Briefings in Bioinformatics, 2024, **25(6)**, bbae459
https://doi.org/10.1093/bib/bbae459
Problem Solving Protocol

dHICA: a deep transformer-based model enables accurate histone imputation from chromatin accessibility

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

[1]School of Software Technology, Dalian University of Technology, Linggong Rd, Liaoning 116024, China
[2]Meinig School of Biomedical Engineering, Cornell University, Weill Hall, Ithaca, NY 14853, United States
[3]Department of Molecular Biology and Genetics, Cornell University, Biotechnology Building, Ithaca, NY 14853, United States
[4]Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Hungerford Hill Rd, Ithaca, NY 14853, United States
[5]Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Tower Rd, Ithaca, NY 14853, United States
[*]Corresponding author. School of Software Technology, Dalian University of Technology, Linggong Rd, Liaoning 116024, China. E-mail: zhongwang@dlut.edu.cn

Abstract

Histone modifications (HMs) are pivotal in various biological processes, including transcription, replication, and DNA repair, significantly impacting chromatin structure. These modifications underpin the molecular mechanisms of cell-type-specific gene expression and complex diseases. However, annotating HMs across different cell types solely using experimental approaches is impractical due to cost and time constraints. Herein, we present dHICA (deep histone imputation using chromatin accessibility), a novel deep learning framework that integrates DNA sequences and chromatin accessibility data to predict multiple HM tracks. Employing the transformer architecture alongside dilated convolutions, dHICA boasts an extensive receptive field and captures more cell-type-specific information. dHICA outperforms state-of-the-art baselines and achieves superior performance in cell-type-specific loci and gene elements, aligning with biological expectations. Furthermore, dHICA's imputations hold significant potential for downstream applications, including chromatin state segmentation and elucidating the functional implications of SNPs (Single Nucleotide Polymorphisms). In conclusion, dHICA serves as a valuable tool for advancing the understanding of chromatin dynamics, offering enhanced predictive capabilities and interpretability.

Keywords: cell-type-specific; histone modifications; histone modification prediction; transformer; deep learning

Introduction

At the core of chromatin architecture are the highly conserved histone proteins—H1, H2A, H2B, H3, and H4—which serve as fundamental building blocks for packaging eukaryotic DNA into repetitive nucleosomal units [1]. These units are subsequently folded into higher-order chromatin fibers [2]. Histone modifications (HMs) significantly influence a broad range of cellular processes, including gene expression, chromatin structure modulation, and DNA repair [3]. To elucidate the genome-wide signals associated with different cell types and tissues, initiatives such as the Encyclopedia of DNA Elements (ENCODE) [4, 5] and the Roadmap Epigenomics Consortiums [6] have made substantial strides in systematically characterizing *in vivo* biochemical signatures across different cell types and tissues, including HMs, chromatin accessibility, and DNA methylation.

Despite these efforts to comprehensively map the epigenomes, significant challenges remain. Due to the high costs and time-consuming nature of experimental work, data have only been collected for a fraction of potential cell type and assay combinations outlined in these projects. Furthermore, considering the myriad developmental stages and environmental conditions, the diversity of possible human cell types is virtually boundless. It is impractical to anticipate collecting exhaustive data for every potential cell type/assay combination. Furthermore, no high-throughput assay is perfectly reproducible, and run-to-run differences in the same experiment may reflect either biological variation in the cells being assayed or experimental variance arising from sample preparation or downstream steps in the protocol.

As a practical solution, the development of *in silico* models to impute unknown epigenomic profiles based on existing data offers a promising alternative to these experimental limitations. Epigenomic imputation methods such as ChromImpute [7] and PREDICTD [8] have been presented to use available data to accurately impute the outcomes of missing experiments, thereby extending our understanding of epigenomic regulation across a more comprehensive spectrum of cell types and conditions. Alongside imputing missing data, Avocado [9] has produced a dense and information-rich representation of the human epigenomes, reducing redundancy, noise, and bias.

Considering the extensive range of tissues and HMs, relying solely on biological experiments to explore underlying mechanisms is somewhat unrealistic. Certain studies have been conducted to identify HM peaks, helping researchers to focus on regions more closely associated with regulatory effects on gene expression. For instance, DeepSEA [10] models regulatory information encoded by the DNA sequence to predict a wide array of epigenomics data, including TF-binding, DNase I sensitivity,

Received: May 16, 2024. **Revised:** July 13, 2024. **Accepted:** September 4, 2024

and HM sites. Both DeepHistone [11] and iHMnBS [12] combine DNA sequence and DNase-seq data to classify multiple HM peaks. Meanwhile, DeepPTM [13] uses TF-binding data and DNA sequences to predict histone posttranslational modifications; however, it is limited to predicting only a single modification marker in the center of the sequence for a given cell line. These methods are trained only in regions with HM peaks, not genome-wide, which may prevent them from capturing features across the entire genome.

Several models have been developed to predict genome-wide, cell-type-specific epigenetic, and transcriptional profiles in large mammalian genomes. For instance, Kelley developed Basenji2 [14], a deep learning model that predicts experimental HMs using DNA sequence alone. Enformer [15] advanced this approach by integrating a transformer architecture into the convolutional blocks, allowing it to process a more extensive range (197-kbp) of DNA sequences through the self-attention mechanism, handling longer sequences than Basenji2 (131-kbp).

However, while DNA sequence encodes regulatory information for cells and tissue types, Enformer falls short in capturing the highly cell-type and developmental stage-specific nature of HMs [16]. dHIT [17] has shown promising outcomes in predicting HMs from GRO-seq data, demonstrating the potential of leveraging single-assay, cell-type-specific features to enhance predictive accuracy. EPCOT [18] incorporates DNase-seq and DNA sequence to predict HM tracks for a given cell type, utilizing a pretraining and fine-tuning framework. However, the resolution of EPCOT (1000 bp) is significantly lower than that of Enformer and Basenji2, both of which are 128 bp, far underperforming the popular chromatin segmentation method (ChromHMM [19]), which uses 200-bp resolution. It should be noted that achieving higher resolution is crucial for various downstream applications, including chromatin segmentation.

Motivated by these insights, we introduce deep Histone Imputation using Chromatin Accessibility (dHICA) to simultaneously predict multiple HM levels using DNA sequence data and chromatin accessibility as inputs. More importantly, incorporating the transformer architecture [20] into our model expands the model's receptive field, allowing it to capture distal information. In cross-cell line and species evaluations, dHICA outperformed other state-of-the-art methods, primarily due to its effective integration of chromatin accessibility data. Chromatin accessibility data, particularly active marks, are crucial for model predictions. Unlike models that depend solely on DNA, dHICA can predict HMs in new cell lines and species without the need for re-training. Furthermore, dHICA's imputed data can be utilized for downstream applications, including segmenting chromatin states and distinguishing histone acetylation quantitative trait loci (haQTLs) from SNPs. dHICA's robust performance and versatility highlight its potential as an innovative tool in genomic research.

Materials and methods

Dataset

This study sourced multiple HMs and sequencing files from ENCODE for ATAC-seq and DNase-seq. The HMs utilized included H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, H3K36me3, H3K9ac, H3K9me3, and H4K20me1. These markers indicate specific functional elements such as enhancers, promoters, and gene bodies. This study also explored less commonly studied modifications such as H3 lysine 122 acetylation (H3K122ac). For the experimental setup, the K562 cell line (chromosomes 1-21) was used for model training and validation. Chromosome 22 of the K562 and other cell lines, such as GM12878 and HCT116, served as the testing ground. Two separate models were optimally trained using DNase-seq and ATAC-seq data, respectively. Tables S2–S5 in the Supplementary material provide further details on the data used.

Data preprocessing

This study excludes mitochondrial DNA from the analysis, focusing only on the autosomes and sex chromosomes. To minimize potential confounding factors, ENCODE blacklist regions are also omitted from the entire genome [21]. To avoid assembly gaps and unmappable regions that are more significant than 1 kb, we extract 131-kb non-overlapping sequences across the chromosomes, which expand to 197 kb on both sides as model inputs. This procedure yielded 22727 intervals for extracting DNA sequences and chromatin accessibility data.

