bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

Ha Vu[1,2], Jason Ernst[1,2,3,4,5,6,7]

[1] Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, 90095, USA.

[2] Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA

[3] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at University of California, Los Angeles, Los Angeles, CA 90095, USA

[4] Computer Science Department, University of California, Los Angeles, Los Angeles, CA 90095, USA

[5] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

[6] Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA

[7] Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Background:

Genome-wide maps of chromatin marks such as histone modifications and open chromatin sites provide valuable information for annotating the non-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.

Results:

Using a version of ChromHMM enhanced for large-scale applications, we applied 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 used 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.

Conclusions:

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [1–5]. 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 [6–12]. 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 [1,8,13–20].

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 [21–23] 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 [3,16,24]. 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 (**Supp. Fig. 1**) [22,25]. 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 [26,27]. 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 (**Supp. Fig. 1**) [22,25]. Instead of learning per-cell-type annotations based on a limited number of datasets per-cell-type, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

stacked modeling approach learns 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 per-cell-type 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' per-cell-type annotations were post-hoc clustered to analyze chromatin dynamics across cell and tissue types [3,16]. 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,

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [28,29]. 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 per-cell-type 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, 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 genome based on data from over 100 cell and tissue types from the Roadmap Epigenomics and ENCODE projects (**Fig. 1**) [14,16]. In total we applied ChromHMM to 1032 datasets for 30 histone modifications, a histone variant (H2A.Z), and DNase I hypersensitivity (**Supp. Fig. 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 [16,22,25]. 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 (**Supp. Fig. 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, 3A, Supp. Fig. 4-7**). The other genomic annotations include previous per-cell-type chromatin state annotations (**Supp. Fig. 8-9**), cell-type-specific gene expression data (**Supp. Fig. 10-11**), and various independent existing genomic annotations (**Fig. 3A**). These independent genomic annotations included annotated gene features, evolutionary constrained elements, and assembly gaps, among others (**Methods**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

These analyses led us to group the 100 full-stack states into 16 groups (**Fig. 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 of all experiments, 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 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 active and weak 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, and 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. 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. 3A**). In previous concatenated models based on the Roadmap Epigenomics data [16], no specific state was associated with assembly gaps likely due to the limited number of input chromatin mark signals compared

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. 3A**). States in the Acet group had a lower average expression of proximal genes compared to states in the enhancer and promoter groups (**Fig. 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, Supp. Fig. 10-11**). Distinctions were seen among these states, for example, in terms of their positional enrichments relative to TES (**Fig. 3A, D, Supp. Fig. 12**). States in the flanking promoter group (PromF1-7) showed 6.5-28 fold enrichment for being within 2kb of annotated TSS, and genes whose TSS regions overlapped these states had higher average gene expression across different cell types (**Fig. 3A, C, Supp. Fig. 10-11**). These states differed among each other in their enrichments with upstream or downstream regions of the TSS (**Fig. 3E, Supp. Fig. 12**). The states in the transcription start site group (TSS1-2) had enrichment values that peaked at the TSS (>= 100 fold enrichment) (**Fig. 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 (**Supp. Fig. 13, Supp. Data File 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 [30] and the constitutively active nature of the two states, respectively. The DNase-specific state DNase1, showed distinct enrichment for CTCF-specific per-cell-type chromatin states defined in six cell types compared to other full-stack states [31] (**Fig. 3F, Supp. Fig. 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 [32–35]. The DNase1 state may correspond to similar roles, particularly where the CTCF-binding is relatively stable across cell types.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Supp. Fig 15**). Among promoter-associated states, those showing stronger enrichments with CpG Islands also show lowest methylation levels (**Fig. 3A, Supp. Fig. 15**), consistent with previous studies [36,37]. Some promoter-associated states (TSS1-2, PromF3-5, BivProm1-2) are most enriched at the center of binding regions of polycomb repressed complex 1 and 2 (PRC1 and 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 (**Supp. Fig. 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 File 1-3. W**e 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, Supp. Fig. 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. 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 [16] and 100-state per-cell-type annotations learned independently in each cell type (**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 (**Supp. Fig. 19-22, Supplementary Data 4)**. One of the exceptions to this was lambin B1 associated domains from Tig3 human lung fibroblasts [38], where six per-cell type 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

per-cell-type 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 per-cell type chromatin state annotation may show better recovery.

Stacked Model Differentiates Cell-Type-Specific from Constitutive Activity

While the major groups of states outlined above can correspond to states from models producing per-cell-type annotations [3,16], 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. 2B,C, Supp. Data. File 1)** and the overlap of per-cell-type chromatin state annotations from different cell types [39] **(Supp. Figure 8-9, Supp. Data 3)**.

Consistent with previous findings that enhancers tend to be relatively cell-type-specific while promoters tend to be shared across cell types [3,40], enhancer states exhibited clearer cell-type-specific associations than those of the promoter states (**Figure 2C, Supplementary Data. File 1,3**). On average, states of enhancer and weak enhancer groups (EnhW1-8, EnhA1-20) show 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 (**Supp. Fig. 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. 1-2, Supp. Fig. 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. 3C, Supp. Fig. 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 show higher emission

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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). Additionally, state HET9 showed strong association as heterochromatin in ESC/iPSC groups, while being mostly quiescent in other cell types based on per-cell-type annotations (**Fig. 2C, Supp. Fig. 26, Supplementary Data File 1,3**). 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 (**Supp. Fig. 27, Supplementary Data 1,3**). In addition, some quiescent states (Quies1-2, Quies4-5) show weak signals of H3K9me3 in specific groups of cell types (**Supplementary Data File 1**). States in the polycomb repressed and bivalent promoter groups (ReprPC1-9, BivProm1-4) also show differences in signals across cell groups, such as state ReprPC9, which showed H3K27me3 signals in only ESC/iPSC cell types (**Supplementary Data**). 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 (**Supp. Fig. 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. 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. 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 [41]. We also observed that individual full-stack states had distinct enrichments for different repeat classes (**Fig. 4C, Supp. Fig. 28**). For example, Acet1, a state associated with various acetylation marks and H3K9me3 had a 23-fold enrichment for simple

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Supp. Fig. 28**). As (CA)n and (TG)n repeats are known to be highly polymorphic in humans [42], 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. 4C, Supp. Fig. 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 (**Supp. Fig. 28**). Moreover, the TSS1-2 states are also most enriched with tRNA class (50-61 fold) (**Supp. Fig. 28**), consistent with tRNAs being short genes [43].

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. 5C**), which had high emission of H3K36me3 (**Fig. 2A-B, Supp. Fig. 4-5**), consistent with previous studies showing that SINEs are more enriched in gene-rich regions and in transcription-related states based on per-cell-type annotations [21,44,45]. In contrast, LINEs are depleted in most transcription-related states, reflecting the negative selection against long-sequence insertions in or near genes [44]. LINEs are most enriched in state HET3 (3.4 folds) (**Supp. Fig. 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 LTRs (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. 4C, Supp. Fig. 28**). LTRs showing strong enrichment with states associated with strong presence of H3K9me3 is consistent with per-cell-type chromatin state

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

analyses [21,45]. We also confirmed that the increased predictive power of the full-stack model over per-cell-type annotations, which was previously seen for repeat elements overall, also held for most of the individual repeat classes (**Supp. Fig. 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 per-cell-type annotations (**Supp. Fig. 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. 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. 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. 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 [46]. 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. 3B, Supp. Fig. 30-31**).

