ChromBERT: Uncovering Chromatin State Motifs in the Human Genome Using a BERT-based Approach

Seohyun Lee[1], Junichi Sakatsume[2], Gina Miku Oba[2], Yuya Nagaoka[2], Che Lin[3], Chien-Yu Chen[4], and Ryuichiro Nakato[2,*]

[1]Interfaculty initiative in Information Studies, Graduate School of Interdisciplinary Information Studies, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.
[2]Laboratory of Computational Genomics, Institute for Quantitative Biosciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan.
[3]Department of Electrical Engineering, Graduate Institute of Communication Engineering, School of Medicine, Center for Advanced Computing and Imaging in Biomedicine, Center for Biotechnology, Taipei 10617, Taiwan.
[4]Department of Biomechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan.

*To whom correspondence should be addressed: rnakato@iqb.u-tokyo.ac.jp

Abstract

Chromatin states, which are defined by specific combinations of histone post-translational modifications, are fundamental to gene regulation and cellular identity. Despite their importance, comprehensive patterns within chromatin state sequences, which could provide insights into key biological functions, remain largely unexplored. In this study, we introduce ChromBERT, a BERT-based model specifically designed to detect distinct chromatin state patterns as "motifs." We pre-trained ChromBERT on 15-state chromatin annotations from 127 human cell and tissue types from the ROADMAP consortium. This pre-trained model can be fine-tuned for various downstream tasks, and obtained high-attention chromatin state patterns are extracted as motifs. To account for the variable-length nature of chromatin state motifs, ChromBERT uses Dynamic Time Warping to cluster similar motifs and identify meaningful representative patterns. In this study, we evaluated the performance of the model on several tasks, including binary and quantitative gene expression prediction, cell type classification, and

three-dimensional genome feature classification. Our analyses yielded biologically grounded results and revealed the associated chromatin state motifs. This workflow facilitates the discovery of specific chromatin state patterns across different biological contexts and offers a new framework for exploring the dynamics of epigenomic states.

Background

The understanding of chromatin organization is essential to reveal the complex mechanisms that govern gene regulation and function in the human genome[1], [2]. Chromatin states, defined by the combined features of histone modifications, including the involvement of DNA methylation and acetylation, are associated with specific functions such as gene activation, repression, or structural organization [3], [4]. For instance, active chromatin states—often marked by histone modifications such as H3K4me3 (trimethylation of lysine 4 on histone H3) and H3K27ac (acetylation of lysine 27 on histone H3) —are generally linked to gene activation and transcriptional initiation [5], [6]. These associations were further refined by Heintzman et al.[7], who distinguished promoter- and enhancer-specific marks, and Creyghton et al.[8], who identified H3K27ac as a marker of active regulatory elements. A deeper understanding of chromatin states and their genomic distribution has been facilitated by the chromatin state annotation tools such as ChromHMM [9] and Segway [10] coupled with the advent of chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) experiments [11], [12], [13]. Such annotation tools process genome-wide ChIP-seq datasets and systematically learn and characterize combinatorial patterns of histone modifications into distinct chromatin states, such as “active promoters,” “genic enhancers” and “heterochromatin.”

Large databases of epigenome information have been generated by international consortia, such as the ROADMAP project [14], the ENCODE project [15], and the International Human Epigenome Consortium (IHEC) [16]. For instance, the ROADMAP project characterized and annotated the chromatin state of 127 distinct human cell and tissue types. The chromatin state dataset can be used to perform various downstream analyses to decipher the hidden intricacies of the chromatin state landscape. As one such attempt, ChromDiff [17] compared a summarized set of chromatin states across different epigenomes (cell and tissue types), offering new insights into tissue-specific or age-dependent chromatin alterations. More recently, CSREP [18], a framework for deriving representative chromatin state maps for groups of samples, was

developed by Vu *et al*. Utilizing an ensemble of multi-class logistic regression classifiers, this method enables the summarization of chromatin states for different groups of epigenomes with high resolution. Additionally, Epilogos aims to visualize the conserved chromatin states across different epigenomes based on the surprisal score measured from the rarity of the chromatin state distribution [19]. A similar method, established with a different metric, the conservation-associated activity score, was suggested by Libbrecht *et al* [20] as a novel annotation strategy that has the advantage of detecting more biologically functional regions.

Recently, deep learning approaches designed for natural language processing have shown great promise in deciphering complex biological patterns from large-scale genomic and epigenomic datasets. Transformer-based architectures have been especially effective in modeling sequential dependencies, leading to successful applications in biological contexts. For instance, DNABERT [21] has been developed for DNA sequence analysis, while Geneformer [22] and EPInformer [23] have demonstrated the potential of deep learning for modeling cellular states and integrating epigenomic data.

The availability of chromatin state annotations for a large dataset, recent advances in deep learning, and the precedent related studies have paved the way for the next important step: identifying patterns of chromatin states as 'motifs'. A 'motif' refers to a recurring sequence that possesses a biological significance, often associated with specific functions such as binding sites in DNA sequences [24]. In the context of our study, we extend this concept to chromatin states to identify distinct patterns that could be key to understanding gene regulation. Instead of examining individual genomic loci separately, analyzing the patterns and combinations of chromatin states in relation to the surrounding regions, which we refer to as chromatin state motifs, could reveal the functional significance and regulatory roles of spatially coordinated epigenomic modifications that were previously unrecognized. By mapping these motifs, we may begin to decode the epigenetic language that potentially modulates gene expression patterns across different cell types and developmental stages. However, since chromatin state patterns are inherently dynamic and varying in length and structure, they may not be fully captured by traditional motif identification methods, such as k-mer-based algorithms, which often rely on simplistic representations of DNA sequence [25], [26]. This underscores the need for strategies that can accurately decipher the motifs of chromatin states and reveal their functional implications in gene regulation and chromatin organization.

In this study, we introduce ChromBERT, which is specifically tailored for the discovery of chromatin state motifs using the Bidirectional Encoder Representations from Transformers (BERT) model [27]. BERT has proven highly effective in natural language processing tasks and provides an important way to analyze sequential patterns in biological data [28]. By integrating the BERT model with the sequence alignment algorithm called Dynamic Time Warping (DTW) [29], ChromBERT effectively analyzes dynamic chromatin states, which extends the concept proposed by DNABERT for DNA sequence analysis [21]. ChromBERT is designed as a general-purpose model capable of learning and representing genome-wide chromatin state patterns across cell types. In this study, we evaluated the performance of ChromBERT on several tasks, including binary and quantitative gene expression prediction, cell type classification, and three-dimensional genome feature classification. Using the chromatin state annotations from 127 human cell and tissue types from the ROADMAP consortium, ChromBERT successfully identified chromatin state motifs, demonstrating its potential for predicting more complex genomic features. The source code of ChromBERT is available at https://github.com/caocao0525/ChromBERT.

Results

ChromBERT framework

Figure 1 provides an overview of the ChromBERT framework, illustrating the key components and workflow involved. We used the 15-state chromatin annotations defined by the ROADMAP project, which are based on a combination of five histone modifications. Initially, numerically annotated chromatin state labels ranging from 1 to 15 are converted into alphabetical codes ranging from A to O, resulting in the generation of a chromatin state sequence (Figure 1(a)). In the training phase, the chromatin state sequences are tokenized and fed into a BERT model for embedding (Figure 1(b)). We tokenized the chromatin state sequences into overlapping 4-mers by sliding one character, resulting in a large vocabulary size of 50,630 including standard special tokens. We also explored tokenization by sliding *N* characters at once (stride *N*, Supplementary Fig. S1). Since chromatin state sequences are more continuous than DNA sequences, we assumed that skipping multiple characters would not significantly affect prediction performance. This approach enables the use of longer input sequences and allows the model to learn long-range dependencies across regulatory regions.

The overall workflow of ChromBERT is captured in Figure 1(c), starting with the preprocessing of chromatin state sequence data. ChromBERT learns the general patterns of chromatin state placement through a pretraining stage. Following this, ChromBERT is fine-tuned for specific binary classification and quantitative regression tasks, such as promoter region classification and gene expression level prediction. As a downstream analysis of the classification task, chromatin state motifs specific to each task are extracted from the attention matrix values.

Motif discovery is a key downstream component of ChromBERT. Unlike DNA sequence motifs, chromatin state motifs lack well-defined reference libraries, making their identification and interpretation more challenging. Additionally, chromatin state sequences are inherently cell-type specific and dynamic. Even chromatin state patterns with similar regulatory functions can vary in length across genomic regions, reflecting both biological variability (e.g., differences in regulatory domain size) and technical variability (e.g., fluctuations in ChIP-seq signal-to-noise ratios). To address these challenges, ChromBERT employs a Dynamic Time Warping (DTW)-based approach to align and cluster motifs based on their structural similarity rather than exact sequence identity. This enables biologically meaningful motif discovery that accounts for the variability in chromatin state patterns. See the Methods section for further technical details.

Pretraining ChromBERT

To learn generalizable representations of chromatin state sequences, ChromBERT was first pretrained using a masked token prediction objective on chromatin state sequences across 127 human cell and tissue types in the ROADMAP dataset. We pretrained ChromBERT in two ways for different use cases. First, ChromBERT was pretrained on promoter regions (from 2 kb upstream to 4 kb downstream of transcription start sites [TSSs]), as illustrated in Figure 2(a). These regions were chosen because they play a central role in gene regulation, integrating signals from nearby and distal cis-regulatory elements such as enhancers, silencers, and insulators [30]. Second, we pretrained the model on chromatin state sequences spanning the entire genome to support broader applications and facilitate flexible fine-tuning for other genomic regions or tasks. In this study, we primarily used the whole-genome–pretrained model.

Throughout the pretraining process, the model's perplexity decreased steadily, indicating successful learning of chromatin sequence structure (Figure 2(b)). For genome-wide pretraining,

perplexity dropped from approximately 4.96 to 1.09. For promoter-specific pretraining, it decreased from approximately 2.75 to 1.24. The relatively low initial perplexity, particularly in promoter regions, likely reflects the repetitive and locally stable nature of chromatin state patterns near regulatory elements. See Methods section for further technical details.

ChromBERT distinguishes between highly and lowly expressed genes based on chromatin state patterns of promoter-proximal regions

To evaluate the ability of ChromBERT to capture regulatory features, we first fine-tuned the model on chromatin state sequences around the TSSs to classify regions associated with genes of different expression levels. We used 57 cell types for which RNA-seq data are available in ROADMAP and evaluated the binary classification accuracy between highly expressed genes and genes with different expression levels. Genes were grouped based on their log-transformed expression values (RPKM). Genes with log-transformed RPKM values greater than 5 were labeled as 'highly expressed,' and we compared them with five groups: non-expressed genes and gene groups whose log-transformed RPKM values fell within 1–2, 2–3, 3–4, or 4–5. Figure 3(a) shows representative chromatin state sequences in promoter regions associated with highly expressed and non-expressed genes. Active states such as "active TSS" ("A"), "strong transcription" ("D"), "weak transcription" ("E"), and "genic enhancers" ("F") frequently appear adjacent to highly expressed genes. In contrast, repressed states, such as Polycomb-associated states ("M" and "N"), dominate non-expressed regions.

We next evaluated how the choice of input region around TSSs affects fine-tuning performance. We tested multiple combinations of upstream and downstream windows and summarized the resulting F1 scores in Figure 3(b) (AUC and accuracy shown in Supplementary Fig. S2). As expected, classification between highly expressed and non-expressed genes was consistently high across all input lengths, while comparisons against genes with higher expression levels were progressively more challenging and yielded lower scores. Across input configurations, we observed modest performance gains when upstream regions were extended from short (50 kb) to moderately long ranges (approximately 90–190 kb). Including downstream regions also improved performance, particularly when combined with longer upstream windows. The highest overall accuracy was obtained with the up190k/down100k configuration (orange arrow). These results indicate that promoter-proximal chromatin states over extended upstream

and downstream regions contribute meaningful information for distinguishing transcriptional activity levels. Overall, genes with high and low expression were classified with high accuracy, suggesting that ChromBERT effectively captures the regulatory patterns underlying transcriptional activity.

