	1.14	1.11	0.92	1.16	1.67	0.00	0.20	0.00
mEnhWk6	0.42	1.25	1.15	1.07	0.85	1.04	1.19	0.93	0.73	1.08	1.36	1.40	0.96	1.24	0.84	1.11	1.23	1.06	0.75	1.49	0.00	0.31	0.01
mEnhWk7	1.37	1.34	1.06	1.11	0.80	0.84	1.15	0.90	0.79	1.14	1.36	1.24	1.00	1.35	1.02	1.11	1.27	1.27	0.89	1.26	5.41	0.22	0.01
mEnhWk8	1.57	1.14	1.08	1.34	0.99	1.01	1.16	0.77	1.13	1.04	1.22	0.87	1.22	1.11	1.20	0.94	1.36	0.88	1.30	0.93	0.00	0.24	0.00
mEnhWk9	0.62	1.16	1.09	1.25	1.04	1.12	1.16	0.80	1.11	1.20	1.22	0.77	1.22	1.11	1.09	1.14	1.30	0.88	1.23	0.83	0.00	0.16	0.00
mEnhWk10	0.73	1.13	1.09	1.12	0.89	0.98	1.29	0.82	1.12	1.21	1.12	1.03	0.95	1.43	0.99	0.99	0.97	1.09	1.42	1.36	0.00	0.23	0.00
mEnhWk11	0.92	1.16	1.15	1.29	0.98	0.96	1.13	0.86	1.11	1.16	1.17	1.03	1.07	1.37	1.15	0.98	0.94	0.91	1.36	1.10	0.00	0.11	0.00
mEnhWk12	0.35	0.93	1.15	0.88	1.05	1.23	1.12	1.17	0.93	1.19	1.04	1.81	0.95	1.05	0.74	1.07	0.91	1.35	0.91	1.68	0.00	0.20	0.00
mEnhWk13	0.90	1.15	1.09	1.13	0.90	1.03	0.99	0.81	0.68	1.10	1.01	1.38	1.12	1.29	1.01	1.15	0.93	1.17	1.00	1.58	0.00	0.64	0.01
mTxEnh1	0.24	0.73	1.30	0.63	1.27	1.39	0.86	1.18	1.27	1.18	0.98	2.10	0.96	0.76	0.87	1.48	0.82	1.44	0.74	1.42	0.00	0.01	0.00
mTxEnh2	0.36	0.89	1.21	0.84	1.22	1.28	0.95	0.98	1.10	1.26	1.01	1.86	0.96	0.95	0.79	1.35	1.01	1.29	0.97	1.21	0.00	0.11	0.00
mTxEnh3	0.75	0.96	1.27	0.95	0.98	1.07	1.05	0.83	0.92	1.30	1.01	1.66	1.06	1.03	0.87	1.36	1.09	1.10	1.11	1.42	0.00	0.28	0.00
mTxEnh4	0.25	0.77	1.31	0.67	1.27	1.35	0.79	1.34	1.04	1.20	0.96	2.28	0.85	0.79	0.78	1.16	0.99	1.38	0.86	1.53	0.00	0.08	0.00
mTxEnh5	0.37	0.78	1.28	0.55	1.17	1.41	0.85	1.31	1.50	1.27	0.99	2.06	1.03	0.60	0.64	1.22	0.89	1.59	0.80	1.41	0.00	0.04	0.00
mTxEnh6	0.19	0.88	1.36	0.72	1.51	1.34	0.89	1.16	1.28	0.94	0.88	1.85	1.01	0.69	0.77	1.26	0.90	1.31	1.23	0.87	0.00	0.10	0.00
mTx1	0.96	1.03	1.24	0.90	1.09	1.14	0.93	0.81	0.90	1.38	0.94	1.49	1.11	1.05	0.93	0.86	1.07	0.93	1.10	1.23	0.00	0.71	0.09
mTx2	1.07	1.04	1.23	0.92	1.16	1.16	0.91	0.97	0.97	1.51	0.95	1.43	1.16	1.01	0.91	0.80	1.12	1.00	1.11	1.11	0.00	0.43	0.04
mTx3	0.96	1.05	1.36	0.84	1.17	1.28	0.87	1.00	0.95	1.55	0.86	1.74	1.02	0.96	0.88	0.84	1.14	0.98	1.08	1.18	0.00	0.24	0.01
mTx4	0.91	0.98	1.36	0.74	1.23	1.48	0.82	1.16	1.11	1.41	0.89	1.91	0.93	0.93	0.71	0.93	1.16	1.00	1.00	1.29	0.00	0.04	0.00
mTx5	0.62	0.86	1.23	0.68	1.23	1.53	0.84	1.32	1.38	1.21	0.93	1.62	1.11	0.74	0.77	1.12	1.07	1.35	0.87	1.21	0.00	0.09	0.00
mTx6	0.44	0.94	1.25	0.83	1.09	1.25	1.00	0.95	1.01	1.40	0.94	1.99	1.00	0.93	0.81	0.95	1.18	1.20	1.03	1.46	0.00	0.15	0.00
mTx7	0.68	0.96	1.23	0.77	1.04	1.24	1.20	1.09	1.26	1.19	1.13	1.50	1.04	0.93	0.88	1.03	1.03	1.33	1.02	1.32	0.00	0.06	0.00
mTx8	0.35	0.94	1.29	0.83	1.21	1.27	1.30	1.34	1.15	1.63	1.03	1.09	1.35	0.75	0.55	1.17	0.92	0.99	1.05	0.74	0.00	0.12	0.00
mTxEx1	0.46	0.54	1.30	0.49	1.39	1.81	0.62	1.62	1.42	1.04	1.08	2.24	0.68	0.50	0.55	1.55	0.79	1.86	0.56	1.36	0.00	0.02	0.00
mTxEx2	0.18	0.53	1.11	0.47	1.21	1.45	0.63	1.71	1.33	1.13	1.52	2.48	0.77	0.40	0.63	1.45	0.72	2.04	0.49	1.64	0.00	0.02	0.00
mTxEx3	0.04	0.57	1.11	0.64	0.96	1.11	0.60	1.88	0.78	1.35	1.50	2.51	0.57	0.64	0.55	1.16	0.48	2.21	0.50	2.53	0.00	0.24	0.02
mTxWk1	2.32	1.11	1.22	1.08	0.95	1.00	1.04	0.76	0.85	1.23	1.10	1.09	1.08	1.29	0.98	0.98	1.06	0.96	1.11	1.32	0.00	0.69	0.06
mTxWk2	1.23	1.10	1.20	1.03	1.00	1.07	1.23	0.89	0.89	1.32	1.16	1.01	1.00	1.27	0.94	0.92	1.04	1.02	1.16	1.19	0.00	0.47	0.01
mBivProm1	0.13	0.58	1.44	0.51	1.32	1.16	1.09	1.22	1.21	0.93	0.91	2.60	0.88	0.99	0.52	1.67	0.53	1.13	0.63	1.93	0.00	0.26	0.00
mBivProm2	0.14	0.80	1.43	0.66	1.26	1.17	0.87	1.51	1.11	1.16	0.90	1.91	0.87	0.94	0.72	1.25	0.58	1.29	0.82	1.64	0.00	0.32	0.00
mBivProm3	0.09	0.99	1.50	0.90	1.13	1.34	0.81	1.69	0.98	1.29	0.80	1.35	0.79	1.12	1.10	0.91	0.49	0.78	0.76	1.10	0.00	0.58	0.00
mPromF1	0.11	0.92	1.07	1.15	1.01	1.04	0.94	1.17	1.14	1.07	1.08	1.28	1.00	0.98	1.06	0.97	0.87	0.90	1.05	0.92	0.00	1.00	0.01
mPromF2	0.19	0.70	1.17	0.77	1.11	1.15	0.82	1.56	1.25	1.22	1.10	2.20	0.73	0.90	0.67	1.14	0.81	1.63	0.68	1.85	0.00	0.15	0.00
mPromF3	0.14	0.73	1.22	0.80	1.24	1.08	0.91	1.47	1.27	1.23	1.34	2.07	0.76	0.72	0.59	1.24	0.75	1.41	0.82	1.64	0.00	0.12	0.00
mPromF4	0.14	0.91	1.07	1.02	0.96	1.04	1.14	0.96	1.10	1.11	1.34	1.33	0.89	1.26	0.76	0.96	1.05	1.07	0.98	1.36	0.00	0.73	0.02
mPromF5	0.14	0.82	1.25	0.79	1.15	1.26	0.85	1.16	1.06	1.23	0.91	1.95	0.83	1.05	0.75	1.01	0.94	1.38	0.80	1.56	0.00	0.49	0.01
mPromF6	0.13	0.86	1.27	0.88	1.16	1.22	0.87	1.16	1.12	1.25	0.92	1.96	0.83	0.95	0.72	1.01	0.94	1.42	0.85	1.45	0.00	0.35	0.00
mPromF7	0.14	1.05	1.05	1.09	1.00	0.90	1.07	1.06	1.07	1.06	1.13	0.95	0.92	1.07	1.00	1.02	0.98	0.90	1.00	1.01	0.00	1.24	0.03
mTSS1	0.15	0.81	1.17	0.92	1.13	1.05	0.96	1.20	1.12	1.20	1.02	1.53	0.86	0.89	0.87	1.13	0.91	1.16	1.11	1.19	0.00	0.78	0.00
mTSS2	0.28	0.80	1.12	0.86	1.08	1.18	0.85	1.33	1.07	1.20	1.09	1.89	0.86	1.00	0.75	1.02	0.94	1.49	0.83	1.59	0.00	0.42	0.00
mTSS3	0.07	0.77	1.07	0.79	1.00	1.09	0.87	1.53	1.17	1.16	1.12	1.73	0.77	1.19	0.62	0.86	0.81	1.44	0.80	1.45	0.00	0.86	0.03
	100.0	7.16	6.67	5.86	5.73	5.56	5.48	5.33	4.74	4.56	4.79	4.47	4.40	4.41	4.57	3.81	3.60	3.48	3.32	2.25	0.00	6.26	3.36

B **States in the top 1 most enriched in with non-primary chromosomes**



State	Genome %	chr1_GL456210_random	chr1_GL456211_random	chr1_GL456212_random	chr1_GL456213_random	chr1_GL456221_random	chr4_GL456216_random	chr4_GL456350_random	chr4_JH584292_random	chr4_JH584293_random	chr4_JH584294_random	chr4_JH584295_random	chr5_GL456354_random	chr5_JH584296_random	chr5_JH584297_random	chr5_JH584298_random	chr5_JH584299_random	chr7_GL456219_random	chrM	chrUn_GL456239	chrUn_GL456359	chrUn_GL456360	chrUn_GL456366	chrUn_GL456367	chrUn_GL456368	chrUn_GL456370	chrUn_GL456372	chrUn_GL456378	chrUn_GL456379	chrUn_GL456381	chrUn_GL456382	chrUn_GL456383	chrUn_GL456385	chrUn_GL456387	chrUn_GL456389	chrUn_GL456390	chrUn_GL456392	chrUn_GL456393	chrUn_GL456394	chrUn_GL456396	chrUn_JH584304	chrX_GL456233_random	chrY_JH584300_random	chrY_JH584301_random	chrY_JH584302_random	chrY_JH584303_random
mGapArtf1	15	2.9	3.4	3.7	3	3.6	1.1	6.6	2.6	6.6	6.3	0	6.1	6.4	6.3	6.6	5.7	6.6	1.9	0.2	1.6	1.7	2.3	2	2.1	2.1	2.4	1.6	2.8	1.2	3.5	1.2	2	0.8	1.1	2.7	0.2	0.8	1	0.3	0	2.4	6.6	6.6	6.6	6.6
mGapArtf3	0	0	4	19	0	0	208	0	226	0	0	1591	0	0	0	0	0	0	29	597	147	0	0	11	0	341	0	121	0	0	0	383	0	679	217	466	222	172	79	225	1160	0	0	0	0	0
mHET4	0.5	0	0	0	0	0	9.7	0	0	0	0	0	0	0	0	0	0	0	75	0	10	3.6	2.4	0	0	1.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0
mHET7	0.2	19	38	43	0	24	4.5	0	6.7	0	0	0	0	0	0	0	0.2	0	0	206	0	0	2.1	17	39	56	31	28	0	285	0	191	0	166	202	141	244	342	95	202	4.35	6.5	0	0	0	0
mHET8	1.1	16	11	9.7	0	11	1.1	0	7.4	0	0	0	1.4	0.5	0.3	0	3.5	0	0	7.8	8	0	2.3	16	9.9	3.4	20	3.5	13	1.4	9.5	7.6	21	0	7.7	0.7	2.3	2.6	17	16	0.16	12	0	0	0	0
mHET9	2.8	7.7	6.6	4.8	0	7.5	1.6	0	11	0	0	0	2	0.9	1.4	0	3.3	0	0	3.2	5.6	2.5	1.4	13	5.3	3.2	4.7	2.9	13	2.2	11	2	11	0.3	4.2	1.5	2.1	1.3	6.2	2.4	0	9.4	0	0	0	0
mHET10	1.8	0.3	0.2	0.1	0	0.1	1.5	0	1.5	0	0	0	0.1	0	0	0	0	0	0	0.3	3	11	18	0	3.9	1.7	0.4	3.6	2	0	0	0	0	0	0	0.5	1	0	0	0.5	0	0	0	0	0	0
mZNF1	0.2	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
mEnhWk13	0.9	0	0	0	5.1	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Supplementary Figure 3. 3: Mouse full-stack states enrichments with different chromosomes.

(A) The first and second columns show the mouse full-stack states and their percent genome coverage. The following columns correspond to different chromosomes. The heatmap shows fold enrichments of each state with each chromosome. Coloring of the heatmap is column-specific. The last row shows the percentage of the genome that each chromosome covers. Certain states in polycomb repressed group (mReprPC4-6) show distinctly high enrichments with chromosome X. **(B)** The first and second columns show mouse full-stack states and their genome coverage, respectively. The following columns correspond to different scaffold chromosomes. Only states that show highest enrichments with at least one scaffold chromosome are shown. Within each column, the highest enrichment values across 100 mouse full-stack states are colored red. States in 'assembly gaps and alignment artifacts' or in heterochromatin groups show highest enrichments with multiple scaffold chromosomes.