These 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 (**Supp. Fig. 30**). Another ConsHMM state, state 28, which is associated with moderate aligning and

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Supp. Fig. 30**). Additionally, this conservation state is consistently the most enriched conservation state for full-stack states associated with flanking and bivalent promoters (**Fig. 3B, Supp. Fig. 30**). 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 [46], 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. 3B, Supp. Fig. 30**). 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. 3A-B, Supp. Fig. 30**). This is consistent with previous analysis using per-cell-type annotations showing that ConsHMM state 100 was most enriched in a ZNF gene-associated chromatin state [46]. 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. 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. 3A-B, Supp. Fig 51**). 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 per-cell-type 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 [47], and focused on the two largest classes of SVs, deletions and duplications. Abel et al., 2020 [47] analysed the enrichments of these deletions and

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

duplications with per-cell-type chromatin states in 127 reference epigenomes [16], 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 in the top 10 maximally enriched states with 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. 5A, Supp. Fig. 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 the SV enrichments (**Supp. Fig. 30**). The other five states included two artifact states (GapArtf2-3) and three quiescent states (Quies1-2,4) (**Fig. 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. 5B**).

The full-stack model was also more predictive of SV than per-cell-type annotations (**Supp. Fig. 33-34, Supplementary Data 4**). Additionally, we verified that the full-stack model 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 [47] (**Methods**, **Supp. Fig. 35**). These results show that the full-stack annotation can uncover enrichment patterns with SVs that are consistent with per-cell-type 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 per-cell-type annotations. We considered a set of 14 different variant prioritization scores that were previously analyzed in the context of conservation state analysis [46]. 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 [48–61]. 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. 5C, Supp. Fig. 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. 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. 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 (**Supp. Fig. 42**) [62]. 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. 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) [54], and few eQTLs are expected to be proximal to genes with limited or no expression in that cell type. Enhancer

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. 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. 5C). FATHMM had the greatest relative enrichment for the primary DNase state associated with CTCF cell type-specific chromatin states (DNase1) (~10 fold), and was the only score for which this state was among the top five most enriched states (Fig. 5C, Supp. Fig. 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 (**Supp. Fig. 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 with these states (**Supp. Fig. 36**). In addition, we analyzed top 1%, 5%, and 10% prioritized variants genome-wide from 12 of the scores (**Methods**) (**Supp. Fig. 37-40**). Compared to the non-coding analysis, we saw a majority scores that have 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 shows that the scores tend to prioritize bases in different epigenetic contexts. As these 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

inconsistent enrichments of different methods for prioritized variants in GapArtf2-3 states (**Supp. Fig. 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 (**Supp. Fig. 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 per-cell-type annotations from a concatenated model for all 14 scores (**Supp. Fig. 33**). Compared to all 100-state per-cell-type annotations from independent models, the full-stack model showed the highest AUROC for 13 out of 14 scores in all 127 cell types (**Supp. Fig. 33**).

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) [62]. Across eleven ranges of MAFs, the state enrichments ranged from a 2-fold enrichment to a 4-fold depletion (**Supp. Fig. 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 (**Supp. Fig. 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**) (**Supp. Fig. 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) (**Supp. Fig. 42**). The high enrichment for states PromF4 and TSS1-2 for rare variants, despite their being

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. 3A, Supp. Fig. 42) [62].** At the same time, purifying selection can have a weaker effect on large-effect rare variants than on large-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 (**Supp. Fig. 42**). This pattern was not observed in most states from other groups such as heterochromatin, polycomb repressed, quiescent, and acetylations only (**Supp. Fig. 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. 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. 3B, Fig. 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. 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 [63] (**Methods**). This revealed six states with at least a two-fold enrichment (**Supp. Fig. 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. 3B**). This observation is consistent with previous results that GWAS catalog variants show enrichments for coding sequence and sequence constrained bases [46,64,65]. The

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

other two states with greater than two-fold enrichment for GWAS catalog variants relative to common variants were two promoter states, PromF2-3 (**Supp. Fig. 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) (**Supp. Fig. 43**). Both Quies5 and ReprPC8 are highly specific to chrX, 18.3 and 19.1 fold enriched respectively (**Supp. Fig. 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 [66]. Specifically, we considered separately the fine mapped variants from two fine-mapping methods, CAVIAR [67] and FINEMAP [68], 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) (**Supp. Fig. 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 in blood and thymus (EnhA9) and an enhancer state in most other cell types except blood (EnhA1), and three highly conserved transcription-associated states (TxEnh4,6, TxEx4) (**Fig. 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. 5E, C**). These results show that there are agreements in the types of chromatin 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 resulted in higher AUROC

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in predicting locations of fine-mapped variants within a background of common variants, compared to the per-cell-type annotations in all cell types (**Supp. Fig. 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 [69]: 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. 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. 5G**). This pattern of H3K9me3-associated states being enriched with somatic mutations in cancer is previously confirmed in multiple studies where H3K9me3 and other repressive epigenetic features showed positive association with mutation density across different types of cancer cells [70–72]. 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 [73,74]. Notably, the GapArtf2-3 states, which had strong satellite repeats enrichments (**Fig. 4C, Supp. Fig. 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. 5F, Supp. Fig. 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-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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, 4C, Supp. Fig. 42**). We note that the presence of cancer variants is better recovered by full-stack annotation as compared to the per-cell-type chromatin state annotations for all four cancer types (**Supp. Fig. 40-41**).

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 [39], however, we chose not to use imputed data as input to the full-stack model 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 held for per-cell-type annotations, yet it provided more

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

detailed stratification of genomic regions into chromatin states with heterogeneous associations with other genomic information. We provide extensive characterizations of full-stack states in **Supplementary Data Files 1-5** that will be resources in future applications of the full-stack annotations.

The full-stack modeling has advantages to commonly used per-cell-type 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 per-cell-type 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 per-cell-type 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.