ChromBERT identifies promoter chromatin state motifs associated with high gene expression

We next examined how ChromBERT distributes attention across the input sequences to identify which regions contribute most strongly to the classification task. As shown in Figure 4(a) and 4(b), attention values were sharply concentrated around the TSS, indicating that ChromBERT consistently prioritizes promoter-proximal features even when provided with extended upstream and downstream contexts. Outside this central peak, we also observed smaller high-attention regions on both sides of the TSS, suggesting that additional regulatory patterns in nearby upstream and downstream regions contribute to the model's decisions. These results highlight that promoter-proximal chromatin states serve as the primary informative signals for distinguishing expression levels, while adjacent regions provide supplementary contextual cues.

To illustrate how the model focuses on individual promoter-proximal sequences, we visualized representative loci by plotting chromatin states with heights proportional to their attention values (Figure 4(c)). In the highly expressed example, the region surrounding the TSS shows high attention on promoter- and transcription-associated chromatin states, including "active TSS" ("A"), "flanking active TSS" ("B"), "strong transcription" ("D"), "weak transcription" ("E"), and with smaller contributions from "enhancers" ("G"). In a moderately expressed gene, attention around the TSS is lower overall and shifts toward transcribed and "ZNF genes + repeat" associated states ("H"), reflecting the chromatin context of this locus. Together with the global distribution of attention values (Figure 4(d)), these examples suggest that promoter- and transcription-associated states provide the primary signals for distinguishing expression levels, while nearby enhancer or repeat-associated states supply additional contextual information.

DTW-based clustering reveals promoter chromatin motifs linked to high gene expression

The previous results suggested that chromatin state sequences tend to be persistent. Additionally, due to biological factors (e.g., different lengths of genes and enhancer regions) or technical factors (e.g., differences in read depth and signal-to-noise ratios across ChIP-seq samples), motif instances may vary in both character and length, even if they have the same biological functions. To summarize overarching patterns among the variable-length instances, we applied DTW and grouped similar chromatin state motifs into coherent families. This grouping uncovers recurring regulatory structures across genomic contexts post-fine-tuning, while still accounting for local flexibility in sequence patterns.

Initially, ChromBERT identifies motifs without imposing constraints on the merging. This approach captures a comprehensive set of chromatin state patterns that potentially signify the regions of interest, with selection based on statistical significance (p-value). ChromBERT then uses DTW to align the similar motifs in the set. Originating from the field of speech recognition, DTW is adept at accommodating variations in the tempo of analogous spoken words [31], making it particularly suitable for our purposes of aligning the different lengths of continuous sequences. DTW computes sequence similarities not by conventional point-to-point Euclidean distances but by minimizing the cumulative distances between motifs. Finally, the agglomerative clustering algorithm [32] is applied to cluster similar patterns into categories based on the sequence similarities, utilizing the hierarchical merging of data points based on their pairwise similarities.

The overall workflow of motif clustering is described in Figure 5(a). First, the pairwise DTW scores of motifs are calculated in both forward-forward and forward-reverse manners because the epigenomic patterns obtained by ChIP-seq do not distinguish between the forward and reverse strands. The lower DTW score representing higher similarities is adopted. Clustered motifs can be identified after determining the optimal number of clusters using a dendrogram. To visualize the clustering results, Figure 5(b) shows a UMAP projection of the motifs in a reduced-dimensional space [33]. This projection enables clearer assessment of how well the clusters are separated and reveals the overall structure of the motif space learned by ChromBERT. Figure 5(c) provides a cluster-wise summary that visualizes the composition and relative size of each cluster, allowing for intuitive interpretation of the distribution and prevalence of chromatin state motif patterns across samples. Together, these visualizations help convey both the diversity and consistency of motif types captured by ChromBERT. A line plot representation of all motif

248 sequences by cluster, which preserves the full chromatin state transitions, is included in
249 Supplementary Figure S3.
250 To estimate the optimal number of clusters, we integrated several evaluative measures,
251 including the dendrogram generated by the agglomerative clustering algorithm, to achieve a
252 judicious balance between detail and overarching pattern recognition. In Figure 5(c), for
253 instance, motifs with similar state transitions but different lengths, such as "BBBBGGG",
254 "BGGGGGG", "GGGBBB", were clustered into a single category, while "GGBAAAAAA,"
255 which has a different transition, was clearly segregated into a separate category. Such transitions
256 like "G-B-A" could represent enhancer regions immediately preceding active TSS regions,
257 potentially acting as a herald for transcription initiation. This suggests a dynamic regulatory
258 setup where enhancer activation might directly influence the onset of transcription at proximal
259 promoter sites. Additionally, sequences such as "OOAAA" and "OOOOOAAA" in other clusters
260 may reflect abrupt changes in chromatin states, possibly serving as precursors to promoter
261 regions. Such patterns suggest the potential complexity of chromatin regulation captured by
262 ChromBERT, and these motifs could provide interesting insights into the orchestrated events
263 leading up to gene expression.
264
265 **ChromBERT predicts gene expression levels from promoter chromatin state sequences**
266 To further expand ChromBERT's utility, we adapted the model for regression tasks and
267 attempted to predict gene expression levels quantitatively from promoter-proximal chromatin
268 state sequences. The model was fine-tuned using log-transformed RPKM values associated with
269 each gene.
270 We tested multiple combinations of upstream and downstream input lengths around TSSs
271 (Figure 6a). Consistent with the binary classification, performance improved when both
272 upstream and downstream regions were extended. The best configuration (100 kb upstream and
273 90 kb downstream) achieved a Pearson correlation of 0.791 between predicted and observed log-
274 transformed RPKM values (Figure 6(b)), indicating that ChromBERT can quantitatively predict
275 gene expression levels from chromatin state sequences. As illustrated in Supplementary Fig. S5,
276 the categorical nature of chromatin state data leads to cases where similar sequence patterns
277 exhibit different expression levels, which may constrain the maximum attainable correlation.

The gene expression density plot (Figure 6(c)) showed that most genes have log-transformed RPKM values below approximately 6, and ChromBERT could accurately estimate their expression levels within this range. In contrast, the model exhibited limitations for genes with log-transformed RPKM values exceeding 10, possibly due to the small number of such genes in the training data. Increasing the amount of training data for these genes would mitigate the imbalance and likely improve prediction accuracy.

ChromBERT identifies cell-type–specific chromatin state signatures

To further evaluate the effectiveness of ChromBERT, we examined whether chromatin state sequences contain recognizable signatures of different cell-type groups and whether the model can classify them. We focused on cis-regulatory module (CRM) regions, which are enriched for regulatory activity. CRM regions were obtained from the SCREEN database [34], and chromatin state sequences were extracted from 20-kb windows centered on each CRM site. Cell-type group annotations followed the ROADMAP classification ("ESC," "iPSC," "ES-derived," "blood-Tcell," "HSC-Bcell," "Brain," "Muscle," "Heart," and "Smooth Muscle"), and we fine-tuned ChromBERT for binary classification across all pairwise combinations (Figure 7a).

The resulting accuracy distribution closely reflected known biological relationships among cell groups. For example, ESCs and iPSCs are highly similar in their epigenomic profiles and accordingly were difficult to distinguish (0.655). In contrast, ESCs and T cells showed more distinct chromatin patterns and were classified with higher accuracy (0.776). These results suggest that each cell group carries characteristic epigenomic signatures. One exception was the Brain vs. Smooth Muscle comparison, which had low accuracy (0.529). One possible reason is a problem with the data quality. According to the ROADMAP QC summary, multiple histone-modification ChIP-seq samples for these two cell groups were flagged as low quality[14]. These findings are consistent with those from the TSS-based gene-expression classification task (see Supplementary Fig. S4).

We next sought to identify motifs contributing to cell-type specificity. Figure 7(b) shows motif clusters enriched in ESCs when compared with T-cells. Interestingly, many ESC-specific motifs contain the character "J," corresponding to the "bivalent/poised TSS" state, a hallmark of pluripotent stem cells [35], [36]. Motifs containing "J" appeared in in ESCs and iPSCs but not in

309 differentiated cell types. A complete list of identified motifs is provided in Supplementary Table
310 S1. These findings demonstrate that ChromBERT can reveal biologically meaningful, cell-type-
311 specific motifs.
312 However, ChromBERT also identified motifs that appeared cell-type-unspecific (e.g.,
313 AAAAA). This observation suggests that differences between cell types may arise not only from
314 biological distinctions but also from technical variability, such as differences in sample
315 preparation or sequencing depth among cell groups.
316
317 **ChromBERT captures chromatin state signatures of 3D genome compartments**
318 Lastly, we evaluated whether ChromBERT could capture features associated with three-
319 dimensional (3D) genome organization. Although the relationship between the epigenome and
320 3D genome structure is still not well understood, if ChromBERT can accurately predict certain
321 3D genome features, it would suggest a strong correspondence between epigenomic states and
322 higher-order genome organization. Using *in situ* Hi-C data from five cell types (GM12878,
323 HMEC, HUVEC, IMR90, and K562) at 25-kb resolution obtained from Rao *et al.* [37], we first
324 examined whether chromatin state sequences could distinguish A/B compartments (Figure 8(a)).
325 See the Methods section for details on the Hi-C analysis. Compartments were defined using the
326 first principal component (PC1) of the observed/expected contact matrix, where positive and
327 negative PC1 values correspond to compartment A (generally active) and compartment B
328 (generally inactive), respectively [38]. We additionally defined "strong compartment A" and
329 "strong compartment B" as genomic bins with extreme PC1 values. As expected from their
330 distinct chromatin compositions, ChromBERT accurately classified compartments A and B, with
331 even higher performance for the strong A/B subsets where epigenomic differences were more
332 pronounced.
333 We next attempted to classify topologically associating domain (TAD) boundaries versus
334 non-TAD boundary regions (Figure 8(b)). TAD boundaries were identified from Hi-C contact
335 maps following standard insulation-score–based procedures [39]. These regions are known to be
336 enriched for the architectural protein CTCF and for housekeeping genes [40], suggesting that
337 characteristic chromatin states might help distinguish them. However, unlike the compartment
338 task, classification accuracy for TAD boundary versus non-boundary regions was modest ($F1 <$
339 0.7), and extending the input window to 200 kb resulted in only limited improvement. Several

factors may contribute to this reduced performance. The number of TAD boundaries is much smaller than the number of compartment bins or gene-associated regions, and only five cell types were available, limiting the training data. Boundary calling also varies across computational tools, introducing technical variability. Moreover, both TAD boundary and non-boundary regions can contain mixtures of active and inactive chromatin states, reducing separability, and compartment information was not incorporated in this analysis.

Overall, these results indicate that large-scale features such as A/B compartments show strong correspondence with chromatin state patterns, whereas finer-scale structures like TAD boundaries are more difficult to predict from epigenomic sequences alone. Nevertheless, ChromBERT provides a useful framework for investigating how local chromatin states relate to 3D genome organization.

Discussion

ChromBERT is a Transformer-based architecture specifically designed for chromatin state sequences. Unlike DNA sequence models such as DNABERT, which operate on static nucleotide sequences, ChromBERT learns from epigenomic annotations that are cell-type–specific and inherently variable in length. While recent deep learning approaches such as Geneformer [22] and EPInformer [23] have modeled regulatory states or integrated multi-omics signals, ChromBERT provides a complementary perspective by learning sequential patterns directly from chromatin state annotation data. By leveraging self-attention mechanisms, ChromBERT learns recurrent chromatin state motifs directly from sequence-level data, capturing structural patterns that are not the focus of annotation summarization tools like CSREP [18], alignment-based methods such as EpiAlign [41], or the conservation-based approach like Epilogos and the conservation-associated activity score [20]. This sequence-level viewpoint offers a data-driven framework for studying local epigenomic organization and its relationship to transcriptional regulation.

