Discriminative histone imputation using chromatin accessibility

Wen Wen[1], Jiaxin Zhong[1], Zhaoxi Zhang[1], Tinyi Chu[5], Nating Wang[2], Charles G. Danko[3,4] and Zhong Wang[1, *]

1 School of Software Technology, Dalian University of Technology, Dalian 116024, China, 2 Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA, 3 Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA, 4 Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA, 5 Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, NY 10021, USA.

*To whom correspondence should be addressed.

Abstract

Motivation: Histone modifications (HMs) play a crucial role in various biological processes, including transcription, replication, and DNA repair, and they have a significant impact on chromatin structure. However, annotating HMs across different cell types using experimental approaches alone is impractical due to cost and time constraints. Therefore, computational techniques are valuable for predicting HMs, complementing experimental methods, and enhancing our understanding of DNA characteristics and gene expression patterns.

Results: In this study, we introduce a deep learning model called discriminative histone imputation using chromatin accessibility (dHICA). dHICA combines extensive DNA sequence information with chromatin accessibility data, utilizing the Transformer architecture. This allows dHICA to have a large receptive field, and more cell-type specific information, which can not only incorporate distal information across the entire genome, but also improve predictive accuracy and interpretability. Comparing performance with other models, we found that dHICA exhibits superior performance in both peak calling and signal predictions, especially in biologically salient regions such as gene elements. dHICA serves as a crucial tool for advancing our understanding of chromatin dynamics, offering enhanced predictive capabilities and interpretability, particularly in critical biological regions such as gene elements.

Availability and implementation: https://github.com/wzhy2000/dHICA and https://dreg.dnasequence.org/ .

Contact: zhongwang@dlut.edu.cn.

1 Introduction

Histones constitute the predominant protein components within chromatin. In eukaryotic organisms, DNA is intricately coiled around octameric structures comprised of core histones, namely H2A, H2B, H3, and H4. The introduction of linker histone H1 further contributes to the assembly of intricate chromatin formations. It is noteworthy that post-translational modifications, including acetylation, phosphorylation, methylation, and ubiquitination, can occur at the N- and C-terminal tails of these histones, imparting an additional layer of complexity and regulation to chromatin structure and function. (Strahl and Allis, 2000). These post-translational modifications can alter the charge and structure of histone tails bound to DNA, inducing changes to chromatin packaging and gene expression (Kouzarides, 2007). HMs have been demonstrated to exert a pivotal influence on a myriad of cellular processes, such as gene expression regulation, chromatin structure modulation, and DNA repair(Bannister and Kouzarides, 2011; Stillman, 2018).

Histone modifications are mainly profiled by high-throughput experimental techniques such as chromatin immunoprecipitation and sequencing(O Geen, et al., 2011) (ChIP-seq) and Cleavage Under Targets and Release Using Nuclease(Skene, et al., 2017) (CUT&RUN). ChIP-seq is performed , using an antibody that recognizes a specific histone modifications or chromatin-bound protein, thus restricting its ability to comprehensively analyze multiple modifications or conditions concurrently. Therefore, relying solely on biological experiments to build a picture of HMs is resource-intensive and time-consuming. To address this challenge, researchers have turned to computational methods, particularly deep learning techniques, to predict histone modifications by measuring its interactions with other genomic and epigenetic signals. These approaches not only complement experimental methodologies but also improve our understanding of the fundamental principles of molecular biology that drive gene expression or other cellular processes.

To decipher the link between histone modifications and DNA sequence, Kelley trained Basenji2(Kelley, 2020), a deep learning model which predicts experimental histone modifications using DNA sequence. Enformer(Avsec, et al., 2021), advanced this approach by seamlessly integrating a transformer architecture into the convolutional block. It can process a larger range of DNA sequences using the self-attention mechanism and handle sequences that are longer than Basenji2 (131,072 base pairs).

However, while the DNA sequence encodes regulatory information for cell and tissue types, Enformer falls short of capturing the highly cell type and developmental stage-specific nature of histone modifications(Karollus, et al., 2023).

Considering this, Benreniste(Benveniste, et al., 2014), highlighted that transcription factor (TF) information leads to more accurate predictions of histone modifications compared to relying solely on DNA sequence data. dHIT(Wang, et al., 2022) demonstrated promising outcomes in predicting histone modifications from GRO-seq (Lopes, et al., 2017) data, showing the potential of leveraging single-assay, cell type-specific features to improve predictive accuracy. On the other hand, researchers also incorporated cell type-specific chromatin state differences into their models, giving rise to hybrid frameworks designed to enhance prediction accuracy. Dipankar et al. introduced DeepPTM(Baisya and Lonardi, 2021), a model that integrates transcription factor binding site information with DNA sequence to further elevate the precision of histone modification predictions. DeepHistone(Yin, et al., 2019), a deep learning architecture, successfully integrated DNA sequence information and chromatin accessibility data, enabling the simultaneous prediction of seven histone modifications. Nonetheless, DeepHistone's scope is limited to detecting histone marker modification types and does not extend to predicting modification levels.