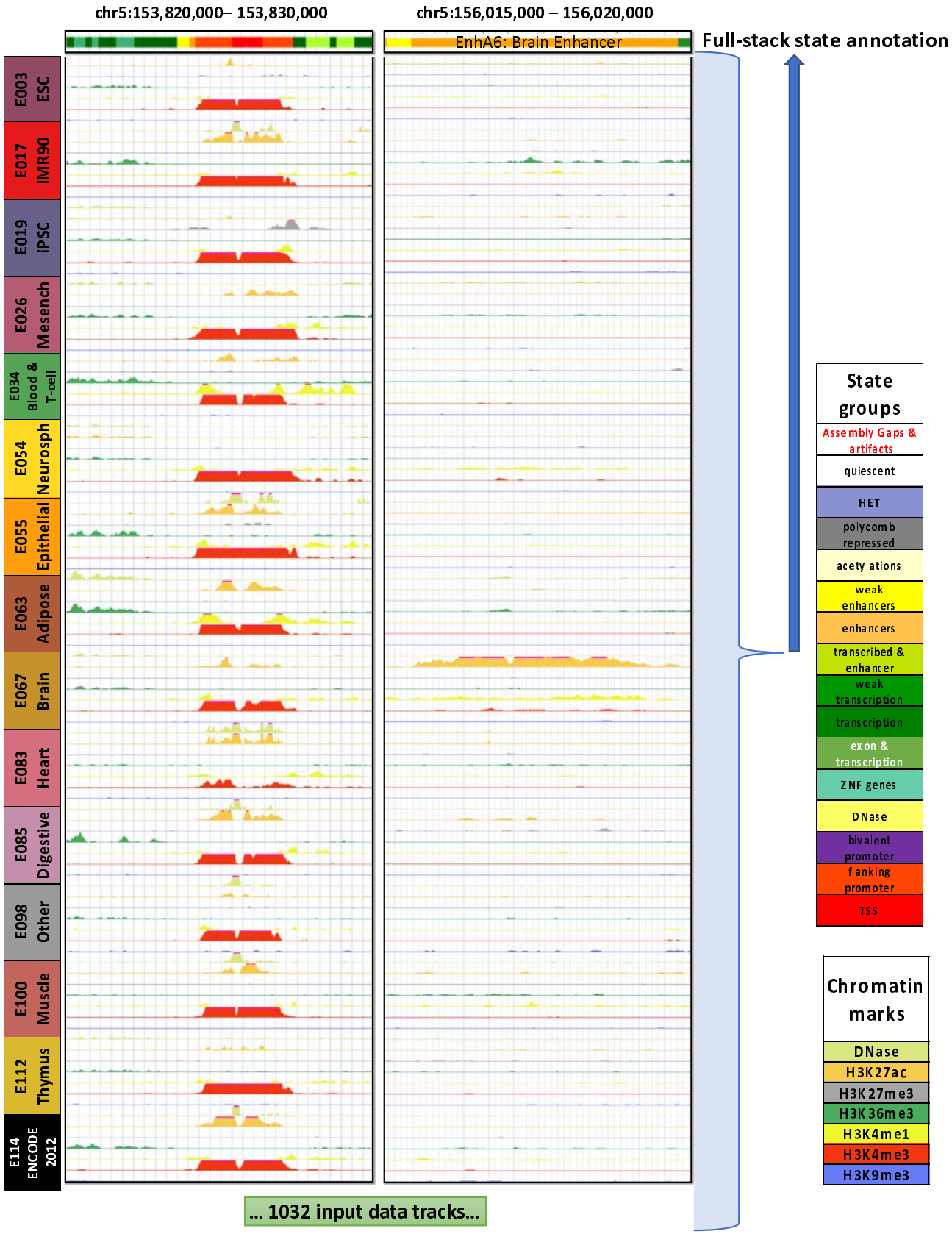
Despite these advantages, there are trade-offs in using the stacked modeling approach, and we emphasize that the stacked modeling approach should be considered a complement to and not a replacement of the per-cell-type annotations. Compared to typical concatenated models, the full-stack model has increased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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the existing model. We also note that post-hoc per-cell-type 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. While per-state analyses using per-cell-type concatenated annotations can be relatively straightforward and informative, they give only partial and potentially oversimplified views of all the data, ignoring distinctions among different per-cell-type states. Whether to use per-cell-type annotations or full-stack annotations will depend on the specific application. Per-cell-type annotations may be preferable when one is interested in studying a specific cell type, while full-stack annotations can 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 [75–77], 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 [18], future work could also learn new stack-models to incorporate such data. The state characterizations (**Supplementary Data File 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.

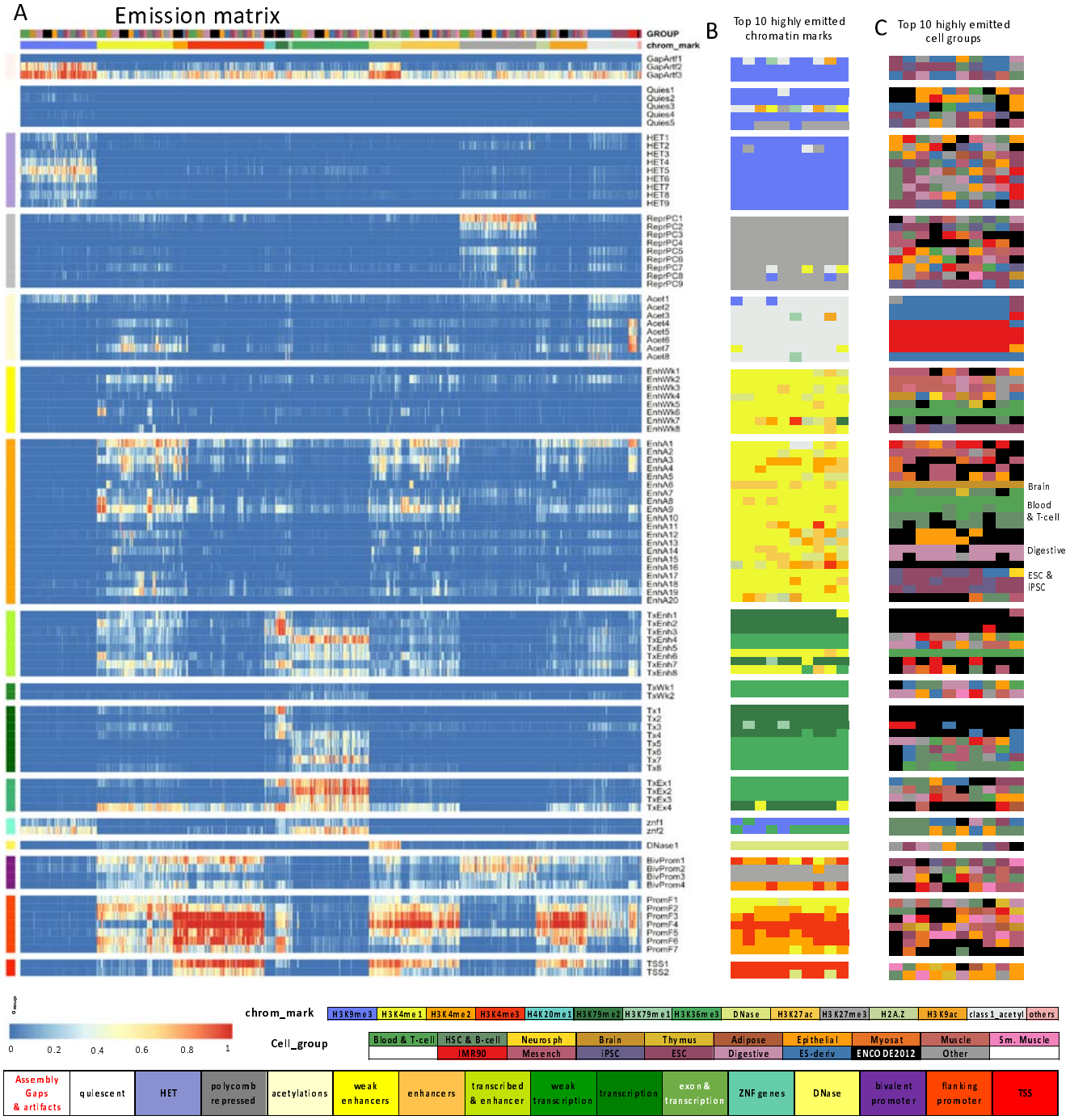
bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 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. Per-cell-type concatenated model annotations for these loci from these and additional reference epigenomes can be found in **Supp. Fig. 24.**