DNA sequences are read using a one-hot encoding scheme, where A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1], and N = [0,0,0,0]. Chromatin accessibility data are extracted directly from fold change bigwig files, and we ensure data integrity by setting any negative or NaN values to zero, without applying any further data transformation. For predicting HM signals, within each interval, we summed coverage estimates in a bin with a length of 128 bp to serve as the signal for the model to predict.

To mitigate the influence of experimental factors such as batch effects and data quality variations, four distinct chromatin accessibility datasets are utilized for the same set of interval partitions. This method not only captures the variability across different experimental conditions but also effectively quadruples the sample size, thereby enhancing the robustness of the data processing. Consequently, the assembled dataset encompasses a comprehensive total of 90908 samples (exactly four-fold the base count of 22727 intervals), with 87868 samples designated for training, 1472 for validation, and 1568 for testing, thereby guaranteeing extensive coverage and reliable model evaluation.

Architecture of the dHICA

Our proposed model, dHICA, builds upon the foundation of Enformer and integrates chromatin accessibility data with DNA sequence to enhance the prediction of HMs, as illustrated in Fig. 1A. The architecture comprises a sophisticated sequence of layers designed for optimal data processing and prediction accuracy. It starts with two separate convolutional blocks, one dedicated to DNA sequences and the other to chromatin accessibility data. Each block is tailored to extract the specific characteristics of its input data type. These features are then processed through a fusion layer, which prepares them for the next critical phase. The Transformer block, a pivotal model component, excels in capturing long-range dependencies and interactions between the DNA sequences and chromatin accessibility, which is crucial for understanding their combined influence on HMs. The processing sequence concludes with a cropping layer followed by a fully connected layer, which together refine and predict the 10 specific types of HMs. This integrated approach ensures that dHICA not only captures the unique aspects of each data but also effectively interprets the complex interdependencies that dictate HM patterns.

Convolutional blocks and fusion layer

We applied convolutional blocks with pooling to distill the input data, DNA and chromatin accessibility data, into fixed-size representations. Precisely, the model dissects the input sequences into 128-base pair bins to achieve the desired resolution.

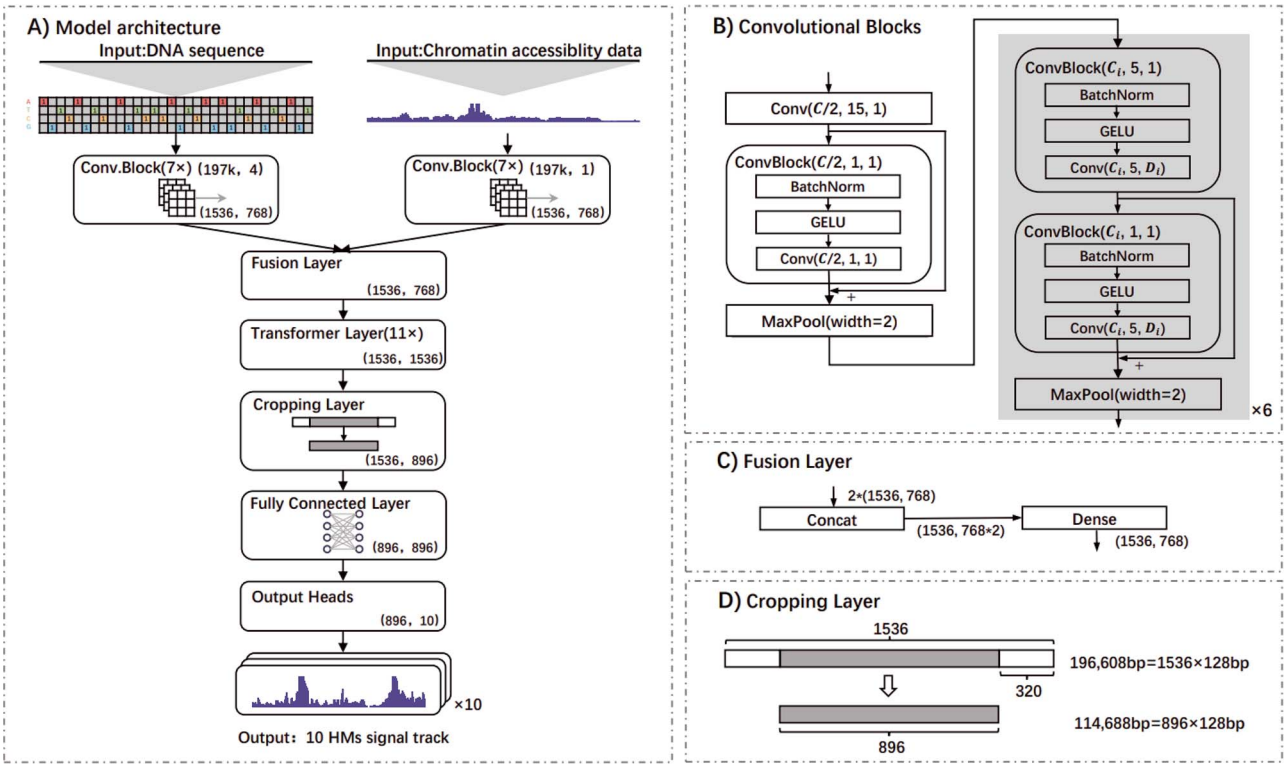

Figure 1. The dHICA framework illustrated; (**A**) overview of the dHICA; (**B**) detailed depiction of the convolutional blocks within dHICA; (**C**) the fusion layer of dHICA integrates features derived from DNA sequence and chromatin accessibility data, and (**D**) the strategy dHICA employs to segment distant genomic regions.

The architecture comprises seven distinct parts within the convolutional blocks, categorized into two primary convolutional blocks as depicted in Fig. 1B. The initial convolutional block effectively condenses the spatial dimension from 196608 bp down to 1536 bins, which ensures that each vector in the sequence symbolizes a 128-bp bin aligning with the resolution parameters set for dHICA.

Following this, six additional blocks utilize dilated convolutions—a technique where the convolutional filters incorporate gaps progressively enlarged by a factor of two in each subsequent layer. This approach allows the model's receptive field to expand exponentially without linearly increasing complexity. A vital feature of this model is the dense connectivity of these layers, whereby each layer utilizes inputs from all preceding layers instead of only the previous one. This design optimizes the number of filters required per layer. It allows for preserving and integrating the rich feature set extracted from the initial convolutional operations through the complex nuances teased out by the dilated convolutions. Each layer can thus concentrate on capturing the residual variation that previous layers have not addressed.

For the dilated convolution layers, we increase the number of channels C_i by a consistent multiplier until we attain the specified channel count C, starting from half that value $C/2$ in the initial six layers. In tandem, the dilation rate D_i is augmented by a factor of 1.5 for each successive layer, with the resulting figure rounded to the nearest whole number. For our setup, we define the channel size C as 1536 and initiate with $C/2$ filters along with a pooling size 2.

To further refine and integrate the features extracted from the DNA and chromatin accessibility data, we employ fusion layers (Fig. 1C). The first dense layer doubles the channel capacity, amplifying the feature space, while the subsequent layer scales it back to the model's baseline number of channels. This methodical expansion and contraction of the channel space facilitate a more nuanced synthesis of the underlying biological signals.

Transformer block

The transformer block is the core component of the model, encompassing 11 distinct layers that each play a pivotal role in interpreting sequence data. This block transforms each position in the input sequence through a computed weighted sum of all position representations, a process known as attention. Here, attention weights are influenced by the embeddings of the positions and their relative distances, enabling the incorporation of spatial context.