Across multiple downstream tasks, including gene expression prediction, cell-type classification, and the prediction of selected 3D genome features, ChromBERT demonstrated that chromatin state sequences encode biologically meaningful patterns. For example, promoter-proximal chromatin states enabled accurate binary and quantitative gene expression prediction,

371 and cell-type–specific CRM regions yielded signatures consistent with known epigenomic
372 programs (e.g., enrichment of the bivalent/poised TSS state "J" in pluripotent cells [35], [36]).
373 Chromatin states also captured large-scale 3D genome organization: A/B compartments were
374 readily classified, with stronger compartment regions showing particularly high separability.
375 These results support the idea that chromatin state annotations retain regulatory information that
376 can be learned directly at the sequence level.
377 A notable strength of ChromBERT is its scalability with respect to pretraining data.
378 Because chromatin state sequences differ across cell types, the number of available training
379 samples naturally increases as more cell types are incorporated. We also pretrained ChromBERT
380 using chromatin annotation data from the IHEC project [16], which provides 1,699 epigenome
381 annotations based on an 18-state model. Once the IHEC database becomes publicly available, the
382 corresponding pretrained model will also be released. However, increased data volume also
383 magnifies technical variation. Public ChIP-seq datasets vary in quality, and including samples
384 with weak signal-to-noise ratios may introduce noise during pretraining. Similarly, many cell
385 types are represented by multiple donors or slightly modified conditions, potentially leading to
386 imbalanced datasets. Careful data curation and quality filtering will therefore be essential for
387 maximizing model performance.
388 From a computational perspective, chromatin state modeling introduces unique
389 challenges. DNABERT's nucleotide alphabet consists of only four bases, resulting in
390 manageable k-mer vocabularies (e.g., 4,101 (6-mers)). In contrast, ChromBERT must represent
391 15 chromatin states, and the resulting k-mer vocabularies expand sharply in absolute size: 3,379
392 (3-mers), 50,629 (4-mers), 759,379 (5-mers), and over 11 million (6-mers). This pronounced
393 difference in vocabulary size is computationally challenging and can lead GPU memory
394 limitations when attempting to pretrain with more than 5-mers, even with a single sample set.
395 This effectively limited us to using 4-mers for the pre-training step. The vocabulary complexity
396 becomes even more substantial when using the 18-state chromatin annotation system from the
397 IHEC dataset.
398 Additionally, the perplexity of the model (a metric that quantifies how well a model
399 predicts a sample) showed a decreasing trend from approximately 4 to 1 over iterations during
400 the pretraining for both promoter regions and broader genic regions. Interestingly, the perplexity
401 was already relatively low at the beginning of training (approx. 4). This suggests that the model

is successfully predicting the masked tokens in the pretraining, even in the early stages. This initially low perplexity can be attributed to several factors. First, the stable, long-lasting nature of chromatin states and their functional bias likely lead to more predictable sequence patterns, easing the model's predictions. Second, our use of 4-mer tokenization further simplified the task for the model, as it only needs to predict patterns within a more limited context compared to larger k-mers.

ChromBERT achieved strong predictive performance across multiple tasks. In gene expression regression, the model reached a Pearson's r of 0.79 and captured promoter-proximal regulatory patterns that reflect relative transcriptional activity. Extending the promoter window improved performance up to a point, indicating that distal regulatory features contribute additional information. Chromatin state sequences also facilitated the discovery of recurring regulatory motifs, as their categorical representation integrates multiple histone marks and enables long effective input windows.

Like other BERT-based architectures, ChromBERT is subject to inherent input-length constraints, limiting the maximum sequence to 510 tokens. However, the locally stable and consecutive nature of chromatin states allows this constraint to be partially mitigated through stride-based tokenization, which enables the model to incorporate a broader genomic context while maintaining meaningful patterns. In addition, while 4-mer tokenization during pretraining does not directly restrict the motif length, its contextual granularity may bias the model toward learning patterns on a similar scale. To detect longer motifs, it may be useful to modify the sequence length constraints or revise the tokenization process to better accommodate the unique properties of chromatin state annotation data. Such efforts could further improve ChromBERT's utility in decoding the complex, hierarchical nature of epigenomic regulation.

Unlike previous methods that use regression-based models to directly predict gene expression from raw histone-modification signals, chromatin state data does not use peak-intensity information. This could result in reduced performance for highly localized predictions. However, chromatin states offer a simplified sequence representation that integrates multiple histone marks, enabling much longer input sequences and allowing the model to focus on the length and combinatorial patterns of state transitions. This representation enables the detection of recurring regulatory patterns across samples and facilitates the discovery of sequence-level chromatin motifs that generalize across cell types. Furthermore, this framework is not limited to

433 gene expression prediction. It can also serve as a foundation for many downstream tasks, such as
434 cell type classification and 3D genome-related analyses.
435 Performance for TAD-boundary prediction was modest, likely reflecting both technical
436 and biological factors. TAD boundaries are relatively few per cell type, vary by caller, and can
437 contain mixtures of active and inactive chromatin states, reducing separability. Incorporating
438 additional modalities, such as CTCF binding [42] or more finely resolved regulatory annotations,
439 may improve the model's ability to identify finer-scale 3D architectural features as multi-omics
440 datasets expand.
441 Future work may explore model architectures with extended context, hierarchical
442 tokenization strategies, or dynamically learned segmentation to better capture long-range
443 epigenomic organization. Improved curation of input datasets, inclusion of additional cell types,
444 and integration of complementary epigenomic modalities may further enhance ChromBERT's
445 generalizability.
446 In summary, ChromBERT provides a scalable, interpretable framework for learning
447 sequence-level patterns within chromatin state annotations. By capturing recurrent motifs,
448 promoter activity signatures, and compartment-level structures, ChromBERT illustrates how
449 language-model approaches can decode the structured complexity of the epigenome and link
450 local chromatin organization to diverse functional genomic outcomes.
451
452
453 **Methods**
454 **1. Data collection**
455 The chromatin annotation BED files are downloaded from the ROADMAP [14] project,
456 which includes 127 distinct epigenomes annotated with 15 different chromatin states at 200-bp
457 intervals using ChromHMM [9] for genome build hg19. The 15 different chromatin states are 1.
458 Active Transcription Start Site (TSS), 2. Flanking Active TSS, 3. Transcription at Gene 5' to 3'
459 Ends, 4. Strong Transcription, 5. Weak Transcription, 6. Genic Enhancers, 7. Enhancers, 8. Zinc
460 Finger Protein (ZNF) Genes & Repeats, 9. Heterochromatin, 10. Bivalent/Poised TSS, 11.
461 Flanking Bivalent TSS/Enhancer, 12. Bivalent Enhancer, 13. Repressed Polycomb, 14. Weak
462 Repressed Polycomb, and 15. Quiescent and low signal. The chromatin annotation BED files

consist of four columns: chromosome number, start position, end position, and the number corresponding to the annotated state.

2. Data preprocessing

To prepare the data for natural language processing techniques, we first converted the chromatin-annotated data, originally represented by numbers 1 to 15, into corresponding alphabetic characters A to O. This transformation was necessary because natural language processing cannot be directly applied to sequences of numerical chromatin states. After changing the chromatin state annotations to alphabetic symbols, we condensed the genomic positions by a factor of 200 since the original annotations were conducted at 200-bp intervals. Consequently, we generated a continuous string of chromatin states, with each alphabet character (A to O) representing the corresponding chromatin state (1 to 15) for a 200-bp segment. Additionally, to focus our analysis on the relevant genomic regions and to exclude potential biases from telomeres, we removed the initial and final 10,000 bps from each chromosome, as these regions generally represent the extreme ends and may not accurately reflect the overall chromatin states. In this manner, all the chromatin annotation data for 127 epigenomes were processed chromosome-wise. Since we adapted BERT architecture, the processed chromatin annotation data were randomly cut to ensure that the length of the data did not exceed 510 but was longer than 5, following their data preparation method. Subsequently, the chromatin state sequences were tokenized into 3-mers, 4-mers, 5-mers, and 6-mers. Any sequences shorter than the length of each k-mer were not included in the dataset.

3. Pretraining

The BERT architecture utilized in ChromBERT was pre-trained on chromatin annotation data using datasets from 127 different cell types provided by the ROADMAP project for the 15 chromatin state case. The chromatin annotation sequences for the genic regions of these epigenomes were preprocessed as described in the data processing section, concatenated, and then tokenized using k-merization. Unlike DNABERT, we focused on testing 3-mers and 4-mers only, as 5-mers and 6-mers have a significantly larger vocabulary size (759,380 and 11,390,630, respectively), making them impractical to run. The vocabulary sizes for 3-mers and 4-mers were 3,379 and 50,630, respectively. Although both 3-mers and 4-mers demonstrated sufficiently low

perplexity, we opted for 4-mers because they comparably capture more complex patterns more effectively, enhancing our analysis. The transformer architecture used for pretraining remained consistent with DNABERT, featuring an intermediate layer size of 3,072 and a hidden layer size of 768. To ensure smooth training without encountering out-of-memory errors, we adjusted the number of train and evaluation batches per GPU to 5 and 3, respectively. For other training conditions, we followed the same strategy as DNABERT, including a linear increase in the learning rate from 0 to 4e-4 during the warm-up phase. The pretraining process was conducted on a GPU machine equipped with two Nvidia GeForce RTX 2080Ti cards, and it took approximately three days to complete.

For the IHEC dataset, chromatin state annotations consist of 18 states (A–R) defined at 200-bp resolution. These sequences were converted into 4-mers to construct the pretraining corpus, yielding approximately 41 GB of data after removing tokens composed entirely of the quiescent state R. Because the full dataset could not be loaded into memory at once, we adopted a sequential pretraining strategy. The data were first randomly shuffled to distribute cell-line–specific patterns uniformly, then split into chunks of 500,000 lines. Each chunk was used in succession, with model weights carried over between chunks, enabling efficient training while avoiding overfitting to contiguous regions from individual cell lines. Pretraining was performed using a DNA-based Transformer architecture with 4-mer tokenization and a vocabulary of 104,980 tokens, representing all valid 4-mers from the 18-state system (excluding the RRRR token) together with the standard special tokens ([UNK], [PAD], [CLS], [SEP], [MASK]).