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 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 experiment. For each state and each experiment, the heatmap gives the probability within the state of observing a binary present call for the experiment's signal. Above the heatmap there are two rows, one indicating the cell or tissue type of the experiment 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**. The model's transition parameters between states can be found in **Supp. Fig. 6.** Columns are ordered such that experiments 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 experiments 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 experiments are colored by the associated cell or tissue type group. We noted on the right the cell or tissue groups of some per-cell-type enhancer states.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

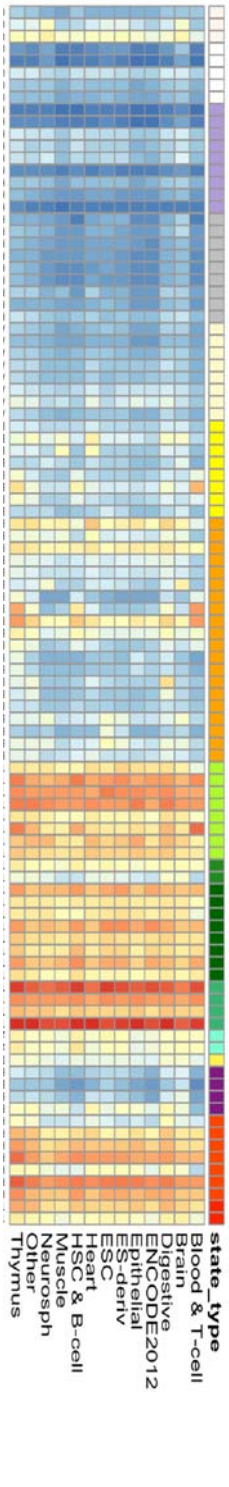
A Enrichments with external annotations

state	%Genome	GencODE_coding	RefSeqExon	RefSeqGene	RefSeqTSS	CpGIslnd	RefSeqTSS	RefSeqTSS2kb	assembly_gap	LamnB1Juds	phastCons_placentalia	pseudogene	repeat_elements	eVT_genes
GrpArt1	11.86	0.23	0.19	0.27	0.31	0.26	0.25	0.27	5.43	0.43	0.21	1.59	0.56	0.21
GrpArt2	0.05	0.26	0.33	0.15	0.42	1.25	1.6	0.8	0.02	0.09	0.35	1.65	1.85	1.44
GrpArt3	0.012	0.69	0.91	0.3	2.03	5.01	4.87	2.23	0.01	0.18	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	1.93	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	2.02	0.58	0.52	1.16	0.29
Quies3	12.23	0.22	0.39	1.01	0.76	0.07	0.14	0.44	0	1.09	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	1.5	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	1.51	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	1.95	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	1.53	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	1.4	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	0.92	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	0.92	0.59	2.52	1.35	1.4
HET6	0.581	0.86	0.67	0.48	0.63	1.28	0.86	0.63	0	1.23	0.46	2.11	1.66	2.58
HET7	1.023	0.36	0.35	0.68	0.4	0.64	0.43	0.41	0	1.36	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	0.93	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	1.2	0.51	3.33	1.49	1.44
ReprPC1	0.191	2.91	3.07	1.05	3.68	8.86	3.04	6.51	0	0.78	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	0.96	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	1.22	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	1.13	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	0.78	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	1.07	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	0.49	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	1.24	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	1.09	1.12	1.26	1.02	0.53
Acte1	0.184	2.58	1.68	1.09	0.47	3.23	1.75	1.1	0	0.69	1.06	1.64	1.24	1.09
Acte2	0.855	0.77	0.59	0.83	0.25	0.11	0.52	0.55	0	1.18	0.81	1.07	1.01	0.37
Acte3	2.649	0.36	0.36	0.83	0.34	0.05	0.24	0.45	0	1.07	0.54	1.02	1.29	0.4
Acte4	0.403	0.76	0.81	0.99	0.56	0.1	0.66	0.93	0	0.19	0.78	0.82	0.93	0.38
Acte5	0.86	0.29	0.4	0.89	0.48	0.01	0.28	0.48	0	0.27	0.75	0.86	1.06	0.32
Acte6	0.428	0.36	0.56	1.04	0.76	0	0.47	0.63	0	0.19	1.13	0.68	0.87	0.46
Acte7	0.285	0.63	0.89	1.13	0.61	0.13	0.98	1.29	0	0.11	1.25	0.58	0.64	0.39
Acte8	0.562	0.53	0.58	0.76	0.53	0.21	0.58	0.81	0	0.95	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	0.8	1.57	0.7	0.81	0.59
EnhWk2	0.352	1.95	2.03	1.66	1.76	0.77	1.85	2.21	0	0.42	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	0.93	1.86	0.57	0.69	0.72
EnhWk4	2.216	0.19	0.38	1.12	0.68	0.02	0.22	0.45	0	1.45	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	1.04	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	0.66	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	0.94	0.84	1.06	0.5
EnhWk8	1.369	0.19	0.35	0.94	0.61	0.01	0.22	0.35	0	1.57	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	0.07	2.32	0.33	0.32	0.73
EnhA2	0.208	0.47	0.94	1.31	1.2	0.13	1.97	1.72	0	0.53	3.36	0.54	0.42	0.8
EnhA3	0.194	0.41	0.87	1.61	1.4	0.01	1.96	2.22	0	0.14	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	0.27	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	0.62	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	1.1	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	0.66	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	0.51	1.27	0.55	0.67	0.71
EnhA9	0.162	1.24	1.78	1.11	2.12	0.39	2.78	4.24	0	0.09	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	0.42	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	0.76	0.81	1	1.14	0.73
EnhA12	0.332	1	1.08	0.92	0.9	0.22	1.76	1.61	0	0.77	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	1.11	1.06	0.76	0.96	0.27
EnhA14	0.366	1.09	1.22	1.14	1.28	0.16	2.84	3.02	0	0.65	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	1.39	1.05	0.95	1.01	0.42
EnhA16	0.645	0.42	0.59	0.91	1.01	0.37	0.69	2.09	0	0.94	1.01	1.34	1.04	0.43
EnhA17	0.526	0.3	0.65	1.36	1	0.05	0.82	1.15	0	1	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	1.45	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	0.59	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	1.01	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	0.06	1.64	0.31	0.65	1.14
TxEnh2	0.391	0.37	0.45	2.31	0.5	0.01	0.43	2.74	0	0.1	1.06	0.27	0.85	1.95
TxEnh3	0.25	1.15	1.14	2.31	0.65	0.58	0.69	1.27	0	0.06	0.86	0.68	1.01	3.15
TxEnh4	0.267	1.42	1.05	2.28	0.55	2.58	1.3	1.62	0	0.06	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	0.2	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	0.22	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	0.03	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	0.09	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	0.62	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	0.63	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	0.12	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	0.23	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	0.13	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	0.13	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	0.29	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	0.41	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	0.26	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	0.37	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	0.06	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	0.14	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	0.18	2.21	0.9	0.83	2.6
TxEx4	0.098	10.7	7.83	2.26	5.26	1.99	4	8.1	0	0.03	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	0.49	0.6	1.33	1.67	20.8
znf2	0.152	10.4	7.13	2.16	3.26	0.92	0.93	0.84	0	0.39	1.16	1.88	1.21	68.8
DNase1	0.201	0.99	1.51	1.07	2.42	3.59	5.59	1.88	0	0.72	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	0.69	3.94	1.02	0.16	0.87
BivProm2	0.159	7.36	6.41	1.32	4.03	54.5	14.8	14.9	0	0.83	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	0.63	2.47	0.8	0.42	0.82
BivProm4	0.13	2.55	3.39	1.74	2.99	3.75	8.11	16.5	0	0.4	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	0.31	1.65	0.63	0.45	0.81
PromF2	0.138	2.4	3.21	1.38	4.35	3.88	8.11	13.9	0	0.11	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	0.12	2.19	0.53	0.31	3.39
PromF4	0.19	10.2	13.5	1.88	2.52	96.7	62.2	26	0	0.17	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	0.37	3.15	0.72	0.16	2.27
PromF6	0.129	3.69	3.25	2.1	3.66	1.55	2.74	26.4	0	0.17	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	0.21	1.53	0.71	0.68	4.41
TSS1	0.123	6.48	12.6	1.44	2.63	10.1	152	25.9	0	0.24	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	0.46	4.46	1.74	0.57	2.18
	100	1.15	2.45	42.1	0.001	0.68	0.001	3.41	7.57	36.9	5.17	1.94	46.7	0.73

