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

Seohyun Lee[1], Che Lin[2], Chien-Yu Chen[3], and Ryuichiro Nakato[4,*]

[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]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.
[3]Department of Biomechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan.
[4]Laboratory of Computational Genomics, Institute for Quantitative Biosciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan.

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

Abstract

Chromatin states, fundamental to gene regulation and cellular identity, are defined by a unique combination of histone post-translational modifications. 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 patterns of chromatin state annotation data sequences. Notably, ChromBERT was pre-trained on promoter regions across a diverse range of epigenomes and subsequently fine-tuned using a dataset from multiple cell lines where RNA-seq data were available, highlighting the model's ability to discern conserved chromatin state patterns within these regions. In addition to its predictive powers across tasks, evidenced by high AUC scores, ChromBERT provides further analysis through the incorporation of motif clustering using Dynamic Time Warping (DTW). This method enhances the model's ability to dissect chromatin state sequence motifs, typically involving transcription and enhancer sites. The introduction of motif clustering with DTW into ChromBERT's workflow is poised to facilitate

the discovery of genomic regions linked to novel biological functions, deepening our understanding of chromatin state dynamics.

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, the presence of active chromatin states, marked by histone modifications such as H3K4me3 (trimethylation of lysine 4 on histone H3) and H3K27ac (acetylation of lysine 27 on histone H3), is generally linked to gene activation and transcriptional initiation [5], [6]. A deeper understanding of chromatin states and their genomic distribution has been facilitated by the chromatin state annotation tools such as ChromHMM [7] and Segway [8] coupled with the advent of next-generation sequencing (NGS) technologies [9], [10], [11]. Such annotation tools systematically learn and characterize combinatorial patterns of histone modifications into distinct chromatin states, such as "active promoters" and "genic enhancers." The algorithm processes genome-wide datasets generated from chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) experiments, inferring the most likely chromatin state at each genomic locus.

Large databases of epigenome information have been generated by international consortia, such as ROADMAP [12], ENCODE [13], and IHEC [14]. For instance, the ROADMAP Epigenomics Project characterized and annotated the chromatin state of 127 distinct human cell and tissue types. The databases can be used to perform various downstream analyses to decipher the hidden intricacies of the chromatin state landscape. As one such attempt, ChromDiff [15] compared a summarized set of chromatin states across different epigenomes, offering new insights into tissue-specific or age-dependent chromatin alterations. More recently, CSREP [16], 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 many epigenomes based on the surprisal score measured from the rarity

of the chromatin state distribution [17]. A similar method, established with a different metric, the conservation-associated activity score, was suggested by Libbrecht *et al* [18] as a novel annotation strategy that has the advantage of detecting more biologically functional regions.

The availability of chromatin state annotations for a large dataset and the precedent related studies have paved the way for the next important step: identifying patterns of chromatin state 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 [19]. In the context of our study, we extend this concept to chromatin state sequences to identify distinct patterns of chromatin states that could be key to understanding gene regulation. The significance of chromatin state motifs extends to their potential role in dictating the structural and functional landscape of the genome, influencing the transcriptional output and cellular identity. By mapping these motifs, we can begin to decode the epigenetic language that modulates gene expression patterns across different cell types and developmental stages, offering insights into the molecular basis of diseases and identifying new targets for therapeutic intervention. However, due to their inherent complexity, these chromatin state patterns may not be fully captured by traditional computational methods, such as k-mer-based motif discovery algorithms [20], [21], which often rely on simplistic representations of DNA sequence motifs and do not consider the combinatorial nature of histone modifications and other chromatin-associated marks. This underscores the need for innovative 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 [22]. BERT has proven highly effective in natural language processing tasks and provides an important way to analyze sequential patterns in biological data [23]. While DNABERT has been proposed for the analysis of DNA sequences [24], our ChromBERT extends the concept fundamentally to the adaptation of chromatin state-annotated human genome sequences by combining it with Dynamic Time Warping (DTW) [25]. Using the chromatin state annotation data for 127 distinct human cell and tissue types obtained from ROADMAP, ChromBERT uncovered previously unrecognized patterns of chromatin states. It has the potential to contribute to a deeper understanding of gene regulation and its implications for

human health and disease. 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 in the study. Initially, numerically annotated chromatin state labeling data from ROADMAP, based on a combination of core histone modifications, are converted into an alphabetic code ranging from A to O. This process is illustrated in Fig. 1(a) where chromatin state sequences are depicted in alphabetical labels following data preprocessing via ChromBERT. In the training phase, as shown in Fig. 1(b), these chromatin state sequences are tokenized and fed into a BERT model for embedding. The comprehensive workflow of ChromBERT is captured in Fig. 1(c), starting with the preprocessing of chromatin state sequence data. Subsequently, ChromBERT undergoes training to learn the general patterns of chromatin state placement through a pretraining stage. Following this, it is fine-tuned using binary classification data, such as distinguishing between regions with complex chromatin state patterns, high gene expression, and promoters, from those with less complex chromatin patterns, low gene expression, or non-promoter regions. The motifs identified from the targeted classes are then clustered using DTW to discern and interpret the representative motifs for specific regions of interest. See the Methods section for further technical details.

Categorization of complex genic regions and less-complex genic regions based on chromatin state switching frequency

Utilizing ChromBERT to classify genic regions based on the characteristics of chromatin state compositions, our initial approach involved fine-tuning for two distinct genic region types: complex and less-complex. The complexity of genic regions was determined through the frequency of annotated chromatin state switching within these regions (see Methods). The fine-tuning for this task was carried out after the pretraining phase was conducted on the entire spectrum of genic regions, concatenated by 10 different epigenomes. In the pretraining process, while DNABERT [24] used longer k-mers (3, 4, 5, and 6-mers) to capture the enduring patterns

of sequences, we chose 4-mers for ChromBERT due to the substantial vocabulary size of chromatin states (15 characters) compared to the DNA sequence (4 characters, see Methods).

Figure 2(a) shows examples of complex and less-complex genic regions represented in chromatin state annotation data. Typically, over 50% of complex genic regions are annotated as either weak transcription (E), strong transcription (D), or active Active Transcription Start Site (TSS) (A). Conversely, more than 50% of the genic regions categorized as less-complex are annotated as quiescent or having low signal (O) (Supplementary Fig. S1). Because longer genes can accommodate more chromatin states, there is a positive correlation between gene length and the number of chromatin state changes, as shown in Fig. 2(b), for a representative cell. It is important to note that the complexity of chromatin states assigned to genes is inversely proportional to their length, indicating a negative relationship (Supplementary Fig. S2). The outcomes of pretraining across all genic regions are detailed in Fig. 2(c). The perplexity of the pretraining decreases over iterations, eventually reaching approximately 1.13 for both 3-mer and 4-mer cases, implying that the general placements of chromatin state annotation can be effectively trained. The relatively low initial value of perplexity is thought to stem from the persistent nature of chromatin states, which tend to remain consistent over extended regions of DNA, thereby simplifying the pretraining process.