This attention-driven mechanism is critical to the model's advanced ability to predict HMs. It leverages information from critical regulatory regions, such as enhancers, essential for gene regulation. A standout feature of the model is its ability to focus attention directly across the entire sequence, facilitating seamless information exchange across potentially distant elements along the DNA strand. As a result, the transformer layers significantly broaden the model's receptive field, capturing regulatory elements up to 100 kb away while preserving the integrity of information crucial for accurate predictions.

The attention mechanism within these blocks is mathematically represented as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + R}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, V represent the queries, keys, and values vectors—each a critical component of the attention calculation; d_k is

the dimensionality of the keys and queries, providing necessary scaling. The softmax function ensures that the attention weights are normalized across the sequence. The term R denotes the relative positional encodings, which integrate spatial context into the model. The mathematical representation of relative positional encoding is as follows:

$$R_{i,j} = \exp\left(-\frac{|i-j|}{\log_2(|i-j|)}\right), \tag{2}$$

where $R_{i,j}$ represents the relative positional encoding between positions i and j. The parameter $\log_2(|i-j|)$ governs the rate of decay for the positional encoding, which enables the model to dynamically adjust the decay rate based on the positional distance, thereby facilitating the capture of long-range dependencies. By incorporating a logarithmic adjustment, dHICA can effectively manage varying sequence lengths, enhancing its ability to capture intricate dependencies within diverse sequences. Unlike static parameters, this adaptive mechanism allows the model to better generalize across different sequence contexts and lengths, thus improving overall performance.

Furthermore, the model employs Multi-Head Attention (MHA) to conduct multiple attention computations in parallel, allowing each head to capture distinct features of the input data independently

$$\begin{aligned} \mathrm{MHA}(Q,K,V) &= \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)W_O \\ \mathrm{head}_i &= \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \tag{3}$$

where W_i^Q, W_i^K, W_i^V are parameter matrices for ith attention head, W_O is the output weight matrix that combines the heads, and the number of heads h is set to 8.

Cropping layer

To tackle the computational challenges associated with analyzing distant regions in genomic sequence data, a cropping layer is employed. This layer excises 320 positions from each end of the sequence, effectively shortening it by 320 × 128 bp as illustrated in Fig. 1D. The process leaves only the central 896 positions. This cropping strategy is crucial because of the model's inherent limitations in capturing and learning effectively from regions distant from the sequencing center. These constraints are due to the model's architecture, which is designed to primarily perceive and analyze regulatory elements facing toward the sequencing center. At the same time, it struggles to detect elements beyond the sequence boundaries.

Following this, a fully connected layer is implemented to predict HMs with a resolution of 128 bp. This step not only refines the model's output but also significantly reduces the computational burden.

Model training

We adopted a supervised learning approach to optimize our dHICA model, utilizing the Mean Squared Error (MSE) loss function to optimize performance. To assess which chromatin accessibility data better complements dHICA, we conducted parallel training using identical model architectures on both ATAC-seq and DNase-seq data. Our training dataset comprised chromosomes 1–20 from the K562 cell line, with chromosome 21 reserved for validation and chromosome 22 for testing.

We configured the MSE loss function with an initial learning rate of 0.0001, incorporating a learning rate decay strategy that reduces the rate by a factor of 1.4 every 5 epochs following the first 10 epochs. Although recent studies have shown that PoissonNLL Loss can outperform MSE Loss in predicting epigenomic signals [22], our observations indicated no significant difference in performance between the two loss functions (Supplementary Text S1), which may be due to the Poisson distribution approximating a normal distribution when the mean parameter is sufficiently large. The model was optimized over 300 epochs, processing batches of 1500 samples each. This rigorous training ensures that dHICA can reliably predict steady-state HMs across various cell types, assuming that the underlying relationships between HMs, DNA, and chromatin accessibility signals remain consistent.

Results

Performance evaluation across different cell lines, tissues, and species

To thoroughly investigate the generalization capability of our model, we applied dHICA on a diverse array of cell lines (GM12878, MCF-7, HeLa-S3, HCT116, HepG2, IMR-90, and A549), along with human (heart and spleen) and mouse (hindbrain, heart, and G1E) tissues, despite its exclusive training on K562.

The imputed HMs exhibit robust correlation with experimental data (Fig. 2A) [23]; particularly noteworthy is the predicted background region data, which exhibits lower noise levels than the experimental data. Among the marks we attempted to model, only the repressive marks H3K9me3 and H3K27me3 showed subpar performance, likely due to their weak correlation with chromatin accessibility signals, low data values, and average sequencing quality [17, 24].

For the evaluation of HM imputations, we employed the method outlined in the ENCODE imputation challenge [25] to compute the Pearson's correlation between imputed and experimental HMs across seven distinct cell lines and four tissues (Fig. 2D and S5). Active marks (H3K4me3, H3K4me2, H3K4me1, and H3K9ac), which are predominantly associated with promoters and enhancers, exhibited consistent performance across holdout cell types, akin to the performance observed in the training cell line K562 (with an average Pearson's correlation exceeding 0.7). There was a slight decrease in performance near repressive regions (H3K9me3, H3K27me3, and H3K20me1). And the imputation performance across different cell lines correlated with the similarity of HMs in the correlation between the predicted and training cell lines K562 (Fig. S6). Therefore, the model showcases robust generalization across all HMs, demonstrating its effectiveness in diverse chromatin contexts.

Moving beyond genome-wide predictive accuracy, we delved into the imputed results near HM peaks and transcription start sites (TSS). The signals in those regions are the most informative [26], enabling a detailed examination of their distribution. While the dHICA's performance varies significantly across different cell lines on a genome-wide scale, there were no significant disparities in accuracy observed across various HM peaks with high signal intensity (Fig. 2C and Supplementary Text S2). And the imputation effectively captured the nuanced distribution of HM signals near the TSS of annotated genes (Fig. 2B, S10, and S11). Within TSS regions, dHICA comprehensively encompasses both active and repressive marks, aligning closely with biological expectations [27]. The substantial correlation between the distribution of imputed and experimental HM signals further bolsters confidence in dHICA's ability to accurately represent biological phenomena, affirming that it goes beyond merely learning average signal intensities of HMs [28].