B Associated consHMM state

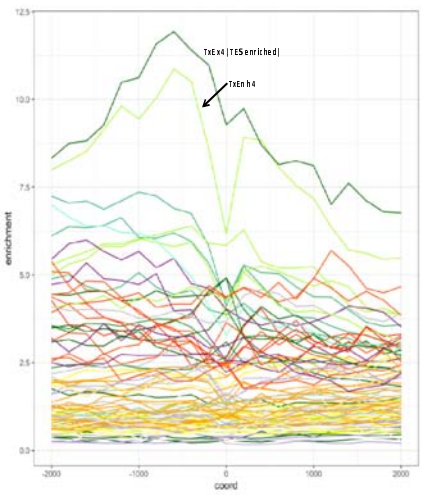
max_conditional consHMM state
9.6- AM_SPrim
9.6- AM_SPrim
9.5- AM_SPrim
3.0- AM_SMam
7.6- AM_Prim
8.0- AM_Prim
8.6- AM_Prim
7.2- AM_Prim
7.6- AM_Prim
8.2- AM_Prim
8.6- AM_Prim
9.3- AM_SPrim
1.00- artifact
9.3- AM_SPrim
9.3- AM_SPrim
8.2- AM_Prim
9.3- AM_SPrim
1.- AM_allVert
5.- AM_Mam
1.00- artifact
6.9- AM_Prim
8.2- AM_Prim
1.00- artifact
2.8- AM_SMam
1.00- artifact
1.00- artifact
8.2- AM_Prim
1.00- artifact
7.5- AM_Prim
1.4- AM_Mam
7.5- AM_Prim
9.- AM_Mam
1.4- AM_Mam
7.5- AM_Prim
5.- AM_Mam
2.8- AM_SMam
5.- AM_Mam
2.- AM_nonMam
7.5- AM_Prim
1.3- AM_Mam
7.7- AM_Prim
2.- AM_nonMam
5.- AM_Mam
2.- AM_nonMam
5.- AM_Mam
5.- AM_Mam
2.- AM_nonMam
2.8- AM_SMam
1.5- AM_Mam
1.4- AM_Mam
6.- AM_Mam
7.7- AM_Prim
1.4- AM_Mam
9.- AM_Mam
1.4- AM_Mam
6.- AM_Mam
1.00- artifact
2.- AM_nonMam
9.4- AM_SPrim
2.- AM_nonMam
5.- AM_Mam
6.- AM_Mam
3.7- AM_SMam
8.2- AM_Prim
1.- AM_allVert
1.- AM_allVert
1.- AM_allVert
4.7- A_SMam
1.- AM_allVert
7.- AM_Mam
1.- AM_allVert
4.4- AM_SMam
7.7- AM_Prim
7.5- AM_Prim
1.- AM_allVert
7.7- AM_Prim
7.- AM_Mam
1.- AM_allVert
3.- AM_nonMam
1.- AM_allVert
1.- AM_allVert
1.- AM_allVert
1.- AM_allVert
7.9- AM_Prim
1.00- artifact
2.8- AM_SMam
2.8- AM_SMam
2.8- AM_SMam
2.8- AM_SMam
2.- AM_nonMam
2.8- AM_SMam
5.- AM_Mam
2.8- AM_SMam
2.8- AM_SMam
2.8- AM_SMam
1.- AM_allVert
6.- AM_Mam
2.8- AM_SMam
2.8- AM_SMam

C Avg. gene exp



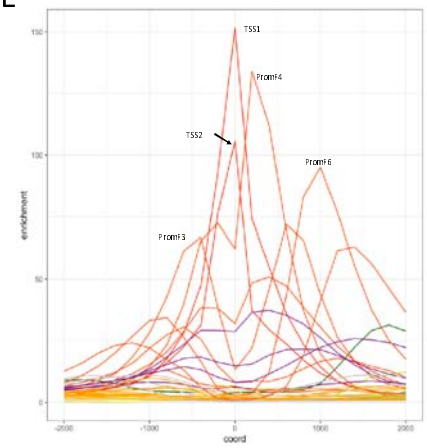
Annotated TES neighborhood enrichments

D

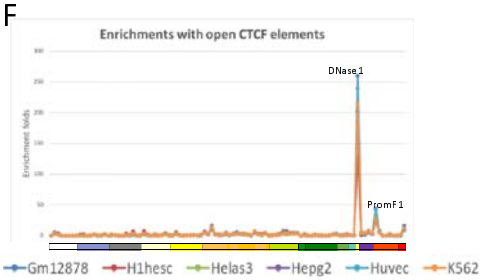


Annotated TSS neighborhood enrichments

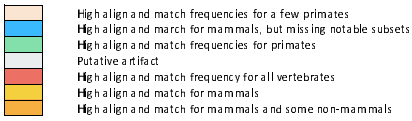
E



F



ConsHMM states



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 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 [46] 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 & Ernst, 2019) are available in **Supplementary Data 6.**