During pretraining, we used a masked language modeling (MLM) objective, following the standard BERT pretraining strategy. Specifically, 15% of tokens were randomly masked, as in DNABERT, and the model was trained to predict these masked tokens using surrounding context. Perplexity was used as the primary evaluation metric to assess learning performance. While chromatin state sequences exhibit structured dependencies, the self-attention mechanism enables ChromBERT to effectively capture both local and long-range relationships. Gradient accumulation and a carefully chosen batch size were used to efficiently manage GPU memory and allow training on large datasets. Model weights generated after each chunk were saved and used as initialization for subsequent chunks, progressively refining the model across the dataset. The training was conducted on a system equipped with an AMD EPYC 7513 32-Core Processor, 251 GiB of RAM, and two NVIDIA RTX 6000 Ada GPUs, which were instrumental in handling

525 the computational demands of this extensive dataset. Perplexity dropped sharply from
526 approximately 4.99 to 1.19 within the first chunk (completed in two days), indicating efficient
527 early-stage learning (Supplementary Fig. S6). After this initial phase, perplexity stabilized
528 around 1.0 for the remaining chunks, suggesting that the model had effectively captured the
529 underlying chromatin state structure early in the training process.
530
531 **4. Fine-tuning for gene expression level classification**
532 We conducted fine-tuning for each downstream application based on the parameters
533 acquired from the pretraining process. ChromBERT follows the BERT-base architecture,
534 consisting of 12 transformer layers, 12 attention heads, and a hidden size of 384. The
535 intermediate layer size was 1536, and we applied the GELU activation function. The model uses
536 a vocabulary size of 50,630, corresponding to all possible 4-mers derived from chromatin state
537 sequences. The maximum sequence length was set to 512, with dropout probabilities of 0.1 for
538 attention and hidden layers.
539 We followed the fine-tuning conditions used in DNABERT, applying AdamW as the
540 optimization method with a weight decay of 0.01. Fine-tuning was conducted using a batch size
541 of 32, a learning rate of 2e-4, and a warm-up proportion of 10% of the total training steps. Each
542 fine-tuning task was trained for 10 epochs, with evaluation performed every 100 steps. Model
543 checkpoints were saved every 4000 steps, and gradient accumulation was used to efficiently train
544 on large-scale datasets.
545 For pairwise binary classification across different gene expression levels, we utilized
546 RNA-seq data from 57 different cell lines provided by the ROADMAP project. A genic region
547 was classified as 'highly expressed' if its log-transformed RPKM (Reads Per Kilobase per
548 Million mapped reads) was greater than 5, while genic regions with an RPKM value of 0 were
549 considered 'non-expressed'. To compare genes with different expression levels, we established
550 five log-transformed RPKM thresholds: 1, 2, 3, 4, and 5. These thresholds were determined
551 based on a preliminary distribution analysis, and they resulted in more balanced gene counts
552 across classes.
553 For promoter-based classification tasks, ChromBERT was fine-tuned using input
554 sequences of various lengths upstream and downstream of the TSSs. Since BERT distinguishes
555 upstream from downstream context, chromatin state sequences corresponding to genes on the

556 reverse strand were reversed. During the fine-tuning step, we tried tokenization by sliding *N*
557 characters at once (stride *N*), in addition to normal tokenization by sliding one character (stride
558 1), to accommodate longer sequences (Supplementary Fig S1). Here we tested strides 2 and 3,
559 allowing us to include up to 290 kbp around the TSS (190 kbp upstream and 100 kbp
560 downstream). For each binary classification task, 30,000 and 6,000 genes were randomly
561 selected for the training and test sets, respectively. The model was trained using a maximum
562 sequence length of 512, batch size of 32, learning rate of 2e-4, and training schedule of 10
563 epochs. The performance was evaluated using F1 score, mean area under the curve (AUC), and
564 accuracy.
565
566 **5. Implementation of a Regression Model for Gene Expression Prediction**
567 The ChromBERT framework was adapted for regression by modifying its processing
568 pipeline to handle continuous RPKM values. Promoter region sequences were paired with their
569 log-transformed RPKM values. A GeneExpressionProcessor, modeled after StsbProcessor, was
570 developed for reading, processing, and tokenizing chromatin state sequences. Since
571 StsbProcessor is designed for sentence-pair regression and requires paired inputs, a custom
572 processor was implemented to streamline data handling, simplify preprocessing, and align with
573 our regression task. This processor was integrated into the task dictionary under the task name
574 "gene_expression", enabling quantitative gene expression prediction rather than categorical
575 classification. The fine-tuning script was updated accordingly to support regression.
576 Unlike classification tasks with balanced RPKM thresholds, regression required handling
577 skewed RPKM distributions. The dataset was dominated by low or zero RPKM values, which
578 could bias loss optimization (e.g., Mean Squared Error) and degrade prediction accuracy for
579 high-expression cases. Sequences composed entirely of the “O” state correlated strongly with
580 non-expression, contributing to a higher fraction of zero RPKM values (12.92%) compared to
581 datasets excluding them (8.31%) (Supplementary Fig. S7). To reduce redundancy and bias, these
582 sequences were excluded from the primary dataset.
583 For gene expression prediction, we fine-tuned a regression model under three TSS-
584 anchored genomic window conditions: 90 kbp upstream and 5 kbp downstream of the TSSs with
585 a stride of 1,190 kb upstream and 5 kb downstream with a stride of 2, and 100 kb upstream and
586 90 kb downstream with a stride of 2. The model was trained with a maximum sequence length of

512, using a batch size of 32, a learning rate of 2e-5, and a training schedule of 10 epochs. The performance was assessed using Pearson correlation coefficients to measure the consistency between predicted and actual gene expression levels.

6. Fine-tuning for cell-type classification

For cell-type classification, we used the cell group annotations defined by the ROADMAP, including "ESC," "iPSC," "ES-derived," "blood-Tcell," "HSC-Bcell," "Brain," "Muscle," "Heart," and "Smooth Muscle." We excluded cell groups with small sample sizes and those with coarse annotations. We used genomic regions of cis-regulatory modules (CRMs) for classification because they are likely to include most functional epigenomic regions. We obtained the reference CRM region lists for genome build hg38 from the SCREEN database 3/17/26 1:55:00 PM and converted them to genomic coordinates for build hg19 using the UCSC LiftOver tool. We removed the CRMs in chromosomes X and Y. The adjacent regions within 200 bp were merged.

Then we extracted the chromatin state sequences from 20-kbp regions around the center of each CRM site for each sample. We merged CRM sequences from all samples for each group. For fine-tuning, we randomly selected 20,000 and 4,000 sequences for training and testing, respectively. Both training and testing data contained an equal number of positive and negative samples corresponding to two cell types. To prevent data leakage due to overlapping genomic regions, training and evaluation datasets were separated by chromosome: chromosomes 1–19 for training and chromosomes 20–22 for evaluation. The model was trained using a batch size of 16, a learning rate of 2e-5, a training schedule of three epochs, and a warm-up proportion of 10% of the total training steps. Model checkpoints were saved every 5,000 steps, and gradient accumulation was applied to enable efficient training on large-scale datasets.

For the cell-type classification using TSSs, we used 20 kbp upstream and 5 kbp downstream of the TSSs with a stride of 1. We used 17,756 expressed genes after excluding non-expressed genes with an RPKM value of 0. To prevent data leakage, we excluded the overlapping regions from the dataset. We then separated the data into chromosomes 1–17 for training and chromosomes 18–22 for evaluation. 30,000 and 5,000 genes were randomly selected for the training and test sets, respectively. The model was trained using a maximum sequence length of 512, the batch size 32, learning rate 2e-5, and training schedule 10 epochs.

618

619 **7. Fine-tuning for 3D genome classification**

620 We used *in situ* Hi-C data of the cell lines GM12878, HMEC, HUVEC, IMR90, and
621 K562 obtained from Rao *et al.* [37]. We downloaded the .hic files for the genome build hg19
622 with MAPQ ≥ 30 from the Gene Expression Omnibus (GEO) database under accession number
623 GSE63525. The .hic were processed using CustardPy [43] to generate genome-wide Hi-C
624 contact matrices with 25-kbp resolution and the square-root vanilla coverage (VC-SQRT)
625 normalization.

626 The compartments A and B were identified by performing a principal component analysis
627 (PCA) on the normalized contact matrices, and genomic coordinates were assigned according to
628 the sign of the first eigenvector (PC1). To capture the strength of compartmentalization, the
629 compartments were further classified into four categories based on their corresponding PC1
630 values: Strong A (PC1 > 75th percentile), Weak A (median < PC1 ≤ 75th percentile), Weak B
631 (25th percentile < PC1 ≤ median), and Strong B (PC1 ≤ 25th percentile). These categories were
632 then segmented into 25-kb genomic bins and used as input data for fine-tuning.

633 TADs were identified using the Arrowhead function of Juicer [44] at a 25-kb resolution.
634 TAD boundaries were defined as regions located within 25 kb of either end of TADs that
635 exhibited an insulation score [45] of ≤ 0.7, which is a quantitative metric of the boundary
636 strength. Non-TAD boundaries were defined as regions located within TADs that were
637 positioned at least 100 kb away from any TAD boundary and exhibiting an insulation score of >
638 0.7. The identified TAD boundaries and non-TAD boundaries were merged across five cell types
639 and used for fine-tuning.

640 For compartment classification, we used the 25 kbp regions with a stride of 1. We used
641 chromosomes 3–22 for training and chromosomes 1 and 2 for evaluation because the later
642 chromosomes show an imbalanced proportion of regions assigned to compartments A and B
643 [37]. We randomly selected 50,000 and 5,000 regions for the training and test sets, respectively.
644 The model was trained using a maximum sequence length of 512, batch size of 32, learning rate
645 of 2e-5, and training schedule of 10 epochs. For TAD boundary classification, we used the 50
646 kbp regions with a stride of 1 and the 200 kbp regions with a stride of 2 centered around the
647 regions. We used chromosomes 3–22 for training and chromosomes 1 and 2 for evaluation. We

648 randomly selected 10,000 and 2,000 regions for the training and test sets, respectively. The
649 parameter set is the same as that used for the compartment analysis.
650
651 **8. Motif finding**
652 Similarly to the definition of DNA motifs in DNABERT, chromatin state motifs in our
653 study are defined as frequently appearing patterns of consecutive chromatin states that are
654 distinctively present in a certain category during binary classification. We followed the
655 parameter conditions used in DNABERT to detect and filter these motifs. Firstly, we captured
656 regions with high attention scores from the output attention matrix. A threshold was defined as a
657 score higher than the mean attention scores of the matrix and higher than 10 times the minimum
658 score to determine regions where motifs are likely included. These regions were registered as
659 candidates, which were further filtered using a false discovery rate (FDR) threshold of 0.05,
660 based on a hypergeometric test followed by a Benjamini–Hochberg correction. This FDR
661 determined whether the motifs were significantly enriched in one class compared to the other.
662 Finally, motifs were identified as regions that are longer than four chromatin state sequences
663 (equivalent to 800 bps) and appear at least three times exclusively in the target class. While in
664 the case of DNA motifs, final motifs are typically compared with existing DNA motif libraries,
665 we collected chromatin state motifs without such a comparison because there is currently no
666 known library available for this task.
667
668 **9. Motif clustering**
669 To further estimate the representative motif chromatin sequence pattern, we utilized
670 DTW and agglomerative clustering to analyze the chromatin state motifs. Our initial dataset
671 included motifs identified after careful screening for p-values, minimum length, and minimum
672 occurrences within predefined regions of interest, avoiding the merging step to allow a detailed
673 examination of chromatin sequences, which are represented alphabetically. Given the potential
674 for variability in these sequences due to factors such as data quality (e.g., the signal-to-noise ratio
675 of the antibodies) or fluctuations in histone modification peaks, we opted for a more flexible
676 approach to motif identification. DTW was selected for its capability to adjust for sequence
677 length variations, accommodating the intrinsic dynamics of chromatin state sequences.

First, to implement DTW, we converted the chromatin state alphabets (“A” to “O”) into numerical values (1 to 15) to facilitate a standardized analysis. We provide two options for handling this numerical translation of chromatin states by offering a ‘categorical’ parameter, which can be set to ‘True’ or ‘False’. If users set it to ‘True’, the numbers are considered categorical, meaning the states A to B and A to O are equally distant. On the other hand, if it is set to ‘False’, which is the default option, the numbers representing the states are considered numerically; thus, the distance from state A to C is twice that of the distance from state A to B. The result of motif clustering when the ‘categorical’ parameter is set to ‘True’ is as shown in Supplementary Figure S8. Next, Shorter sequences were padded with ‘NaN’ (Not a Number) to match the maximum length dictated by the window size, ensuring consistency across all entries. For DTW calculation, the NaN values in shorter sequences are, by default, converted to the number representing the nearest state according to the 'ffill' filling method. Since chromatin state annotations are typically strand-agnostic and assigned without reference to gene orientation, we considered both forward–forward and forward–reverse comparisons to account for potential reverse-orientation patterns in chromatin state motifs. Consequently, the DTW matrix was composed of the minimum score between these two comparisons. Using the ‘tslearn’ Python package, particularly the ‘dtw’ module, we calculated similarity scores between sequences by focusing on the minimum cumulative distances in both directions, allowing for a detailed comparison of motifs.