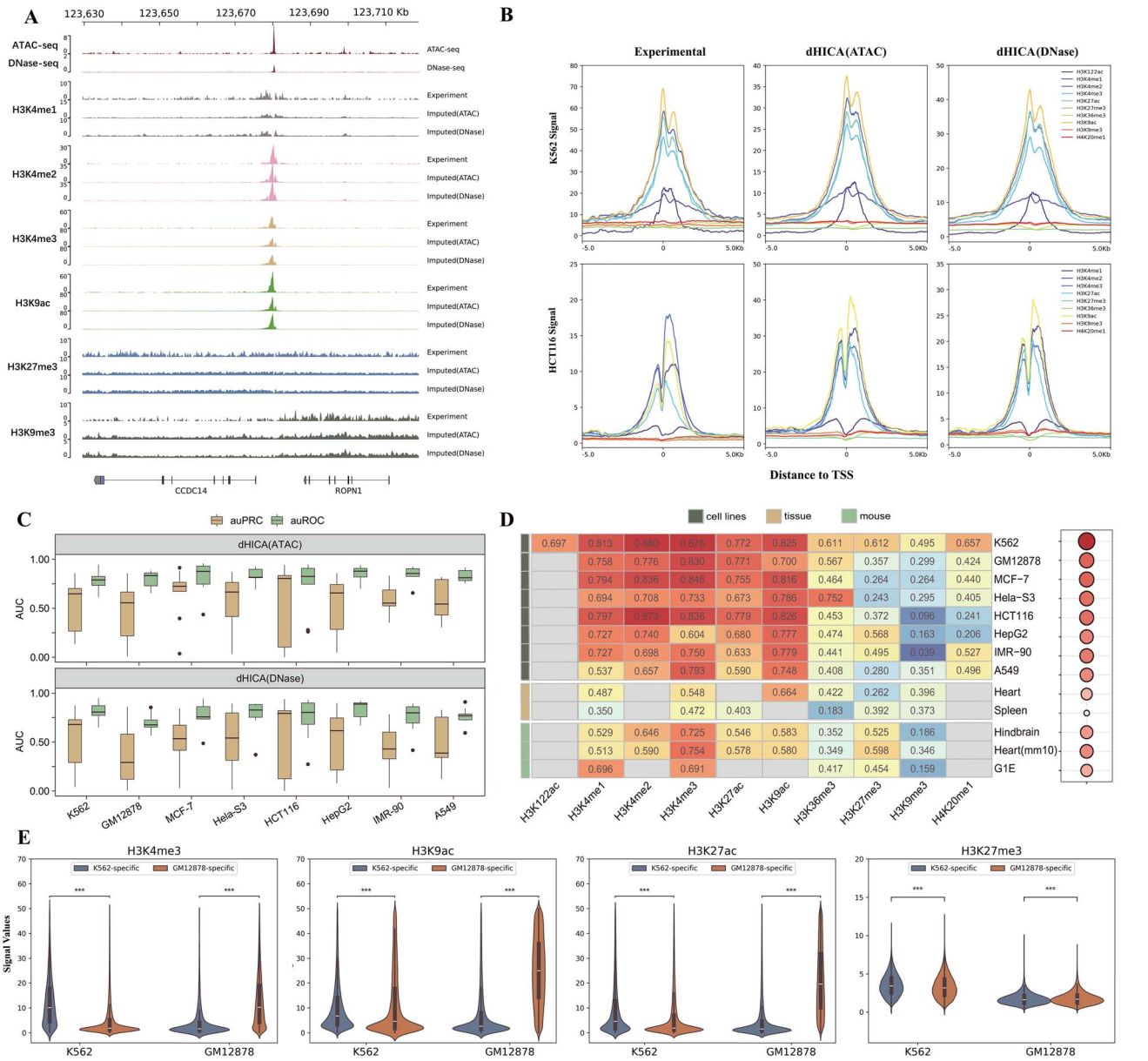

Figure 2. Comprehensive analysis of cross-cell-type and cross-species imputations; (**A**) genome browser comparison between experimental and predicted histone marks near gene CCDC14 and ROPN1 in GM12878, and (**B**) comparison of dHICA's imputed signals and experimental data proximal to the TSS in K562 and HCT116; (**C**) evaluation of dHICA's performance across HM peak regions; (**D**) comparison of Pearson's correlation of ATAC model across cell lines, tissues, and species; empty cells indicate that no experimental data are available for comparison in the cell type shown, and (**E**) distribution of cell-type-specific HM imputations by dHICA, indicating the effective capability of dHICA in distinguishing cell-type-specific features.

Additionally, our model, originally trained on the human K562 cell line, extended to HM signals in mouse cell and tissue types (Fig. 2D). What surprised us is that dHICA achieved even higher accuracy in mouse tissues compared with human tissues, which was due to data quality issues. The result exhibits distinct levels of similarity with the training cell line without retraining the model, which could help explore the general features of HMs that are shared across mammalian cell types. This capability unveils significant potential applications in annotating genomes of less-explored mammalian species.

Inspired by CEMIG [29], we employed dHICA to discern cell-type-specific HM peak sites in K562 and GM12878 cells. For each HM, we delineated K562-specific, GM12878-specific, and shared peaks (detailed in Supplementary Text S3). The variation in loci between different cell types reflects distinct regulatory mechanisms and gene expression patterns [30]. We investigate dHICA's ability to identify cell-type-specific features by comparing the distribution of signal values in different cell-type-specific peak regions, measured in raw counts imputed by dHICA (Fig. 2E). Ideally, the K562 imputation should show significantly higher signal values at K562-specific peaks than at GM12878-specific peaks, with a reciprocal pattern expected for GM12878 imputations. A one-sided Wilcoxon rank sum test against the null hypothesis that the signal values in different cell-type-specific peaks are identical yielded a p-value significantly less than 0.01, supporting the conclusion that dHICA effectively distinguishes cell-type-specific features. Despite being trained solely on the K562, dHICA accurately identifies specific regions, even in GM12878 cells that have not been trained.

Given dHICA's ability to accurately predict epigenomic features across cell lines, for further evaluation, we compare its performance with other baseline methods, including EPCOT,

Table 1. The comparison of auPRC between different models across HM peak regions

Methods	Input	Cell line	H3K4me3	H3K9ac	H3K27ac	H3K27me3	Average
dHICA	DNA+ATAC	K562	0.919	0.708	0.718	0.575	0.730
dHICA	DNA+DNase	K562	0.871	0.649	0.647	0.360	0.632
EPCOT	DNA+ATAC	K562	0.887	0.528	0.482	0.258	0.539
EPCOT	DNA+DNase	K562	0.857	0.531	0.637	0.359	0.596
Enformer	DNA	K562	0.865	0.578	0.530	0.160	0.533
deepPTM	DNA+TF	H1	0.903	0.719	0.554	0.240	0.604
deepSEA	DNA	H1	0.737	0.563	0.534	0.440	0.569
ChromImpute	HM(existing)	H1	0.617	0.688	0.200	0.788	0.573

Enformer, deepPTM, deepSEA, and ChromImpute. However, due to the diverse nature of prediction tasks (binary models for classifying HM peaks versus quantitative models for imputing signal tracks), making direct fair comparisons is challenging [22]. To address this issue, we divided the comparison into two main aspects: focusing on performance in HM peak regions and assessing genome-wide performance.

To compare the performance of state-of-the-art methods in HM peak regions, and due to data imbalance, we calculate the auPRC of imputations rather than auROC from different baseline methods, as shown in Table 1 and Figs S7 and S8. Given that deepPTM only imputes H3K4me3, H3K9ac, H3K27ac, and H3K27me3, we select these four HM markers for comparison, as they correlate with both active and repressive regions in the genome. This selection ensures that the subsets of HMs are reasonable. And dHICA using ATAC-seq achieved the best performance, with an average auPRC higher than 0.7.

For genome-wide HM imputation comparison, we compute Pearson's correlation, Spearman's correlation, and Root Mean Square Error (RMSE) between experimental signals and imputation from dHICA, EPCOT, and Enformer (Supplementary Text S4). From Fig. 3A, we easily discern that dHICA outperforms other baselines in all performance metrics. We further compare dHICA with EPCOT, as it is the most similar and comparable to our model. Like dHICA, it incorporates DNA sequences and chromatin accessibility data for the predictive tasks. However, EPCOT used four cell lines (K562, MCF-7, GM12878, and HepG2) as training datasets, whereas dHICA only used K562. To make the comparison fairer, we calculate the correlation of HMs between different cell lines (Fig. S6), and we ultimately select five cell lines that are most similar to the K562, along with K562 itself for evaluation between EPCOT and dHICA. Among the six cell lines (K562, GM12878, MCF-7, HCT116, HeLa-S3, and IMR-90) used for comparison, dHICA only used one cell line, K562, for training, while EPCOT used three cell lines (K562, GM12878, MCF-7) for model training. Though EPCOT is generally considered to be effective, dHICA has consistently shown superior performance in predicting HM markers (Fig. 3B). This is true regardless of whether the chromatin accessibility signals used by the model come from ATAC-seq or DNase-seq.

Contribution of DNA and chromatin accessibility data

Most baselines rely solely on DNA sequence inputs, potentially lacking cell-type-specific features. In contrast, dHICA integrates one-hot encoded DNA sequences and cell-type-specific chromatin accessibility signals. These cell-type-specific signals, represented by raw sequencing reads without data transformation, comprehensively impute HM signals. For cell-type-specific inputs, we opt for ubiquitous chromatin accessibility profiles from DNase-seq and ATAC-seq due to their profound implications in gene regulation and chromatin organization.