(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).**

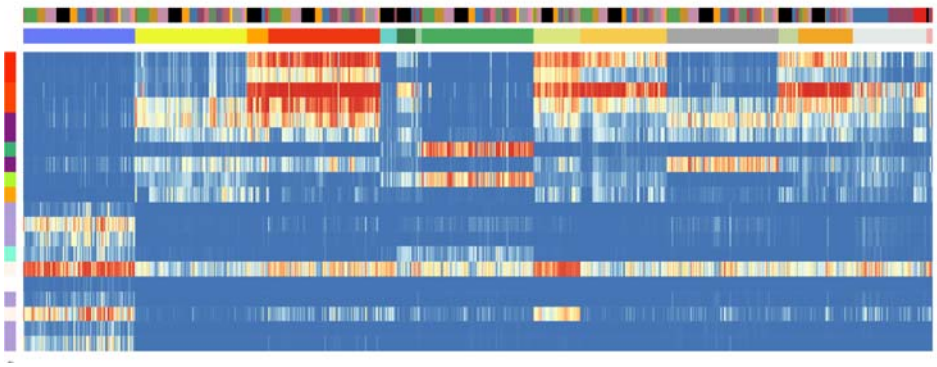
(F) Enrichments of full-stacks states with per-cell-type chromatin states associated with CTCF and open chromatin, but limited histone modifications in six cell types [31] (**Methods**). The six cell types are indicated along the bottom of the figure. States are displayed horizontally in the same order as **(A)**. The DNase1 state showed the strongest enrichment for the per-cell-type chromatin states associated with CTCF and open chromatin in all six cell types.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A

state	% genome	repeat_elements	PhastCons elements
TSS1	0.1	0.3	4.9
TSS2	0.1	0.6	4.5
PromF4	0.2	0.1	4.1
PromF5	0.1	0.2	4.0
BivProm1	0.1	0.2	3.9
BivProm4	0.1	0.3	3.9
TxEx2	0.6	0.5	3.7
BivProm2	0.2	0.2	3.6
TxEnh4	0.3	0.4	3.4
EnhA2	0.3	0.4	3.4
HET9	1.0	1.5	0.5
HET5	0.2	1.6	0.6
HET6	0.6	1.7	0.5
znf1	0.4	1.7	0.6
GapArtf3	0.0	1.7	0.7
Quies4	4.5	1.7	0.3
HET7	1.0	1.8	0.3
GapArtf2	0.1	1.8	0.4
HET3	1.4	1.9	0.3
HET4	0.6	1.9	0.3

B



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 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. 2A** for the subset of states highlighted in panel (A). The colors are the same in **Fig. 2A**.

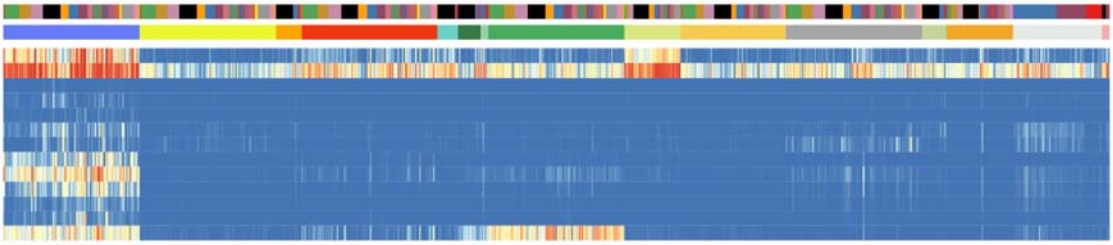
(C) Fold enrichments of full-stack states with different repeat classes (**Methods**). Rows correspond to states and columns to different repeat classes. Only states that are most enriched with at least one repeat class are shown. Fold enrichment values that are maximal for a given are shown in dark red. Other fold enrichments greater than one are shaded light red.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

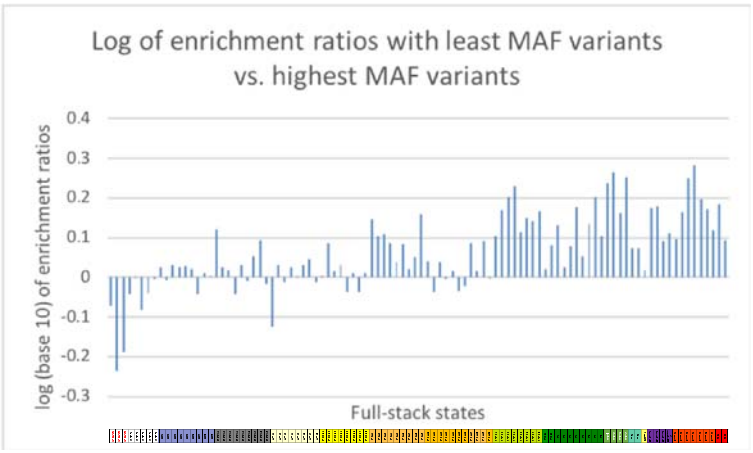


Rank of most enriched states	1	2	3	4	5	6-10	>10

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 TE5; per-ct state TxEnh3p 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

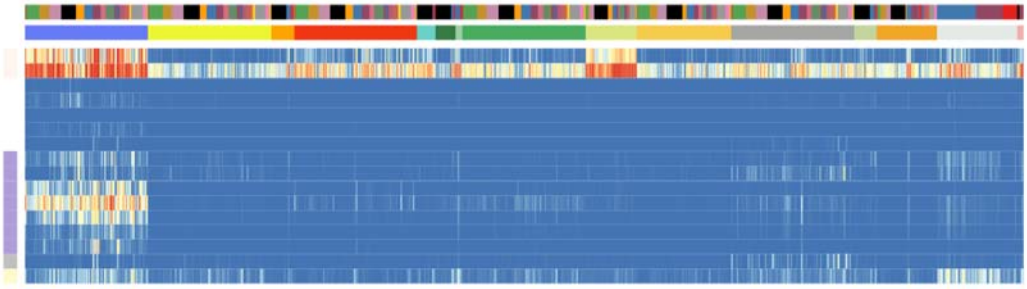
state	% background	caviar_lead_snps	finemap_lead_snps
EnhA1	0.17	2.92	2.93
EnhA9	0.15	2.72	2.72
TxEnh4	0.23	2.93	2.92
TxEnh6	0.17	2.65	2.65
TxEx4	0.08	3.36	3.36
BivProm4	0.12	2.86	2.86
PromF2	0.12	3.05	3.04
PromF3	0.13	3	3.01
PromF4	0.19	3.08	3.08
PromF5	0.13	2.88	2.89

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



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

(A) Enrichments of full-stack states with duplications and deletions from [47]. 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. 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 [46]. 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 [62] with the lowest MAFs (< 0.0001) vs. GNOMAD variants with the highest MAFs (0.4-0.5). States are ordered as in **Fig. 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 [67] or FINEMAP [68] (**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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. 2A**. The coloring is the same as **Fig. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 experiments and their corresponding input control data. The experiments 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 experiments, 979 were of ChIP-seq data targeting 31 different epigenetic marks and 53 were of DNase-seq (**Sup Fig. 2**). For each of the 127 reference epigenomes there was a single ChIP-seq input control experiment. For the 53 reference epigenomes that had a DNase-seq experiment 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 experiments.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To merge data from the 104 subdirectories from the previous step into files containing binarized data of all experiments, 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.

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 [46], 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 chromatin state 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 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 has ChromHMM not to 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

needed. The asymptotic worst-case time and memory of the model learning is discussed in the S**upplementary Information**.

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) [78] and Bayesian Information Criterion (BIC) [79] value based on a subset of the genome (**Supp. Fig. 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 [25] 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 models (**Supp. Fig. 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 [16], 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 (**Supp. Fig. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Supp. Fig. 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 with per-cell-type annotations [3,16,21], 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 per-cell-type annotations inferred. 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 [80] 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) (**Supp. Fig. 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 (**Supp. Fig.**

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18). The sources of external annotations from hg38 are outlined in section "External annotation sources" below.