Following this similarity assessment, we employed agglomerative clustering via the ‘AgglomerativeClustering’ module from the same package, grouping the motifs into clusters based on their similarities. Additionally, we provide a function to generate a dendrogram, which displays the hierarchy and helps users determine the optimal number of clusters when the number of clusters is not predetermined. While the dendrogram was our primary tool for selecting the optimal number of clusters, users are encouraged to consider alternative methods for determining the cluster count. For visualization purposes, we provide two functions: the first displays the UMAP, which reduces the dimensionality of the data for easier interpretation and visualization of the clustering results; the second visualizes the final result of clustering with actual motif entries. This approach facilitates the identification of representative sequential patterns among chromatin states, thereby offering deeper insights into genomic regulatory landscapes.

Data availability

The pretraining weights of ChromBERT have been submitted to Zenodo (https://doi.org/10.5281/zenodo.15518584).

Code availability

The source code of ChromBERT is available at https://github.com/caocao0525/ChromBERT.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

We thank all the members of the Nakato and Lin laboratories for their helpful discussions. We also thank the IHEC project for allowing us to use reprocessed and harmonized epigenomic data from a large collection of human cell types in this study.

Authors' contributions

S.L. developed ChromBERT and performed all analyses in this study. S.L. and J.S. performed a quantitative prediction of gene expression. G.M.O. and R.N. performed the cell-type classification. S.L., Y.N., and R.N. performed the 3D genome classification. S.L. and R.N. performed the binary classification of gene expression. R.N. conceived and designed the study. S.L. and R.N. contributed to the drafting of the manuscript. C.L. and C.Y.C. supervised the deep learning process and provided suggestions for improving the analysis and the manuscript.

Funding

This work was supported by a Grant-in-Aid for Scientific Research under grant number 23H02466, Grant-in-Aid for Early-Career Scientists under grant number 25K18447, the Japan Agency for Medical Research and Development under grant number JP23gm6310012h0004, and the JST FOREST Program under grant number JPMJFR224Y.

Competing interests

The authors declare no competing interests.

Reference

[1] The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.

[2] S. S. P. Rao *et al.*, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014, doi: 10.1016/j.cell.2014.11.021.

[3] C. D. Allis and T. Jenuwein, "The molecular hallmarks of epigenetic control," *Nat Rev Genet*, vol. 17, no. 8, pp. 487–500, Aug. 2016, doi: 10.1038/nrg.2016.59.

[4] C. L. Peterson and M.-A. Laniel, "Histones and histone modifications," *Current Biology*, vol. 14, no. 14, pp. R546–R551, Jul. 2004, doi: 10.1016/j.cub.2004.07.007.

[5] L. Siggens and K. Ekwall, "Epigenetics, chromatin and genome organization: recent advances from the ENCODE project," *J Intern Med*, vol. 276, no. 3, pp. 201–214, Sep. 2014, doi: 10.1111/joim.12231.

[6] J. Ernst *et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, no. 7345, pp. 43–49, May 2011, doi: 10.1038/nature09906.

[7] N. D. Heintzman *et al.*, "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome," *Nat Genet*, vol. 39, no. 3, pp. 311–318, Mar. 2007, doi: 10.1038/ng1966.

[8] M. P. Creyghton *et al.*, "Histone H3K27ac separates active from poised enhancers and predicts developmental state," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, no. 50, pp. 21931–21936, Dec. 2010, doi: 10.1073/pnas.1016071107.

[9] J. Ernst and M. Kellis, "Chromatin state discovery and genome annotation with ChromHMM," *Nat Protoc*, vol. 12, no. 12, pp. 2478–2492, Dec. 2017, doi: 10.1038/nprot.2017.124.

[10] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation," *Nat Methods*, vol. 9, no. 5, pp. 473–476, May 2012, doi: 10.1038/nmeth.1937.

[11] E. Meaburn and R. Schulz, "Next generation sequencing in epigenetics: Insights and challenges," *Seminars in Cell & Developmental Biology*, vol. 23, no. 2, pp. 192–199, Apr. 2012, doi: 10.1016/j.semcdb.2011.10.010.

[12] I. Arora and T. O. Tollefsbol, "Computational methods and next-generation sequencing approaches to analyze epigenetics data: Profiling of methods and applications," *Methods*, vol. 187, pp. 92–103, Mar. 2021, doi: 10.1016/j.ymeth.2020.09.008.

[13] R. Nakato and T. Sakata, "Methods for ChIP-seq analysis: A practical workflow and advanced applications," *Methods*, vol. 187, pp. 44–53, Mar. 2021, doi: 10.1016/j.ymeth.2020.03.005.

[14] Roadmap Epigenomics Consortium *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, Feb. 2015, doi: 10.1038/nature14248.

[15] The ENCODE Project Consortium *et al.*, "Expanded encyclopaedias of DNA elements in the human and mouse genomes," *Nature*, vol. 583, no. 7818, pp. 699–710, Jul. 2020, doi: 10.1038/s41586-020-2493-4.

[16] H. G. Stunnenberg *et al.*, "The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery," *Cell*, vol. 167, no. 5, pp. 1145–1149, Nov. 2016, doi: 10.1016/j.cell.2016.11.007.

[17] A. Yen and M. Kellis, "Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type," *Nat Commun*, vol. 6, no. 1, p. 7973, Aug. 2015, doi: 10.1038/ncomms8973.

[18] H. Vu, Z. Koch, P. Fiziev, and J. Ernst, "A framework for group-wise summarization and comparison of chromatin state annotations," *Bioinformatics*, vol. 39, no. 1, p. btac722, Jan. 2023, doi: 10.1093/bioinformatics/btac722.

[19] Meuleman, "Epilogos: Information-theoretic navigation of multi-tissue functional genomic annotations," 2019, [Online]. Available: https://epilogos.altius.org

[20] M. W. Libbrecht, O. L. Rodriguez, Z. Weng, J. A. Bilmes, M. M. Hoffman, and W. S. Noble, "A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types," *Genome Biol*, vol. 20, no. 1, p. 180, Dec. 2019, doi: 10.1186/s13059-019-1784-2.

[21] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, Aug. 2021, doi: 10.1093/bioinformatics/btab083.

[22] C. V. Theodoris *et al.*, "Transfer learning enables predictions in network biology," *Nature*, vol. 618, no. 7965, pp. 616–624, Jun. 2023, doi: 10.1038/s41586-023-06139-9.

[23] J. Lin, R. Luo, and L. Pinello, "EPInformer: A Scalable Deep Learning Framework for Gene Expression Prediction by Integrating Promoter-enhancer Sequences with Multimodal Epigenomic Data," Aug. 01, 2024, *Genomics*. doi: 10.1101/2024.08.01.606099.

[24] P. D'haeseleer, "What are DNA sequence motifs?," *Nat Biotechnol*, vol. 24, no. 4, pp. 423–425, Apr. 2006, doi: 10.1038/nbt0406-423.

[25] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," *Bioinformatics*, vol. 27, no. 12, pp. 1696–1697, Jun. 2011, doi: 10.1093/bioinformatics/btr189.

[26] M. Menzel, S. Hurka, S. Glasenhardt, and A. Gogol-Döring, "NoPeak: k-mer-based motif discovery in ChIP-Seq data without peak calling," *Bioinformatics*, vol. 37, no. 5, pp. 596–602, May 2021, doi: 10.1093/bioinformatics/btaa845.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. Accessed: Mar. 25, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805

[28] D. Lee, J. Yang, and S. Kim, "Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer," *Nat Commun*, vol. 13, no. 1, Art. no. 1, Nov. 2022, doi: 10.1038/s41467-022-34152-5.

[29] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978, doi: 10.1109/TASSP.1978.1163055.

[30] G. A. Maston, S. K. Evans, and M. R. Green, “Transcriptional Regulatory Elements in the Human Genome,” *Annu. Rev. Genom. Hum. Genet.*, vol. 7, no. 1, pp. 29–59, Sep. 2006, doi: 10.1146/annurev.genom.7.080505.115623.

[31] J. Aach and G. M. Church, “Aligning gene expression time series with time warping algorithms,” *Bioinformatics*, vol. 17, no. 6, pp. 495–508, Jun. 2001, doi: 10.1093/bioinformatics/17.6.495.

[32] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14, pp. 281–297, 1967.

[33] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Sep. 17, 2020, *arXiv*: arXiv:1802.03426. Accessed: May 07, 2024. [Online]. Available: http://arxiv.org/abs/1802.03426

[34] The ENCODE Project Consortium *et al.*, “Expanded encyclopaedias of DNA elements in the human and mouse genomes,” *Nature*, vol. 583, no. 7818, pp. 699–710, Jul. 2020, doi: 10.1038/s41586-020-2493-4.

[35] B. E. Bernstein *et al.*, “A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells,” *Cell*, vol. 125, no. 2, pp. 315–326, Apr. 2006, doi: 10.1016/j.cell.2006.02.041.

[36] M. Sachs, C. Onodera, K. Blaschke, K. T. Ebata, J. S. Song, and M. Ramalho-Santos, “Bivalent Chromatin Marks Developmental Regulatory Genes in the Mouse Embryonic Germline In Vivo,” *Cell Reports*, vol. 3, no. 6, pp. 1777–1784, Jun. 2013, doi: 10.1016/j.celrep.2013.04.032.

[37] S. S. P. Rao *et al.*, “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping,” *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014, doi: 10.1016/j.cell.2014.11.021.

[38] E. Lieberman-Aiden *et al.*, “Comprehensive mapping of long-range interactions reveals folding principles of the human genome,” *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, doi: 10.1126/science.1181369.

[39] J. A. Da Costa-Nunes and D. Noordermeer, “TADs: Dynamic structures to create stable regulatory functions,” *Current Opinion in Structural Biology*, vol. 81, p. 102622, Aug. 2023, doi: 10.1016/j.sbi.2023.102622.

[40] J. R. Dixon, D. U. Gorkin, and B. Ren, “Chromatin Domains: The Unit of Chromosome Organization,” *Molecular Cell*, vol. 62, no. 5, pp. 668–680, Jun. 2016, doi: 10.1016/j.molcel.2016.05.018.

[41] X. Ge, H. Zhang, L. Xie, W. V. Li, S. B. Kwon, and J. J. Li, “EpiAlign: an alignment-based bioinformatic tool for comparing chromatin state sequences,” *Nucleic Acids Research*, vol. 47, no. 13, pp. e77–e77, Jul. 2019, doi: 10.1093/nar/gkz287.

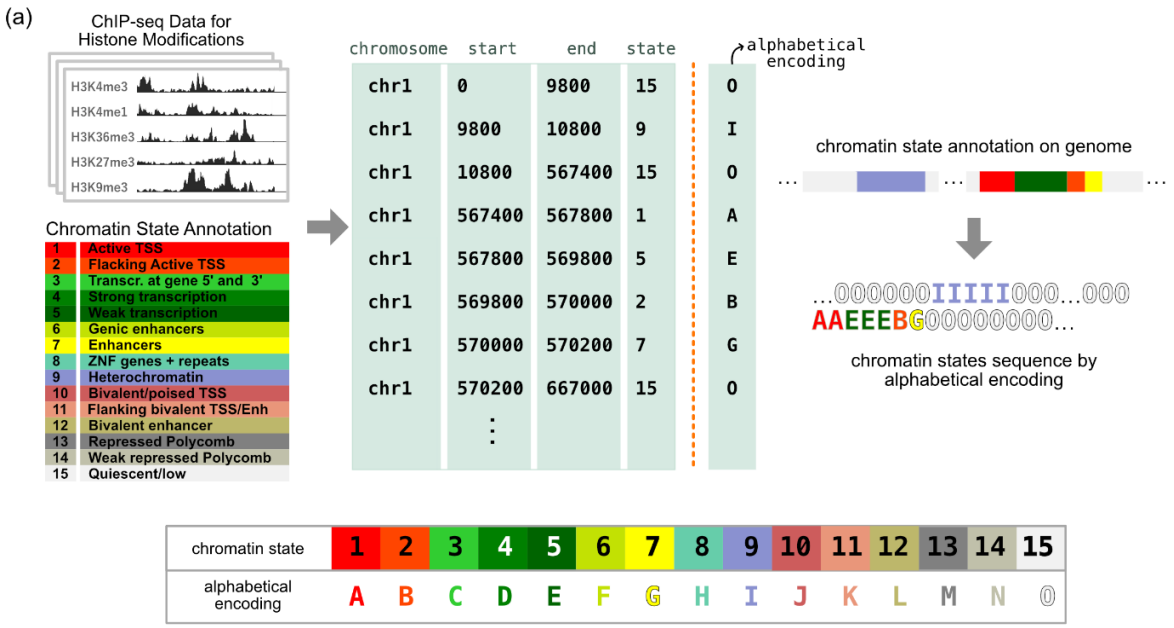
[42] J. Wang and R. Nakato, “HiC1Dmetrics: framework to extract various one-dimensional features from chromosome structure data,” *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab509, Jan. 2022, doi: 10.1093/bib/bbab509.