Motivated by these insights, we introduce a deep learning framework named discriminative histone imputation using chromatin accessibility (dHICA) to predict multiple histone modification levels simultaneously by integrating DNA sequence data and chromatin accessibility. The incorporation of the Transformer structure (Vaswani, et al., 2017) into our model expands its receptive field, allowing for the identification of distant gene elements and the capture of their interactions. In this study, we demonstrate the exceptional performance of dHICA across multiple dimensions, including cross-cell and cross-species analyses. Furthermore, we compare dHICA with DeepSEA and DeepHistone, encompassing aspects like peak calling, signal prediction, and genetic components. The results consistently highlight dHICA's superior performance over DeepSEA and DeepHistone in all aspects, underlining its significant potential as an innovative tool in genomics research.

2 Materials and methods

2.1 Data collection

In this study, multiple HMs for different cell lines, as well as ATAC-seq and DNase-seq sequencing files, were obtained from ENCODE(ENCODE Project Consortium, 2004). Since the HM sequencing files used the human genome version GRCh37, we used LiftOver to convert any GRCh38 data to GRCh37(Luu, et al., 2020). See Table S1 and Table S2 for data details.

The HMs employed in this study include H3K122ac, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, H3K36me3, H3K9ac, H3K9me3, and H4K20me1. These modifications are known to serve as markers for specific functional elements such as enhancers, promoters, and gene bodies. For instance, H3K4me3 is commonly found near the transcription start site (TSS) of active genes along with H3K27ac. Both H3K4me1 and H3K27ac are frequently observed in active enhancer sequences (Zhu, et al., 2013) , while H3K27me3 and H3K9me3 are associated with transcriptional repression (Mikkelsen, et al., 2007).

In this study, the K562 cell line (chr1-21) was utilized for model training, while the chromosome 22 of the K562 cell line and other cell lines such as GM12878 and HCT116 were used for testing. Two separate optimal models were trained using DNase-seq and ATAC-seq data, respectively.

2.2 Data preprocessing

For a given cell line, the mitochondrial genome is removed, leaving only the autosomes, X and Y chromosome. ENCODE blacklist regions are also excluded across the entire genome to eliminate potential confounding factors(Amemiya, et al., 2019). Following this, a uniform sampling approach is employed using 114,688 bp as the base length. To provide more information to the model, each 114,688 bp interval is expanded on both sides to 196,608 bp, allowing for the retrieval of chromatin accessibility sequencing signals and DNA sequences corresponding to that interval. DNA sequences are read using a one-hot encoding scheme (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1], N = [0,0,0,0]). Simultaneously, chromatin accessibility information is directly extracted. In cases where the regions fall outside the defined correct boundaries, they are supplemented with 0 values. To retrieve the predicted histone modification information, a window size of 128 bp was used to read the 10 types of histone modification signals corresponding to 896 window regions centered within each interval. Within each 128 bp window, the sequencing data was subjected to a summation operation.

2.3 Design of dHICA

Fig. 1 depicts the architecture of our presented model dHICA, which combines chromatin accessibility data with DNA sequencing data for histone modification prediction. The model is composed of four parts: a convolutional layer with two distinct convolutional blocks for extracting DNA sequences and chromatin accessibility information, a Transformer block, a crop layer, and a fully connected layer for predicting 10 histone modifications. Two separate convolutional blocks in the first layer capture specific characteristics inherent to DNA sequences and chromatin accessibility. The combination of the extracted features from DNA sequences and chromatin accessibility feed to the Transformer block. The Transformer architecture is adept at capturing long-range dependencies and interactions, which are crucial in understanding how DNA sequences and chromatin accessibility jointly contribute to histone modifications.

The model takes one-hot encoded DNA sequences as input with a length of 196,608 bp. The convolutional layer extracted and incorporated the DNA sequencing data and its corresponding chromatin accessibility data. The convolutional blocks with pooling first reduce the spatial dimension from 196,608 bp to 1,536 so that each sequence position vector represents 128 bp (although the convolutions do observe nucleotides in the adjacent pooled regions). The two sets of features extracted by convolutional blocks were integrated via fully connected layers as the newly synthesized feature.

The Transformer block comprises 11 Transformer layers, which enable the transformation of each position in the input sequence by employing a weighted sum of representations from all other positions. The attention weights between

positions depend on their respective embeddings and their spatial proximity. This mechanism allows the model to improve predictions of HMs by incorporating information from relevant regions, such as enhancers involved in gene regulation. Through direct attention to all positions, the model facilitates effective information flow between distant elements. The integration of Transformer layers significantly expands the model's receptive field, enabling it to capture distal regulatory elements up to 100 kb away while maintaining efficient information integration for prediction purposes.