Summary sets of experiments

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 Methods.** 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 [16], 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 experiments. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For each state, we looked into the emission probabilities of experiments associated with six chromatin marks strongly associated with promoter and enhancer activities (DNase, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac). We grouped these experiments based on their associated chromatin mark and tissue groups, and calculated the average emission probabilities of experiments 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\#SX$: 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 per-cell-type chromatin state annotations

We obtained per-cell type chromatin state annotations based on a 25-state ChromHMM model learned using concatenated approach for 127 reference epigenomes, which we will refer to as cell types for ease of presentation, from the Roadmap Epigenomics project [16,39]. This model was trained based on observed and 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 per-cell-type state in CT-25-state model that is maximally enriched, resulting in a 100-by-127 table (**Supp. Fig. 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 File 3** to serve as a resource for future applications. We also reported, for each of the 100 full-stack states and each per-cell-type-25-state, the maximum and median values of fold enrichments across 127 cell types (**Supplementary Data File 3**).

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 [16]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Supp. Fig. 9**). These values, along with detailed comments about patterns of these overlap probabilities for each full-stack state, are available in **Supplementary Data File 3**.

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).

Per-cell-type 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 per-cell type annotations was the 18-state ChromHMM from Roadmap Epigenomic Project [16], 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 per-cell type annotations were based on models learned independently in each of the 127 cell types. In learning these

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 per-cell-type 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 per-cell-type 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 [66]. Specifically we obtained posterior probabilities of variants being causal based on two fine-mapping methods, FINEMAP [68] and CAVIAR [67], 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [81].

Computing structural variant enrichments

To compute enrichment of the full-stack states for structural variant enrichments, we obtained data of structural variants from [47]. We used the B38 call set, which was in hg38 and used for the analysis presented in [47]. We filtered out structural variants that did not pass the quality control criteria of [47]. 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 and per-cell-type concatenated 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. per-cell-type chromatin state annotations' power in recovering structural variants.

To compare the power of full-stack state annotations vs. per-cell-type state annotation frequency, we utilized the 15-state per-cell-type chromatin state annotation for 127 cell types (reference epigenomes) from Roadmap Epigenomics Consortium. We followed the analysis outlined in [47], for each of the 15 per-cell-type 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 per-cell-type state to compare the predictive power of the state's annotation frequency against the full-stack annotation.

Computing enrichments with cancer-associated variants

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We obtained data of somatic mutations associated with different types of cancer from COSMIC non-coding variants dataset v.88 in hg38 [69]. 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) [82], 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 [83] 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 gene ontology enrichments of genes being in proximity to each full-stack state annotation using GREAT [84]. For each full-stack state, we reported the top-GO terms with lowest FDR-corrected p-values. If multiple GO terms showed the same minimum p-values, we reported all of them. A full list of GO-terms associated with full-stack states are available in **Supplementary Data File 1**.

External annotations sources

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The sources for external annotations for enrichments analyses, not given above, were as follows (all download links are listed in **Supplementary Data File 5**):

- 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 [38].
- 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 [83]. 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 [83].
- 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 [83].
- Annotations of repeat elements were obtained from UCSC genome browser RepeatMasker hg19 tracks.
- Per-cell-type concatenated ChromHMM chromatin state annotations were obtained from the Roadmap Epigenomics Consortium through http://compbio.mit.edu/roadmap [16]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- CTCF- per-cell-type 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 [22,31]. 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 [82].
- ConsHMM conservation state annotations for human (hg19) were those from [46].
- Annotations of human genetic variants and their allele frequency were from GNOMAD v2.1.1 [62]. The dataset includes 229 million SNVs and 33 million indels from 15,708 genomes of unrelated individuals, which are aligned against the GRCg37/hg19 reference.
- GWAS catalog variants were obtained from the NHGRI-EBI Catalog, accessed on December 5th, 2016 [63].
- 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 [85]. Download links are listed in **Supplementary Data File 5.**

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 [16]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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:

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

We also calculated for each state the average and coefficient of variation of these averages across reference epigenomes. We used the BEDTools [86] *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 group as presented in **Fig. 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 [3]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [16]. 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 [46]. We obtained coordinates of bases containing prioritized variants based on

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

14 different methods as processed and described in [46]. The scores were Eigen and Eigen-PC version 1.1, funSeq2 version 2.1.6, and CADD v1.4, REMM, FIRE, fitCons, CDTS, LINSIGHT, FATHMM-XF, GERP++, phastCons, phyloP and DANN [48,50–61]. 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 [46], 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 is available at https://github.com/ernstlab/full_stack_ChromHMM_annotations. An updated 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 File**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We thank Adriana Arneson for helping us in collecting data of analyses involving prioritized variants, repeat classes and GWAS catalog variants. We thank Hector Corrada Bravo, Amin Haghani, and Steve Horvath, Caesar Li, and Ake Lu for helpful discussions related to the manuscript. 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.

Ethics Declarations

The authors announce no conflicts of interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. Elsevier; 2007;129:823–37.

2. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. Elsevier; 2008;132:311–22.

3. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. Nature Publishing Group; 2011;473:43–9.

4. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. Nature Publishing Group; 2012;489:75–82.

5. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. Elsevier; 2013;153:1134–48.

6. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, et al. FTO obesity variant circuitry and adipocyte browning in humans. N Engl J Med. Mass Medical Soc; 2015;373:895–907.

7. Gjoneska E, Pfenning AR, Mathys H, Quon G, Kundaje A, Tsai L-H, et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature. Nature Publishing Group; 2015;518:365–9.

8. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. Cold Spring Harbor Lab; 2013;23:800–11.

9. Lay FD, Triche TJ, Tsai YC, Su S-F, Martin SE, Daneshmand S, et al. Reprogramming of the human intestinal epigenome by surgical tissue transposition. Genome Res. Cold Spring Harbor Lab; 2014;24:545–53.

10. Lee J, Krivega I, Dale RK, Dean A. The LDB1 complex co-opts CTCF for erythroid lineage-specific long-range enhancer interactions. Cell Rep. Elsevier; 2017;19:2490–502.

11. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. Cold Spring Harbor Lab; 2014;24:1421–32.

12. Varshney A, Scott LJ, Welch RP, Erdos MR, Chines PS, Narisu N, et al. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc Natl Acad Sci. National Acad Sciences; 2017;114:2301–6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

13. Consortium EP. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. nature. Nature Publishing Group; 2007;447:799.

14. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. Nature Publishing Group; 2012;489:57–74.

15. Fernández AF, Bayón GF, Urdinguio RG, Toraño EG, García MG, Carella A, et al. H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. Genome Res. Cold Spring Harbor Lab; 2015;25:27–40.

16. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. Nature Publishing Group; 2015;518:317–30.

17. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. Nature Publishing Group; 2007;448:553–60.

18. Stunnenberg HG, Abrignani S, Adams D, de Almeida M, Altucci L, Amin V, et al. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell. Elsevier; 2016;167:1145–9.