When we evaluated 1,000 random test instances—comprising 502 chromatin state segments from complex genic regions and 498 from less-complex genic regions, sourced from a sample cell type (E003, H1 cell line)—ChromBERT effectively predicted each case, achieving an F1 score of 0.98 (Fig. 2(d)). For this case, the attention score matrix exhibited relatively high scores towards the beginning, which is likely due to the location of the TSS in the initial segment of the genic region. This interpretation is further supported by similar trends observed in other classification tasks, such as those distinguishing genic regions from intergenic regions and complex genic regions from intergenic regions (Supplementary Fig. S3).

However, following the motif finding algorithm used in DNABERT [24] (see Methods), no distinct motifs in the complex genic regions were detected when compared to the less-complex regions. Upon relaxing the conditions for motif detection—defining high attention scores as values exceeding the third quartile without a minimum condition—only a series of active TSS states, represented as ‘AAA,’ emerged as a motif. This suggests no distinctive

154 chromatin state patterns in complex genic regions compared to less-complex ones. When
155 comparing complex genic regions with intergenic regions, several simple motifs, such as
156 'AAAAB,' 'EEDDD,' 'EEEEE,' and 'DDDDD,' were detected using the standard motif
157 detection algorithm. Because 'D' and 'E' respectively represent strong and weak transcriptions,
158 this outcome is consistent with the understanding that genic regions are defined as 'transcribed'
159 regions. The example motifs found in complex genic regions, along with their attention scores,
160 are shown in Fig. 2(e). Note that not all areas with high attention scores qualify as motifs. For an
161 area to be considered a motif, it is required to appear at least three times (see Methods).
162

163 **Classification of highly-expressed genic regions and non-expressed genic regions**

164 In another task, we deployed ChromBERT to categorize genic regions based on their
165 expression levels. Since this task involved the same genic regions, we used the parameters
166 obtained from the pretraining on the whole genic regions, which was initially conducted for
167 distinguishing between complex and less-complex genic regions. We defined genic regions with
168 RPKM values exceeding 10, 20, 30, and 50 as 'highly expressed' and conducted pairwise fine
169 tunings for each against 'expressed genic regions (RPKM>0)' and 'non-expressed genic regions
170 (RPKM=0)'.

171 Figure 3(a) displays the prediction performances, which are represented as the average of
172 accuracy, AUC (Area Under the Curve), and F1 score calculated as the final score after training
173 for each fine-tuning task. Although there are slight differences among the classification tasks for
174 the different gene expression levels, an increase was observed corresponding to higher gene
175 expression. For example, the average AUC of the final ten data points for the classification
176 between RPKM>10 versus >0 produced 0.77, while an AUC of 0.895 was obtained for the
177 classification between RPKM>50 versus >0. When the averages of ten final data points of AUC
178 are compared, the classification between RPKM>30 versus =0 yielded the highest score, 0.904.

179 In addition, similar to the previous task, the high attention scores appeared to be
180 relatively concentrated near the start of the genic regions compared to other areas, as shown in
181 Fig. 3(b). Such a pattern suggests that transcription start sites play a crucial role in distinguishing
182 regions that are expressed. This is consistent with the understanding that the chromatin state
183 around the transcription start sites is essential for estimating gene expression levels.

Figure 3(c) and (d) respectively illustrate the representative motifs in expressed genic regions, comparing the case of RPKM>30 versus RPKM>0 and RPKM>30 versus RPKM=0. The Word Cloud visualization shows that the motifs with the larger fonts represent those that appear more frequently. For example, “FFFFB”, representing an 800-bps long genic enhancer (“F”) followed by a 200-bps long flanking active TSS (“B”), is the most frequent chromatin state motif for the first case. Upon comparing the characteristics among the classifications at different gene expression levels, specifically RPKM>10 and RPKM>20, the most frequent motifs predominantly consist of a sequence of weak genic regions (“E”) and strong genic regions (“D”), represented by sequences like “EEEEE” or “DDDDE”. In contrast, at higher expression levels, such as RPKM>30 and RPKM>50, the motifs primarily composed of genic enhancers (“F”) and enhancers (“G”) appear, such as “FFFFF” or “GGGBB” (Supplementary Table S1). This result suggests the importance of a large number of enhancer-like regions for the high-level expression. Intriguingly, complex motifs involving more than three distinct chromatin states or showing frequent switching more than three times were observed in the classifications between RPKM>30 and RPKM>0 and between RPKM>30 and RPKM=0 (Supplementary Table S1). These classifications presented the motifs “DEFFF” and “GBBBG”, respectively. Here, “DEFFF” signifies the presence of a short, weak transcription region (“E") positioned between a strong transcription region (“D”) and genic enhancer (“F”), while “GBBBG” suggests that flanking active TSS (“B”) is positioned between enhancers (“G”), effectively flanked on both ends. It should be noted that these motif patterns frequently appear in many genes from multiple epigenomes and are therefore not derived from chance observation or technical issues such as read depth or variance in the signal-to-noise ratio of the antibodies. These observations are particularly noteworthy as they reveal distinct chromatin state patterns in highly expressed genic regions compared to those in non-expressed or minimally expressed genic regions.

Identification of strong promoters based on gene expression level

After the initial categorization task for genic regions, we further employed ChromBERT to identify chromatin state motifs around promoter regions. While various lengths can be defined for promoter regions, we chose to use data spanning from 20 kbps upstream to 40 kbps downstream of genes for the 57 cell lines that had available RNA-seq data in ROADMAP [12]. We made this choice because these regions are often crucial for gene regulation, containing cis-

regulatory elements like enhancers, silencers, and insulators that influence gene expression [26]. Using criteria similar to those defined in the previous task, promoter regions adjacent to genic regions with RPKM values higher than 10, 20, 30, and 50 were labeled as 'highly expressed'. Promoter regions adjacent to genes with RPKM > 0 were labeled 'expressed', and those with RPKM = 0 were termed 'non-expressed'. For pretraining purposes, only the promoter regions from all target cell lines were utilized. From the pairwise fine-tuning of the above combinations, we observed a clear tendency: the promoter regions adjacent to genes with higher RPKM values exhibited higher AUC values in classification tasks, especially when compared to promoter regions near genes with low (RPKM > 0) or no expression (RPKM = 0), as shown in Figure 4(a).