Enrichment of mouse full-stack states with classes of repeat elements

State	Genome %	DNA	DNA?	LINE	LINE?	LTR	LTR?	Low_complexity	Other	RC	RC?	RNA	SINE	SINE?	Satellite	Simple_repeat	Unknown	rRNA	scRNA	snRNA	srpRNA	tRNA
mGapArtf1	15.10	0.23	0.08	2.14	0.00	1.17	0.06	0.66	0.93	0.20	0.09	0.07	0.23	0.05	1.48	0.57	3.63	0.46	0.40	0.32	0.07	0.15
mGapArtf2	0.25	0.51	0.00	0.52	0.00	2.89	0.56	0.42	2.00	0.00	0.00	2.21	2.08	0.00	1.11	0.63	0.21	8.62	0.56	0.78	3.58	0.93
mGapArtf3	0.04	0.47	0.00	0.23	0.00	1.03	0.84	0.84	0.99	0.00	0.00	0.00	2.56	0.00	22.47	1.21	0.37	95.49	0.99	2.29	0.00	1.79
mQuies1	2.79	1.82	2.33	0.44	1.39	0.82	2.16	1.25	0.50	1.91	3.24	1.00	0.93	3.20	0.86	1.23	0.37	0.78	1.08	1.10	0.74	1.08
mQuies2	8.01	1.66	0.88	0.87	0.61	0.97	1.87	1.87	0.76	1.99	1.17	0.47	1.41	0.43	1.13	2.03	0.28	0.96	1.44	1.67	1.00	1.27
mQuies3	16.39	1.10	1.01	1.23	1.51	0.70	1.47	1.69	0.52	1.27	1.14	0.54	0.55	0.73	0.65	1.48	0.20	0.56	0.66	0.96	0.39	0.44
mHET1	0.81	0.53	0.13	1.04	0.00	1.81	0.27	0.26	17.53	0.51	0.00	2.55	0.79	0.00	2.30	1.20	0.11	1.62	0.56	0.35	1.44	1.05
mHET2	2.41	1.19	1.05	0.74	0.00	1.38	0.79	0.70	2.08	1.14	1.73	2.09	1.00	0.18	1.52	1.23	0.18	0.99	0.72	0.67	1.18	0.74
mHET3	3.56	0.66	0.47	1.64	0.26	1.49	0.55	0.42	1.29	0.62	0.74	2.03	0.82	0.64	0.52	0.60	0.15	0.82	0.73	1.01	1.81	0.68
mHET4	0.52	0.72	0.54	1.24	0.00	1.57	0.38	0.31	0.90	1.75	0.00	3.87	1.10	1.75	0.72	0.55	0.29	0.80	0.80	0.71	3.13	0.50
mHET5	1.23	0.40	0.06	1.55	0.00	3.01	0.42	0.33	2.02	0.08	0.00	1.48	0.72	0.00	0.54	0.41	0.31	1.48	0.57	0.96	0.81	0.66
mHET6	0.74	0.27	0.07	1.49	0.00	3.17	0.14	0.23	1.55	0.04	0.00	0.64	0.59	0.00	1.37	0.29	2.31	3.39	0.57	0.75	0.00	0.83
mHET7	0.20	0.23	0.00	1.06	0.00	3.25	0.64	0.21	1.51	0.00	0.00	3.88	0.72	0.00	7.56	0.41	4.91	9.91	0.49	0.43	3.07	2.05
mHET8	1.10	0.29	0.00	2.07	0.00	2.24	0.01	0.40	0.75	0.00	0.00	0.15	0.55	0.00	1.21	0.48	2.72	2.23	0.53	1.21	0.27	0.60
mHET9	2.80	0.32	0.07	2.72	0.00	1.48	0.09	0.54	0.63	0.12	0.00	0.46	0.44	0.12	0.92	0.58	1.53	0.48	0.52	1.24	0.38	0.38
mHET10	1.77	0.59	0.21	1.83	0.59	2.35	0.53	0.70	0.58	0.73	0.00	0.40	0.53	0.41	0.50	0.66	0.71	0.44	0.54	0.93	1.00	0.48
mZNF1	0.24	0.65	0.00	1.08	0.00	2.30	0.30	0.38	0.40	0.00	0.00	1.14	1.51	0.00	4.16	0.55	14.29	0.49	0.72	1.01	1.25	0.83
mZNF2	0.13	0.60	0.85	0.62	0.00	1.99	0.40	0.24	0.94	0.00	0.00	0.88	1.84	0.00	19.89	0.36	35.11	3.10	0.81	0.63	2.83	1.25
mReprPC1	0.76	1.04	0.49	0.28	0.00	1.04	0.40	0.71	1.51	0.19	0.00	2.91	1.79	0.00	1.29	0.79	0.35	1.61	1.47	0.78	1.19	1.69
mReprPC2	2.65	1.25	0.26	0.51	0.00	1.36	0.65	0.69	1.46	0.77	0.61	2.74	1.85	0.07	1.28	1.04	0.34	1.07	1.47	1.42	1.67	1.59
mReprPC3	0.94	1.21	1.52	0.26	1.16	0.86	1.41	0.84	0.97	2.09	2.58	1.15	1.17	1.76	1.39	0.91	0.25	0.82	0.79	0.54	1.00	0.68
mReprPC4	0.10	1.19	0.00	0.39	0.00	0.61	1.25	0.44	0.66	0.00	0.00	1.64	2.11	0.00	0.35	0.44	0.24	0.38	1.98	1.16	0.90	1.43
mReprPC5	1.21	1.68	1.73	0.89	0.83	0.97	1.38	1.49	0.99	1.63	0.00	2.36	1.36	1.18	1.20	1.26	0.36	1.05	1.33	1.12	0.90	1.27
mReprPC6	0.49	1.53	0.67	0.54	0.48	0.99	1.85	0.88	1.64	0.58	5.94	1.74	1.80	0.93	0.93	0.77	0.45	2.38	1.39	1.71	3.21	1.44
mReprPC_openC1	0.15	0.61	3.66	0.08	0.00	0.19	1.31	0.55	0.21	0.00	0.00	0.00	0.62	0.00	0.31	0.44	0.20	0.31	0.96	0.21	0.00	0.69
mReprPC_openC2	0.38	0.97	2.30	0.14	16.59	0.50	0.82	0.67	0.97	0.13	0.00	1.32	1.06	1.54	0.95	0.67	0.24	0.43	1.04	0.15	1.64	0.97
mReprPC_openC3	0.22	0.74	0.00	0.11	0.00	0.42	0.00	0.53	0.89	0.00	0.00	0.08	1.34	0.00	0.70	0.55	0.27	1.24	1.22	0.98	0.31	1.65
mReprPC_openC4	0.22	0.65	1.66	0.10	0.00	0.25	0.57	0.68	0.22	0.00	0.00	0.77	1.14	2.73	0.73	0.78	0.37	0.67	1.57	1.60	0.58	1.84
mOpenC1	0.46	0.41	0.58	0.10	0.00	0.52	0.06	0.26	5.60	0.37	0.00	2.45	1.04	0.00	0.47	0.41	0.11	2.37	0.94	0.65	2.18	0.66
mOpenC2	1.15	0.82	0.23	0.24	0.00	1.00	0.62	0.37	5.24	0.33	0.00	2.71	1.75	0.00	0.75	0.75	0.22	1.57	1.34	0.74	1.08	1.45
mOpenC3	0.43	1.41	1.53	0.19	0.00	0.73	2.25	0.42	0.88	2.33	5.92	1.06	0.93	2.20	0.36	0.59	0.40	1.25	0.84	0.34	1.26	1.02
mOpenC4	0.90	1.83	2.04	0.30	0.00	0.80	1.74	0.77	0.63	2.10	2.35	1.68	1.00	5.09	0.62	0.79	0.42	0.86	1.06	0.84	2.15	1.22
mOpenC5	0.24	0.61	0.23	0.10	0.00	0.49	0.67	0.48	2.31	2.07	0.00	1.33	1.06	0.79	0.81	0.58	0.20	0.76	0.87	0.57	1.74	2.18
mOpenC6	0.14	0.76	0.00	0.08	0.00	0.25	0.34	0.54	0.83	0.00	0.00	2.56	0.92	0.26	0.73	0.62	0.28	6.47	0.56	5.64	4.24	5.47
mOpenC7	0.20	1.10	0.00	0.17	0.00	0.50	0.03	0.90	1.16	1.85	0.00	0.81	2.40	0.00	0.87	0.93	0.36	4.79	0.98	0.73	0.96	1.34
mOpenC8	0.22	1.29	0.78	0.26	0.00	0.82	0.15	1.18	7.07	0.00	2.92	0.82	1.91	2.12	4.42	2.40	0.31	7.70	1.47	6.01	4.05	4.15
mOpenC9	1.38	1.63	1.63	0.39	3.58	0.69	2.44	0.74	0.41	2.46	4.70	1.69	0.83	4.83	0.53	0.84	0.37	1.05	0.69	1.11	0.60	0.82
mEnhA1	0.26	1.22	2.28	0.13	9.39	0.34	3.64	0.37	0.39	4.70	0.00	1.11	0.52	4.21	0.32	0.46	0.72	0.38	0.84	0.57	0.00	1.09
mEnhA2	0.24	0.69	3.26	0.07	0.00	0.19	1.34	0.46	0.20	0.00	0.00	0.65	0.51	3.38	0.31	0.43	0.23	0.00	0.54	0.14	0.97	0.86
mEnhA3	0.15	0.46	4.19	0.05	0.00	0.13	0.71	0.40	0.06	2.15	0.00	0.64	0.52	0.00	0.31	0.38	0.13	0.36	0.63	0.47	0.00	0.95
mEnhA4	0.28	0.84	1.84	0.09	0.00	0.27	2.10	0.42	0.13	0.12	0.00	0.69	0.71	1.26	0.36	0.42	0.21	0.23	0.47	0.33	1.40	0.55
mEnhA5	0.28	0.97	0.84	0.12	0.00	0.59	0.28	0.44	0.71	1.08	0.59	0.88	1.19	0.63	0.61	0.52	0.26	0.97	1.71	0.44	0.27	1.51
mEnhA6	0.40	1.55	2.21	0.20	0.00	0.44	1.26	0.80	0.19	1.76	2.81	0.61	1.16	2.44	0.95	0.72	0.34	1.11	1.87	1.03	1.55	1.71
mEnhA7	0.38	0.96	3.14	0.11	9.47	0.40	1.88	0.64	0.25	3.34	1.17	0.47	0.77	4.18	0.61	0.57	0.32	1.06	0.57	0.83	2.48	0.79
mEnhA8	0.67	1.28	6.40	0.21	12.61	0.52	2.19	0.89	0.20	4.15	0.60	1.71	0.82	4.50	1.06	0.89	0.39	0.73	0.82	0.53	2.89	1.22
mEnhA9	0.36	1.06	9.95	0.12	11.26	0.29	4.15	0.57	0.16	2.60	2.04	1.68	0.51	5.63	0.74	0.63	1.15	0.82	0.49	0.24	1.57	0.82
mEnhA10	0.20	0.69	5.02	0.07	0.00	0.18	2.41	0.61	0.06	0.83	6.60	0.36	0.38	6.04	0.35	0.53	0.87	1.06	0.67	0.00	0.00	0.32
mEnhA11	0.13	0.56	3.01	0.06	0.00	0.17	0.16	0.72	0.04	0.00	0.00	0.00	0.44	11.60	0.25	0.51	0.20	0.20	0.90	0.97	0.75	0.67
mEnhA12	0.18	0.66	0.00	0.09	0.00	0.35	2.08	0.42	0.38	1.04	0.00	1.95	1.09	0.00	0.49	0.40	0.39	1.40	1.22	0.29	0.85	0.66
mEnhA13	0.28	1.04	0.47	0.17	0.00	0.72	0.88	0.53	0.76	0.67	0.00	1.49	1.29	0.00	0.49	0.65	0.17	1.76	2.06	0.49	4.76	1.39
mEnhA14	0.46	1.14	0.00	0.22	0.00	0.91	0.51	0.62	1.16	0.53	0.97	1.51	1.86	0.76	1.10	1.02	0.30	2.21	1.45	1.05	1.61	2.17
mEnhA15	0.34	1.00	0.46	0.16	0.00	0.50	1.37	0.49	0.70	0.00	0.00	1.66	1.14	2.24	0.72	0.60	0.28	0.32	1.15	0.89	1.45	1.73
mEnhA16	0.40	1.17	5.04	0.13	8.37	0.46	0.95	0.58	0.38	1.30	0.00	1.28	0.75	7.63	0.63	0.55	0.48	0.05	0.47	0.43	0.94	0.98
mEnhA17	0.23	1.08	0.00	0.18	0.00	0.63	0.26	0.57	1.31	3.02	0.00	0.71	1.73	0.74	0.57	0.73	0.31	2.26	1.80	0.70	0.35	3.01
mEnhA18	0.31	1.53	0.00	0.37	0.00	0.96	1.04	1.17	1.03	1.35	0.00	2.02	2.19	0.00	1.07	1.25	0.35	0.91	2.34	1.49	0.75	1.82
mEnhWk1	0.80	0.92	0.47	0.41	2.33	2.38	0.95	0.98	0.68	0.54	0.00	1.09	1.89	1.29	0.92	1.00	0.42	0.98	1.92	1.14	0.59	1.37
mEnhWk2	0.45	0.90	0.92	0.20	0.00	1.37	0.59	0.86	1.80	0.25	2.10	1.99	1.56	0.00	0.55	0.75	0.29	1.94	1.41	0.58	3.69	1.37
mEnhWk3	0.47	1.11	1.72	0.16	0.00	0.52	3.31	0.30	0.72	3.02	0.00	1.73	0.83	0.82	0.46	0.49	0.56	1.17	0.63	0.39	0.64	0.79
mEnhWk4	0.41	1.29	1.05	0.19	0.00	0.91	1.19	0.64	0.69	1.58	4.99	2.69	1.67	0.52	0.94	0.85	0.30	1.59	1.61	0.38	1.43	1.93
mEnhWk5	0.87	1.49	0.85	0.35	2.52	1.24	1.25	1.00	0.71	2.86	0.94	0.76	1.68	0.00	0.99	1.33	0.41	1.34	1.58	1.23	2.12	1.72
mEnhWk6	0.42	1.58	0.36	0.32	1.73	0.96	1.13	1.07	0.46	0.73	0.00	1.57	1.38	1.15	0.98	1.02	0.37	2.61	1.78	1.24	2.08	1.52
mEnhWk7	1.37	1.59	0.23	0.46	0.35	1.08	0.84	1.29	0.78	1.04	1.16	0.70	1.63	0.57	1.71	1.51	0.34	0.99	1.75	1.48	1.15	1.14
mEnhWk8	1.57	1.49	8.96	0.32	9.14	0.52	2.88	1.15	0.39	2.77	3.28	0.69	0.73	4.29	1.10	1.22	0.57	0.58	0.76	0.67	0.37	0.79
mEnhWk9	0.62	1.63	8.46	0.22	3.47	0.40	4.18	0.56	0.18	1.86	2.03	1.87	0.57	7.81	0.50	0.64	1.25	0.26	0.52	0.79	0.57	0.80
mEnhWk10	0.73	1.47	1.08	0.34	3.14	0.79	2.16	0.91	0.59	1.28	0.00	1.83	1.29	1.25	1.18	0.96	0.30	1.41	1.33	0.90	1.90	1.43
mEnhWk11	0.92	1.59	4.56	0.23	5.35	0.68	2.21	0.86	0.33	1.82	5.27	1.17	0.91	11.64	0.96	0.81	0.57	0.43	0.88	0.59	0.42	1.01
mEnhWk12	0.35	1.07	0.23	0.22	0.00	1.10	0.77	0.64	2.31	0.38	0.00	2.07	2.11	0.56	0.96	0.80	0.49	2.12	2.03	1.27	1.81	1.62
mEnhWk13	0.90	1.54	0.95	0.53	0.00	1.20	1.04	1.37	0.85	0.99	3.68	1.92	1.85	1.09	0.97	1.36	0.34	0.92	1.84	1.56	1.66	1.75
mTxEnh1	0.24	0.91	0.00	0.13	0.00	0.38	0.15	0.24	0.83	0.25	3.83	1.51	1.44	1.94	0.45	0.39	0.31	1.63	1.54	0.58	3.40	1.57
mTxEnh2	0.36	1.45	0.72	0.25	0.00	0.56	0.69	0.52	0.64	2.25	1.46	2.34	2.48	1.20	0.38	0.59	0.31	1.42	2.74	1.32	2.55	2.77
mTxEnh3	0.75	1.80	0.50	0.39	0.00	0.60	0.55	0.94	0.48	1.20	1.08	1.64	2.42	1.16	0.46	0.92	0.34	0.75	2.44	2.39	2.23	2.37
mTxEnh4	0.25	0.83	0.00	0.14	0.00	0.52	0.56	0.23	0.47	0.00	0.00	1.87	2.29	0.00	0.23	0.29	0.32	0.92	2.28	1.48	2.40	2.83
mTxEnh5	0.37	0.57	0.17	0.10	0.00	0.42	0.71	0.25	1.82	0.50	0.00	1.93	1.22	0.61	0.53	0.35	0.14	1.12	0.83	0.36	2.19	1.76
mTxEnh6	0.19	0.79	3.11	0.11	0.00	0.29	0.52	0.25	0.34	0.00	0.00	0.00	0.95	0.27	0.47	0.40	0.28	0.76	0.78	0.00	0.00	0.99
mTx1	0.96	1.43	0.31	0.48	0.00	0.51	0.83	1.55	0.30	0.00	0.19	0.69	2.01	0.39	0.61	1.14	1.09	0.52	2.18	1.85	1.05	1.75
mTx2	1.07	1.23	0.30	0.37	0.00	0.60	0.47	0.88	0.37	0.49	1.31	1.60	2.52	0.44	0.75	0.80	0.54	0.68	2.66	2.35	2.23	2.48
mTx3	0.96	1.16	0.03	0.25	0.00	0.54	0.70	0.63	0.32	0.24	2.47	1.26	2.73	0.86	0.60	0.51	0.61	1.50	2.46	2.19	2.44	2.11
mTx4	0.91	0.77	0.00	0.14	0.00	0.49	0.31	0.24	0.37	0.51	0.00	2.78	2.44	0.33	0.69	0.26	0.66	0.78	2.12	1.31	2.74	2.30
mTx5	0.62	0.74	0.16	0.16	0.00	0.57	0.00	0.29	0.93	0.56	0.00	1.95	1.86	0.00	0.70	0.47	0.30	1.03	1.51	0.92	1.24	1.56
mTx6	0.44	1.18	1.29	0.27	0.00	0.73	0.89	0.53	0.47	0.94	0.00	1.47	2.64	1.79	0.33	0.57	0.45	1.37	2.26	2.30	3.80	2.44
mTx7	0.68	1.01	0.56	0.21	0.00	0.67	0.55	0.38	1.16	0.29	2.83	0.93	1.87	0.00	0.62	0.70	0.21	1.48	1.61	1.07	1.77	2.22
mTx8	0.35	0.91	0.46	0.22	0.00	0.56	0.52	0.36	0.48	1.17	0.00	2.38	1.45	0.00	1.29	0.64	0.39	0.64	1.23	0.98	1.06	1.21
mTxEx1	0.46	0.28	0.09	0.05	0.00	0.23	0.00	0.14	0.86	0.00	0.00	2.73	0.96	0.00	0.36	0.19	0.10	0.54	0.59	0.74	1.96	0.81
mTxEx2	0.18	0.34	0.35	0.04	0.00	0.16	0.39	0.17	0.68	0.64	0.00	2.60	0.87	0.00	0.41	0.20	0.09	0.25	0.37	0.00	2.19	0.77
mTxEx3	0.04	0.31	0.00	0.04	0.00	0.15	0.00	0.27	0.04	0.00	0.00	2.27	1.22	0.00	8.11	0.23	0.13	49.46	0.89	4.23	6.66	1.29
mTxWk1	2.32	1.58	0.39	0.71	0.00	0.55	0.84	1.98	0.32	0.20	0.85	0.10	1.51	0.48	0.66	1.52	0.79	0.81	1.69	1.56	1.31	1.35
mTxWk2	1.23	1.33	0.93	0.51	0.00	0.83	0.52	0.64	0.51	0.92	1.14	0.87	1.83	0.19	0.82	0.87	0.63	0.92	1.55	1.26	1.65	1.65
mBivProm1	0.13	0.19	0.26	0.03	0.00	0.07	0.00	0.68	0.10	0.00	0.00	0.00	0.41	0.00	0.45	0.47	0.21	2.75	0.15	0.33	0.00	4.26
mBivProm2	0.14	0.10	0.18	0.01	0.00	0.01	0.28	1.84	0.07	0.00	0.00	0.00	0.10	4.22	0.11	0.50	0.06	3.28	0.20	0.47	0.00	1.66
mBivProm3	0.09	0.24	0.45	0.03	0.00	0.05	0.00	1.05	0.00	0.00	0.00	0.00	0.26	0.00	0.37	0.76	0.21	0.00	0.38	0.00	0.00	0.13
mPromF1	0.11	0.28	0.03	0.05	0.00	0.14	0.00	1.90	0.01	0.00	0.00	0.00	0.28	1.45	0.53	1.04	0.01	0.91	0.39	0.13	0.00	1.42
mPromF2	0.19	0.34	0.00	0.04	0.00	0.11	0.00	0.50	0.05	0.00	0.00	0.00	0.83	1.76	0.18	0.34	0.07	0.33	0.70	1.30	0.00	1.16
mPromF3	0.14	0.42	0.68	0.06	0.00	0.21	1.21	0.53	0.07	0.96	0.00	1.05	0.88	0.00	0.46	0.45	0.21	0.23	0.48	0.47	0.79	2.15
mPromF4	0.14	0.85	2.49	0.10	0.00	0.37	0.32	1.35	0.14	0.00	0.00	0.00	0.92	6.32	0.92	1.13	0.43	2.60	1.25	0.78	0.00	4.06
mPromF5	0.14	0.91	0.00	0.08	0.00	0.19	0.00	0.67	0.05	1.70	0.00	0.52	1.65	0.00	0.20	0.55	0.32	0.77	1.87	1.68	1.82	0.99
mPromF6	0.13	1.27	0.00	0.14	0.00	0.30	0.00	0.62	0.13	0.19	0.00	2.20	2.55	0.00	0.18	0.57	0.52	0.90	2.53	1.18	6.92	3.07
mPromF7	0.14	0.78	0.38	0.12	0.00	0.22	0.00	1.41	0.03	0.00	0.00	0.00	0.77	0.00	0.60	1.23	0.10	0.95	1.75	0.57	0.94	0.83
mTSS1	0.15	0.07	0.00	0.01	0.00	0.04	0.00	3.59	0.01	0.00	0.00	0.00	0.06	0.00	0.15	0.83	0.00	5.20	0.19	0.00	0.00	0.38
mTSS2	0.28	0.08	0.14	0.01	0.00	0.03	0.00	3.75	0.01	0.00	0.00	0.00	0.16	0.00	0.08	0.83	0.04	1.06	0.36	0.96	0.00	3.18
mTSS3	0.07	0.26	1.07	0.03	0.00	0.09	0.00	3.02	0.00	0.00	0.00	0.00	0.40	0.00	0.27	1.09	0.01	0.63	2.54	13.79	13.90	54.07
	100.00	1.08	0.01	20.08	0.0003	11.68	0.01	0.75	0.30	0.002	0.0004	0.004	7.48	0.001	0.19	2.17	0.10	0.01	0.02	0.01	0.002	0.01