To evaluate the impact of DNA information and chromatin accessibility data on enhancing the model's performance in predicting HM signals, we separately trained the dHICA using only DNA or chromatin accessibility data on the K562. Subsequently, we assessed the performance of these individual models on the HCT116 and compared them with the standard dHICA model. Initially, we computed genome-wide Pearson's correlation from imputations generated by different individual models. As depicted in Fig. 4A, using both DNA and chromatin accessibility data consistently yielded higher Pearson's correlation than using either component alone. For active marks, relying solely on chromatin accessibility data outperformed the use of DNA sequence data alone, whereas this was not the case for repressive marks.

Considering the pivotal role of gene elements and their intricate interactions with HMs [31], along with the use of HM signals near various gene elements by many computational methods to predict gene expression [32–34], we are particularly interested in the predictive performance of HMs around gene elements. We finally selected five gene elements: promoter, enhancer, insulator, gene body, and PolyA, and calculated the Pearson's correlation in these regions (Fig. 4B). Aligned with the intricate interaction between HMs and gene elements, HMs exhibit significant performance, particularly in regions where they closely associate with specific gene elements—for instance, H3K4me3 in promoters, H3K4me1 in enhancers, and H3K36me3 in gene bodies.

To delve deeper into the primary impact of DNA information and chromatin accessibility data in gene elements, we conducted a comparative analysis of Pearson's correlation of the individual models around these gene elements, as depicted in Fig. 4C. Consistent with the genome-wide conclusion, for markers associated with enhancers and promoters, chromatin accessibility data play a more crucial role than DNA. Conversely, DNA assumes greater importance for marks associated with transcription and repressive regions. However, the incorporation of chromatin accessibility data also contributes significantly, as evidenced by the superior performance of the standard dHICA model compared with individual models using DNA alone genome-wide.

dHICA enables precise peak calls and chromatin state imputation for landscape insights

Since dHICA can impute the signal track of HM markers, we asked whether it can be used to identify HM peak regions. We apply LanceOtron [35] to call peaks from HM signals generated by dHICA, EPCOT, and ENCODE data (Table S4). We then compared the imputed peak regions with those from ENCODE by computing the Jaccard correlation, Recall, Precision, and F1 score (Fig. 5A and S9).

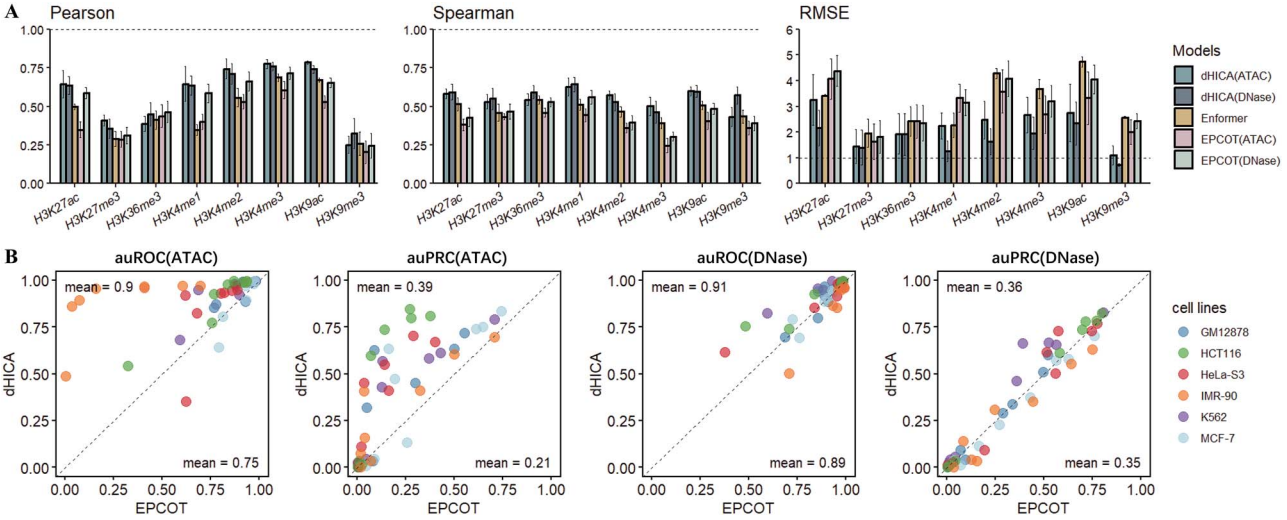

Figure 3. Multidimensional performance assessment; (**A**) aggregate metrics, including Pearson's correlation, Spearman's correlation, and RMSE, were evaluated over six cell lines (K562, GM12878, HCT116, Hela-S3, MCF-7, and IMR-90), and (**B**) genome-wide comparison of EPCOT and dHICA methodologies across multiple cell lines.

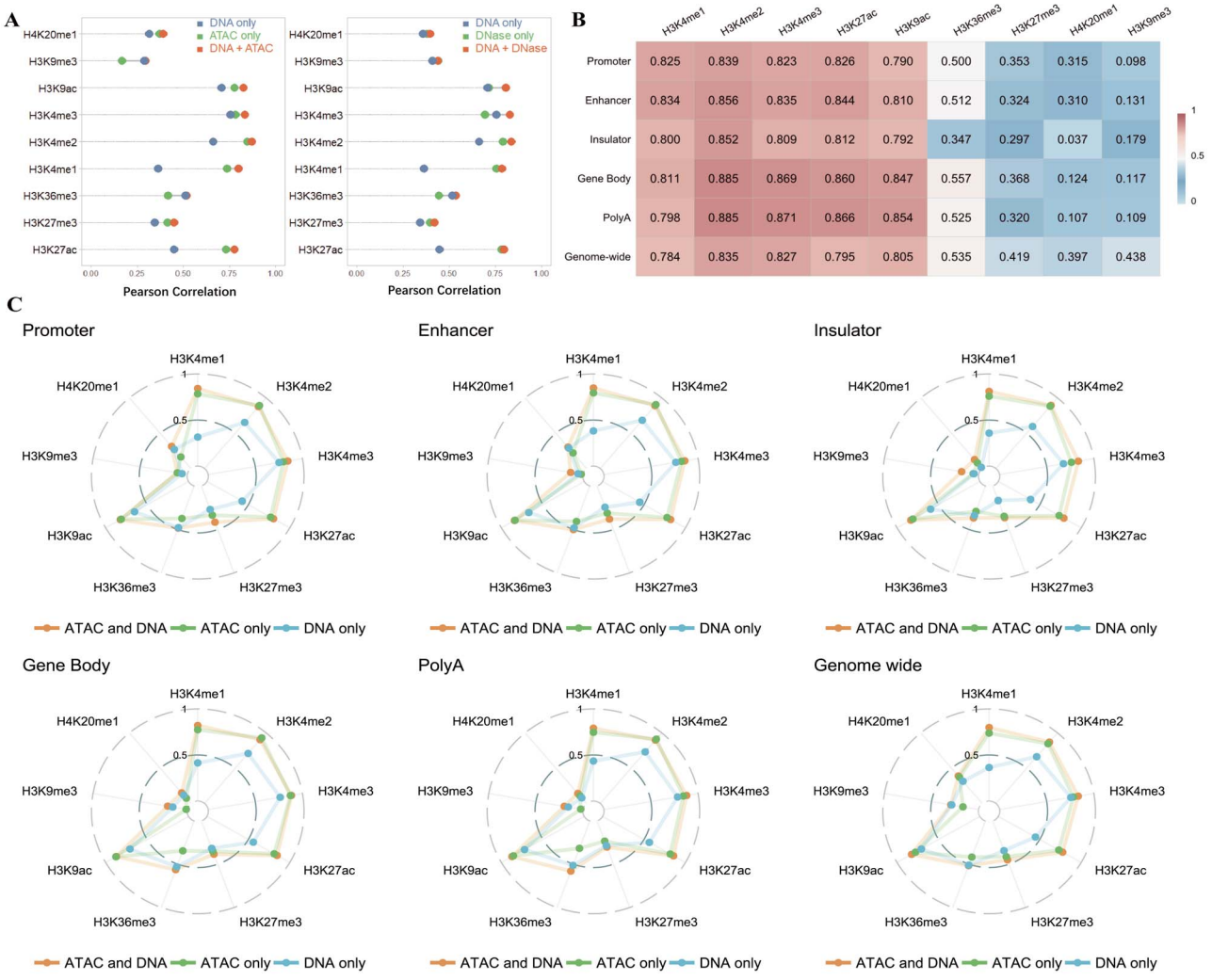

Figure 4. Analysis of the contribution of DNA and chromatin accessibility data; (**A**) Pearson's correlation of dHICA genome-wide using DNA or chromatin accessibility data in the HCT116; (**B**) the performance of dHCIA's imputations across different gene elements in the HCT116; (**C**) the contribution of DNA and chromatin accessibility data across different gene elements in the HCT116.