The motifs identified in the promoter regions adjacent to highly expressed genes predominantly consisted of active markers, such as active TSS ("A"), flanking active TSS ("B"), strong transcription regions ("D"), weak transcription regions ("E"), and genic enhancers ("F"), as described in Figures 4(b) and (c). This aligns with conventional biology understanding where a strong promoter encompasses an active TSS to activate gene expression. The most common motifs comprised one or two of the aforementioned chromatin states in sequential order, such as "ABBBB", "BBBBB", "AAABB", "DDEEE", or "EEEEE". This pattern mirrors observations from previous tasks, underscoring that chromatin state sequences tend to persist rather than switch frequently. Motifs with a relatively complex composition were also identified, especially when comparing promoter regions adjacent to genes with RPKM > 30 and RPKM > 50 to those associated with non-expressed genes (RPKM=0). For instance, sequences such as "BABA", "GBAA", and "GBBBA" were distinctive in the promoter regions of these highly expressed genes. This suggests that these specific patterns in chromatin state sequences could be precursors or indicators of highly expressed genes.

Clustering the chromatin state motifs using Dynamic Time Warping

The previous results suggested that chromatin state sequences tend to be persistent. Additionally, the observed patterns (both in terms of characters and length) may vary slightly due to the fact that some chromatin states are correlated with each other (e.g., strong and weak transcription) and due to technical issues, such as different read depths and signal-to-noise ratios of the antibodies across samples. Therefore, chromatin state motif analysis requires 'collapsing' the similar patterns of chromatin states into a single representative motif. This approach helps to

understand the overarching patterns of chromatin state motifs within specific genomic regions post-fine-tuning.

To accomplish this, we have introduced a clustering step utilizing DTW. 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 [27], 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 [28] is applied to cluster similar patterns into categories, 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 manner because DNA transcription involves both forward and reverse directions. Then, the DTW score matrix is generated using the lower scores, which represent higher similarities. Clustered motifs can be identified after determining the optimal number of clusters using a dendrogram. Figure 5(b) visualizes the clustered motifs of a representative case study on promoter regions associated with highly expressed genes. Chromatin state motifs were compiled from those identified across three distinct fine-tuning thresholds, comparing highly expressed genes (with RPKM values exceeding 50, 30, and 20, respectively) against non-expressed genes (RPKM=0). We used a window size of 12, a minimum motif length of 5, and a minimum occurrence frequency of 2. The x-axis denotes the position of the chromatin state within a motif, whereas the y-axis denotes the chromatin state symbols, represented alphabetically. After applying agglomerative clustering followed by DTW, the motifs have been categorized into 11 clusters in this example. In this visualization, motifs categorized as similar share the same color and line style, facilitating easy identification of motifs that exhibit similar state transitions, prevalence, and patterns of stability and variability. Figure 5(c) represents the detected motif clusters plotted in the UMAP analysis [29]. ChromBERT includes a visualization function that summarizes the clustered chromatin

state motifs, as shown in Figure 5(d). This feature helps users to understand the composition and relative size of each cluster.

In our analysis to identify the optimal number of clusters, we integrated several evaluative measures, including insights from the dendrogram, to achieve a judicious balance between detail and overarching pattern recognition. We tested different numbers of clusters, such as 9 and 11, and found that DTW was effective in improving the accuracy of our clustering process. With 11 clusters, for instance, motifs with similar state transitions but different lengths, such as “GBBBB”, “GBBAA”, “GBAAAA”, and “GBBBA”, were clustered into a single category, while “GABBBB,” which has a different transition, was clearly segregated into a separate category. Such transitions like “G-B-A” could represent enhancer regions immediately preceding active TSS regions, potentially acting as a herald for transcription initiation. This suggests a dynamic regulatory setup where enhancer activation might directly influence the onset of transcription at proximal promoter sites. Additionally, sequences such as “OAAA” and “OOOAAA” in other clusters may reflect abrupt changes in chromatin states, possibly serving as precursors to promoter regions. Such patterns suggest the potential complexity of chromatin regulation captured by ChromBERT, and these motifs could provide interesting insights into the orchestrated events leading up to gene expression.

Discussion

In ChromBERT, we applied the BERT algorithm to accommodate the unique characteristics of chromatin state data. While several tools, such as CSREP [16] and EpiAlign [30], have been developed to analyze the same genomic regions of multiple samples, our approach aims to obtain significant chromatin state patterns within different regions of multiple samples. This approach has the potential to identify unexplored functional genomic regions and their epigenomic features.

Unlike DNA sequences focused by DNABERT, which are comprised of 4 base nucleotides, the number of chromatin states in this study was 15. This distinction significantly expands the resulting vocabulary size. Therefore, the substantially increased vocabulary size influenced the complexity of the k-merization process, a critical step for tokenizing the data prior to pretraining. While DNABERT only necessitated vocabulary sets of 68 (3-mers), 260 (4-mers), 1028 (5-mers), and 4101 (6-mers), ChromBERT required 3,379, 50,629, 759,379, and

11,390,629 respectively. This pronounced difference in vocabulary size is computationally challenging and causes GPU memory overflows when attempting to pretrain with more than 5-mers, even with a single sample set. This effectively limited us to using 4-mers for the pre-training step.

Additionally, the perplexity of the model, a metric that quantifies how well a model predicts a sample, showed a decreasing trend from approximately 4 to 1 over iterations during the pretraining. Interestingly, the perplexity 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 early in the phase. This initially low perplexity can be comprehended by considering several factors. The stable, long-lasting nature of chromatin states and their functional bias likely lead to more predictable sequence patterns, easing the model's predictions. Additionally, our choice of 4-mer tokenization further simplifies the task for the model, as it only needs to predict patterns within a more limited context compared to larger k-mers.

Despite the computational constraints that led to the use of 4-mers, the pre-trained parameters have proved to be highly effective in fine-tuning the model for various downstream tasks. This includes binary classification of complex genes versus less-complex genes and highly expressed genes versus non-expressed genes. Furthermore, the model was able to identify several meaningful chromatin state motifs, suggesting its effectiveness in detecting unknown epigenomic patterns from large-scale data.

However, an intriguing aspect of our research emerged when ChromBERT was applied to broader genic regions. The model's performance was notably better when pre-trained on a narrower set of 10 cell lines and then fine-tuned on data from 57 cell lines than when the inverse training scheme was employed. This performance discrepancy might arise from overfitting to the narrower set of 10 cell lines. It is conceivable that chromatin states in a larger set of cell lines, which may contain a certain amount of moderate-quality data, introduce complexities that challenge motif detection. In contrast, promoter regions, which are functionally critical for transcription initiation, may exhibit more conserved chromatin state patterns, enhancing ChromBERT's accuracy in motif detection in such regions.