Supplementary Figure 3. 4: Mouse full-stack states enrichments with different classes of repeats *(Smit* et al.*, 2015)*.

The first and second columns show the mouse full-stack states and their genome coverage. The following columns correspond to different repeat classes. The following columns correspond to different classes of repeat elements (with elements named with '?' excluded). The heatmap shows fold enrichments of each state with each repeat class. Coloring of the heatmap is column-specific. The last row shows the percentage of the genome that each repeat class covers.

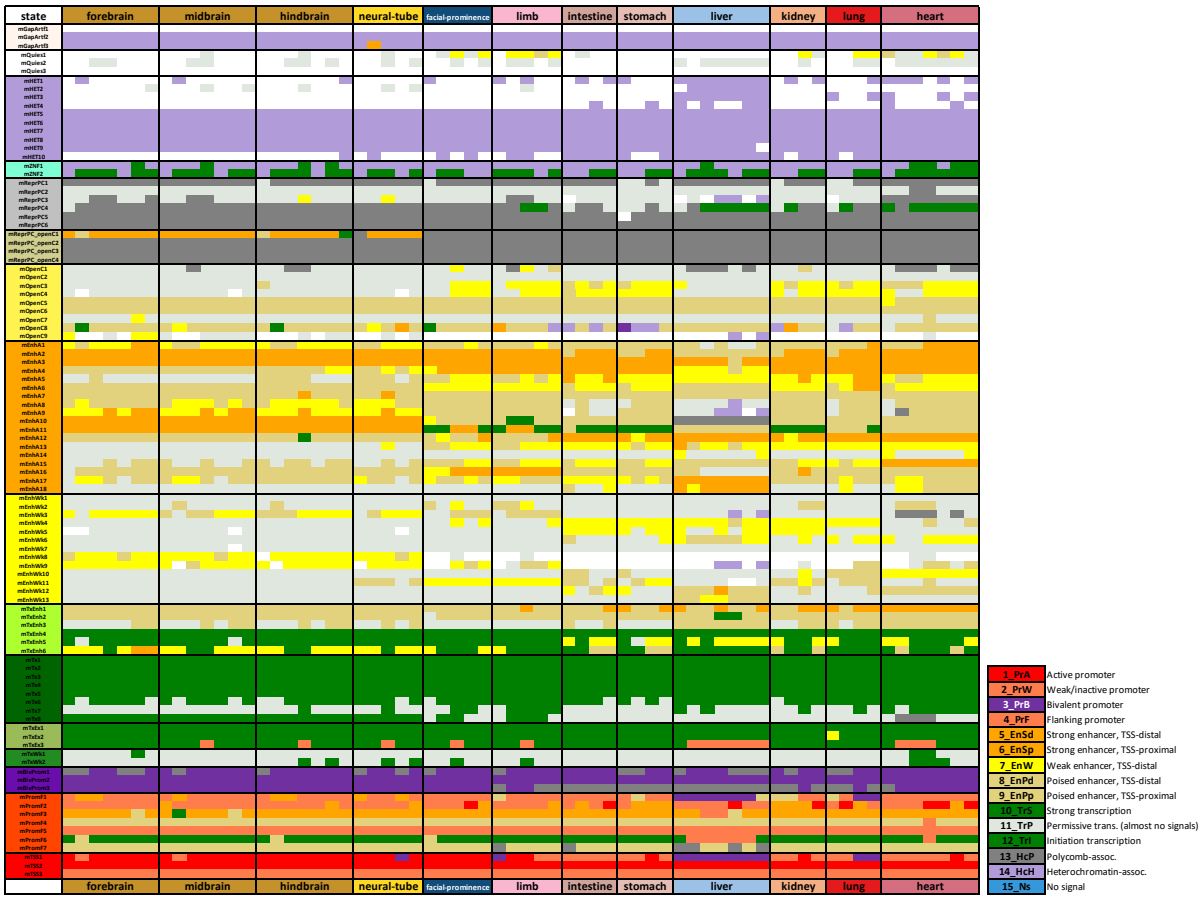
Top mouse full-stack states most enriched with classes of repeat elements

State	Genome %	snRNA	srpRNA	tRNA	Low_complexity	SINE	scRNA	RC	Simple_repeat	DNA	Unknown	LINE	LTR	RNA	Other	Satellite	rRNA
mGapArtf3	0.04	2.29	0.00	1.79	0.84	2.56	0.99	0.00	1.21	0.47	0.37	0.23	1.03	0.00	0.99	22.47	95.49
mHET1	0.81	0.35	1.44	1.05	0.26	0.79	0.56	0.51	1.20	0.53	0.11	1.04	1.81	2.55	17.53	2.30	1.62
mHET7	0.20	0.43	3.07	2.05	0.21	0.72	0.49	0.00	0.41	0.23	4.91	1.06	3.25	3.88	1.51	7.56	9.91
mHET9	2.80	1.24	0.38	0.38	0.54	0.44	0.52	0.12	0.58	0.32	1.53	2.72	1.48	0.46	0.63	0.92	0.48
mZNF2	0.13	0.63	2.83	1.25	0.24	1.84	0.81	0.00	0.36	0.60	35.11	0.62	1.99	0.88	0.94	19.89	3.10
mOpenC4	0.90	0.84	2.15	1.22	0.77	1.00	1.06	2.10	0.79	1.83	0.42	0.30	0.80	1.68	0.63	0.62	0.86
mOpenC8	0.22	6.01	4.05	4.15	1.18	1.91	1.47	0.00	2.40	1.29	0.31	0.26	0.82	0.82	7.07	4.42	7.70
mEnhA1	0.26	0.57	0.00	1.09	0.37	0.52	0.84	4.70	0.46	1.22	0.72	0.13	0.34	1.11	0.39	0.32	0.38
mTxEnh2	0.36	1.32	2.55	2.77	0.52	2.48	2.74	2.25	0.59	1.45	0.31	0.25	0.56	2.34	0.64	0.38	1.42
mTx3	0.96	2.19	2.44	2.11	0.63	2.73	2.46	0.24	0.51	1.16	0.61	0.25	0.54	1.26	0.32	0.60	1.50
mTSS2	0.28	0.96	0.00	3.18	3.75	0.16	0.36	0.00	0.83	0.08	0.04	0.01	0.03	0.00	0.01	0.08	1.06
mTSS3	0.07	13.79	13.90	54.07	3.02	0.40	2.54	0.00	1.09	0.26	0.01	0.03	0.09	0.00	0.00	0.27	0.63

Highest enr. across states

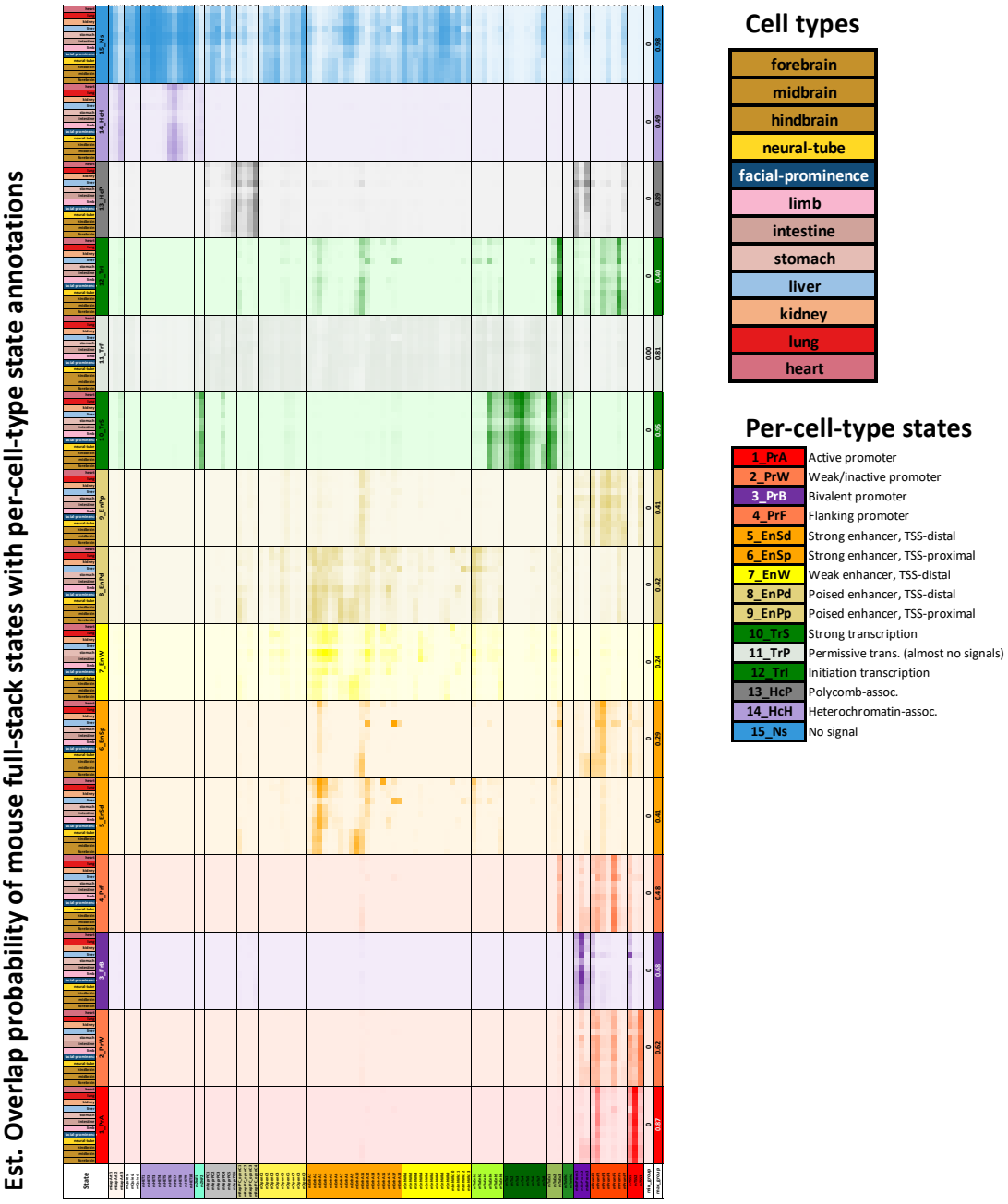
Supplementary Figure 3. 5: Enrichment of select mouse full-stack states with different classes of repeat elements *(Smit* et al.*, 2015)*.
The first and second columns show mouse full-stack states and their genome coverage, respectively. The following columns correspond to different classes of select repeat elements. Only states that show highest enrichments with at least one repeat class are shown. Within each column, the highest enrichment values across 100 mouse full-stack states are colored red.