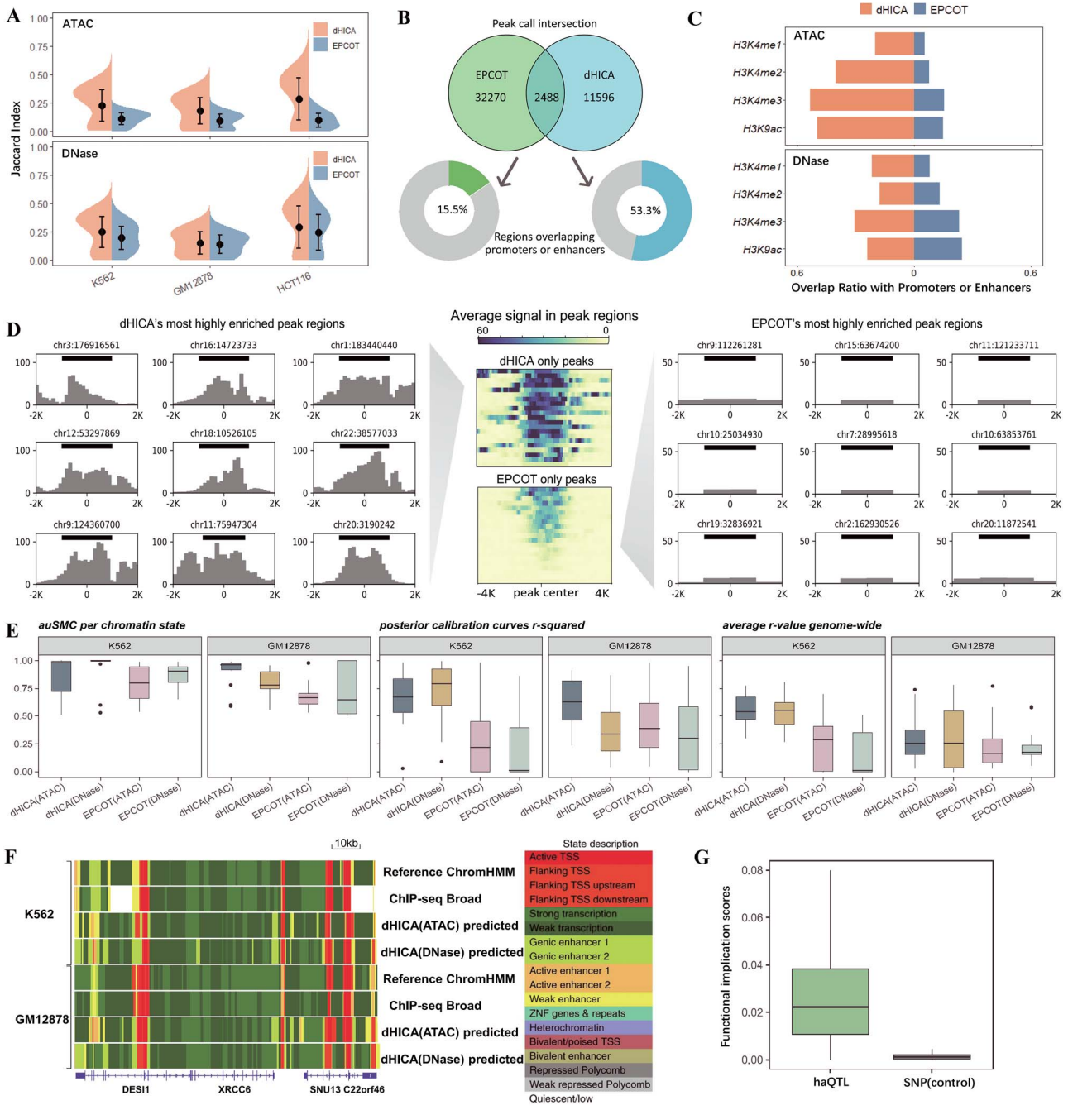

Figure 5. Downstream application and explanation of dHCIA; (**A**) Jaccard scores for dHICA and EPCOT across different cell lines (K562, GM12878, HCT116) by either ATAC (top) or DNase (bottom) model; (**B**) venn diagram of peak calls in H3K4me3 (HCT116) from LanceOtron using data from EPCOT and dHICA; regions that did not intersect were assessed for overlap with promoters or enhancers, and (**C**) bar plot of peak calls in the HCT116 from LanceOtron using data from EPCOT and dHICA by either ATAC (top) or DNase (bottom) model, showing the ratio of non-intersecting regions that overlap with promoters or enhancers; (**D**) thumbnail images from the most highly enriched regions called by LanceOtron from either dHICA (left) or EPCOT (right), and the center panel shows the average coverage of the peak regions from either dHICA (top) or EPCOT (bottom) for H3K4me3 (HCT116); (**E**) performance assessment for chromatin state segmentation using ChromHMM based on HM signals imputed by dHICA and EPCOT, and (**F**) 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); (**G**) FISs of haQTLs and nearby SNPs within 500 bp.

We further investigated the differences between dHICA and EPCOT peak calls. As shown in Fig. 5B, for H3K4me3 in the HCT116, we identified 32270 EPCOT-only peaks and 11596 dHICA-only peaks. Notably, 15.5% of peaks exclusively called by EPCOT overlapped with promoters or enhancers, whereas 53.3% of dHICA-only peak calls showed such overlap. This trend was consistent across all active markers (Fig. 5C). When visualizing the top enriched peaks called using different HM imputations, dHICA's peaks demonstrated significantly higher signal than those from EPCOT (Fig. 5D). This pattern was also observed when inspecting the average signal of the peak calls; EPCOT-only peaks were typically found in areas with less surrounding signal, containing narrower peaks with very low enrichment compared with dHICA-only peaks. It seems that EPCOT-only peaks represent a sporadic

sampling of numerous peaks near noise levels throughout the genome, while the peaks missed by EPCOT and identified by dHICA are relatively strongly enriched.

Chromatin state segmentation and genome annotations are essential for various genomic tasks, including the identification of active regulatory elements and the interpretation of disease-associated genetic variations across different cell types and in human diseases [36, 37]. Given the robust performance of dHICA in the vicinity of gene elements, we investigated whether chromatin states defined by ChromHMM could be inferred using HM markers imputed by dHICA [38]. We used the pretrained reported ChromHMM model that defined 18 distinct chromatin states based on six marks for which we trained imputation models (H3K4me3, H3K27ac, H3K4me1, H3K36me3, H3K9me3, and H3K27me3) [19]. Examination through the Integrative Genomics Viewer showed that chromatin states were highly consistent [39], regardless of whether they were defined using ENCODE data or dHICA's imputation (Fig. 5F). This highlights dHICA's ability to accurately represent the underlying epigenetic landscape, successfully extrapolating complex chromatin configurations from integrated datasets. The consistency of the imputed states with those derived from ENCODE data suggests that this can be a viable alternative for predicting chromatin states, particularly in contexts where ChIP-seq data are unavailable or when detailed analysis of chromatin states is required due to the extensive variability in human chromatin states [40, 41].