A current limitation of ChromBERT pertains to the detection of chromatin state pattern motifs, specifically in terms of their length. As we follow the data preprocessing steps outlined in DNABERT, which likely adheres to the sequence length constraints of the original BERT

338 architecture, the input sequence length for both training and testing is capped at 510 tokens. This
339 limitation may have affected the ability of ChromBERT to detect longer motifs, as the majority
340 of identified motifs were approximately 5 or 6 states long.
341 Another potential reason for the shorter length of the identified motifs could be related to
342 the k-merization process. Although the 4-mer tokenization used for the pretraining primarily
343 helps to learn general patterns within the data and does not directly limit the length of the motifs,
344 it could shape the initial context in which the model learns to predict chromatin states. This
345 might make the model more sensitive to patterns of similar length to the 4-mers, even when a
346 longer window size is used during the motif-finding stage.
347 Therefore, an important avenue for future research will be to devise methodologies that
348 allow the capture of longer motifs, potentially those spanning extended segments of DNA. This
349 might involve modifying the sequence length constraints or adjusting the tokenization process to
350 better accommodate the unique properties of chromatin state annotation data. By doing so, we
351 anticipate a further enhancement of ChromBERT's utility in understanding the complex
352 landscape of chromatin states.
353
354 **Conclusion**
355 In this study, we introduced ChromBERT, a model specifically designed to detect
356 distinctive patterns within chromatin state annotation data sequences. By adapting the BERT
357 algorithm as utilized in DNABERT, we pretrained the model on the complete set of genic
358 regions using 4-mer tokenization. While ChromBERT demonstrated its proficiency in tasks such
359 as differentiating between complex-genic regions characterized by recurrent chromatin state
360 changes and regions with varying gene expression levels determined by RPKM thresholds, its
361 most notable achievement was in the fine-tuning for promoter regions. This implies that
362 promoter regions might encapsulate more conserved chromatin state patterns, enhancing their
363 discernibility, particularly in more expansive datasets. Furthermore, ChromBERT identified
364 chromatin state sequence motifs in these promoter-focused evaluations, often characterized by
365 strong and weak transcription sites along with enhancer regions, predominantly spanning 5 states
366 across 1000 bps with minimal internal variations. To provide deeper insights into these motifs,
367 we added a motif clustering step using DTW, offering a nuanced view for estimating the motif
368 patterns of chromatin state sequences. This addition enriches our analysis, setting the stage for

ChromBERT's potential in distinguishing specialized regions based on chromatin states, thereby aiding in the discovery of genomic regions associated with novel biological functions.

Methods

1. Data collection

The chromatin annotation BED files are downloaded from the ROADMAP [12] project, which includes 127 distinct epigenomes annotated with 15 different chromatin states at 200-bp intervals using ChromHMM [7]. The 15 different chromatin states are 1. Active Transcription Start Site (TSS), 2. Flanking Active TSS, 3. Transcription at Gene 5' to 3' Ends, 4. Strong Transcription, 5. Weak Transcription, 6. Genic Enhancers, 7. Enhancers, 8. Zinc Finger Protein (ZNF) Genes & Repeats, 9. Heterochromatin, 10. Bivalent/Poised TSS, 11. Flanking Bivalent TSS/Enhancer, 12. Bivalent Enhancer, 13. Repressed Polycomb, 14. Weak 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 (NLP) 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 NLP 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 DNABERT, the processed chromatin annotation data were randomly cut to ensure that

400 the length of the data did not exceed 510 but was longer than 5, following their data preparation
401 method. Subsequently, the chromatin state sequences were tokenized into 3-mers, 4-mers, 5-
402 mers, and 6-mers. Any sequences shorter than the length of each k-mer were not included in the
403 dataset.
404
405 **3. Definition of complexity for complex genic regions and less-complex genic regions**
406 For conducting the binary classification between complex genic regions and less-complex
407 genic regions, the entire genic regions were investigated to define the complexity of the specific
408 data segments. To assess the complexity of the chromatin annotation data processed, as
409 explained in the previous section, we utilized two factors: the length of the data and the number
410 of chromatin state switches within a single data segment. Complexity can be defined as the
411 proportion of state symbols that are not adjacent to identical symbols. For example, in the case of
412 the data segment "AABBBCCAABBBCCC", the complexity is 6/15. For a sample epigenome,
413 the mean complexity of the chromatin annotation data for genic regions is approximately 0.11.
414 As we define complex genic regions as those with complexity higher than this threshold, there
415 were approximately 13,192 complex genic regions, while the number of genic regions defined as
416 less-complex was approximately 30,068 (Supplementary Fig. S3).
417
418 **4. Definition of 'highly expressed', 'expressed', and 'non-expressed' genic regions for**
419 **selective comparisons across various gene expression levels**
420 In the case of pairwise binary categorization across various gene expression levels, such
421 as 'highly expressed', 'expressed', and 'non-expressed', we utilized the RNA-seq data from 57
422 different cell lines provided by the ROADMAP project. A genic region is defined as 'expressed'
423 if the RPKM (Reads Per Kilobase per Million mapped reads) is greater than 0, while genic
424 regions with an RPKM value of 0 are considered 'non-expressed'. For 'highly expressed' genic
425 regions, we established four different RPKM thresholds: 10, 20, 30, and 50. These thresholds
426 were chosen based on our preliminary analysis, with an RPKM value of 10 serving as a
427 reasonable starting point for high expression. We selected an upper limit of 50 due to the
428 significantly reduced data quantity at this threshold, which was approximately a tenth of the data
429 available at an RPKM value of 10.
430

5. Pretraining

The BERT architecture that is utilized in the DNABERT model was pretrained for chromatin annotation data, using multiple datasets acquired from different cell types, except for brain cells. The epigenomes used for the pretraining included H1 cell line (E003), H1 BMP4 derived mesendoderm cultured cells (E004), H1 derived mesenchymal stem cells (E006), hESC derived CD184+ endoderm cultured cells (E011), HUES64 cell line (E016), breast myoepithelial cells (E027), CD4 memory primary cells (E037), mobilized CD34 primary cells female (E050), adult liver (E066) and lung (E096). The sequence of chromatin annotation data for the genic regions of these ten epigenomes was preprocessed as described in the data processing section and then concatenated, followed by tokenization 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.

6. Fine-tuning

We conducted fine-tuning for each downstream application based on the parameters acquired from the pretraining process. We followed the fine-tuning conditions used in DNABERT, using AdamW as the optimization method with a fixed weight decay of 0.10 and a dropout probability of 0.1 for the hidden layer. Notably, in each fine-tuning iteration, we prioritized results obtained from 4-mer sequences. This focus was predicated on our intent to intensify our efforts on longer pattern recognition within chromatin state sequences and their respective motifs, going beyond the 3-mer sequences used during pretraining.