Associated per-cell-type state with each mouse full-stack state, by tissue type



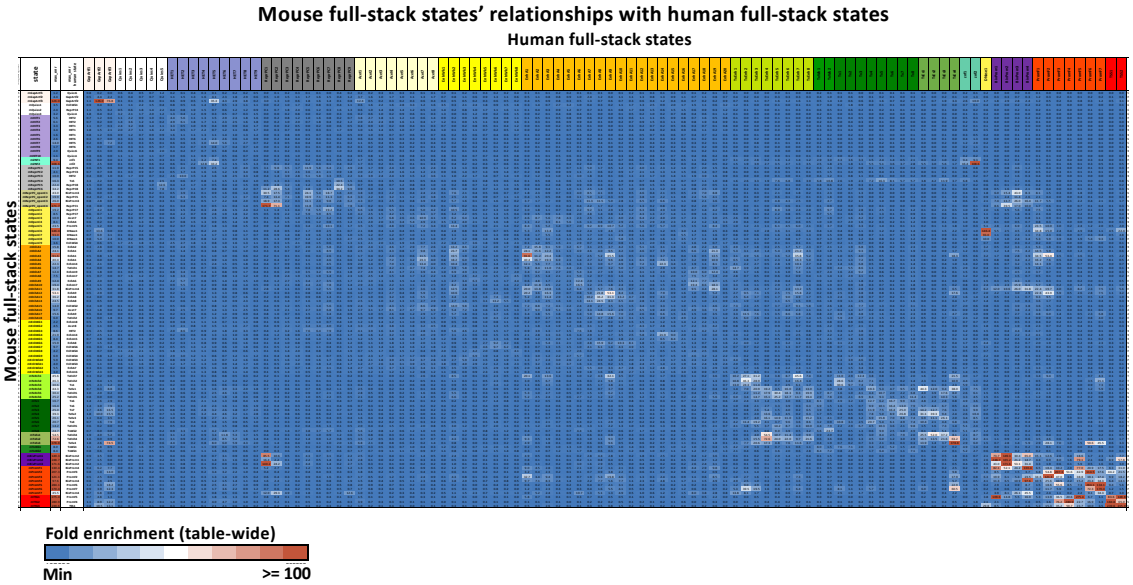
Supplementary Figure 3. 6: Full-stack states maximum-enrichments with annotated concatenated-model chromatin states in 66 mouse reference epigenomes *(Gorkin* et al.*, 2020)*.

Each row corresponds to one of 100 mouse full-stack state (**Methods**). Each column corresponds to a reference epigenome, grouped by the associated cell types as colored at the top and bottom. Each color entry corresponds to a reference epigenome and mouse full-stack state combination. The color corresponds to the chromatin state from the concatenated 15-state model annotating the respective mouse reference epigenome that is most enriched with the respective mouse full-stack state. Description of states in the per-cell-type 15-state concatenated model is in the bottom (Gorkin et al., 2020). The figure highlights how some mouse full-stack states are maximally enriched with the same concatenated-model chromatin states across all the reference epigenomes; for example, states mTx1-5 are maximally enriched with the strong transcription state in all 66 reference epigenomes' 15-state concatenated annotation. Other mouse full-stack states are enriched for distinct concatenated states in different cell types, for example state mEnhA17-- characterized as an enhancer state in liver, spleen and bone marrow based on emission probabilities of enhancer associated marks-- is most enriched with an active enhancer in liver cell types, while being most enriched with poised/weak enhancer states in others. Detailed description of each mouse full-stack state enrichment patterns with concatenated states can be found in **Supplementary Data 3.4**.



Supplementary Figure 3. 7: Estimated probabilities of per-cell-type concatenated-model chromatin states overlapping with mouse full-stack states.
The figure shows estimated probabilities of per-cell-type chromatin state annotations overlapping with mouse full-stack states observed in different cell groups *(Gorkin* et al.*, 2020)*. This figure is also provided as an excel file in **Supplementary Data 3.4**. The figure is based on a 15-state per-cell type chromatin state model trained on 66 mouse reference epigenomes from 12 cell groups *(Gorkin* et al.*, 2020)*. Each row corresponds to a combination of per-cell type state (among 15 states) and cell group, as denoted in the first two columns and legends on the right and matching with the colors in **Supplementary Figure 3.5**. We note that we changed here the concatenated-model no-signal state from white to blue for better visibility. Rows corresponding to the same per-cell-type model state are grouped together (into 15 bigger rows). The 100 following columns

correspond to 100 mouse full-stack states. Values in the heatmap correspond to the estimated probability a genomic position annotated as a mouse full-stack state (column) is also annotated as a concatenated-model state in a reference epigenome from the corresponding cell group (row) (**Methods**). The last two columns show the minimum and maximum probabilities observed for each per-cell type state for any combination of tissue group and mouse full-stack state. The heatmap colors correspond to the 15-state's colors and are scaled such that the maximum probability value in each row block is colored darkest (as seen in the right most column). The figure complements **Supplementary Figure 3.5** in providing information on how each full-stack state can correspond to different per-cell-type states, hence stratifying mouse full-stack states' characteristics in more details. For example, mouse full-stack state mTSS1 shows high probabilities of overlapping bivalent promoter state in liver cells, and moderate probabilities of overlapping the flanking/weak promoter state in other cell groups.



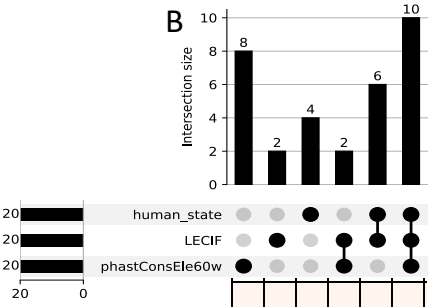
Supplementary Figure 3. 8: Enrichments of mouse full-stack states with human full-stack states (Vu and Ernst, 2022).
The first three columns show mouse full-stack states, the maximum fold enrichment with a human full-stack state (across all human states) and the corresponding human state, respectively. The following columns show the overlap enrichments with of each mouse state (rows) with each human state (columns). Across all pairs of states, the smallest enrichment values are colored blue and enrichment values >= 100 are colored red. This figure is also provided in **Supplementary Data 3.3.**

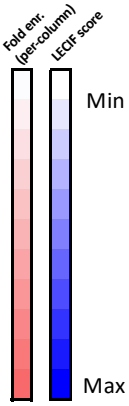
Mouse full-stack states' relationships with functional assay conservation (LECIF), sequence conservation (PhastCons) and human full-stack states

A

State	avg_lecif	Enr. With phastConsEle60w	Enr. With human state	Most enr. Human state
mGapArtf1	0.05	0.34	1.23	Quies5
mGapArtf2	0.10	0.72	4.42	GapArtf2
mGapArtf3	0.18	1.62	125.97	GapArtf2
mQuies1	0.05	1.25	3.50	EnhWk1
mQuies2	0.05	0.76	2.40	ReprPC4
mQuies3	0.05	0.70	3.21	Quies1
mHET1	0.08	0.82	5.62	HET2
mHET2	0.06	0.65	6.08	HET2
mHET3	0.05	0.47	3.32	HET1
mHET4	0.05	0.76	3.41	HET1
mHET5	0.06	0.42	1.60	HET1
mHET6	0.06	0.52	2.60	HET6
mHET7	0.08	0.75	14.00	HET5
mHET8	0.05	0.45	1.70	HET6
mHET9	0.04	0.37	1.99	Quies5
mHET10	0.04	0.38	1.14	Quies4
mZNF1	0.15	0.91	9.40	znf1
mZNF2	0.19	1.85	108.64	znf2
mReprPC1	0.21	0.77	11.92	ReprPC5
mReprPC2	0.08	0.64	4.14	ReprPC6
mReprPC3	0.11	0.94	10.02	HET2
mReprPC4	0.16	2.37	16.32	TxS
mReprPC5	0.08	1.06	22.39	ReprPC8
mReprPC6	0.22	1.17	34.15	ReprPC8
mReprPC_openC1	0.41	2.49	43.00	BivProm3
mReprPC_openC2	0.24	1.36	19.89	ReprPC5
mReprPC_openC3	0.38	1.74	25.99	BivProm3
mReprPC_openC4	0.60	1.90	171.74	ReprPC1
mOpenC1	0.27	1.64	19.20	ReprPC7
mOpenC2	0.14	0.95	8.50	ReprPC7
mOpenC3	0.12	1.72	10.86	Acet7
mOpenC4	0.08	1.54	8.15	EnhA4
mOpenC5	0.28	1.47	23.50	PromF1
mOpenC6	0.43	1.51	146.36	DNase1
mOpenC7	0.26	1.18	93.01	DNase1
mOpenC8	0.12	1.05	9.97	DNase1
mOpenC9	0.06	1.58	3.62	EnhWk4
mEnhA1	0.20	2.78	15.84	EnhA2
mEnhA2	0.36	2.68	24.06	EnhA1
mEnhA3	0.53	2.63	91.61	EnhA1
mEnhA4	0.34	2.14	40.42	EnhA1
mEnhA5	0.26	1.15	22.28	EnhA14
mEnhA6	0.19	1.64	14.66	TxEnh1
mEnhA7	0.21	2.35	12.37	EnhA19
mEnhA8	0.12	2.53	7.50	EnhA17
mEnhA9	0.20	4.02	12.38	EnhA6
mEnhA10	0.35	4.44	19.36	EnhA17
mEnhA11	0.45	4.00	31.83	BivProm4
mEnhA12	0.43	1.54	53.11	EnhA9
mEnhA13	0.27	1.09	36.20	EnhA8
mEnhA14	0.16	0.76	24.46	EnhA8
mEnhA15	0.21	1.37	13.98	EnhWk2
mEnhA16	0.17	2.42	9.73	Acet7
mEnhA17	0.27	1.04	15.61	EnhA9
mEnhA18	0.15	0.77	5.99	TxEnh2
mEnhWk1	0.06	0.61	2.17	EnhA18
mEnhWk2	0.11	0.87	4.53	Acet8
mEnhWk3	0.16	1.54	8.64	HET2
mEnhWk4	0.15	0.87	21.92	EnhA14
mEnhWk5	0.07	0.75	8.39	EnhA15
mEnhWk6	0.15	1.02	15.95	EnhA8
mEnhWk7	0.08	0.75	9.74	EnhWk6
mEnhWk8	0.07	2.38	4.17	EnhWk4
mEnhWk9	0.09	3.50	6.91	EnhWk4
mEnhWk10	0.10	0.98	5.94	EnhWk3
mEnhWk11	0.08	2.27	4.01	EnhWk3
mEnhWk12	0.14	0.73	5.54	EnhA7
mEnhWk13	0.07	0.73	2.99	EnhA16
mTxEnh1	0.47	1.07	45.64	TxEnh7
mTxEnh2	0.44	0.87	46.25	TxEnh2
mTxEnh3	0.26	0.90	30.61	Tx1
mTxEnh4	0.35	1.81	44.30	TxEx1
mTxEnh5	0.35	2.13	30.36	TxEnh5
mTxEnh6	0.34	2.44	25.24	TxEnh5
mTx1	0.23	1.53	16.73	Tx6
mTx2	0.24	1.55	16.38	Tx6
mTx3	0.25	1.81	25.03	Tx7
mTx4	0.30	2.54	39.28	TxEx2
mTx5	0.29	1.99	26.06	TxEx3
mTx6	0.27	1.16	20.74	Tx4
mTx7	0.22	1.36	14.22	TxEnh5
mTx8	0.34	1.95	24.72	TxWk2
mTxEx1	0.39	4.06	54.46	TxEnh4
mTxEx2	0.48	3.54	72.57	TxEnh4
mTxEx3	0.63	3.93	378.83	TxEx4
mTxWk1	0.14	1.19	7.95	TxWk1
mTxWk2	0.15	1.18	8.41	TxWk1
mBivProm1	0.61	3.13	149.26	BivProm2
mBivProm2	0.69	4.59	248.67	BivProm1
mBivProm3	0.71	3.79	171.56	BivProm2
mPromF1	0.57	4.20	130.43	BivProm4
mPromF2	0.60	2.26	207.34	PromF3
mPromF3	0.52	2.40	121.14	PromF2
mPromF4	0.40	2.28	97.56	BivProm4
mPromF5	0.67	1.51	284.05	PromF6
mPromF6	0.63	1.29	170.37	PromF7
mPromF7	0.43	2.88	45.46	BivProm4
mTSS1	0.65	4.63	275.41	PromF5
mTSS2	0.67	3.92	280.90	PromF4
mTSS3	0.63	3.16	239.61	TSS1