To achieve a more quantitatively robust and principled evaluation of chromatin state segmentation, we applied SAGAconf [42] to compare the annotations derived from imputation with ENCODE ChIP-seq data in both K562 and GM12878. We included EPCOT for comparison, as it closely resembles our model, as illustrated in Fig. 5E. We calculated the area under the scaled min-max curve (auSMC), posterior calibration curves *r*-squared, and correlation coefficients (*r*-values) between the ENCODE data and the imputations generated by both dHICA and EPCOT, with detailed metrics provided in the Supplementary Text S5. For each of the metrics, dHICA outperformed EPCOT, regardless of whether ATAC-seq or DNase-seq data were used. Additionally, although GM12878 serves as a test cell type for dHICA and a training cell type for EPCOT, where EPCOT should presumably perform better, dHICA still excelled over EPCOT in segmentation and genome annotations tasks.

dHICA explains functional implications of SNPs

Genome-wide association studies (GWAS) have successfully identified numerous genetic variants associated with complex traits and diseases [43, 44]. However, elucidating these associations' biological mechanisms is challenging, as most SNPs are non-coding, and their regulatory roles remain unclear [45, 46]. The genotype-independent signal correlation and imbalance (G-SCI) test [47], leveraging ChIP-seq assays on H3K27ac, has streamlined the identification of histone acetylation quantitative trait loci (haQTLs), thereby aiding in pinpointing causal variants within GWAS loci and advancing our understanding of their functional implications.

Inspired by the G-SCI method and studies on cell-type-specific haQTLs [48, 49], we employed the dHICA, which can impute cell-type-specific and tissue-specific HM signals, to analyze the SNPs identified by G-SCI, which demonstrates dHICA's potential to enhance understanding of the functional implications of these SNPs. Following DeepHistone [11], we identified a set of 6925 SNPs (haQTLs) specific to H3K27ac in the GM12878 from the 1000 Genomes Project [50]. We also created a negative control set with equivalent SNPs, each ∼500-bp away from a corresponding haQTL. Using the formula $\Delta p = |p_{\text{ref}} - p_{\text{alt}}|$, where p_{ref} denotes the signal intensity associated with the reference allele, and p_{alt} represents the signal intensity of the alternative allele post-mutation, as defined in the study [51], we calculated functional implication scores (FISs) for these SNPs. The results in Fig. 5G indicate that haQTLs have significantly higher scores than control SNPs. This was substantiated by a one-sided Wilcoxon rank sum test, which revealed a marked difference in the median scores between the two groups, with a *p*-value of 3.651E-320. Consequently, our analysis supports the biological understanding that haQTLs are more likely to impact the function of the lymphoblastoid epigenome and, in turn, influence phenotypic traits [52]. This demonstrates that dHICA can effectively distinguish SNPs potentially responsible for specific phenotypes from their neighboring genetic variants.

Discussion

In this study, we presented dHICA, a deep learning framework that integrates chromatin accessibility information and DNA sequences to predict cell-type-specific HM signals accurately. By incorporating the transformer structure alongside dilated convolutions, dHICA significantly expands the model's receptive field, enabling it to capture long-range interactions between genomic elements.

dHICA outperformed other state-of-the-art methods across various cell lines and species, primarily due to its integration of chromatin accessibility data, which provides cell-type-specific features, particularly active gene elements. Moreover, dHICA's imputed data can be utilized for downstream tasks such as chromatin state segmentation and distinguishing haQTLs from SNPs. Given its ability to predict HMs in new cell lines and species without re-training, dHICA holds significant potential for refined and personalized analysis, provided that users can supply the necessary chromatin accessibility data.

Considering the good performance of dHICA trained solely on K562, we further explored applying other cell lines for model training to enhance the accuracy and robustness of dHICA's HM predictions. Like EPCOT, we trained dHICA-multiple, incorporating data from the K562, GM12878, HepG2, and MCF-7 cell lines (Fig. S3). To maintain consistency in dataset size for each marker, we excluded the H3K122ac marker, as it is only sequenced in K562. As illustrated in Fig. S4, integrating data from multiple cell lines improved the performance in the training cell lines (GM12878, HepG2, and MCF-7), compared with dHICA, but no significant enhancements in test cell lines.

While dHICA demonstrates superior performance, there are opportunities for further refinement. (i) Data normalization and quality improvement: dHICA processes raw counts directly from bigWig files without any data transformation or normalization, which may introduce noise and variability. Implementing normalization strategies for GC% [53], employing denoising tools [54], or exploring data transformation techniques such as arcsinh-transformed epigenomic feature signals could enhance data quality. (ii) Improving the performance of multi-cell models: although we have initially integrated data from other cell lines for model training to capture multi-cell features, the performance improvements have been limited. This may require improvements in data preprocessing and model architecture to extract features across multiple cell lines and enhance overall performance. (iii) Prediction of gene expression: given the established link between HMs and gene expression [55], and considering dHICA's

capability in accurately imputing HMs, there is potential for the model to predict gene expression. Ideally, this capability would significantly broaden dHICA's applications and contribute to a deeper understanding of biological mechanisms.

Key Points

- This study proposes a deep learning framework, dHICA, which integrates chromatin accessibility information and DNA sequences to predict cell-type-specific HM signals accurately.
- dHICA largely benefits from the chromatin accessibility data.
- Across various cell lines, dHICA consistently outperforms other state-of-the-art methods.
- dHICA facilitates precise peak calls and chromatin state segmentation, providing deeper insights into the genomic landscape.
- dHICA aids in elucidating the functional implications of SNPs.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

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).

Conflict of interest: No competing interest is declared.

Code availability

The data sources used in this paper are reported in the Supplementary data. The code for dHICA is available on GitHub at https://github.com/wzhy2000/dHICA. We also offer a cloud computing service at https://dreg.dnasequence.org/. Through this platform, users can upload their data to use GPU resources and obtain imputations performed by dHICA.

References

1. Talbert PB, Meers MP, Henikoff S. Old cogs, new tricks: the evolution of gene expression in a chromatin context. *Nat Rev Genet* 2019;**20**:283–97. https://doi.org/10.1038/s41576-019-0105-7.
2. Brouwer T, Pham C, Kaczmarczyk A. *et al.* A critical role for linker dna in higher-order folding of chromatin fibers. *Nucleic Acids Res* 2021;**49**:2537–51. https://doi.org/10.1093/nar/gkab058.
3. Stillman B. Histone modifications: insights into their influence on gene expression. *Cell* 2018;**175**:6–9. https://doi.org/10.1016/j.cell.2018.08.032.
4. ENCODE Project consortium. et al. An integrated encyclopedia of dna elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247.
5. Moore JE, Purcaro MJ, Pratt HE. *et al.* Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature* 2020;**583**:699–710. https://doi.org/10.1038/s41586-020-2493-4.
6. Kundaje A, Meuleman W, Ernst J. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30. https://doi.org/10.1038/nature14248.
7. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* 2015;**33**:364–76. https://doi.org/10.1038/nbt.3157.
8. Durham TJ, Libbrecht MW, Jeffry Howbert J. *et al.* Predictd parallel epigenomics data imputation with cloud-based tensor decomposition. *Nat Commun* 2018;**9**:1402. https://doi.org/10.1038/s41467-018-03635-9.
9. Schreiber J, Durham T, Bilmes J. *et al.* Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol* 2020;**21**:1–18.
10. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015;**12**:931–4. https://doi.org/10.1038/nmeth.3547.
11. Yin Q, Mengmeng W, Liu Q. *et al.* Deephistone: a deep learning approach to predicting histone modifications. *BMC Genomics* 2019;**20**:11–23.
12. Li Y, Quan L, Zhou Y. *et al.* Identifying modifications on dna-bound histones with joint deep learning of multiple binding sites in dna sequence. *Bioinformatics* 2022;**38**:4070–7. https://doi.org/10.1093/bioinformatics/btac489.
13. Dipankar Ranjan Baisya and Stefano Lonardi. Prediction of histone post-translational modifications using deep learning. *Bioinformatics* 2020;**36**:5610–7. https://doi.org/10.1093/bioinformatics/btaa1075.
14. Kelley DR. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* 2020;**16**:e1008050. https://doi.org/10.1371/journal.pcbi.1008050.
15. Avsec ž, Agarwal V, Visentin D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021;**18**:1196–203. https://doi.org/10.1038/s41592-021-01252-x.
16. Karollus A, Mauermeier T, Gagneur J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol* 2023;**24**:56. https://doi.org/10.1186/s13059-023-02899-9.
17. Wang Z, Chivu AG, Choate LA. *et al.* Prediction of histone post-translational modification patterns based on nascent transcription data. *Nat Genet* 2022;**54**:295–305. https://doi.org/10.1038/s41588-022-01026-x.
18. Zhang Z, Feng F, Qiu Y. *et al.* A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. *Nucleic Acids Res* 2023;**51**:5931–47. https://doi.org/10.1093/nar/gkad436.
19. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with chromhmm. *Nat Protoc* 2017;**12**:2478–92. https://doi.org/10.1038/nprot.2017.124.
20. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017;30.
21. Amemiya HM, Kundaje A, Boyle AP. The encode blacklist: identification of problematic regions of the genome. *Sci Rep* 2019;**9**:9354.
22. Toneyan S, Tang Z, Koo PK. Evaluating deep learning for predicting epigenomic profiles. *Nat Mach Intell* 2022;**4**:1088–100. https://doi.org/10.1038/s42256-022-00570-9.
23. Lopez-Delisle L, Rabbani L, Wolff J. *et al.* Pygenometracks: Reproducible plots for multivariate genomic datasets. *Bioinformatics* 2021;**37**:422–3. https://doi.org/10.1093/bioinformatics/btaa692.
24. Kong X, Wei G, Chen N. *et al.* Dynamic chromatin accessibility profiling reveals changes in host genome organization in response to baculovirus infection. *PLoS Pathog* 2020;**16**:e1008633. https://doi.org/10.1371/journal.ppat.1008633.
25. Schreiber J, Boix C. *et al.* The encode imputation challenge: A critical assessment of methods for cross-cell type imputation of epigenomic profiles. *Genome Biol* 2023;**24**:79.