462

463 **7. Motif finding**

464 Similarly to the definition of DNA motifs in DNABERT, chromatin state motifs in our
465 study are defined as frequently appearing patterns of consecutive chromatin states that are
466 distinctively present in a certain category during binary classification. We followed the
467 parameter conditions used in DNABERT to detect and filter these motifs. Firstly, we captured
468 regions with high attention scores from the output attention matrix. A threshold was defined as a
469 score higher than the mean attention scores of the matrix and higher than 10 times the minimum
470 score to determine regions where motifs are likely included. These regions were registered as
471 candidates, which were further filtered using a p-value of 0.05. This p-value determined whether
472 the motifs were significantly enriched in one class compared to the other. Finally, motifs were
473 identified as regions that are longer than four chromatin state sequences (equivalent to 800 bps)
474 and appear at least three times exclusively in the target class. While in the case of DNA motifs,
475 final motifs are typically compared with existing DNA motif libraries, we collected chromatin
476 state motifs without such a comparison because there is currently no known library available for
477 this task.

478

479 **8. Motif clustering**

480 To further estimate the representative motif chromatin sequence pattern, we utilized
481 DTW and agglomerative clustering to analyze the chromatin state motifs. Our initial dataset
482 included motifs identified after careful screening for p-values, minimum length, and minimum
483 occurrences within predefined regions of interest, avoiding the merging step to allow a detailed
484 examination of chromatin sequences, which are represented alphabetically. Given the potential
485 for variability in these sequences due to factors such as data quality (e.g., the signal-to-noise ratio
486 of the antibodies) or fluctuations in histone modification peaks, we opted for a more flexible
487 approach to motif identification. DTW was selected for its capability to adjust for sequence
488 length variations, accommodating the intrinsic dynamics of chromatin state sequences.

489 First, to implement DTW, we converted the chromatin state alphabets ("A" to "O") into
490 numerical values (1 to 15) to facilitate a standardized analysis. We provide two options for
491 handling this numerical translation of chromatin states by offering a 'categorical' parameter,
492 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 S5. 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. Because the DNA sequences involve both forward and reverse directions, we considered both the forward-forward and forward-reverse strands. 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.

Acknowledgements

We thank all the laboratory members for their helpful discussions.

Authors’ contributions

SL developed ChromBERT and performed all analyses in this study. RN conceived and designed the study. SL and RN contributed to the drafting of the manuscript. CL and CYC 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, 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," *Curr. Biol.*, 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] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.
[26] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional Regulatory Elements in the Human Genome," *Annu. Rev. Genomics Hum. Genet.*, vol. 7, no. 1, pp. 29–59, Sep. 2006, doi: 10.1146/annurev.genom.7.080505.115623.
[27] 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.
[28] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. Fifth Berkeley Symp. Math. Stat. Probab.*, vol. 1, no. 14, pp. 281–297, 1967.
[29] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv, Sep. 17, 2020. Accessed: May 07, 2024. [Online]. Available: http://arxiv.org/abs/1802.03426
[30] 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 Res.*, vol. 47, no. 13, pp. e77–e77, Jul. 2019, doi: 10.1093/nar/gkz287.

FIGURE LEGEND

Figure 1. (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. (a) Examples of chromatin state sequences found in complex genic and less complex genic regions. (b) Correlation between the number of switches in chromatin state in genic regions and gene length. (c) Results of pretraining for genic regions using 10 different epigenomes. (d) Confusion matrix obtained from the binary classification of chromatin state patterns between complex genic regions and intergenic regions. (e) Example motifs found in a complex genic region compared to intergenic regions. The colored blocks represent the attention scores of the chromatin state, with motifs presented in bold and large font.

Figure 3. (a) A graph of average AUC, F1, and accuracy scores from the fine-tuning stage of various binary classifications of genic regions at different gene expression levels. (b) A partial attention score matrix representing a case of genic regions classification between RPKM>30 ('highly expressed') and RPKM>0 ('expressed'), with rows indicating the number of each genic region segment and columns showing the position of the chromatin state (200-bps per state). (c-d) Word Cloud visualizations of motifs, with font size indicating frequency of occurrence, for the cases of RPKM>30 vs. RPKM>0 (c) and RPKM>30 vs. RPKM=0 (d).

Figure 4. (a) The fine-tuning results for promoter regions in terms of mean AUC, accuracy (ACC), and F1 score. (b) The composition of chromatin states within detected motifs across various tasks in promoter regions. (c) Representative motifs were identified for different tasks within promoter regions.

Figure 5. (a) Workflow of motif clustering using DTW. (b) Graph representation of motif clustering achieved through DTW. A total of 116 entries are visually differentiated through various colors and line styles. (c) UMAP representation of motifs clustered using agglomerative clustering. (d) Visualization of chromatin state motif clusters, with each cluster's size proportional to its number of elements.