19. Wang Q, Yu G, Ming X, Xia W, Xu X, Zhang Y, et al. Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. Nat Genet. Nature Publishing Group; 2020;52:828–39.

20. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell. Elsevier; 2013;152:642–54.

21. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. Nature Publishing Group; 2010;28:817–25.

22. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. Nature Publishing Group; 2012;9:215–6.

23. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. Nature Publishing Group; 2012;9:473.

24. Libbrecht MW, Rodriguez OL, Weng Z, Bilmes JA, Hoffman MM, Noble WS. A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. Genome Biol. Springer; 2019;20:180.

25. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. Nature Publishing Group; 2017;12:2478.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26. Biesinger J, Wang Y, Xie X. Discovering and mapping chromatin states using a tree hidden Markov model. BMC Bioinformatics. Springer; 2013. p. S4.

27. Zhang Y, An L, Yue F, Hardison RC. Jointly characterizing epigenetic dynamics across multiple human cell types. Nucleic Acids Res. Oxford University Press; 2016;44:6721–31.

28. Chronis C, Fiziev P, Papp B, Butz S, Bonora G, Sabri S, et al. Cooperative binding of transcription factors orchestrates reprogramming. Cell. Elsevier; 2017;168:442–59.

29. Mortazavi A, Pepke S, Jansen C, Marinov GK, Ernst J, Kellis M, et al. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. Genome Res. Cold Spring Harbor Lab; 2013;23:2136–48.

30. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. Elsevier; 2006;125:315–26.

31. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. Oxford University Press; 2013;41:827–41.

32. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. Nature Publishing Group; 2012;485:376–80.

33. McArthur E, Capra JA. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. Am J Hum Genet. Elsevier; 2021;108:269–83.

34. Phillips JE, Corces VG. CTCF: Master Weaver of the Genome. Cell. 2009;137:1194–211.

35. Wang Y, Liu Y, Xu Q, Xu Y, Cao K, Deng N, et al. TAD boundary and strength prediction by integrating sequence and epigenetic profile information. Brief Bioinform. 2021;

36. Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. Science. American Association for the Advancement of Science; 2001;293:1068–70.

37. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet. Nature Publishing Group; 2007;39:457–66.

38. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. Nature Publishing Group; 2008;453:948–51.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

39. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. Nature Publishing Group; 2015;33:364–76.

40. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. Nature Publishing Group; 2007;39:311–8.

41. Becker JS, Nicetto D, Zaret KS. H3K9me3-dependent heterochromatin: barrier to cell fate changes. Trends Genet. Elsevier; 2016;32:29–41.

42. Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature. Nature Publishing Group; 1996;380:152–4.

43. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. Oxford University Press; 1997;25:955–64.

44. Elbarbary RA, Lucas BA, Maquat LE. Retrotransposons as regulators of gene expression. Science. American Association for the Advancement of Science; 2016;351.

45. Pehrsson EC, Choudhary MN, Sundaram V, Wang T. The epigenomic landscape of transposable elements across normal human development and anatomy. Nat Commun. Nature Publishing Group; 2019;10:1–16.

46. Arneson A, Ernst J. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. Commun Biol. Nature Publishing Group; 2019;2:1–14.

47. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. Mapping and characterization of structural variation in 17,795 human genomes. Nature. Nature Publishing Group; 2020;583:83–9.

48. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat Methods. Nature Publishing Group; 2010;7:250–1.

49. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. Public Library of Science; 2010;6:e1001025.

50. Di Iulio J, Bartha I, Wong EH, Yu H-C, Lavrenko V, Yang D, et al. The human noncoding genome defined by genetic diversity. Nat Genet. Nature Publishing Group; 2018;50:333–7.

51. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. BioMed Central; 2014;15:1–15.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

52. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. Nature Publishing Group; 2015;47:276–83.

53. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. Nature Publishing Group; 2017;49:618–24.

54. Ioannidis NM, Davis JR, DeGorter MK, Larson NB, McDonnell SK, French AJ, et al. FIRE: functional inference of genetic variants that regulate gene expression. Bioinformatics. Oxford University Press; 2017;33:3895–901.

55. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. Nature Publishing Group; 2016;48:214.

56. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. Cold Spring Harbor Lab; 2010;20:110–21.

57. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. Oxford University Press; 2015;31:761–3.

58. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. Oxford University Press; 2019;47:D886–94.

59. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. Oxford University Press; 2018;34:511–3.

60. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. Cold Spring Harbor Lab; 2005;15:1034–50.

61. Smedley D, Schubach M, Jacobsen JO, Köhler S, Zemojtel T, Spielmann M, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am J Hum Genet. Elsevier; 2016;99:595–606.

62. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. Nature Publishing Group; 2020;581:434–43.

63. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. Oxford University Press; 2014;42:D1001–6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

64. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci. National Acad Sciences; 2009;106:9362–7.

65. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. Nature Publishing Group; 2011;478:476–82.

66. Wang J, Huang D, Zhou Y, Yao H, Liu H, Zhai S, et al. CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. Nucleic Acids Res. Oxford University Press; 2020;48:D807–16.

67. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics. Genetics Soc America; 2015;200:719–36.

68. Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. Oxford University Press; 2016;32:1493–501.

69. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. Oxford University Press; 2019;47:D941–7.

70. Parker SC, Gartner J, Cardenas-Navia I, Wei X, Abaan HO, Ajay SS, et al. Mutational signatures of de-differentiation in functional non-coding regions of melanoma genomes. PLoS Genet. Public Library of Science; 2012;8:e1002871.

71. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. Nature Publishing Group; 2015;518:360–4.

72. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. nature. Nature Publishing Group; 2012;488:504–7.

73. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. American Association for the Advancement of Science; 2015;349:1483–9.

74. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. Nature Publishing Group; 2015;521:81–4.

75. Arneson A, Haghani A, Thompson MJ, Pellegrini M, Kwon SB, Vu HT, et al. A mammalian methylation array for profiling methylation levels at conserved sequences. Biorxiv. Cold Spring Harbor Laboratory; 2021;

76. Horvath S, Haghani A, Peng S, Hales EN, Zoller JA, Raj K, et al. DNA methylation aging and transcriptomic studies in horses. Biorxiv. Cold Spring Harbor Laboratory; 2021;

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.17.387134; this version posted September 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

77. Li CZ, Haghani A, Robeck TR, Villar D, Lu AT, Zhang J, et al. Epigenetic predictors of maximum lifespan and other life history traits in mammals. bioRxiv. Cold Spring Harbor Laboratory; 2021;

78. Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statistics. Dordr Neth Reidel. Taylor & Francis; 1986;81:26853.

79. Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. Wiley Interdiscip Rev Comput Stat. Wiley Online Library; 2012;4:199–203.

80. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. Cold Spring Harbor Lab; 2002;12:996–1006.

81. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC genome browser database: 2021 update. Nucleic Acids Res. Oxford University Press; 2021;49:D1046–57.

82. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. Sci Rep. Nature Publishing Group; 2019;9:1–5.

83. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. Cold Spring Harbor Lab; 2012;22:1760–74.

84. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. Nature Publishing Group; 2010;28:495–501.

85. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. Oxford University Press; 2018;46:D794–801.

86. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. Oxford University Press; 2010;26:841–2.