B



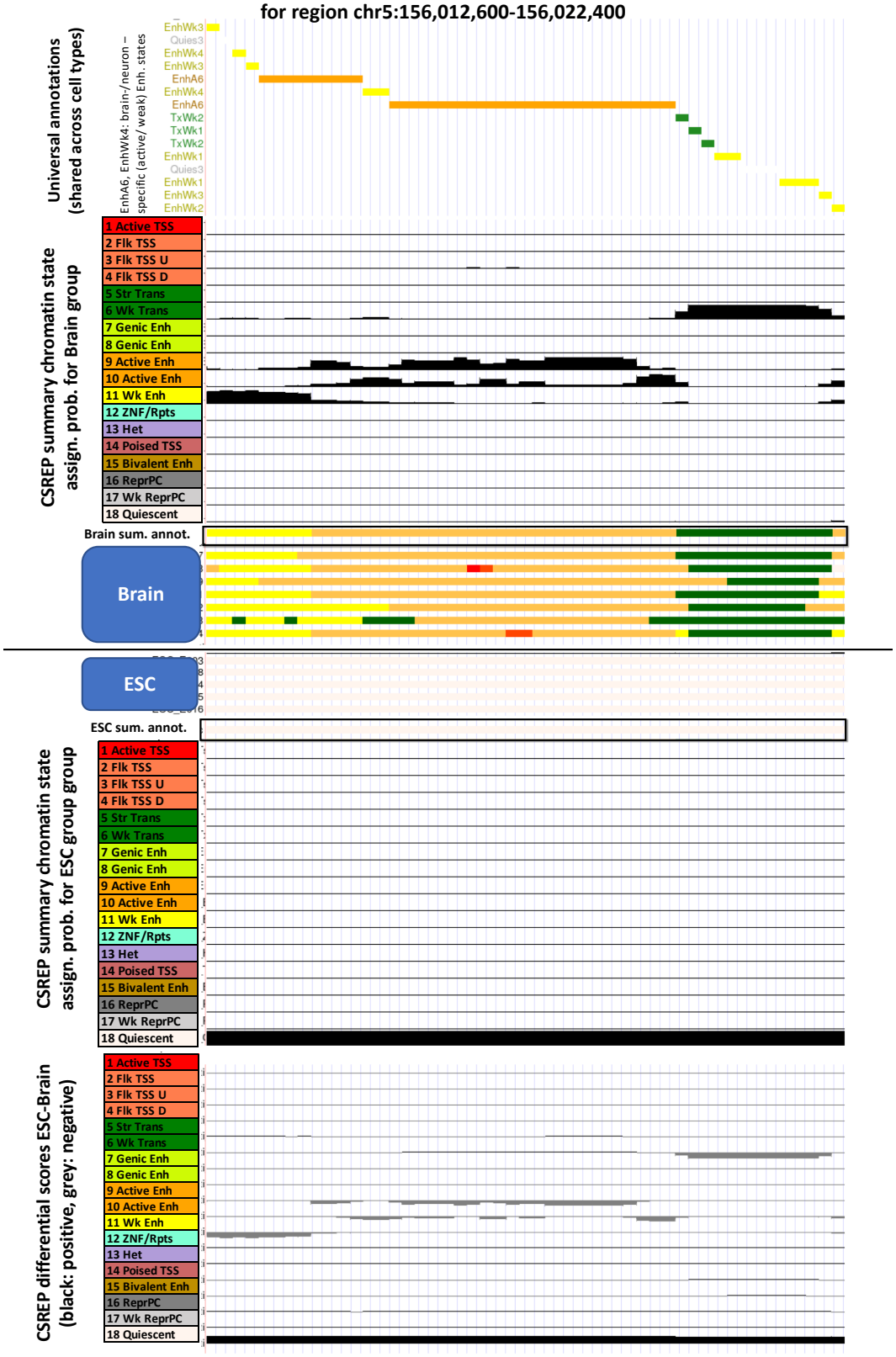


Supplementary Figure 3. 9: Mouse full-stack states' relationship with LECIF scores, human full-stack states and phastCons elements.
LECIF scores were developed to measure the level of evidence of human-mouse conservation at functional/epigenomic levels, with higher score (maximum of 1 and minimum of 0) implies higher evidence of conservation (Kwon and Ernst, 2021). PhastCons elements correspond to genomic regions showing strong 60-way multi-species sequence alignment conservation *(Siepel* et al*., 2005)*. Human full-stack states were learned from >1,000 Chip-seq/DNase-seq datasets in human, and provide annotation of the human genome that is shared across cell/tissue types (Vu and Ernst, 2022). **(A)** The heatmap shows mouse full-stack states (rows)' average LECIF scores, enrichments with phastCons elements and the maximum enrichments with human full-stack states. The first and second columns show the mouse full-stack states, and the percentage of the genome that each state covers. Coloring of the next 3 columns is column specific, as specified in legend. The last row shows the percentage of the genome that each LECIF score range covers. **(B)** Upset plot showing the number of states that are among the top 20 states with either (1) highest average LECIF score, or (2) hig`hest enrichments with PhastCons elements or (3) highest maximal enrichments with human full-stack states. Within each category, the column below the upset plot lists states that are in the top 20 most associated (as measured by average LECIF scores or fold enrichments) with the combination enrichment contexts.

Supplementary figures for chapter 4

A framework for group-wise summarization and comparison of chromatin state annotations

Visualization of ESC and Brain sample's input chromatin state maps and CSREP's output for region chr5:156,012,600-156,022,400



Supplementary Figure 4. 1: Visualization of ESC and Brain sample's input chromatin state maps and CSREP's output for region chr5:156,012,600-156,022,400, hg19.
All the sections of tracks are annotated in the legend on the left. The first section of tracks shows the universal chromatin state annotation that can annotate the epigenome across cell types, with states EnhA6 and EnhWk4 previously characterized as active and weak enhancer states specifically in the Brain/neuron, respectively (Vu and Ernst, 2022). The following three sections of tracks show the CSREP summary state assignment probabilities for the Brain group, the CSREP summary state annotation for Brain group, and the Brain input samples' chromatin state maps from Roadmap Epigenomics. The last four sections of tracks show ESC input samples' chromatin state maps from Roadmap Epigenomics, CSREP summary state annotation and assignment probabilities for the ESC group and the differential scores of ESC-Brain annotations. This figure shows input and CSREP output data for a similar genomic region as in **Figure 4.1C**.

CSREP empirical run time

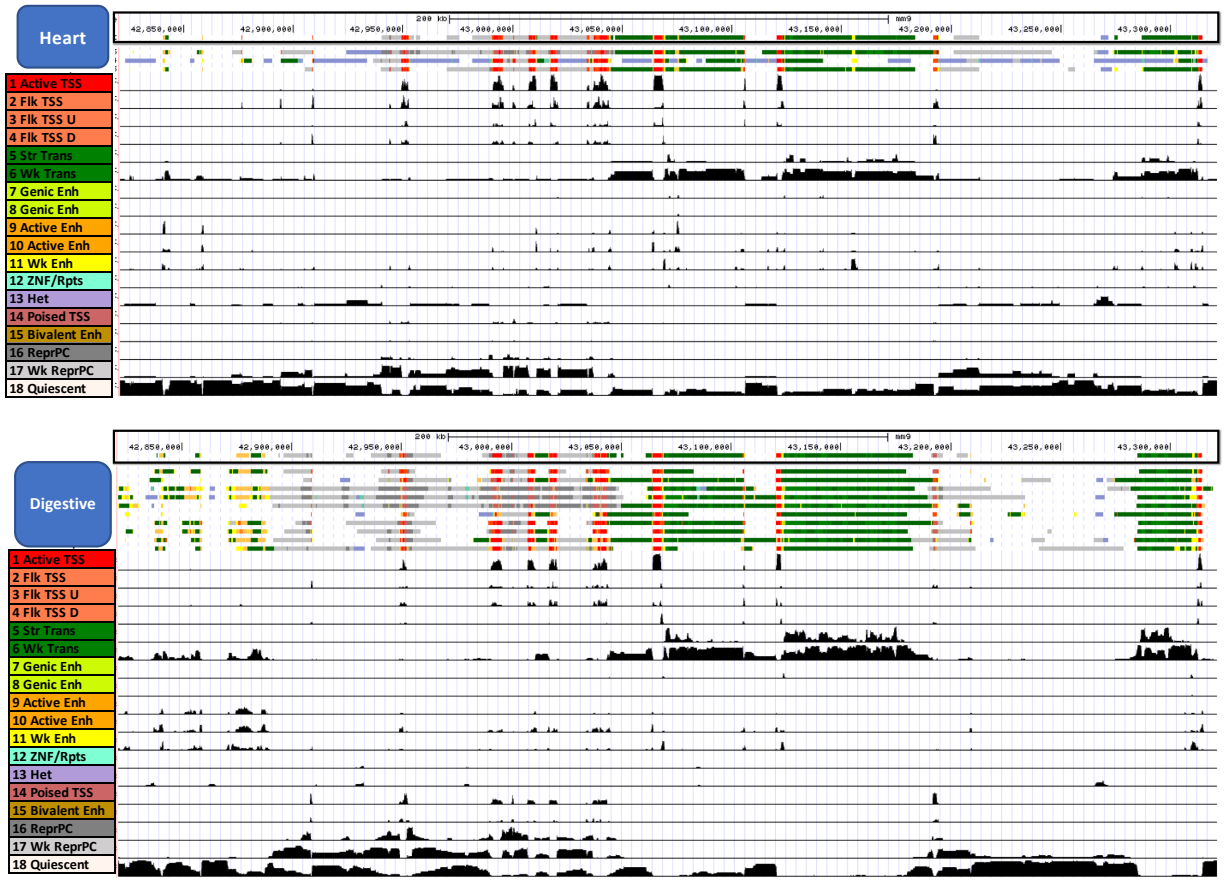
(input data preprocessing time for all 64 samples: 0:42:14)

group	num_sample	Run time (hh:mm:ss)
HSC & B-cell	3	0:56:19
Heart	3	1:32:11
Muscle	4	1:14:12
Sm_Muscle	4	1:22:40
iPSC	4	1:22:04
ESC	5	1:24:22
Epithelial	5	2:48:27
Brain	7	1:41:12
ES-deriv	7	1:41:26
Digestive	10	1:53:54
Blood & T-cell	12	2:26:14

Supplementary Figure 4. 2: Empirical run time of CSREP for summarizing chromatin annotations from 11 groups (of 64 samples in total) from Roadmap.

We ran CSREP on the high-performance computing cluster, where each job was allocated 4 cores with 4GB of memory per core. snakemake *(Köster and Rahmann, 2012; Mölder* et al.*, 2021)* parallelizes the steps in input data preprocessing across all 64 input samples. Additionally, snakemake parallelizes the training process to predict chromatin state maps in individual samples in each group. The total runtime includes data preprocessing time shared across all groups of samples (~42 minutes), and prediction time that is specific to each group and denoted in the table. The prediction run time reported in the table include (1) the maximum time span of one job that outputs predictions for one sample, out of all samples in each group, and (2) the time span for averaging the predictions across samples, to obtain the group-wide summary chromatin state maps.

Visualization of input samples' chromatin state maps and CSREP's summary state assignment probabilities tracks for region chr5: 42821109-43321109

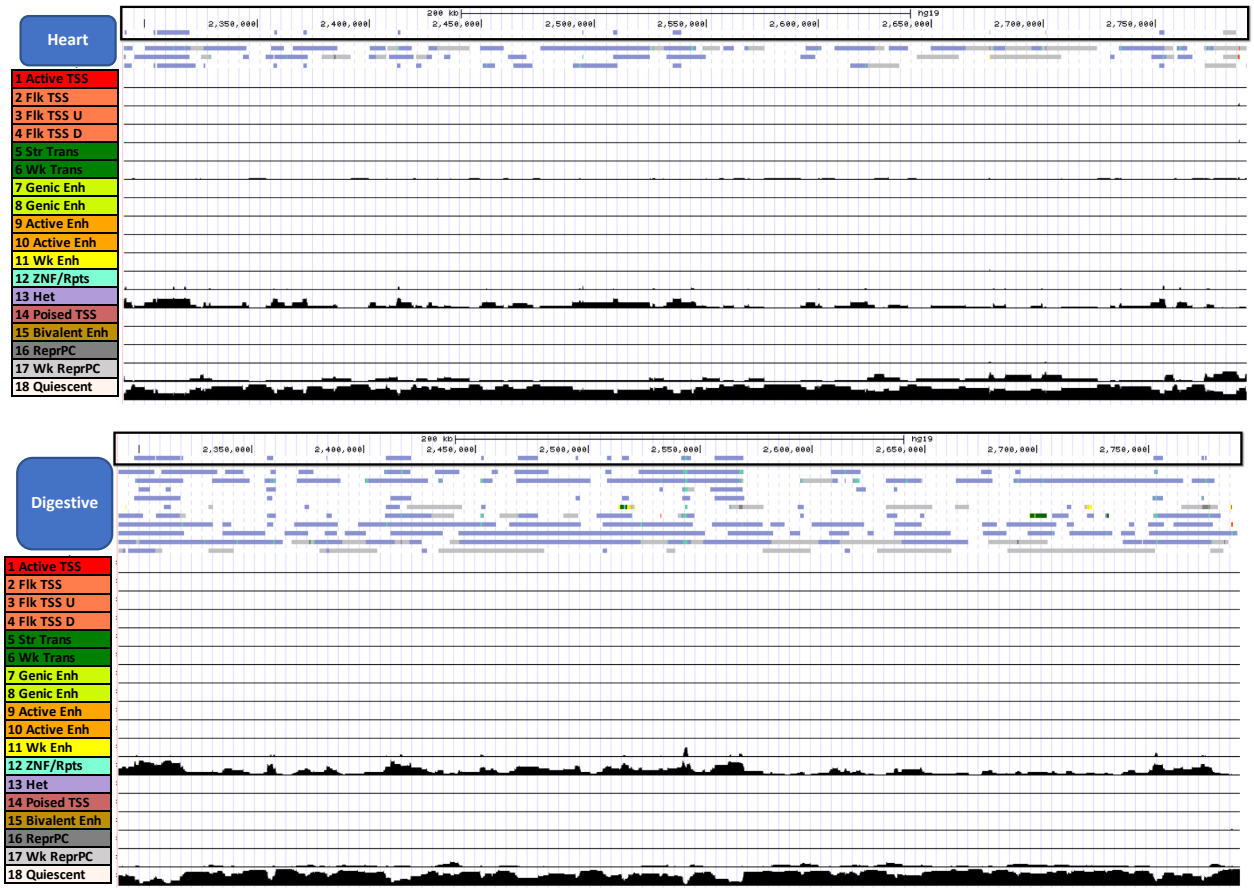


Supplementary Figure 4. 3: Visualization of CSREP's input and output data for an arbitrary 500-kb genomic window (chr5: 42,821,109-43,321,109, hg19).
The visualization shows CSREP's strong agreement with the chromatin state maps from 10 input samples from Digestive and 3 samples from Heart tissue groups from Roadmap Epigenomics. In each panel, the first track shows the summary chromatin state map based on CSREP. The following 3 (Heart) and 10 (Digestive) tracks show input samples' chromatin state maps. States are colored based on legend on the left. In the following 18 tracks, each track shows the probabilities of assignment for one of 18 states. This region is the same as in **Figure 4.2A**.