Figure 1

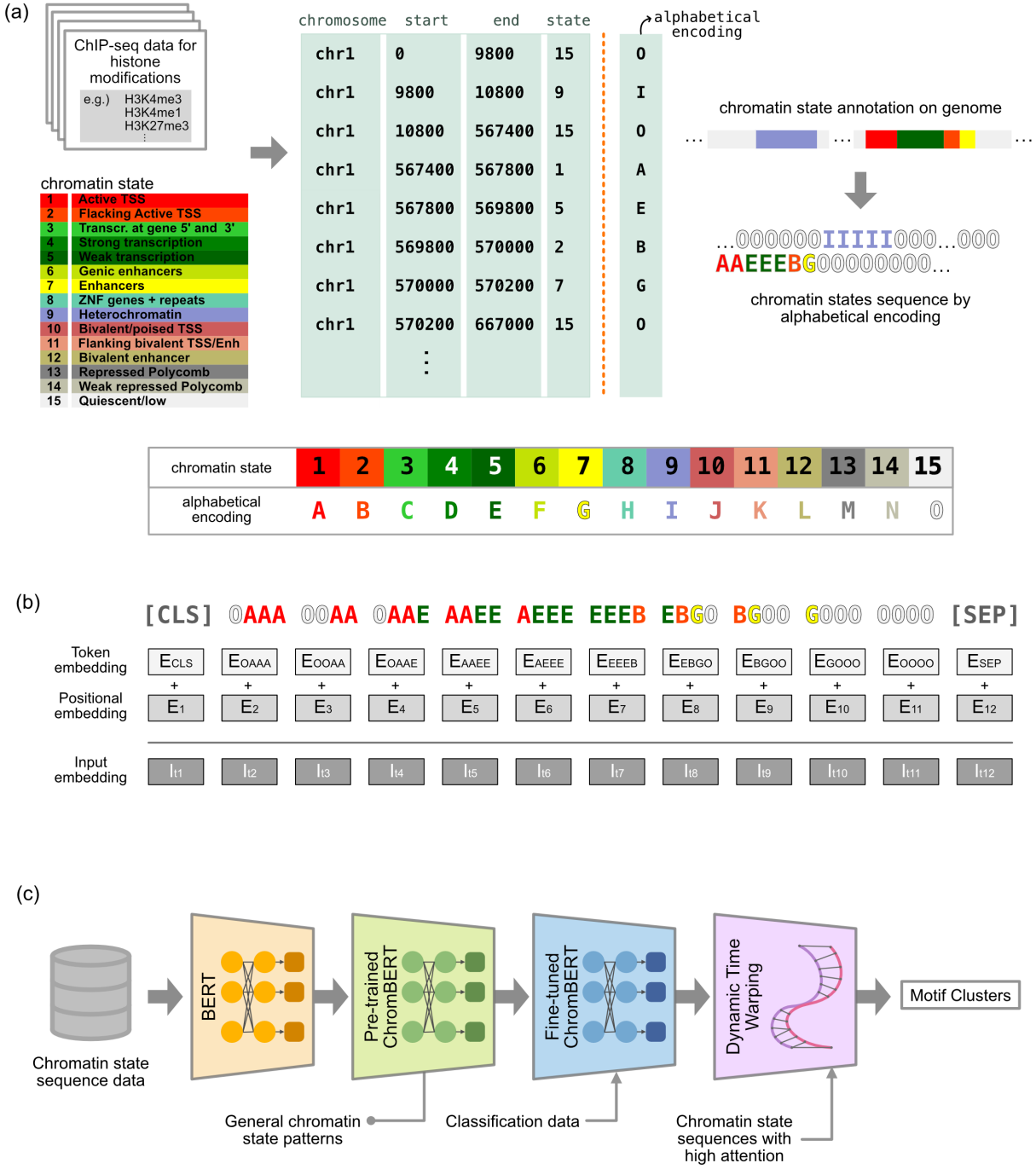

(a)
ChIP-seq data for histone modifications
e.g.) H3K4me3
H3K4me1
H3K27me3
chromatin state
1 Active TSS
2 Flacking Active TSS
3 Transcr. at gene 5' and 3'
4 Strong transcription
5 Weak transcription
6 Genic enhancers
7 Enhancers
8 ZNF genes + repeats
9 Heterochromatin
10 Bivalent/poised TSS
11 Flanking 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)
[CLS] OAAA OOAA OAAE AAEE AEEE EEEB EBGO BGOO GOOO OOOO [SEP]
Token embedding
Positional embedding
Input embedding
(c)
Chromatin state sequence data
BERT
Pre-trained ChromBERT
Fine-tuned ChromBERT
Dynamic Time Warping
Motif Clusters
General chromatin state patterns
Classification data
Chromatin state sequences with high attention

Figure 2

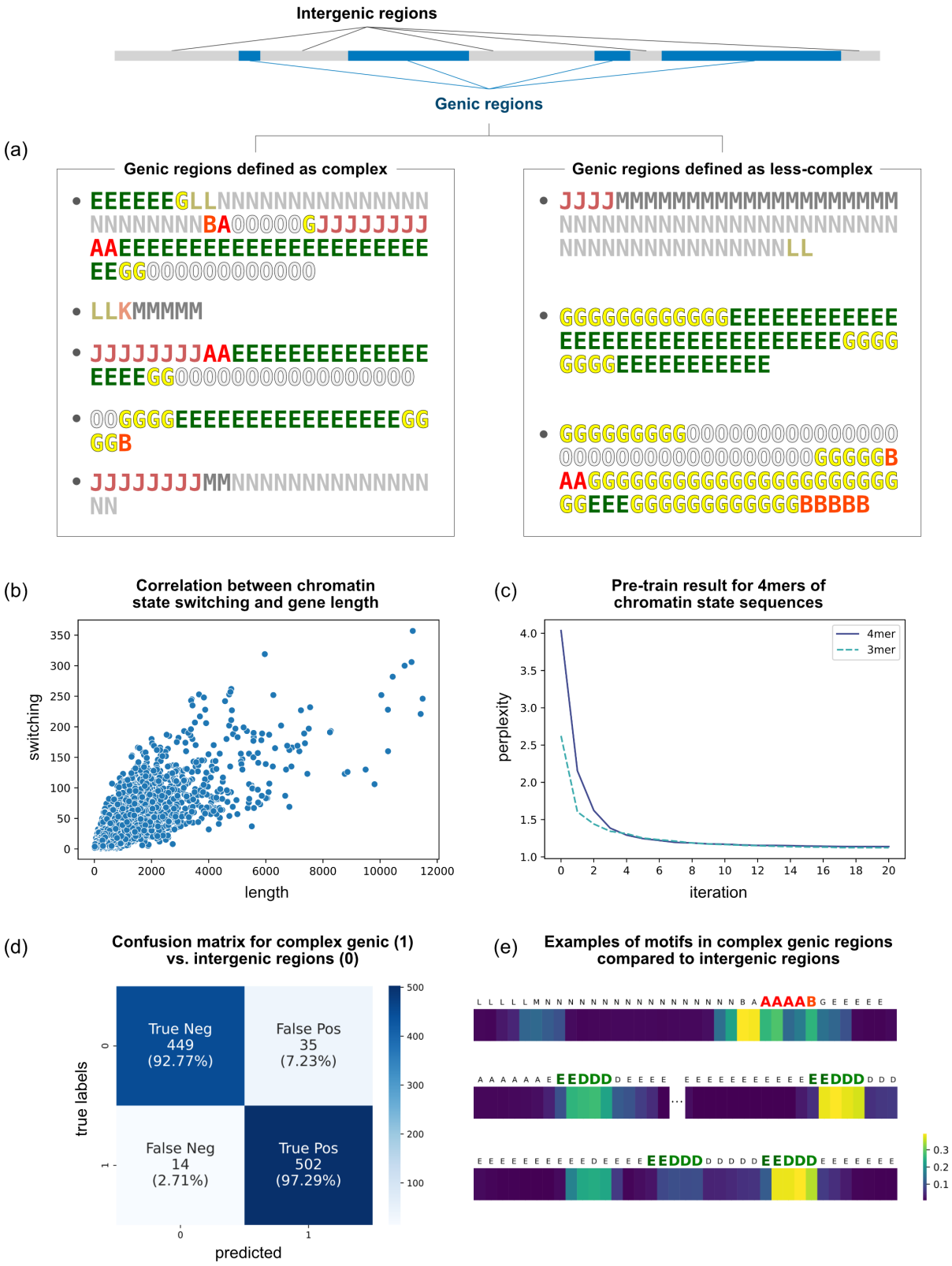

Intergenic regions
Genic regions
(a)
Genic regions defined as complex
Genic regions defined as less-complex
(b)
Correlation between chromatin state switching and gene length
switching
length
(c)
Pre-train result for 4mers of chromatin state sequences
perplexity
iteration
4mer
3mer
(d)
Confusion matrix for complex genic (1) vs. intergenic regions (0)
True Neg 449 (92.77%)
False Pos 35 (7.23%)
False Neg 14 (2.71%)
True Pos 502 (97.29%)
true labels
predicted
(e)
Examples of motifs in complex genic regions compared to intergenic regions