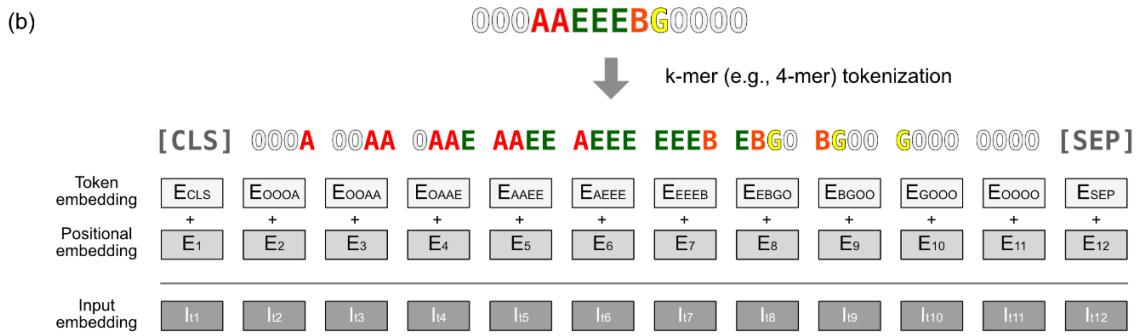
[43] R. Nakato *et al.*, “Context-dependent perturbations in chromatin folding and the transcriptome by cohesin and related factors,” *Nat Commun*, vol. 14, no. 1, p. 5647, Sep. 2023, doi: 10.1038/s41467-023-41316-4.

[44] N. C. Durand *et al.*, “Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments,” *Cell Systems*, vol. 3, no. 1, pp. 95–98, Jul. 2016, doi: 10.1016/j.cels.2016.07.002.

[45] E. Crane *et al.*, “Condensin-driven remodelling of X chromosome topology during dosage compensation,” *Nature*, vol. 523, no. 7559, pp. 240–244, Jul. 2015, doi: 10.1038/nature14450.

FIGURES


(a)
ChIP-seq Data for Histone Modifications
H3K4me3
H3K4me1
H3K36me3
H3K27me3
H3K9me3
Chromatin State Annotation
1 Active TSS
2 Flacking Active TSS
3 Transcr. at gene 5' and 3'
4 Strong transcription
5 Weak transcription
6 Genic enhancers
7 Enhancers
8 ZNF genes + repeats
9 Heterochromatin
10 Bivalent/poised TSS
11 Flanking bivalent TSS/Enh
12 Bivalent enhancer
13 Repressed Polycomb
14 Weak repressed Polycomb
15 Quiescent/low
chromosome start end state
chr1 0 9800 15
chr1 9800 10800 9
chr1 10800 567400 15
chr1 567400 567800 1
chr1 567800 569800 5
chr1 569800 570000 2
chr1 570000 570200 7
chr1 570200 667000 15
alphabetical encoding
O
I
O
A
E
B
G
O
chromatin state annotation on genome
...OOOOOOIIIIIOOO...OOO
AAEEEBGOOOOOOOOO...
chromatin states sequence by alphabetical encoding
chromatin state 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
alphabetical encoding A B C D E F G H I J K L M N O


(b)
OOOAAEEEBGOOOO
k-mer (e.g., 4-mer) tokenization
[CLS] OOOA OOAA OAAE AAEE AEEE EEEB EBGO BGOO GOOO OOOO [SEP]
Token embedding
Positional embedding
Input embedding

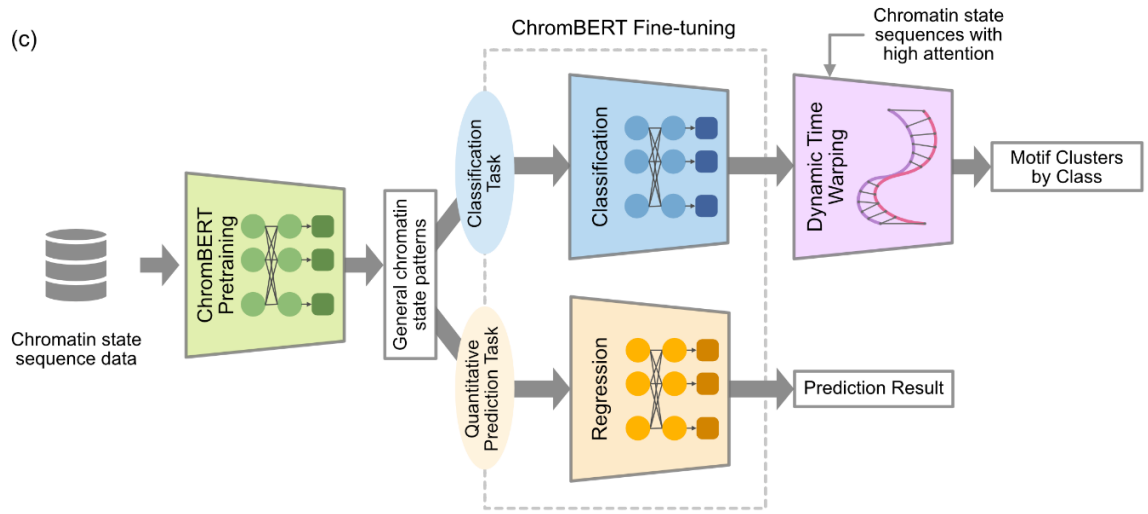

(c)
ChromBERT Fine-tuning
Chromatin state sequences with high attention
Chromatin state sequence data
ChromBERT Pretraining
General chromatin state patterns
Classification Task
Classification
Dynamic Time Warping
Motif Clusters by Class
Quantitative Prediction Task
Regression
Prediction Result

Figure 1. The ChromBERT workflow. (a) Conversion of numerical chromatin state annotations (from a BED file) into alphabetical encoding, followed by concatenation for generating a sequence of chromatin states (upper). The alphabetical encoding for each chromatin state (lower). (b) Creation of input embeddings for the BERT model through the combination of token and positional embeddings. (c) The comprehensive process of ChromBERT, from pretraining on general data to fine-tuning for task-specific applications.

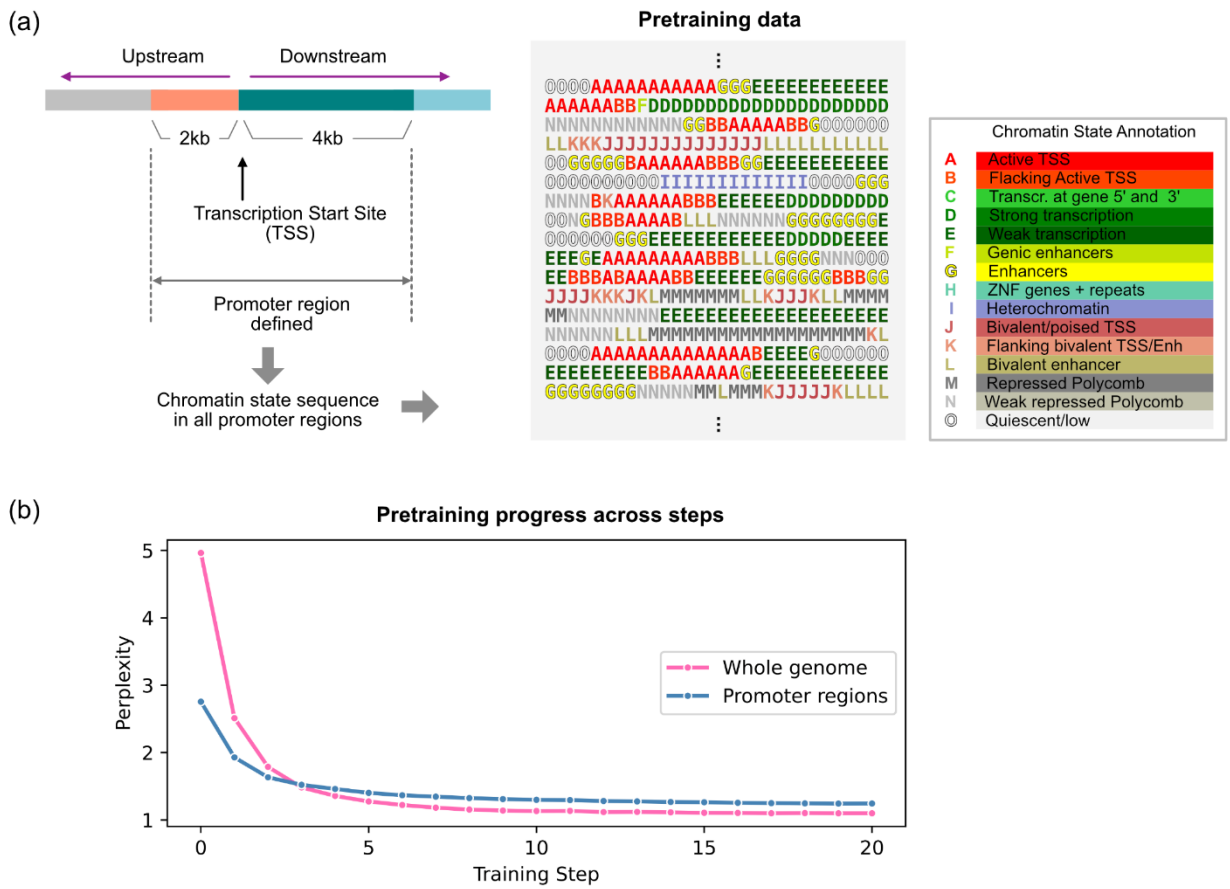


Figure 2. Overview of ChromBERT pretraining for promoter regions. (a) Definition of promoter regions used for pretraining. Chromatin state sequences were extracted from regions spanning 2 kb upstream to 4 kb downstream of TSSs across 127 cell types. The right panel shows an example of chromatin state sequences and a full legend of the 15-state annotation used in this study. The chromatin state annotation panel is from ROADMAP project [14] (b) Pretraining performance on chromatin state sequences from whole-genome and promoter-specific regions, measured by perplexity over training steps.