Visualization of input samples' chromatin state maps and CSREP's summary state assignment probabilities tracks for region chr12:79237500-79737500

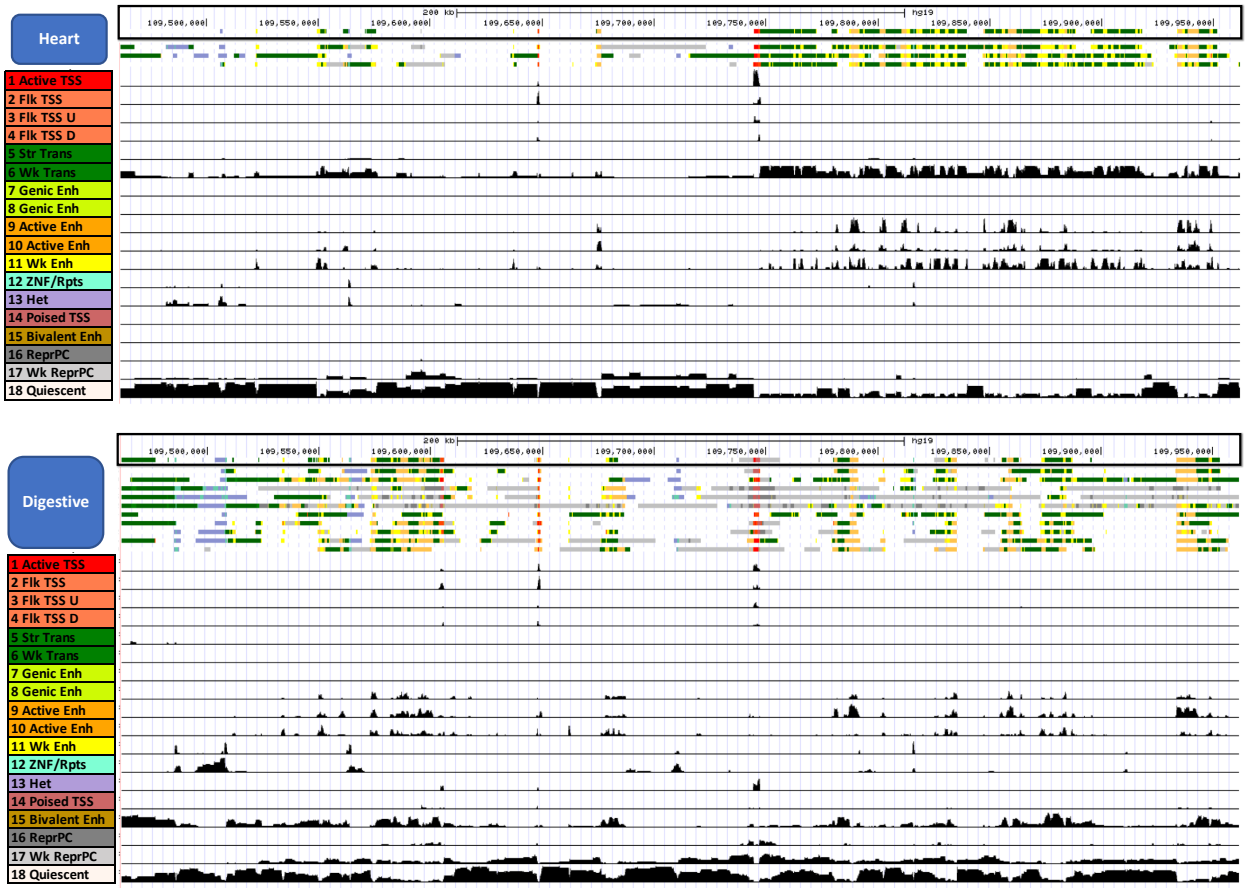
Supplementary Figure 4. 4: **Visualization of CSREP's input and output data for an arbitrary 500-kb genomic window (chr12:79,237,500-79,737,500, hg19).**
Similar to **Supplementary Figure 4.3**, for genomic region chr12:79,237,500-79,737,500, hg19.

Visualization of input samples' chromatin state maps and CSREP's summary state assignment probabilities tracks for region chr10:2290673-2790673



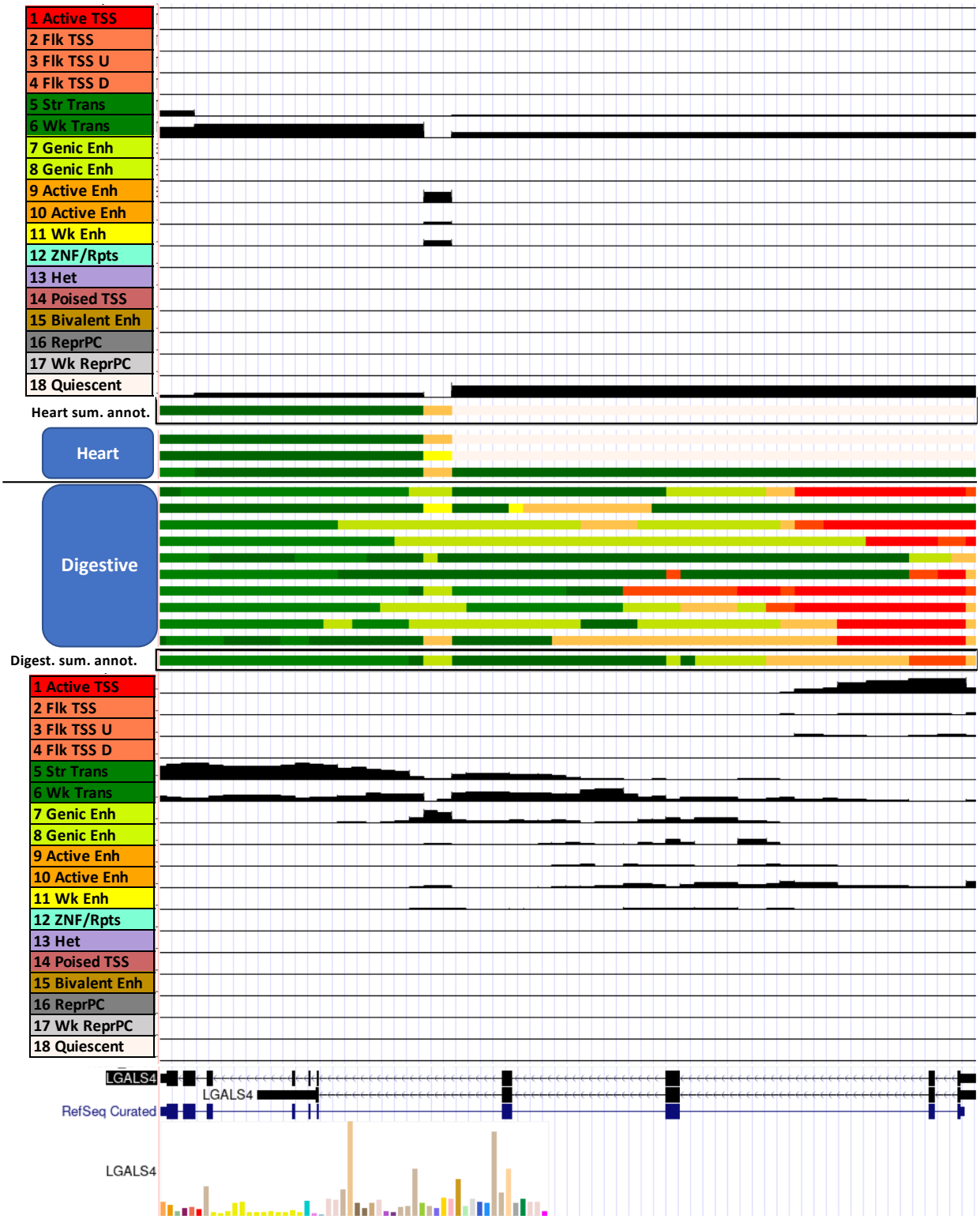
Supplementary Figure 4. 5: Visualization of CSREP's input and output data for an arbitrary 500-kb genomic window (chr10:2,290,673-2,790,673, hg19).
Similar to **Supplementary Figure 4.3**, for genomic region chr10:2,290,673-2,790,673, hg19.

Visualization of input samples' chromatin state maps and CSREP's summary state assignment probabilities tracks for region chr2:109461695-109961695



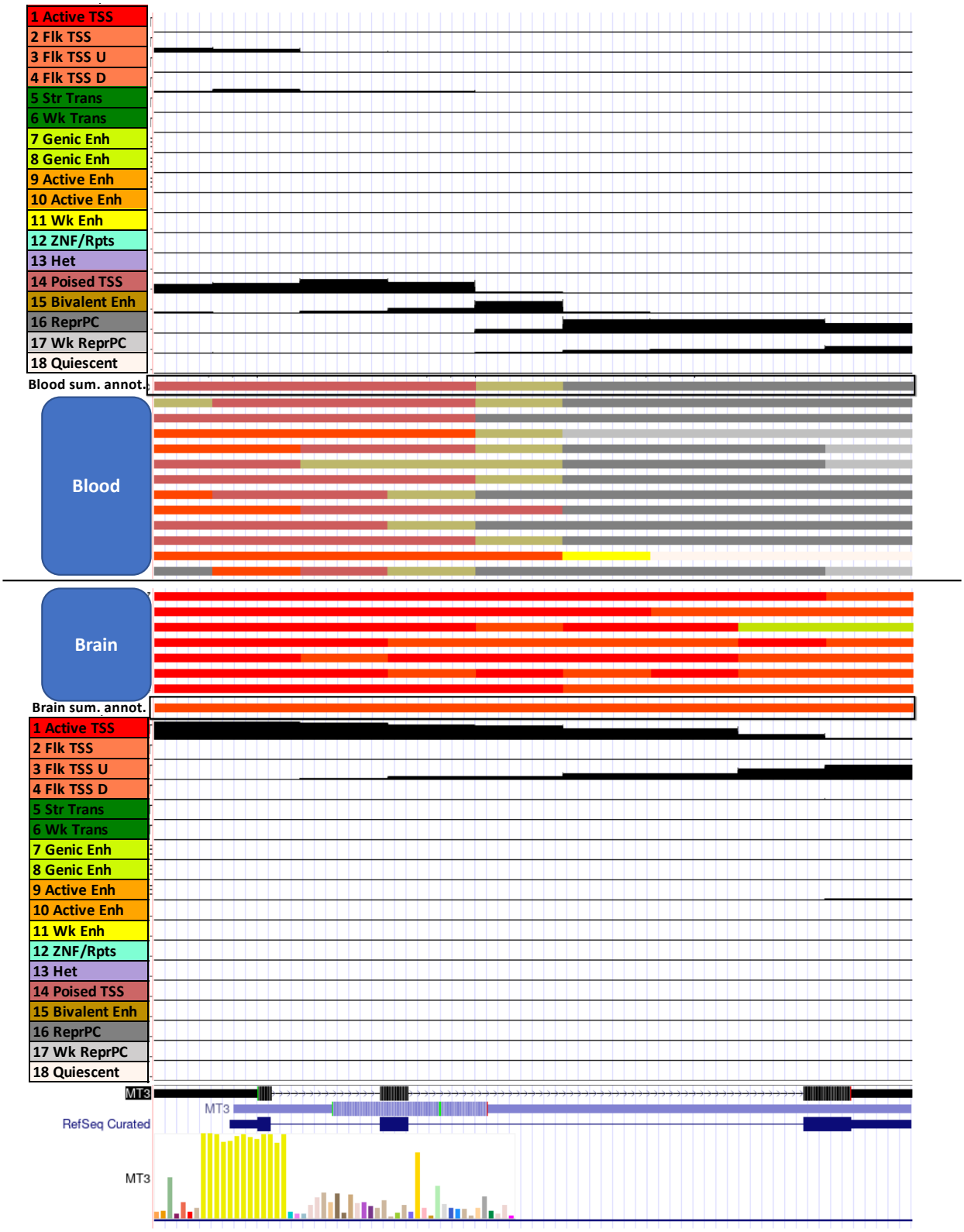
Supplementary Figure 4. 6: Visualization of CSREP's input and output data for an arbitrary 500-kb genomic window (chr2:109,461,695-109,961,695, hg19).
Similar to **Supplementary Figure 4.3**, for genomic region chr2:109,461,695-109,961,695, hg19.

Visualization of input samples' chromatin state maps and CSREP's summary state assignment probabilities tracks for gene LGALS4 (highly expressed in colon and intestine) (chr19:39,292,311-39,303,740)



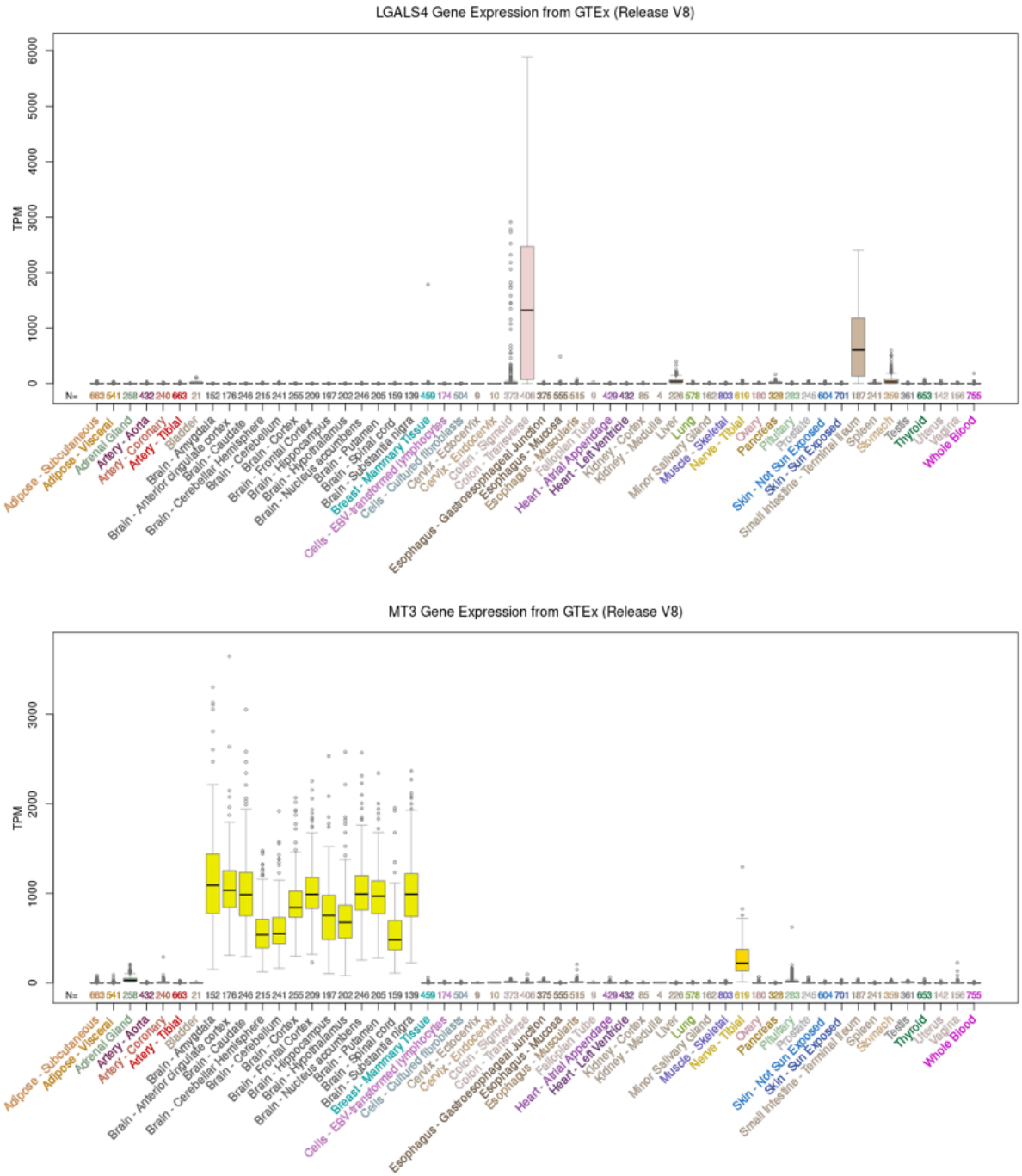
Supplementary Figure 4. 7: Visualization of CSREP's input and output data for a genomic window overlapping the LGALS4 gene (chr19:39,292,311-39,303,740, hg19).
Gene LGALS4 shows the distinctly higher expression in cell types of the Digestive system, with gene expression profile across cell types in **Supplementary Figure 4.9**. The visualization shows UCSC Genome browser view of CSREP's output summary chromatin state maps for 10 input samples from Digestive and 3 samples from Heart tissue groups from Roadmap Epigenomics Consortium. The first 18 tracks show the summary probability assignment of chromatin states for Heart samples, based on CSREP. The following track shows the summary chromatin state annotation for Heart samples. The following 3 (Heart) and 10 (Digestive) tracks show input samples' chromatin state maps. The following tracks show the summary chromatin state map for sample of the Digestive cell groups, followed by the individual states' summary state assignment probabilities. States are colored based on legend on the left.