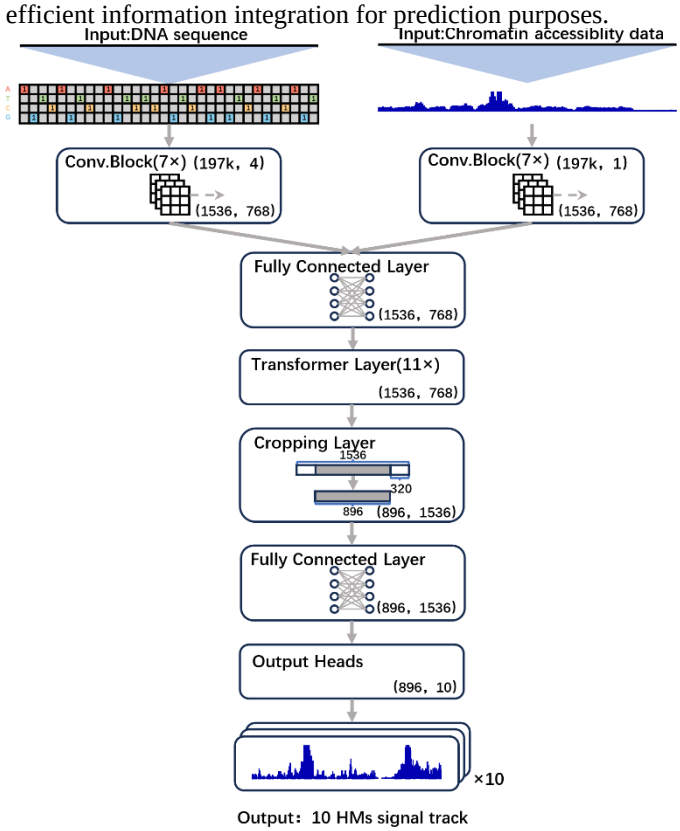

Fig. 1. dHICA model architecture

To mitigate the computational challenges associated with processing distant regions in the genomic sequence data, a cropping layer is utilized. This layer trims 320 positions from both ends of the interval, effectively shortening the sequence by 320×128 base pairs (bp). The justification for this cropping strategy lies in the inherent limitations of the model to effectively capture and learn from regions located far from the sequence center. This limitation arises from the model's inherent capacity to only perceive and analyze one side (towards the sequence center) of the regulatory elements, while lacking the ability to observe the other side (beyond the sequence boundaries). Finally, a fully connected layer is employed to predict a histone modification track with a resolution of 128 bp.

We utilize a supervised learning strategy to optimize our dHICA. The Mean Squared Error loss (MSELoss) function is adopted as the loss function.

2.4 Model training

To figure out which chromatin accessibility data is better suited for dHICA, we employed the same model architecture to train separate models for both ATAC-seq and DNase-seq data. The training dataset consists of chromosomes 1-20 from the K562 cell line, while chromosome 21 and chromosome 22 are designated as the validation and testing sets, respectively. To account for experimental factors such as batch effects and variation in data quality, four different chromatin accessibility datasets are read for the same interval partitions. This approach not only captures diverse experimental conditions but also expands the sample size by a factor of four. The final number of samples used for training, validation, and testing are 87,868, 1,472, and 1,568, respectively.

The MSE loss function is employed, with a learning rate of 0.0001. A learning rate decay strategy is implemented, dividing the learning rate by 1.4 every 5 rounds after 10 epochs. The optimal model is obtained after 300 epochs, with each training round consisting of 1,500 samples.

2.5 Evaluation of model performance

The assessment revolves around quantifying the correlation between predicted results generated by the dHICA model and the corresponding experimental data to gauge model performance. A higher correlation indicates a higher degree of accuracy in the predictions. The dHICA model was employed to predict the histone modification information across the entire genome track. Pearson and Spearman correlation coefficients were calculated using sliding windows of 128 bp, 1000 bp, and 10000 bp, respectively. By assessing correlations at different resolutions, we can gain a comprehensive understanding of the model's performance across various scales and evaluate its ability to capture the underlying patterns and trends in the data. These correlation coefficients act as crucial metrics for evaluating the effectiveness of the model.

The predictions of HMs, with a resolution of 128 bp, will be stored in the bigWig format, allowing for intuitive visualization using Integrative Genomics Viewer (IGV) to check the predicted signals (Thorvaldsdóttir, et al., 2013). Additionally, that resolution facilitates the calculation of correlation coefficients between the predicted results and experimental data at different resolutions, aiding in the evaluation of model performance.

3 Results

3.1 Prediction of HMs across different cell lines

dHICA exhibit excellent performance in human cell lines (Fig. 2). The imputed HMs using dHICA demonstrate a strong correlation with experimental data. Notably, the differences between the imputed and experimental signals were often observed in background regions with low signal intensity, indicating the presence of technical variation in the ChIP-seq background signal that were not accounted for by chromatin accessibility data.