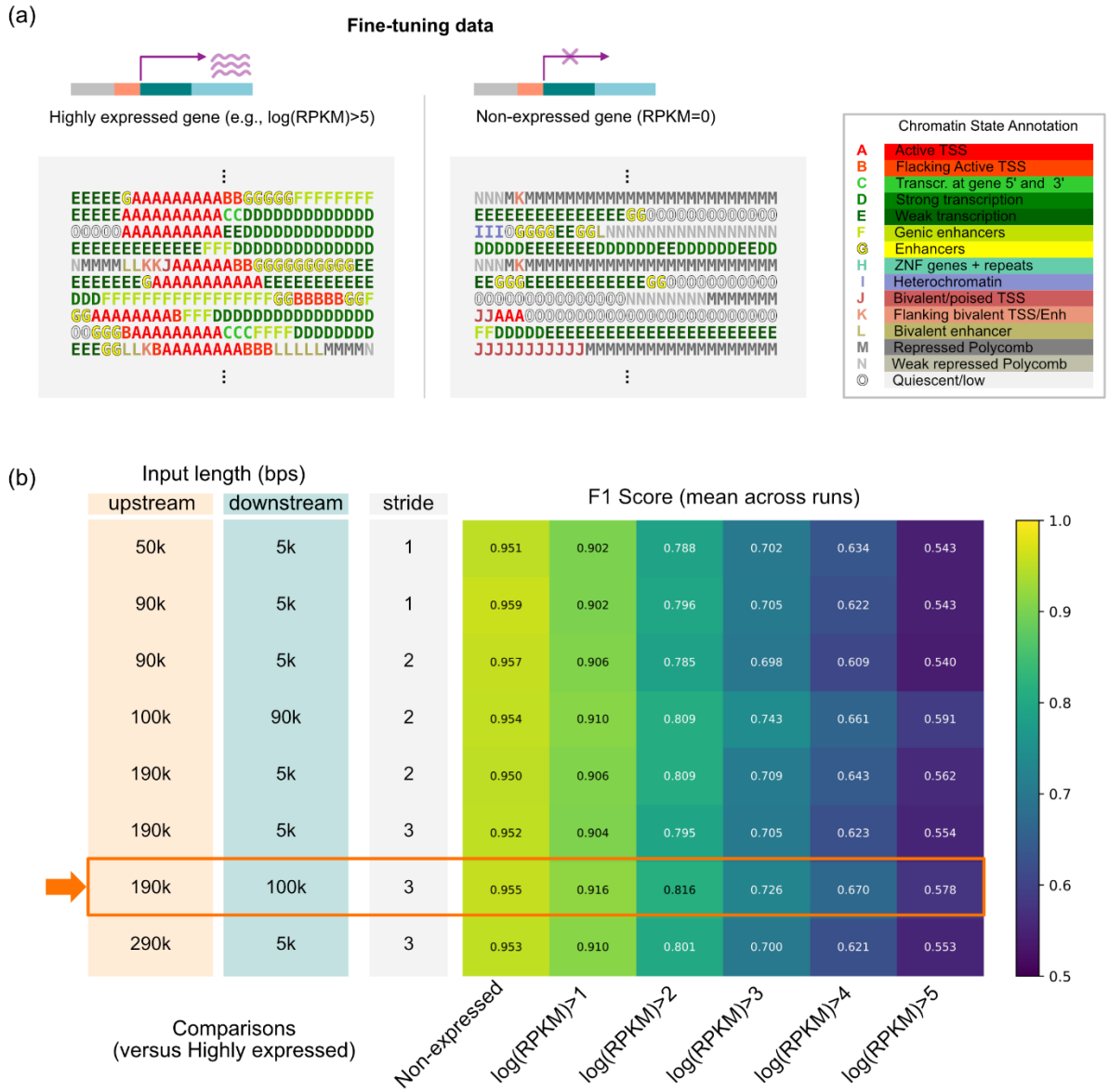


Figure 3. Fine-tuning ChromBERT on promoter regions for gene expression classification. (a) Example chromatin state sequences associated with highly expressed genes (e.g., log (RPKM) > 5, left) and non-expressed genes (RPKM = 0, right) used in fine-tuning. Color labels follow the 15-state annotation in ROADMAP (legend consistent with Figure 2). (b) Summary of the input-region configurations evaluated during fine-tuning. Upstream and downstream windows of varying lengths (50–290 kb upstream; 5–100 kb downstream) were tested using stride values of 1–3. The heatmap displays the mean F1 scores across runs for classification tasks comparing

960 highly expressed genes with gene groups defined by different log-transformed RPKM thresholds.
961 The configuration highlighted with an orange arrow corresponds to the up190k/down100k
962 window. AUC and accuracy matrices for the same configurations are provided in Supplementary
963 Fig. S2.
964

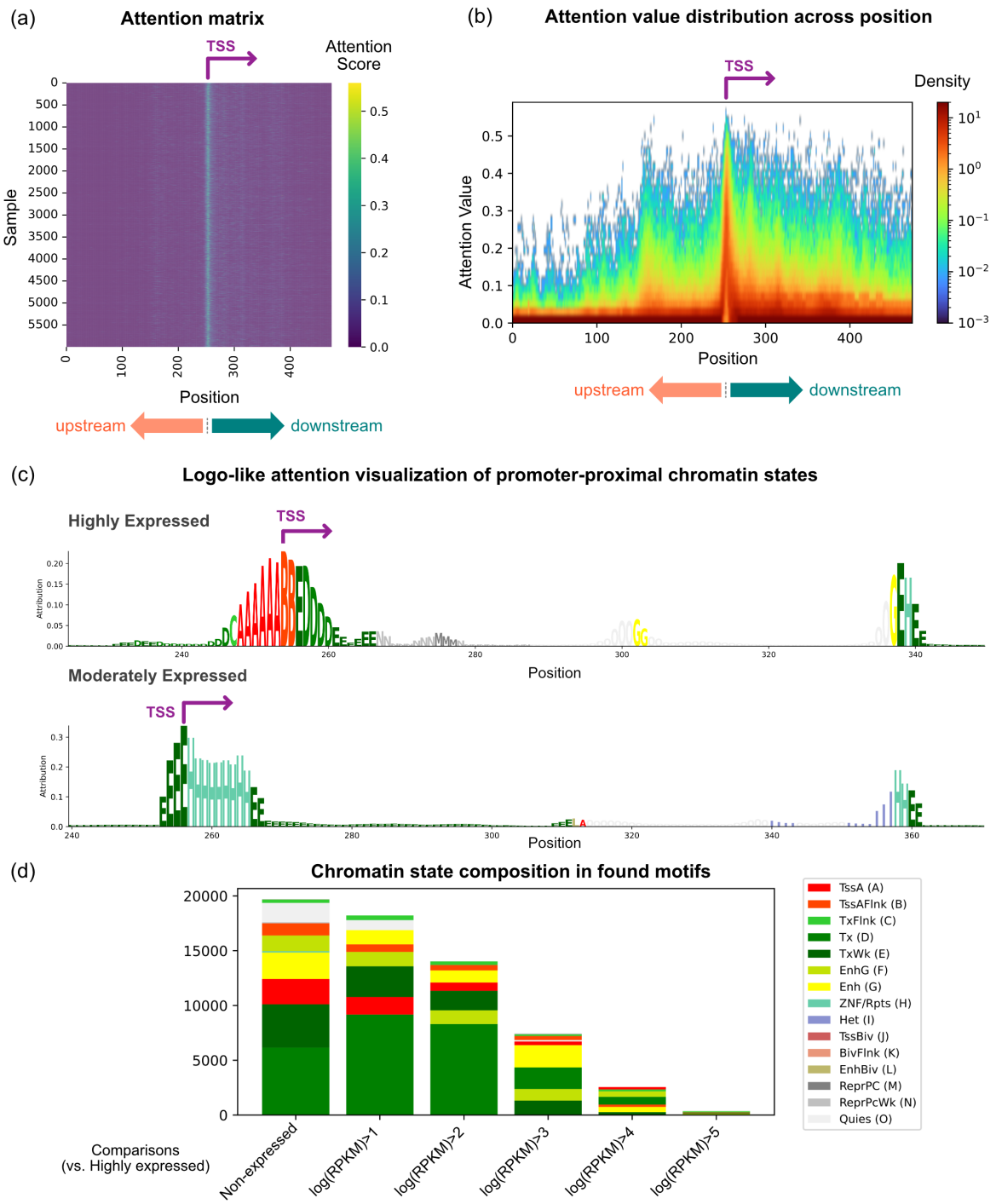
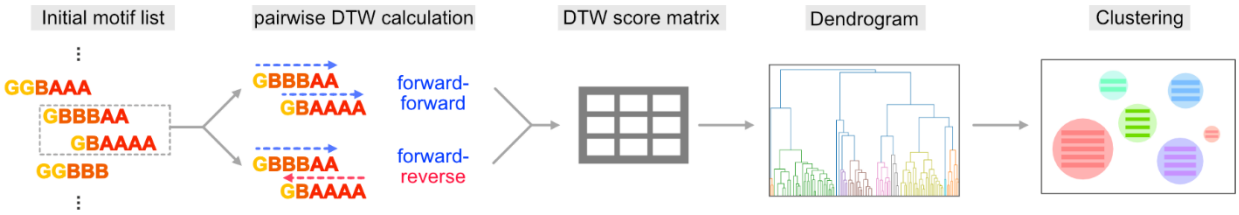


Figure 4. Attention-based interpretation of promoter-proximal chromatin states. (a) Attention matrix for representative samples from the fine-tuned model using a 100 kb upstream / 90 kb downstream input window around the TSS. Color values indicate normalized attention weights. (b) Positional density of attention values across upstream and downstream regions, aggregated over the evaluation set. (c) Logo-like visualization of promoter-proximal chromatin states. Bar

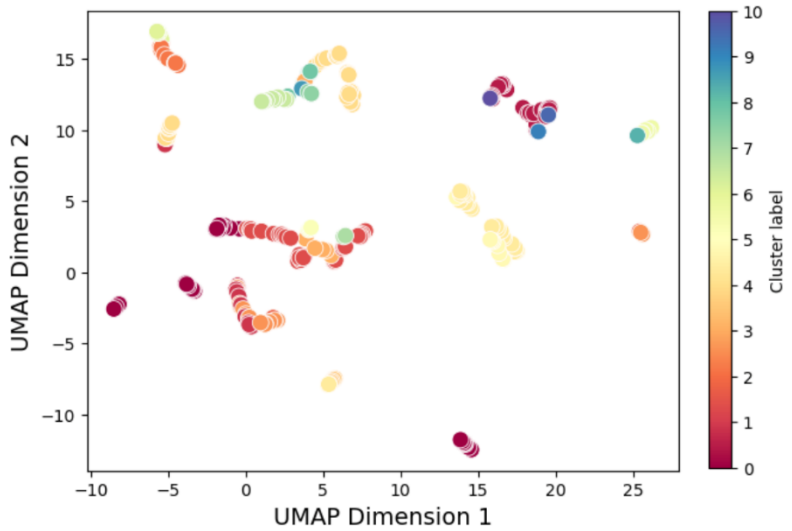
heights are proportional to attention contributions. Examples are shown for a highly expressed and a moderately expressed gene. (d) Chromatin state composition of high-attention motifs for each comparison group, colored according to the 15-state annotation scheme.