Figure 3

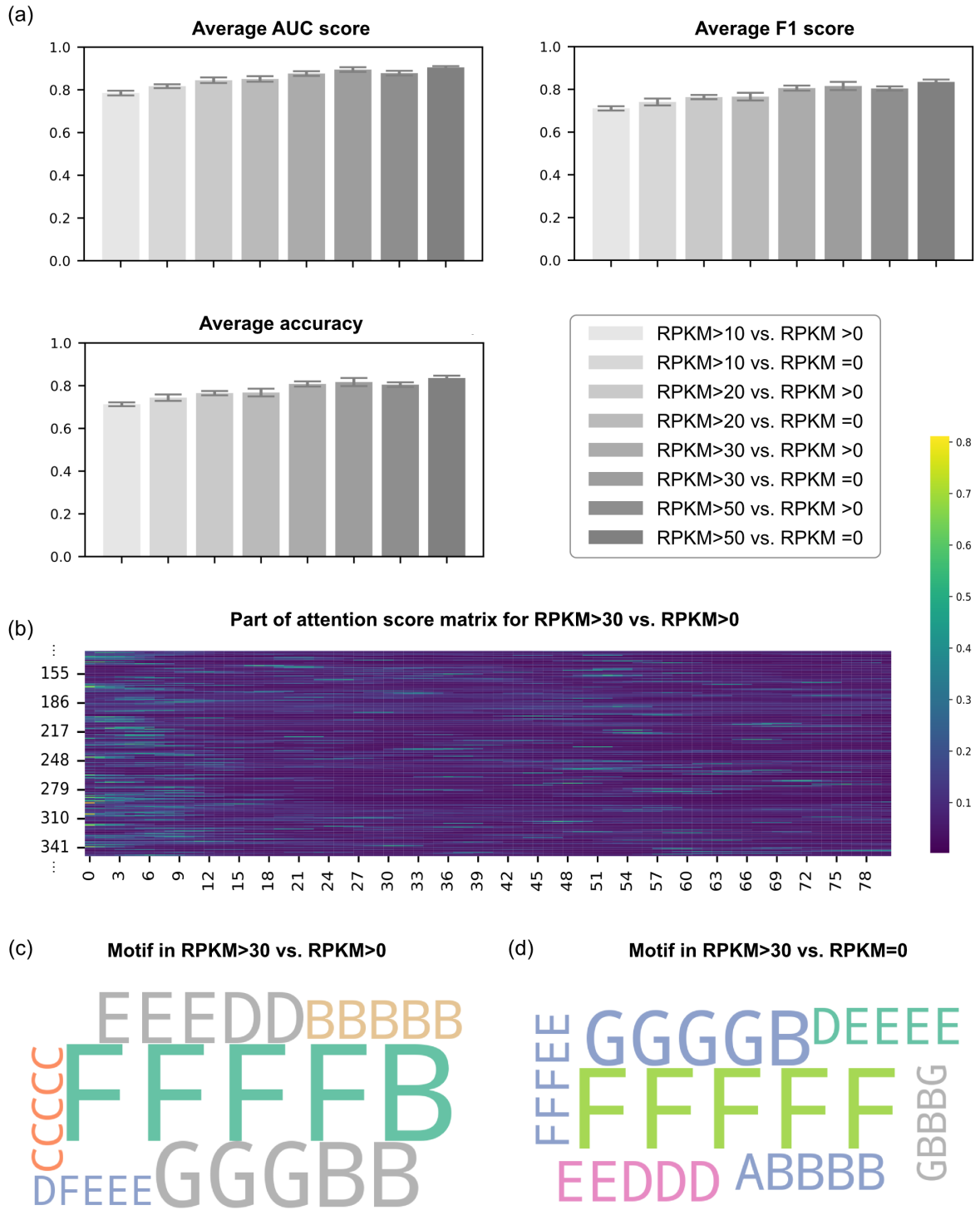

(a)
Average AUC score
Average F1 score
Average accuracy
RPKM>10 vs. RPKM >0
RPKM>10 vs. RPKM =0
RPKM>20 vs. RPKM >0
RPKM>20 vs. RPKM =0
RPKM>30 vs. RPKM >0
RPKM>30 vs. RPKM =0
RPKM>50 vs. RPKM >0
RPKM>50 vs. RPKM =0
(b)
Part of attention score matrix for RPKM>30 vs. RPKM>0
(c)
Motif in RPKM>30 vs. RPKM>0
(d)
Motif in RPKM>30 vs. RPKM=0