Visualization of input samples' chromatin state maps and CSREP's summary state assignment probabilities tracks for gene MT3 (highly expressed in brain; chr16:56,623,267-56,625,000)



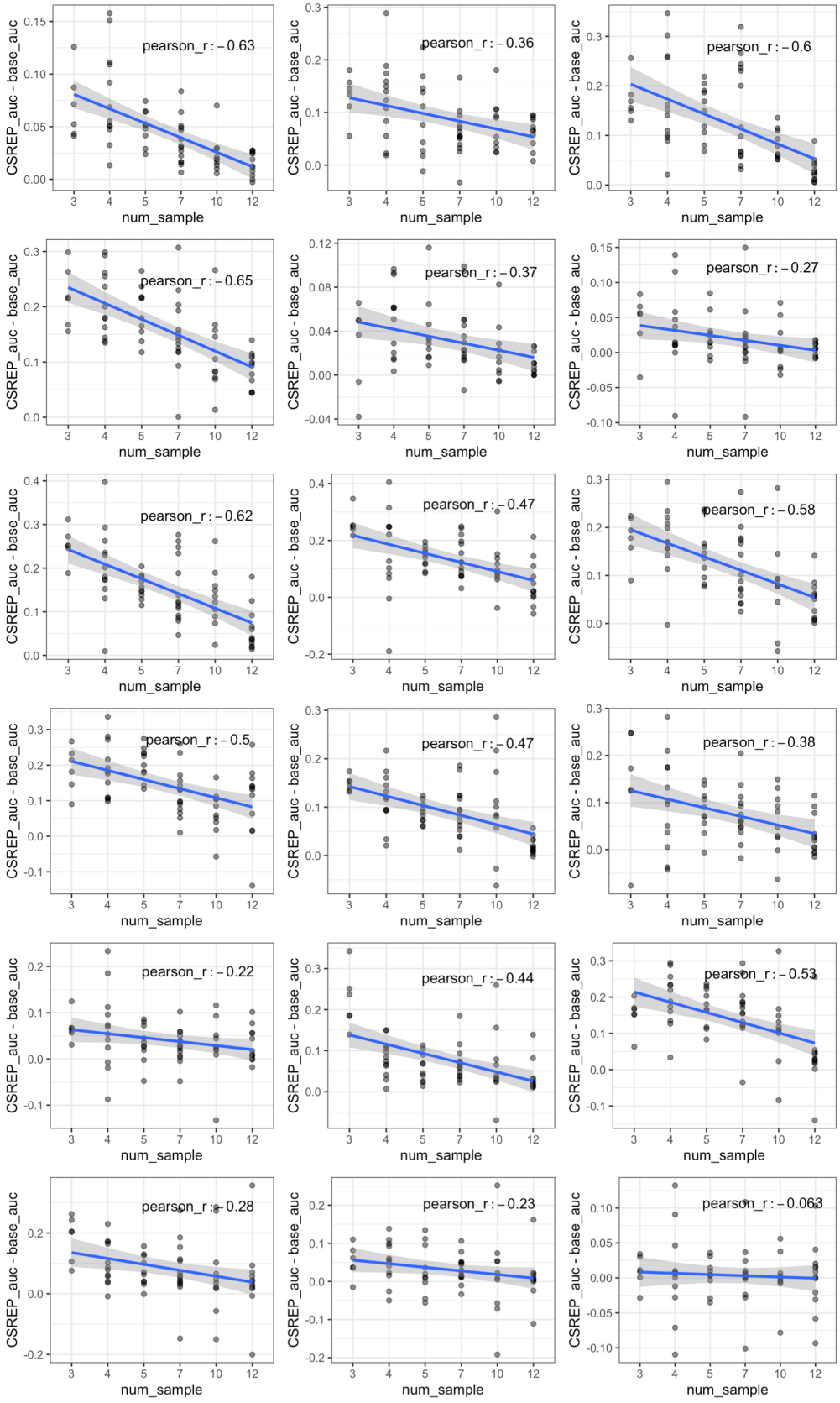
Supplementary Figure 4. 8: Visualization of CSREP's input and output data for a genomic window overlapping the MT3 gene (chr16:56,623,267-56,625,000, hg19).
Gene MT3 shows the distinctly higher expression in Brain cell types, with gene expression profile across cell types in **Supplementary Figure 4.9**. The visualization shows UCSC Genome browser view of CSREP's output summary chromatin state maps for 12 input samples from Blood and 7 samples from Brain tissue groups from Roadmap Epigenomics. The first 18 tracks show the summary probability assignment of chromatin states for Blood samples, based on CSREP. The following track shows the summary chromatin state annotation for Blood samples. The following 12 (Blood) and 7 (Brain) tracks show input samples' chromatin state maps. The following tracks show the summary chromatin state map for sample of the Brain cell groups, followed by the individual states' summary state assignment probabilities. States are colored based on legend on the left.

Gene expression for genes LGALS4 and MT3, outputted by UCSC Genome Browser



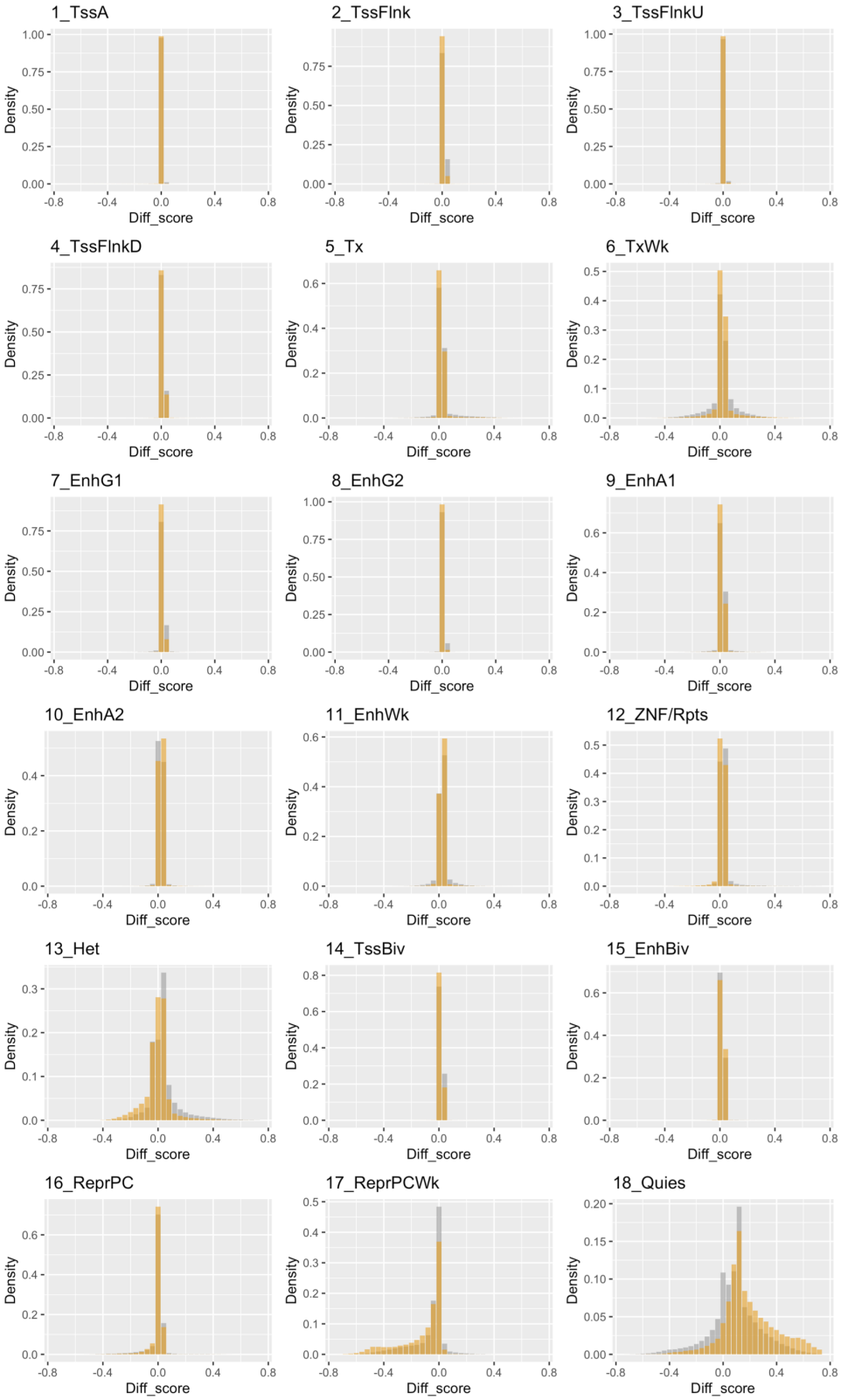
Supplementary Figure 4. 9: Gene expression profile for genes LGALS4 (top) and MT3 (bottom), as shown on UCSC Genome Browser.

Difference of AUROCs between CSREP and base_count stratified by number of input samples



Supplementary Figure 4. 10: Relationship between the number of samples and AUROCs from using summary chromatin state map to predict genomic locations of individual chromatin states.
We conducted cross-validation analysis for each group of samples (**Methods**), and for each group, we calculate ROC curve of CSREP's summary probabilistic chromatin state map in recovering genomic positions of individual chromatin states in a held-out sample. Each panel corresponds to a chromatin state, and shows the difference between CSREP's AUROCs and base_count's AUROCs for predicting locations of the chromatin state in left-out samples. Each dot corresponds to one sample. Y-axis shows the difference of AUROCs between the two methods (positive y-axis means CSREP results in higher AUROCs and vice-versa). X-axis shows the number of input samples for the group, not to scale, but the reported Pearson correlation is based on actual number of samples.

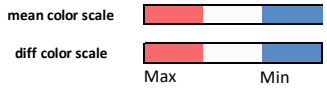
Histogram of *Male - Female* CSREP differential scores in autosomes and chrX



Supplementary Figure 4. 11: Histogram of CSREP differential chromatin scores between Male and Female groups of samples, in autosomes and in chromosome X.
Each subpanel shows the histograms of one state's CSREP Male - Female differential scores, bounded between -1 and 1, in autosomes and chromosome X.

Mean and variance of CSREP Male-Female differential scores in chrX and autosomes

chrom	stats	1_TssA	2_TssFlnk	3_TssFlnkU	4_TssFlnkD	5_Tx	6_TxWk	7_EnhG1	8_EnhG2	9_EnhA1	10_EnhA2	11_EnhWk	12_ZNF/Rpts	13_Het	14_TssBiv	15_EnhBiv	16_ReprPC	17_ReprPCWk	18_Quies
chrX	mean	-2.2E-03	-1.2E-04	-1.0E-03	-4.4E-05	5.0E-03	7.4E-04	-7.7E-04	-4.2E-04	-5.6E-04	2.5E-04	6.2E-04	-2.6E-03	-2.3E-02	-4.3E-05	-1.6E-04	-1.6E-02	-1.4E-01	1.8E-01
	variance	8.2E-04	5.6E-04	1.3E-04	1.5E-04	1.9E-03	5.0E-03	7.3E-05	4.6E-05	2.7E-04	2.3E-04	6.0E-04	1.2E-03	8.6E-03	6.9E-05	9.2E-05	2.6E-03	2.4E-02	4.1E-02
autosomes	mean	-3.0E-03	-3.2E-04	-2.2E-03	-8.9E-05	9.6E-03	-1.1E-03	-2.5E-03	-1.5E-03	3.2E-04	2.5E-04	3.1E-03	8.9E-03	1.6E-02	4.9E-04	7.3E-04	-1.6E-02	-8.3E-02	7.0E-02
	variance	1.1E-03	5.0E-04	5.1E-04	4.2E-04	3.8E-03	8.1E-03	4.9E-04	2.9E-04	9.3E-04	3.4E-04	1.8E-03	4.0E-03	1.2E-02	2.0E-04	3.6E-04	4.9E-03	1.4E-02	3.4E-02
mean(X) - mean(auto)		8.3E-04	1.9E-04	1.2E-03	4.4E-05	-4.6E-03	1.8E-03	1.8E-03	1.1E-03	-8.8E-04	-4.0E-06	-2.5E-03	-1.1E-02	-3.9E-02	-5.3E-04	-9.0E-04	-2.2E-04	-5.4E-02	1.1E-01



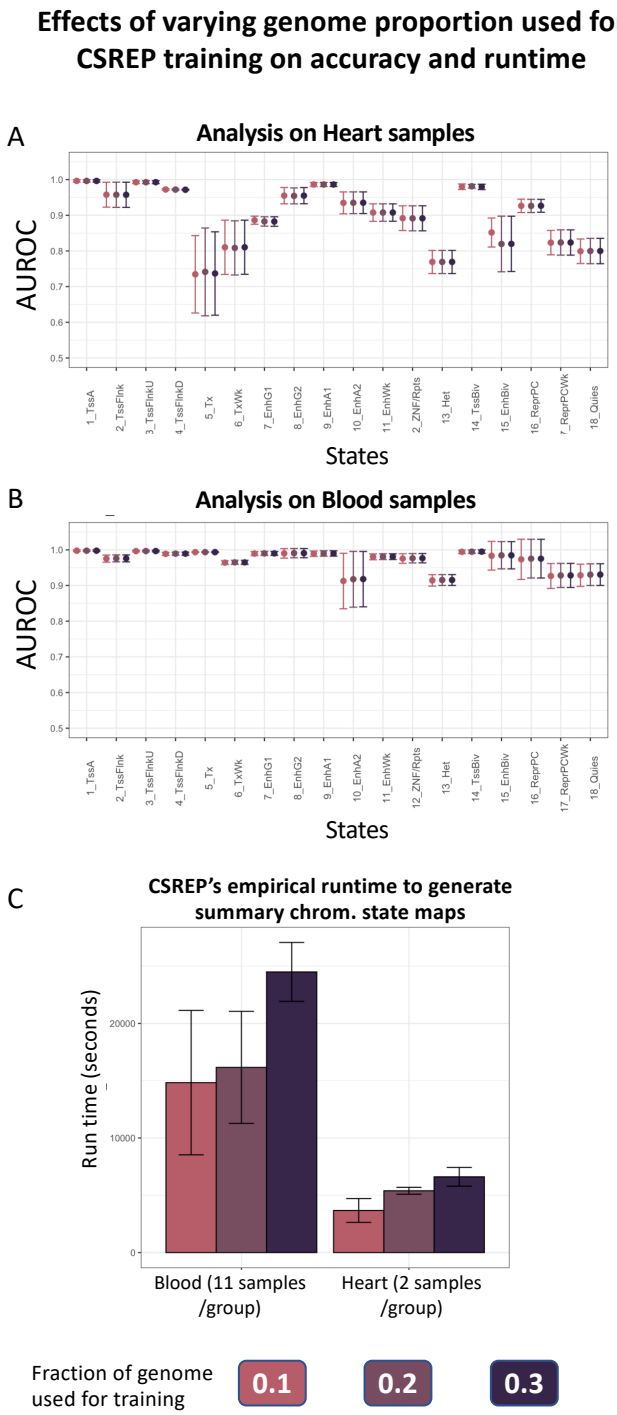
Supplementary Figure 4. 12: Mean and variance of the CSREP differential scores between Male and Female groups of samples, in autosomes and in chromosomes.
The mean differential scores for each state in either chromosome X or autosomes are reported for each state and share the same color scale as in bottom legend. The difference between the mean scores for chromosome X and the autosomes for each state is reported on the bottom row and colored as in bottom legend. Three states with largest-magnitude difference in mean scores are 13_Het, 17_ReprPCWk, 18_Quies.