We evaluated the generalization capacity of our method to new cell lines by training dHICA on K562 and testing it on other cell lines. In Fig. 3 and Fig. S1, we summarized dHICA's performance across cell lines, presenting Pearson correlation heatmaps for ATAC and DNase models at a 1000 bp resolution. Both ATAC-seq and DNase-seq assays are key in assessing chromatin accessibility, providing strong signals for accessible regions like promoters and enhancers but encountering challenges in closed or suppressed regions(Buenrostro, et al., 2013; Corces, et al., 2016). Aligned with biological

expectations, dHICA excels in predicting open chromatin regions linked to promoters and enhancers (H3K4me3, H3K27ac and H3K9ac), with a slightly reduced performance near repressive regions (H3K27me3 and H3K36me3). Nevertheless, the model showcases robust generalization across all histone marks, emphasizing its effectiveness in diverse chromatin contexts.

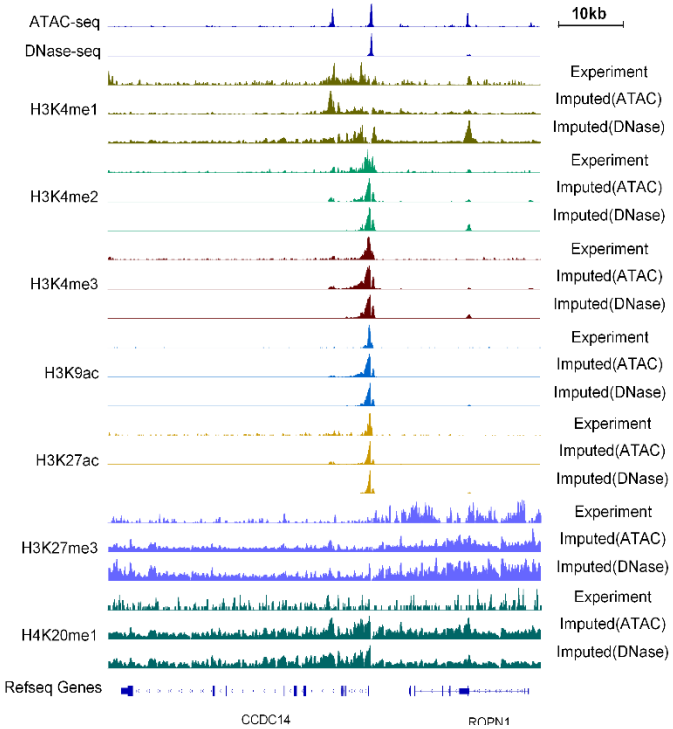

Fig. 2. Genome browser comparison between experimental and predicted histone marks near gene CCDC14 and ROPN1 in GM12878

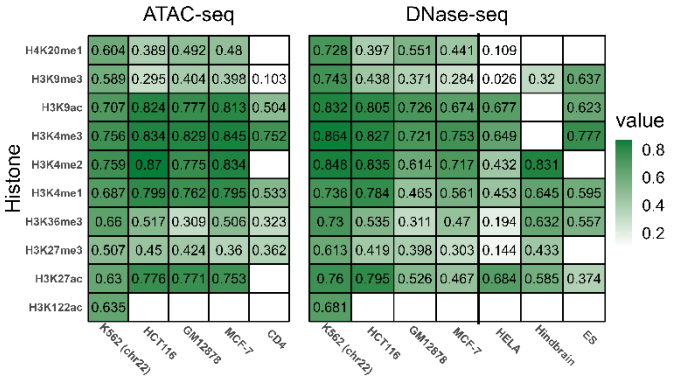

Fig. 3. Comparison of Pearson correlation of ATAC and DNase models across cell lines and species. Empty cells indicate that no experimental data are available for comparison in the cell type shown.

3.2 Performance of the DNA and sequencing data

To assess the significance of DNA information and chromatin accessibility data in the model, we conducted experiments using different combinations of data. Specifically, we evaluated the model's performance using only DNA information, only ATAC-seq data, only DNase-seq data, as well as combinations of DNA and ATAC-seq data, and DNA and DNase-seq data.

As illustrated in Fig. 4, the model's Pearson correlation is consistently higher when utilizing both DNA and chromatin accessibility data, compared to using either component alone. In general, for all HMs, the performance of using chromatin accessibility data alone outweighed that of DNA sequence information, which can be attributed to the fact that chromatin accessibility data offer cell-type specific information, facilitating more accurate predictions. The model's predictive performance is better when using ATAC-seq data as input compared to using DNase-seq data. In contrast, as DNA sequences are shared across all cells, they provide less discriminatory power. These findings underscore the greater contribution of chromatin accessibility information to the predictive performance of the dHICA, and combining both types of information yields further improvements in predictive accuracy.