Figure 4

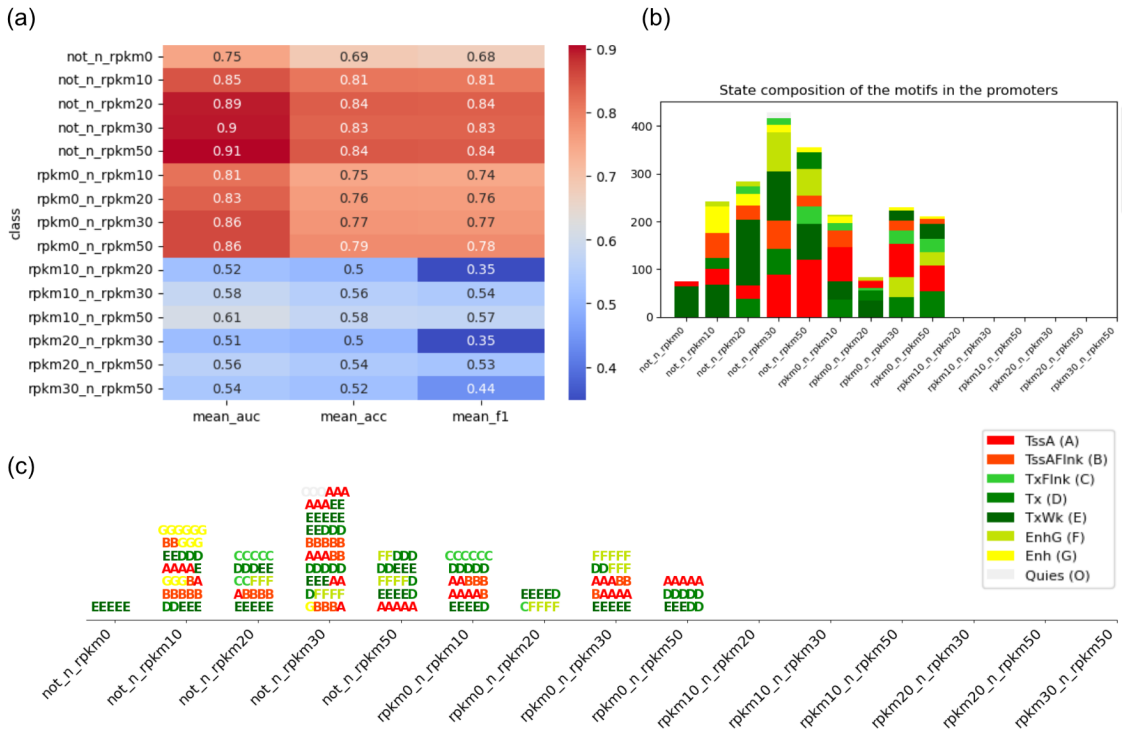
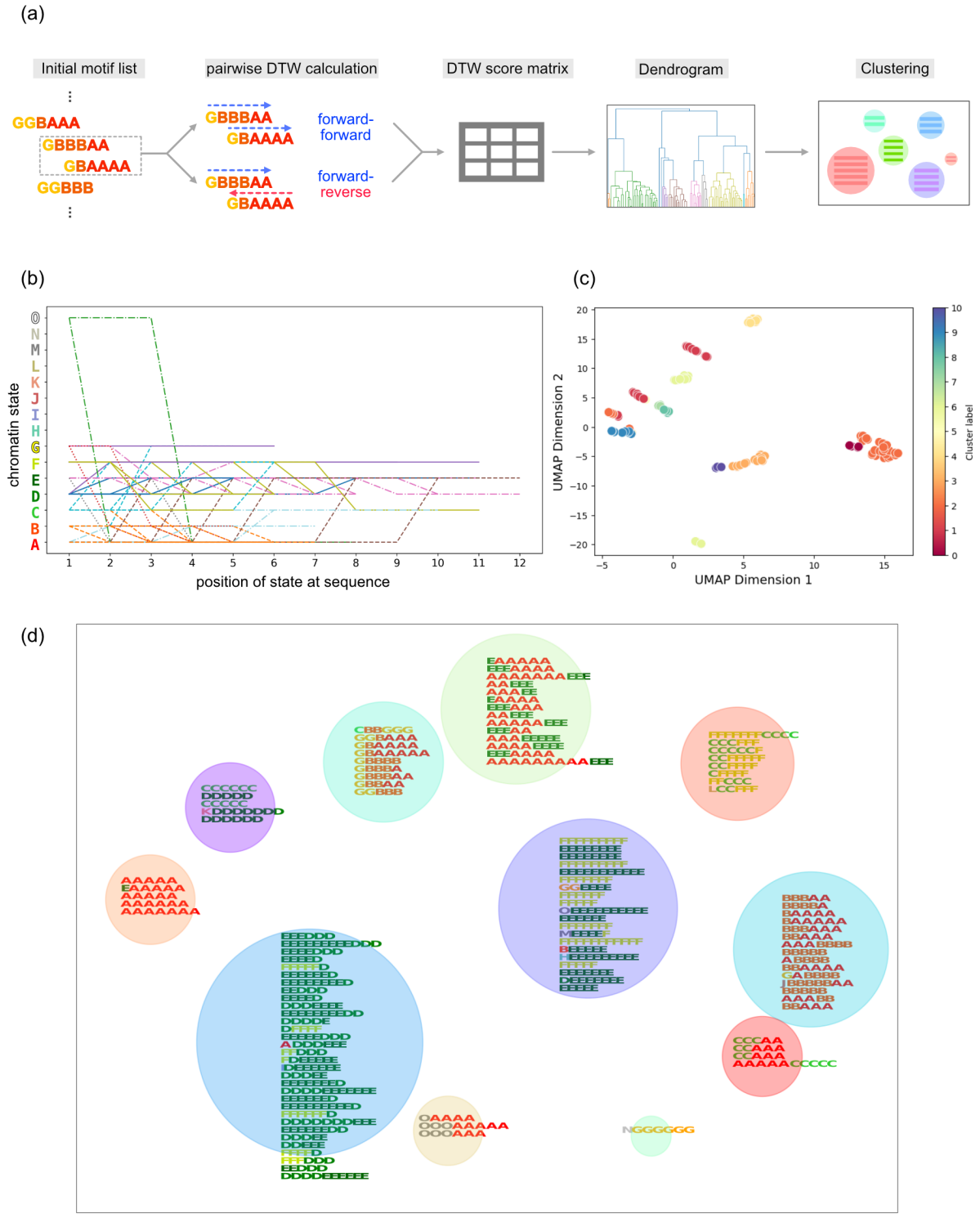

(a)
class
not_n_rpkm0 0.75 0.69 0.68
not_n_rpkm10 0.85 0.81 0.81
not_n_rpkm20 0.89 0.84 0.84
not_n_rpkm30 0.9 0.83 0.83
not_n_rpkm50 0.91 0.84 0.84
rpkm0_n_rpkm10 0.81 0.75 0.74
rpkm0_n_rpkm20 0.83 0.76 0.76
rpkm0_n_rpkm30 0.86 0.77 0.77
rpkm0_n_rpkm50 0.86 0.79 0.78
rpkm10_n_rpkm20 0.52 0.5 0.35
rpkm10_n_rpkm30 0.58 0.56 0.54
rpkm10_n_rpkm50 0.61 0.58 0.57
rpkm20_n_rpkm30 0.51 0.5 0.35
rpkm20_n_rpkm50 0.56 0.54 0.53
rpkm30_n_rpkm50 0.54 0.52 0.44
mean_auc
mean_acc
mean_f1
0.9
0.8
0.7
0.6
0.5
0.4
(b)
State composition of the motifs in the promoters
400
300
200
100
0
(c)
TssA (A)
TssAFlnk (B)
TxFlnk (C)
Tx (D)
TxWk (E)
EnhG (F)
Enh (G)
Quies (O)

Figure 5


(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)
chromatin state
position of state at sequence
(c)
UMAP Dimension 2
UMAP Dimension 1
Cluster label
(d)