995

(a)


Initial motif list
pairwise DTW calculation
DTW score matrix
Dendrogram
Clustering
GGBAAA
GBBBAA
GBAAAA
GGBBB
GBBBAA
GBAAAA
forward-forward
GBBBAA
GBAAAA
forward-reverse

(b)


UMAP Dimension 2
UMAP Dimension 1
Cluster label

(c)

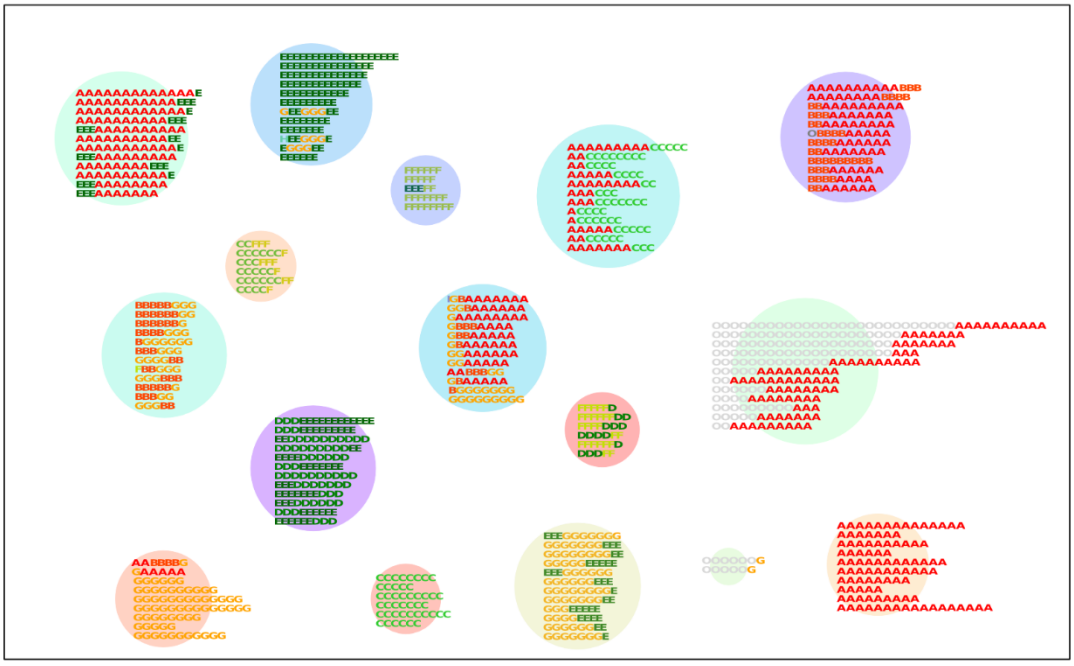


Figure 5. Chromatin state motif clustering with DTW. (a) Workflow of motif clustering using DTW. (b) UMAP representation of motifs extracted from the fine-tuned model using a 50 kb upstream / 5 kb downstream input window for the log(RPKM) > 5 vs. log(RPKM) > 1 classification task, clustered using agglomerative clustering. (c) Visualization of chromatin state motif clusters shown in (b), with each cluster's size proportional to its number of elements.

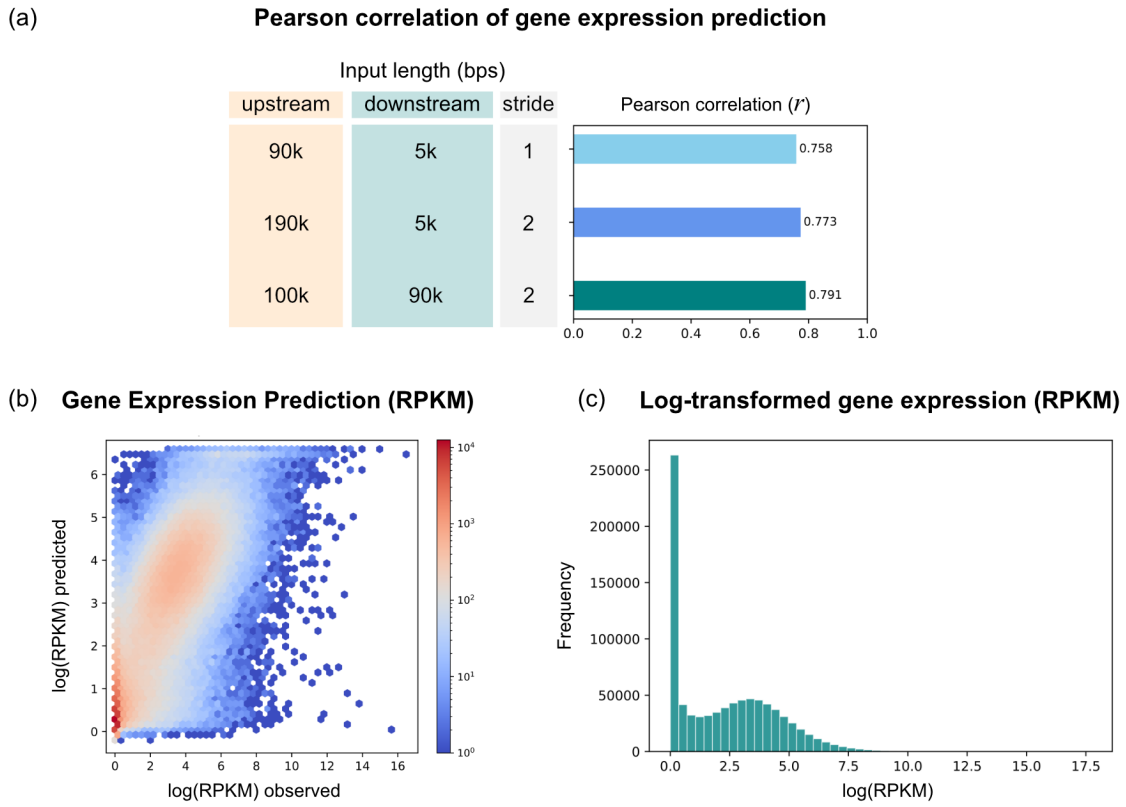


Figure 6. (a) Pearson correlation between predicted and observed log-transformed RPKM values for models trained with different upstream and downstream input windows around the TSS. (b) Hexbin plot showing predicted versus observed log(RPKM) values for the model using the 100 kb upstream / 90 kb downstream configuration. (c) Distribution of log-transformed RPKM values used for training and evaluation.

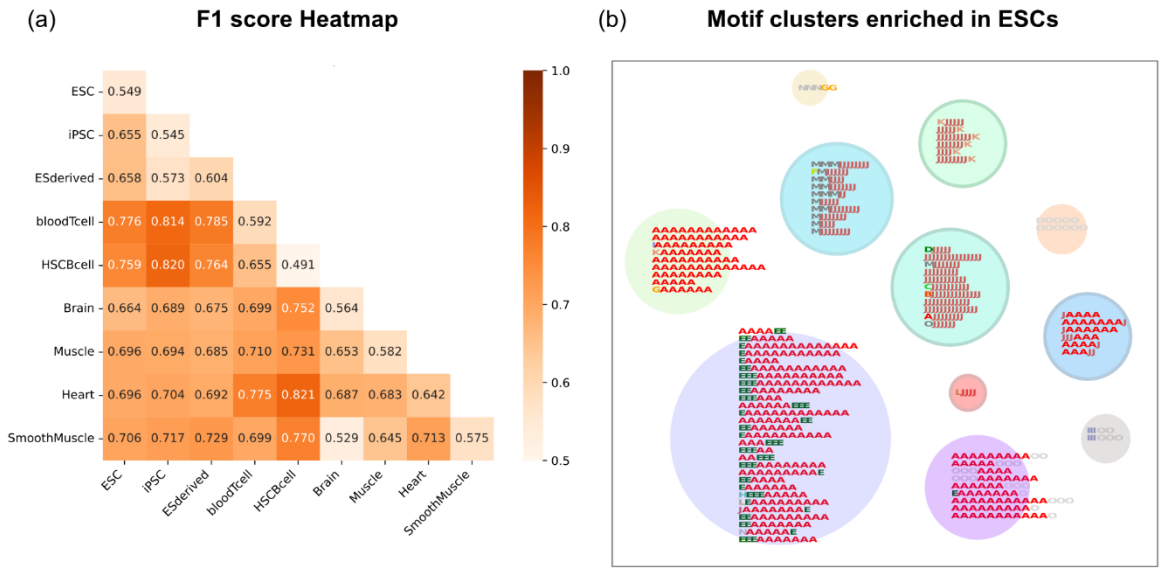


Figure 7. Cell-type classification and motif analysis across CRM regions. (a) F1 score heatmap for pairwise binary classifications of cell-type groups. Chromatin state sequences were extracted from 20-kb windows centered on CRMs obtained from the SCREEN database, and ChromBERT was fine-tuned for each pairwise comparison following the ROADMAP cell-type group definitions. (b) Motif clusters enriched in ESCs relative to T-cells, obtained from DTW-based clustering of promoter-proximal chromatin state motifs. Circle sizes are proportional to the number of motifs in each cluster.

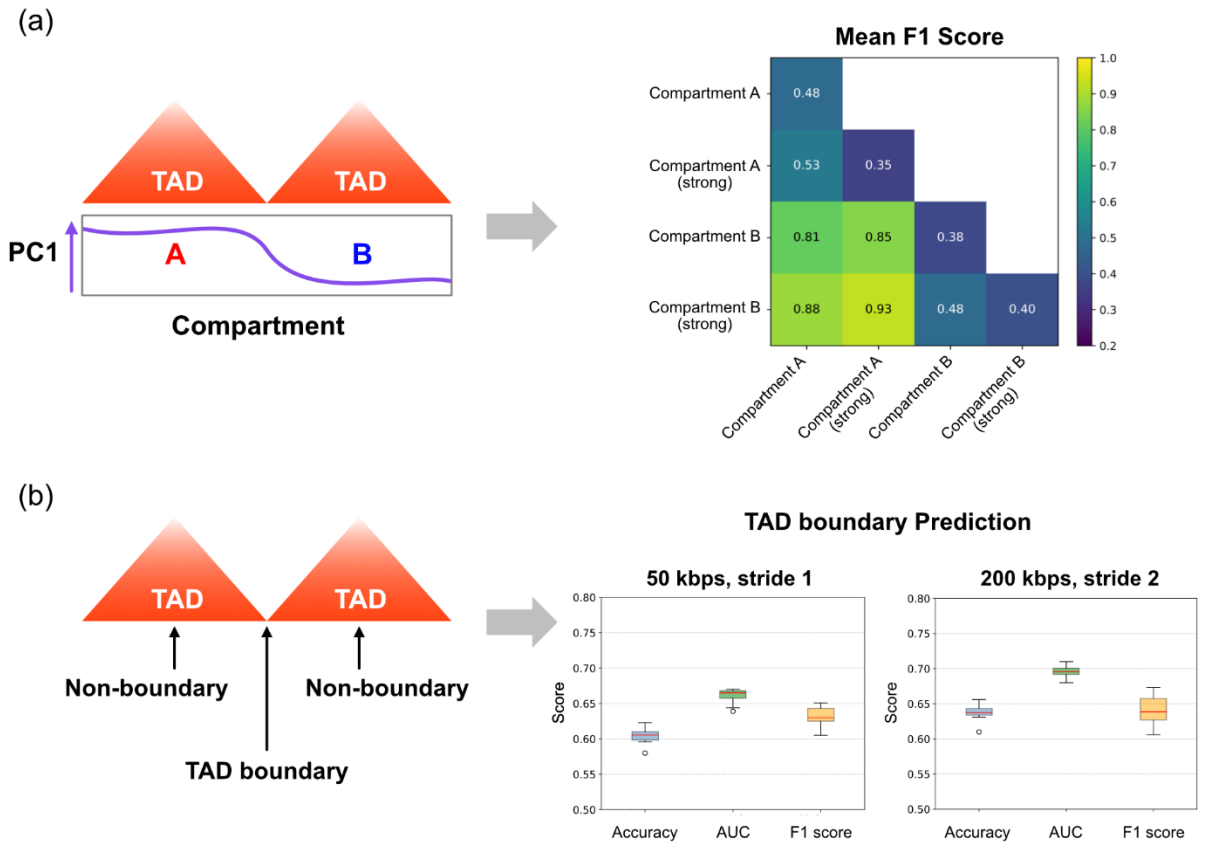


Figure 8. Classification of 3D genome features using ChromBERT. (a) Mean F1 scores for pairwise classification of A/B compartments and strong A/B compartments. Chromatin state sequences were extracted from genomic bins defined by the first principal component (PC1) of the observed/expected Hi-C contact matrix (25-kb resolution). (b) Classification performance for TAD boundary versus non-boundary regions. TAD boundaries were identified using insulation-score–based calling, and chromatin state sequences around boundary and non-boundary sites were used to train and evaluate ChromBERT with different input window sizes (50 kb and 200 kb).