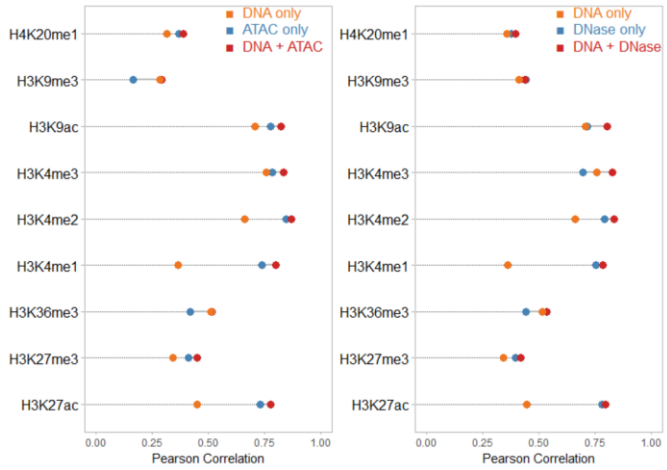

Fig. 4. Pearson correlation of dHICA using DNA or chromatin accessibility data in HCT116(resolution 1000bp)

Notably, for HMs closely associated with promoters and enhancers, including H3K4me3, H3K27ac, and H3K4me1, the performance using chromatin accessibility signals alone significantly surpassed that of DNA sequence information, exhibiting a more pronounced disparity compared to other HMs. Conversely, for HMs with weaker associations to promoters and enhancers, although chromatin accessibility signals still contributed more significantly, the disparity in performance with DNA sequence information alone was less pronounced. This discrepancy may arise from the stronger DNA sequence signals in proximity to promoters and enhancers, such as TATA box, and CpG island, which lack cell type specificity(Carninci, et al., 2006). Conversely, the chromatin accessibility signal furnishes the model with cell-type-specific information and consequently enhances its performance.

3.3 Correlations of HMs near gene element

Informed by extensive biological studies, compelling evidence highlights the intricate interplay of histone modifications, sculpting a diverse chromatin landscape that orchestrates gene expression within specific cellular contexts (Wang et al., 2008; Kouzarides, 2007). While our primary focus remains on genome-wide predictions, the quantitative

abundance of histone modifications in transcriptional regulatory elements (TREs) also holds significant importance. To address this, we evaluated the model's performance in the vicinity TREs. Transcriptional regulatory element intervals were obtained from ENCODE, and the model predicted histone modification signals within these intervals. Pearson correlation coefficients were then computed between the predicted signals and experimental data.

Table 1. Pearson Correlation of dHICA using ATAC-seq and DNase-seq across transcriptional regulatory elements in HCT116.

Histones	Promoter		Enhancer		Insulator		Gene Body		PolyA		Genome-wide	
	ATAC	DNase	ATAC	DNase	ATAC	DNase	ATAC	DNase	ATAC	DNase	ATAC	DNase
H3K4me1	0.843	0.825	0.848	0.834	0.811	0.800	0.821	0.811	0.791	0.798	0.799	0.784
H3K4me2	0.877	0.839	0.890	0.856	0.895	0.852	0.904	0.885	0.898	0.885	0.870	0.835
H3K4me3	0.845	0.823	0.859	0.835	0.836	0.809	0.883	0.869	0.879	0.871	0.834	0.827
H3K27ac	0.803	0.826	0.821	0.844	0.791	0.812	0.847	0.860	0.856	0.866	0.776	0.795
H3K27me3	0.415	0.353	0.379	0.324	0.364	0.297	0.385	0.368	0.285	0.320	0.450	0.419
H3K36me3	0.484	0.500	0.501	0.512	0.338	0.347	0.531	0.557	0.514	0.525	0.517	0.535
H3K9ac	0.823	0.790	0.842	0.810	0.834	0.792	0.871	0.847	0.873	0.854	0.824	0.805
H3K9me3	0.056	0.098	0.019	0.131	0.057	0.179	0.011	0.117	0.024	0.109	0.295	0.438
H4K20me1	0.270	0.315	0.289	0.310	0.079	0.037	0.117	0.124	0.142	0.107	0.389	0.397
Average	0.602	0.596	0.605	0.606	0.556	0.547	0.597	0.604	0.585	0.593	0.639	0.648

The results presented in Table 1, Table S3 and Table S4, which include data for HCT116 and GM12878, illustrated that for histone modifications associated with promoters and enhancers, such as H3K4me1, H3K4me3, and H3K27ac, the predictive performance near TREs exceeded that of predictions made across the entire genome. This superiority is particularly prominent in regions closely associated with the corresponding TREs. For instance, H3K4me1 (ATAC-seq, HCT116), predominantly found near enhancers, exhibited consistently superior predictive performance within all transcriptional regulatory element intervals. In contrast, HMs associated with repressive chromatin (such as H3K27me3 and H3K9me3) showed lower predictive performance near transcriptional regulatory elements compared to genome-wide predictions.