26. Chen Y-H, Keegan S, Kahli M. *et al.* Transcription shapes dna replication initiation and termination in human cells. *Nat Struct Mol Biol* 2019;**26**:67–77. https://doi.org/10.1038/s41594-018-0171-0.
27. van der Velde, Fan K, Tsuji J. *et al.* Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun Biol* 2021;**4**:239. https://doi.org/10.1038/s42003-021-01756-4.
28. Schreiber J, Singh R, Bilmes J. *et al.* A pitfall for machine learning methods aiming to predict across cell types. *Genome Biol* 2020;**21**:1–6. https://doi.org/10.1186/s13059-020-02177-y.
29. Wang Y, Li Y, Wang C. *et al.* Cemig: Prediction of the cis-regulatory motif using the de bruijn graph from atac-seq. *Brief Bioinform* 2024;**25**:bbad505.
30. Kim-Hellmuth S, Aguet F, Oliva M. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* 2020;**369**:1317–23. https://doi.org/10.1126/science.aaz8528.
31. Deniz Ö, Frost JM, Branco MR. Regulation of transposable elements by dna modifications. *Nat Rev Genet* 2019;**20**:417–31. https://doi.org/10.1038/s41576-019-0106-6.
32. Liu Y, Wang Z, Lv J. *et al.* DeepChrom: a diffusion-based framework for long-tailed chromatin state prediction. *Pattern Recognition and Computer Vision* 2024, 188–99. https://doi.org/10.1007/978-981-99-8435-0_15.
33. Lee D, Yang J, Kim S. Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nat Commun* 2022;**13**:6678. https://doi.org/10.1038/s41467-022-34152-5.
34. Schmidt F, Kern F, Schulz MH. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics Chromatin* 2020;**13**:4. https://doi.org/10.1186/s13072-020-0327-0.
35. Hentges LD, Sergeant MJ, Cole CB. *et al.* Lanceotron: A deep learning peak caller for genome sequencing experiments. *Bioinformatics* 2022;**38**:4255–63. https://doi.org/10.1093/bioinformatics/btac525.
36. Boix CA, James BT, Park YP. *et al.* Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 2021;**590**:300–7. https://doi.org/10.1038/s41586-020-03145-z.
37. Filion GJ, van Bemmel, Braunschweig U. *et al.* Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell* 2010;**143**:212–24. https://doi.org/10.1016/j.cell.2010.09.009.
38. Ernst J, Kellis M. Chromhmm: automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**:215–6. https://doi.org/10.1038/nmeth.1906.
39. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;**14**:178–92. https://doi.org/10.1093/bib/bbs017.
40. Gershman A, Sauria MEG, Guitart X. *et al.* Epigenetic patterns in a complete human genome. *Science* 2022;**376**:eabj5089. https://doi.org/10.1126/science.abj5089.
41. Grubert F, Srivas R, Spacek DV. *et al.* Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* 2020;**583**:737–43. https://doi.org/10.1038/s41586-020-2151-x.
42. Shahraki MF, Farahbod M, Libbrecht MW. Robust chromatin state annotation. *Genome Res* 2024;**34**:469–83. https://doi.org/10.1101/gr.278343.123.
43. Claussnitzer M, Cho JH, Collins R. *et al.* A brief history of human disease genetics. *Nature* 2020;**577**:179–89. https://doi.org/10.1038/s41586-019-1879-7.
44. Buniello A, MacArthur JAL, Cerezo M. *et al.* The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**:D1005–12. https://doi.org/10.1093/nar/gky1120.
45. French JD, Edwards SL. The role of noncoding variants in heritable disease. *Trends Genet* 2020;**36**:880–91. https://doi.org/10.1016/j.tig.2020.07.004.
46. Yan J, Qiu Y, Andre M. *et al.* Systematic analysis of binding of transcription factors to noncoding variants. *Nature* 2021;**591**:147–51. https://doi.org/10.1038/s41586-021-03211-0.
47. Cruz-Herrera R, del Rosario, Poschmann SL. *et al.* Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nat Methods* 2015;**12**:458–64. https://doi.org/10.1038/nmeth.3326.
48. Hou L, Xiong X, Park Y. *et al.* Multitissue h3k27ac profiling of gtex samples links epigenomic variation to disease. *Nat Genet* 2023;**55**:1665–76. https://doi.org/10.1038/s41588-023-01509-5.
49. Tan WLW, Anene-Nzelu CG, Wong E. *et al.* Epigenomes of human hearts reveal new genetic variants relevant for cardiac disease and phenotype. *Circ Res* 2020;**127**:761–77. https://doi.org/10.1161/CIRCRESAHA.120.317254.
50. Gibbs RA, Boerwinkle E, Doddapaneni H. *et al.* A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
51. Liu Q, Xia F, Yin Q. *et al.* Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics* 2018;**34**:732–8. https://doi.org/10.1093/bioinformatics/btx679.
52. Sheng T, Ho SWT, Ooi WF. *et al.* Integrative epigenomic and high-throughput functional enhancer profiling reveals determinants of enhancer heterogeneity in gastric cancer. *Genome Med* 2021;**13**:1–25. https://doi.org/10.1186/s13073-021-00970-3.
53. Browne PD, Nielsen TK, Kot W. *et al.* Gc bias affects genomic and metagenomic reconstructions, underrepresenting gc-poor organisms. *GigaScience* 2020;**9**:giaa008. https://doi.org/10.1093/gigascience/giaa008.
54. Lal A, Chiang ZD, Yakovenko N. *et al.* Deep learning-based enhancement of epigenomics data with atacworks. *Nat Commun* 2021;**12**:1507. https://doi.org/10.1038/s41467-021-21765-5.
55. Morgan MAJ, Shilatifard A. Reevaluating the roles of histone-modifying enzymes and their associated chromatin modifications in transcriptional regulation. *Nat Genet* 2020;**52**:1271–81. https://doi.org/10.1038/s41588-020-00736-4.