AUROCs for predicting Brain-/ESC- specific peaks of chromatin marks

	DNase								H3K27ac								H3K9ac							
	Brain-specific				ESC-specific				Brain-specific				ESC-specific				Brain-specific				ESC-specific			
state	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher	csrep	base	mann_whitney	fisher
1 Active TSS	0.54	0.51	0.50	0.51	0.59	0.50	0.51	0.51	0.58	0.52	0.52	0.52	0.70	0.51	0.51	0.51	0.59	0.53	0.53	0.52	0.66	0.55	0.55	0.55
2 Flk TSS	0.54	0.51	0.51	0.51	0.60	0.51	0.52	0.51	0.55	0.51	0.53	0.52	0.66	0.50	0.51	0.51	0.53	0.52	0.53	0.52	0.65	0.55	0.58	0.57
3 Flk TSS U	0.55	0.51	0.51	0.51	0.62	0.51	0.52	0.52	0.61	0.52	0.52	0.52	0.72	0.53	0.52	0.52	0.63	0.53	0.54	0.52	0.68	0.54	0.54	0.54
4 Flk TSS D	0.54	0.51	0.51	0.51	0.65	0.51	0.52	0.52	0.58	0.51	0.52	0.52	0.73	0.51	0.52	0.52	0.57	0.52	0.53	0.51	0.69	0.52	0.53	0.53
5 Str Trans	0.51	0.50	0.49	0.49	0.45	0.49	0.49	0.49	0.58	0.49	0.53	0.52	0.49	0.50	0.51	0.51	0.51	0.48	0.53	0.53	0.41	0.48	0.51	0.50
6 Wk Trans	0.49	0.48	0.52	0.51	0.54	0.50	0.54	0.52	0.56	0.53	0.61	0.59	0.58	0.51	0.59	0.57	0.44	0.44	0.59	0.57	0.45	0.46	0.52	0.51
7 Genic Enh	0.55	0.50	0.50	0.50	0.50	0.50	0.51	0.51	0.72	0.53	0.53	0.53	0.55	0.52	0.52	0.52	0.68	0.53	0.54	0.53	0.50	0.51	0.52	0.51
8 Genic Enh	0.57	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.72	0.51	0.51	0.51	0.56	0.51	0.50	0.51	0.71	0.52	0.52	0.51	0.53	0.51	0.51	0.51
9 Active Enh	0.57	0.53	0.53	0.53	0.58	0.55	0.57	0.56	0.74	0.58	0.58	0.58	0.71	0.62	0.62	0.62	0.72	0.62	0.64	0.60	0.61	0.58	0.59	0.58
10 Active Enh	0.56	0.54	0.54	0.54	0.59	0.52	0.57	0.55	0.73	0.64	0.64	0.65	0.74	0.61	0.61	0.61	0.70	0.61	0.63	0.58	0.62	0.53	0.56	0.55
11 Wk Enh	0.54	0.52	0.57	0.56	0.64	0.61	0.65	0.64	0.69	0.58	0.62	0.61	0.73	0.66	0.69	0.68	0.62	0.54	0.64	0.59	0.58	0.57	0.60	0.58
12 ZNF/Rpts	0.52	0.51	0.49	0.50	0.47	0.50	0.50	0.50	0.52	0.51	0.49	0.50	0.46	0.49	0.49	0.49	0.52	0.51	0.49	0.49	0.43	0.49	0.49	0.49
13 Het	0.47	0.50	0.47	0.48	0.40	0.48	0.46	0.47	0.41	0.50	0.45	0.47	0.31	0.48	0.45	0.46	0.45	0.51	0.45	0.47	0.34	0.48	0.45	0.46
14 Poised TSS	0.56	0.50	0.50	0.50	0.53	0.51	0.52	0.52	0.62	0.49	0.51	0.50	0.53	0.50	0.50	0.50	0.62	0.49	0.51	0.51	0.58	0.54	0.56	0.55
15 Bivalent Enh	0.56	0.50	0.50	0.50	0.53	0.51	0.52	0.52	0.62	0.50	0.51	0.50	0.53	0.50	0.50	0.50	0.61	0.50	0.51	0.51	0.48	0.49	0.53	0.52
16 ReprPC	0.52	0.50	0.50	0.50	0.47	0.49	0.52	0.51	0.51	0.49	0.51	0.50	0.47	0.49	0.50	0.49	0.52	0.49	0.51	0.51	0.40	0.45	0.54	0.53
17 Wk ReprPC	0.49	0.50	0.52	0.51	0.46	0.47	0.52	0.50	0.34	0.45	0.48	0.46	0.41	0.47	0.49	0.47	0.34	0.46	0.48	0.50	0.39	0.44	0.53	0.52
18 Quiescent	0.44	0.45	0.50	0.49	0.43	0.42	0.48	0.47	0.32	0.35	0.42	0.42	0.30	0.29	0.44	0.42	0.34	0.39	0.52	0.54	0.39	0.38	0.55	0.55
sciddo	0.52				0.55				0.56				0.53				0.56				0.59			

Supplementary Figure 4. 13: Evaluation of recovery of differential chromatin marks signals between ESC and Brain.

The table is an extension to **Figure 4.4**, and shows AUROCs for differential scores' predictions of genomic regions associated with differential peak signals for one chromatin mark, from left to right: DNase, H3K27ac and H3K9ac. For each chromatin mark, it shows the AUROCs of predicting signal peaks observed in Brain and ESC exclusively (Brain-spec and ESC-spec). Differential scores outputted by CSREP, base-count, Mann-Whitney U test (used by ChromDiff) and Fisher's exact test (used by EpiCompare) are shown for each chromatin state (rows). In each category of comparisons (a chromatin mark in either ESC or Brain), the top three scores that show highest AUROCs are highlighted in green. Along the bottom is the AUROC for SCIDDO. The differential scores for states that are not related to active promoter and enhancer activities tend to show AUROCs near or lower than 0.5, which is expected since these states are not associated with DNase, H3K27ac or H3K9ac.



Supplementary Figure 4. 14: Effects of varying genome proportion used for training in CSREP on accuracy and runtime.
We conducted leave-one-out analysis to evaluate CSREP's accuracy in predicting a held-out sample's chromatin state map, given varying fractions of the genome used for training (**Methods**). We applied the procedure on data from Roadmap Epigenomics, and reported the AUROC for 3 and 12 samples from Heart **(A)** and Blood **(B)** groups, respectively. The empirical runtime for CSREP to generate the summary chromatin state maps for a group of 2 Heart samples or 11 Blood samples in the leave-one-out analysis are reported in **(C)**.

References

Abel,H.J. *et al.* (2020) Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, **583**, 83–89.

Aitman,T.J. *et al.* (2011) The future of model organisms in human disease research. *Nat. Rev. Genet.*, **12**, 575–582.

Amemiya,H.M. *et al.* (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.*, **9**, 1–5.

Arneson,A. *et al.* (2021) A mammalian methylation array for profiling methylation levels at conserved sequences. *Biorxiv*.

Arneson,A. and Ernst,J. (2019) Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Commun. Biol.*, **2**, 1–14.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Becker,J.S. *et al.* (2016) H3K9me3-dependent heterochromatin: barrier to cell fate changes. *Trends Genet.*, **32**, 29–41.

Benner,C. *et al.* (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, **32**, 1493–1501.

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

Biesinger,J. *et al.* (2013) Discovering and mapping chromatin states using a tree hidden Markov model. In, *BMC bioinformatics*. Springer, p. S4.

Bogu,G.K. *et al.* (2015) Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. *Mol. Cell. Biol.*, **36**, 809–819.

Boix,C.A. *et al.* (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, **590**, 300–307.

Boyle,A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

Chen,W. *et al.* (2015) Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics*, **200**, 719–736.

Chronis,C. *et al.* (2017) Cooperative binding of transcription factors orchestrates reprogramming. *Cell*, **168**, 442–459.

Claussnitzer,M. *et al.* (2015) FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.*, **373**, 895–907.

Consortium,E.P. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *nature*, **447**, 799.

Cooper,G.M. *et al.* (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.

Creyghton,M.P. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.*, **107**, 21931–21936.

Dale,R.K. *et al.* (2011) Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, **27**, 3423–3424.

Davis,C.A. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

Davydov,E.V. *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, **6**, e1001025.

Di Iulio,J. *et al.* (2018) The human noncoding genome defined by genetic diversity. *Nat. Genet.*, **50**, 333–337.

Dib,C. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Ebert,P. and Schulz,M.H. (2021) Fast detection of differential chromatin domains with SCIDDO. *Bioinformatics*, **37**, 1198–1205.

Elbarbary,R.A. *et al.* (2016) Retrotransposons as regulators of gene expression. *Science*, **351**.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ernst,J. ChromHMM v.1.18.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.

Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.*, **33**, 364–376.

Fernández,A.F. *et al.* (2015) H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. *Genome Res.*, **25**, 27–40.

Frankish,A. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.

Fu,Y. *et al.* (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 1–15.

Ge,X. *et al.* (2019) EpiAlign: an alignment-based bioinformatic tool for comparing chromatin state sequences. *Nucleic Acids Res.*, **47**, e77–e77.

Gjoneska,E. *et al.* (2015) Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*, **518**, 365–369.

Gorkin,D.U. *et al.* (2020) An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature*, **583**, 744–751.

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

Guelen,L. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948–951.

Gulko,B. *et al.* (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.

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

Hastie,T. *et al.* (2009) The elements of statistical learning: data mining, inference, and prediction Springer.

He,Y. and Wang,T. (2017) EpiCompare: an online tool to define and explore genomic regions with tissue or cell type-specific epigenomic features. *Bioinformatics*, **33**, 3268–3275.

Heintzman,N.D. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.*, **106**, 9362–9367.

Hoffman,M.M. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.

Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473.

Hon,G.C. *et al.* (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.*, **45**, 1198–1206.

Horvath,S. *et al.* (2021) DNA methylation aging and transcriptomic studies in horses. *Biorxiv*.

Huang,Y.-F. *et al.* (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.

Ioannidis,N.M. *et al.* (2017) FIRE: functional inference of genetic variants that regulate gene expression. *Bioinformatics*, **33**, 3895–3901.

Ionita-Laza,I. *et al.* (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214.

Jessa,S. and Kleinman,C.L. (2018) Chromswitch: a flexible method to detect chromatin state switches. *Bioinformatics*, **34**, 2286–2288.

Ji,H. *et al.* (2013) Differential principal component analysis of ChIP-seq. *Proc. Natl. Acad. Sci.*, **110**, 6789–6794.

Jones,P.A. and Takai,D. (2001) The role of DNA methylation in mammalian epigenetics. *Science*, **293**, 1068–1070.

Karczewski,K.J. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

Kazachenka,A. *et al.* (2018) Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*, **175**, 1259-1271.e13.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kheradpour,P. *et al.* (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.*, **23**, 800–811.

Kimura,H. (2013) Histone modifications for human epigenome analysis. *J. Hum. Genet.*, **58**, 439–445.

Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Kwon,S.B. and Ernst,J. (2021) Learning a genome-wide score of human–mouse conservation at the functional genomics level. *Nat. Commun.*, **12**, 1–14.

Lay,F.D. *et al.* (2014) Reprogramming of the human intestinal epigenome by surgical tissue transposition. *Genome Res.*, **24**, 545–553.

Lee,J. *et al.* (2017) The LDB1 complex co-opts CTCF for erythroid lineage-specific long-range enhancer interactions. *Cell Rep.*, **19**, 2490–2502.

Li,C.Z. *et al.* (2021) Epigenetic predictors of maximum lifespan and other life history traits in mammals. *bioRxiv*.

Libbrecht,M.W. *et al.* (2019) A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol.*, **20**, 180.

Libbrecht,M.W. *et al.* (2021) Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. *PLoS Comput. Biol.*, **17**, e1009423.

Lindblad-Toh,K. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *nature*, **462**, 315–322.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Lu,A.T. *et al.* (2022) Universal DNA methylation age across mammalian tissues. 2021.01.18.426733.

Martincorena,I. and Campbell,P.J. (2015) Somatic mutation in cancer and normal cells. *Science*, **349**, 1483–1489.

McArthur,E. and Capra,J.A. (2021) 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.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Meuleman,W. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Mikkelsen,T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Mölder,F. *et al.* (2021) Sustainable data analysis with Snakemake. *F1000Research*, **10**.

Mortazavi,A. *et al.* (2013) Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Res.*, **23**, 2136–2148.

Navarro Gonzalez,J. *et al.* (2021) The UCSC genome browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.

Neath,A.A. and Cavanaugh,J.E. (2012) The Bayesian information criterion: background, derivation, and applications. *Wiley Interdiscip. Rev. Comput. Stat.*, **4**, 199–203.

Parker,S.C. *et al.* (2012) Mutational signatures of de-differentiation in functional non-coding regions of melanoma genomes. *PLoS Genet*, **8**, e1002871.

Pehrsson,E.C. *et al.* (2019) The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.*, **10**, 1–16.

Perlman,R.L. (2016) Mouse models of human disease: An evolutionary perspective. *Evol. Med. Public Health*, **2016**, 170–176.

Phillips,J.E. and Corces,V.G. (2009) CTCF: Master Weaver of the Genome. *Cell*, **137**, 1194–1211.

Polak,P. *et al.* (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, **518**, 360–364.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Quang,D. *et al.* (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.

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

Rentzsch,P. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.

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

Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes | Nature.

Rogers,M.F. *et al.* (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, **34**, 511–513.

Rosenbloom,K.R. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.

Rugg-Gunn,P.J. *et al.* (2010) Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 10783–10790.

Sakamoto,Y. *et al.* (1986) Akaike information criterion statistics. *Dordr. Neth. Reidel*, **81**, 26853.

Schuster-Böckler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *nature*, **488**, 504–507.

Shen,Y. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Smedley,D. *et al.* (2016) A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.*, **99**, 595–606.

Smit,A.F.A. *et al.* (2015) RepeatMasker Open-4.0. 2013–2015.

Soboleva,T.A. *et al.* (2014) Histone variants at the transcription start-site. *Trends Genet.*, **30**, 199–209.

Stamatoyannopoulos,J.A. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.*, **13**, 1–5.

Stunnenberg,H.G. *et al.* (2016) The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell*, **167**, 1145–1149.

Sugathan,A. and Waxman,D.J. (2013) Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver. *Mol. Cell. Biol.*, **33**, 3594–3610.

Supek,F. and Lehner,B. (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, **521**, 81–84.

Taberlay,P.C. *et al.* (2014) Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res.*, **24**, 1421–1432.

Tate,J.G. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.

The ENCODE Project Consortium *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.

Thurman,R.E. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

Tsai,H.-W. *et al.* (2009) Sex differences in histone modifications in the neonatal mouse brain. *Epigenetics*, **4**, 47–53.

Vanhooren,V. and Libert,C. (2013) The mouse as a model organism in aging research: usefulness, pitfalls and possibilities. *Ageing Res. Rev.*, **12**, 8–21.

Varshney,A. *et al.* (2017) Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci.*, **114**, 2301–2306.

van der Velde,A. *et al.* (2021) Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun. Biol.*, **4**, 1–15.

Vu,H. and Ernst,J. (2021a) Full-stack chromHMM model state characterization archival.

Vu,H. and Ernst,J. (2021b) Full-stack chromHMM state characterization.

Vu,H. and Ernst,J. (2022) Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biol.*, **23**, 1–37.

Wang,J. *et al.* (2020) CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Res.*, **48**, D807–D816.

Wang,Q. *et al.* (2020) Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. *Nat. Genet.*, **52**, 828–839.

Wang,Y. *et al.* (2021) TAD boundary and strength prediction by integrating sequence and epigenetic profile information. *Brief. Bioinform.*

Weber,M. *et al.* (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Wutz,A. (2011) Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat. Rev. Genet.*, **12**, 542–553.

Xie,W. *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**, 1134–1148.

Yen,A. and Kellis,M. (2015) Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat. Commun.*, **6**, 1–13.

Yue,F. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.

Zhang,Y. *et al.* (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.*, **44**, 6721–6731.

Zhu,C. *et al.* (2021) Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods*, **18**, 283–292.

Zhu,J. *et al.* (2013) Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, **152**, 642–654.