Notably, despite the generally stronger signal in promoter regions compared to enhancer regions, the predictive performance of H3K4me1 was more favorable in the enhancer region due to its higher signal intensity in enhancers. Additionally, regardless of the transcriptional regulatory element type, HMs highly associated with promoters and enhancers consistently outperformed those associated with repressive marks. This was evident in both ATAC and DNase predictions, where H3K27ac (associated with promoters and active enhancers) exhibited substantially higher predictive performance compared to H3K27me3 (associated with domains of transcriptional repression).

Furthermore, to gain a clearer understanding of the differences in HM signals, we plotted their enrichment near the TSS (Fig. S2). It's obvious that the shapes of the enrichment plots for each histone modification are quite similar, but there is a noticeable difference in signal values. However, this disparity is not only present between experimental data and dHICA model predictions but is also evident across experimental data from different cell lines. For instance, the signal enrichment near the TSS in the GM12878 cell line is only half that of K562, a situation likely associated with variations in the measurement processes during biological experiments.

3.4 Chromatin states imputed by ChromHMM

In light of the robust performance of dHICA in the vicinity of TREs, we employed the dHICA to impute chromatin states, which defined by ChromHMM (Ernst and Kellis, 2012). Remarkably, the chromatin states inferred from dHICA's imputed data exhibited substantial concordance with those delineated by direct ChIP-seq data (Fig. 5). This underscores dHICA's capability to accurately reflect the underlying epigenetic landscape, as it successfully extrapolates complex chromatin configurations from integrated datasets. The consistency of the imputed states with those derived from ENCODE data suggests that dHICA can be a viable alternative for predicting chromatin states, particularly in contexts where ChIP-seq data is unavailable.

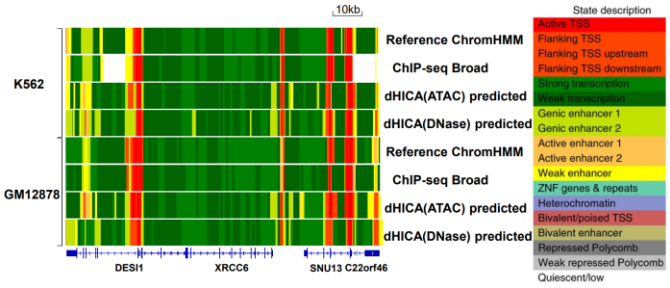

Fig. 5. Genome browser in K562 and GM12878 cells shows the 18-state ChromHMM model using ChIP-seq data used to train the model (Broad) or based on imputation (dHICA predicted)

3.5 Prediction of HMs in mouse

To assess the performance of the model in cross-species prediction of HMs, we utilized a DNase model trained in human cells to predict HMs in mouse's hindbrain and embryonic stem (ES) cells. The data preprocessing was consistent with the human genome. Table S5 and Table S6 present the correlation coefficients between the predicted and experimental data, computed at a resolution of 10k bp.

The DNase model achieved a Pearson correlation coefficient of 0.651 for predicting results in the HCT116 cell line, while it achieved a performance of 0.517 in predicting results in mouse ES cells. The relatively small difference between these two correlations suggests that the model exhibits good performance in cross species prediction. This suggests that the model effectively predicts histone modification patterns across species, opening up significant

potential applications in annotating genomes of less-explored mammalian species.

3.6 Prediction comparison with other methods

We compare dHICA to other two methods, DeepSEA (Zhou and Troyanskaya, 2015) and DeepHistone (Yin, et al., 2019). There's a fundamental diference in the windows between baselines. For both the DeepSEA and DeepHistone models, the genome was splited into 200-bp bins and subsequently extended to 1000 bp. The focus was exclusively on the set of 200-bp bins exhibiting at least one TF binding event or histone modification, resulting in coverages of 17% and 60% of the entire genome, respectively. In contrast, dHICA focuses on the entire genome, with a receptive field of 197 kb. Consequently, suitable window partitioning becomes essential for model comparisons.

Given the disparity in interval lengths between other baselines (1,000 bp) and dHICA (197 kb), we selected the H1 ES cell line for model comparison. Initially, we identified 2,606 dense histone modification windows spanning 197 kb each, serving as the dataset for dHICA's five-fold cross-validation. Then these windows were further partitioned into 1,000 bp segments, utilized as samples for other models. Leveraging the web services provided by DeepSEA, we directly incorporated its predictions for specific windows. In the case of deepHistone, we employed the further partitioned 1,000 bp windows as data for the five-fold cross-validation. Since DeepHitone combined DNA sequence and DNase-seq for HMs predictions, we choose our DNase model for comparsion.

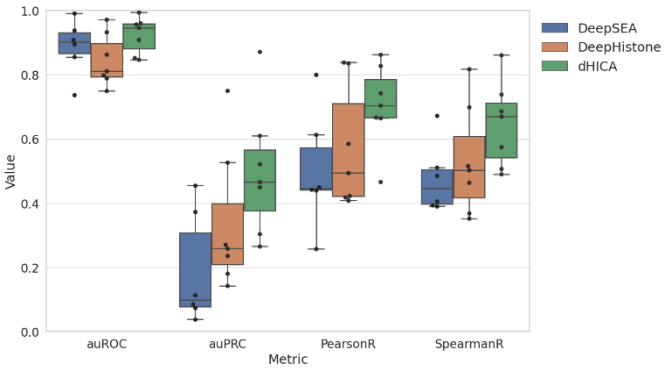

Fig. 6. Comparison of performance between dHICA and other models.

We compare the performance of DeepSEA, DeepHistone, and dHICA in H1 ES cells, as shown in Fig. 6. Comparing the prediction results, we note that models (DeepHistone and dHICA) exhibit higher performance when utilizing both DNA and DNase-seq data compared to the model (DeepSEA) relying solely on DNA sequence. Building upon the earlier discussions, the integration of chromatin accessibility sequencing data proves advantageous for predicting histone modifications. Moreover, we found that dHICA surpasses DeepHistone, acknowledging the limitation that DeepHistone solely employs CNN for predicting histone modifications, while dHICA incorporates a transformer block to capture long-range dependencies and interactions. This addition is pivotal for comprehending how DNA sequences and chromatin accessibility collaboratively contribute to histone modifications, consequently enhancing performance.

Our evaluation focuses not just on the model's genome-wide predictive capability, but with particular scrutiny on its accuracy around promoters and enhancers. In these critical genomic regions, dHICA demonstrates superior predictive performance when compared with other models (Table S7 and Table S8). For a visual comparison, we used deepTools (Ramírez, et al., 2014) to generate heatmaps that delineate signal enrichment of dHICA against other models in proximity to TSSs (Fig. 7). we observe that dHICA's predicted signals closely match the target, with enhanced signal intensity and no false positives, as indicated by the consistent and deep coloration in the heatmaps. In contrast, DeepSEA shows notable inaccuracies, including false positives at the TSS center, leading to an overestimation of signal intensity. DeepHistone's predictions are less pronounced for the H3K27ac marker, underrepresenting the signal intensity relative to experimental data. Such performance disparities likely arise from the broad receptive field of dHICA, which skillfully ensnares the dynamics between distal TREs, endowing dHICA with the finesse to tease out subtle features and thus sharpen its predictive edge near TREs.

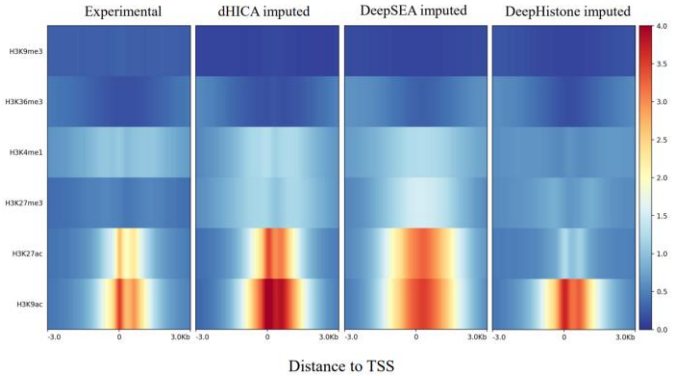

Fig. 7. Comparison of enrichment between dHICA and other models near TSSs.

4 Discussion and conclusion

In this study, we presented dHICA, a deep learning framework that integrates chromatin accessibility information and DNA sequences to accurately predict histone modification signals. By incorporating the transformer structure alongside traditional convolution, dHICA significantly expands the model's receptive field, enabling it to capture long-range interactions between genomic elements. The results of our comprehensive comparisons reveal several key findings: (1) Chromatin accessibility information contributes more significantly to histone modification signal prediction than DNA sequences; (2) dHICA demonstrates superior predictive performance near transcriptional regulatory elements compared to the entire genome; (3) dHICA outperforms in predicting histone modifications associated with promoter and enhancer correlation compared to those related to repressed chromatin; (4) The integration of DNA sequences and chromatin

accessibility information enhances dHICA's prediction accuracy, surpassing the DeepHistone and DeepSEA models, particularly in promoters and enhancers(with the caveats listed in Section 3.6).

Moreover, we plan to broaden dHICA's utility by adapting it to predict diverse types of sequencing data, such as MNase-seq, which profiles nucleosome positioning, and CUT&Tag, which characterizes protein-DNA interactions with high resolution. The expansion into these methodologies promises to deepen our comprehension of chromatin organization and its implications for cellular function.

Additionally, considering the consistent correlation patterns among histone modifications observed across different cell types(Singh, et al., 2016), we are inspired to integrate multi-cell line sequencing data into our model. This integration aims to diversify the training data, thereby enhancing the robustness and versatility of dHICA. By doing so, we anticipate providing a more comprehensive landscape of histone modification patterns, which will be invaluable for cross-cell type comparisons and for understanding the epigenetic underpinnings of cell identity and transformation.

Lastly, the exploration of dHICA's potential in the context of machine learning interpretability stands as a compelling avenue for future research. By delving into the model's decision-making processes, we can potentially uncover novel insights into the genomic features that drive histone modifications, further bridging the gap between predictive performance and biological significance. This could lead to breakthroughs in identifying new biomarkers for diseases and understanding the complex mechanisms underlying various biological processes and disorders.

Funding

This work was supported by the LiaoNing Revitalization Talents Program (No. XLYC2002010) and the Fundamental Research Funds for the Central Universities (No.DUT20RC(3)074).

References

Amemiya, H.M., Kundaje, A. and Boyle, A.P. (2019) The ENCODE Blacklist: Identification of Problematic Regions of the Genome, *SCI REP-UK*, **9**, 9354.

Avsec, }., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P. and Kelley, D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions, *NAT METHODS*, **18**, 1196-1203.

Baisya, D.R. and Lonardi, S. (2021) Prediction of histone post-translational modifications using deep learning, *BIOINFORMATICS*, **36**, 5610-5617.

Bannister, A.J. and Kouzarides, T. (2011) Regulation of chromatin by histone modifications, *CELL RES*, **21**, 381-395.

Benveniste, D., Sonntag, H., Sanguinetti, G. and Sproul, D. (2014) Transcription factor binding predicts histone modifications in human cell lines, *Proceedings of the National Academy of Sciences*, **111**, 13367-13372.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position, *NAT METHODS*, **10**, 1213-1218.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., Forrest, A.R.R., Alkema, W.B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S.M., Wells, C.A., Orlando, V., Wahlestedt, C., Liu, E.T., Harbers, M., Kawai, J., Bajic, V.B., Hume, D.A. and Hayashizaki, Y. (2006) Genome-wide analysis of mammalian promoter architecture and evolution, *NAT GENET*, **38**, 626-635.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., Majeti, R. and Chang, H.Y. (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution, *NAT GENET*, **48**, 1193-1203.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project, *SCIENCE*, **306**, 636-640.

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

Karollus, A., Mauermeier, T. and Gagneur, J. (2023) Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers, *GENOME BIOL*, **24**, 56.

Kelley, D.R. (2020) Cross-species regulatory sequence activity prediction, *PLOS COMPUT BIOL*, **16**, e1008050.

Kouzarides, T. (2007) Chromatin Modifications and Their Function, *CELL*, **128**, 693-705.

Lopes, R., Agami, R. and Korkmaz, G. (2017) GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression. In Napoli, S. (ed), *Promoter Associated RNA: Methods and Protocols*. Springer New York, New York, NY, 45-55.

Luu, P., Ong, P., Dinh, T. and Clark, S.J. (2020) Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data, *NAR GENOM BIOINFORM*, **2**, a54.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T., Koche, R.P., Lee, W., Mendenhall, E., O Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S. and Bernstein, B.E. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *NATURE*, **448**, 553-560.

O Geen, H., Echipare, L. and Farnham, P.J. (2011) Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. In Tollefsbol, T.O. (ed), *Epigenetics Protocols*. Humana Press, Totowa, NJ, 265-286.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. and Manke, T. (2014) deepTools: a flexible platform for exploring deep-sequencing data, *NUCLEIC ACIDS RES*, **42**, W187-W191.

Singh, R., Lanchantin, J., Robins, G. and Qi, Y. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications, *BIOINFORMATICS*, **32**, i639-i648.

Skene, P.J., Henikoff, S. and Reinberg, D. (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites, *ELIFE*, **6**, e21856.

Stillman, B. (2018) Histone Modifications: Insights into Their Influence on Gene Expression, *CELL*, **175**, 6-9.

Strahl, B.D. and Allis, C.D. (2000) The language of covalent histone modifications, *NATURE*, **403**, 41-45.

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *BRIEF BIOINFORM*, **14**, 178-192.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, A. and Polosukhin, I. (2017) Attention is all

you need, *Advances in neural information processing systems*, **30**.

Wang, Z., Chivu, A.G., Choate, L.A., Rice, E.J., Miller, D.C., Chu, T., Chou, S., Kingsley, N.B., Petersen, J.L., Finno, C.J., Bellone, R.R., Antczak, D.F., Lis, J.T. and Danko, C.G. (2022) Prediction of histone post-translational modification patterns based on nascent transcription data, *NAT GENET*, **54**, 295-305.

Yin, Q., Wu, M., Liu, Q., Lv, H. and Jiang, R. (2019) DeepHistone: a deep learning approach to predicting histone modifications, *BMC GENOMICS*, **20**, 193.

Zhou, J. and Troyanskaya, O.G. (2015) Predicting effects of noncoding variants with deep learning－based sequence model, *NAT METHODS*, **12**, 931-934.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., Bennett, D.A., Houmard, J.A., Muoio, D.M., Onder, T.T., Camahort, R., Cowan, C.A., Meissner, A., Epstein, C.B., Shoresh, N. and Bernstein, B.E. (2013) Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues, *CELL*, **152**, 642-654.