frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 June 2024
DOI 10.3389/fgene.2024.1424085

OPEN ACCESS

EDITED BY
Fa Zhang,
Beijing Institute of Technology, China

REVIEWED BY
Shiwei Sun,
Chinese Academy of Sciences (CAS), China
Bingbo Wang,
Xidian University, China
Han Wang,
Northeast Normal University, China

*CORRESPONDENCE
Bingqiang Liu,
bingqiang@sdu.edu.cn
Duanchen Sun,
dcsun@sdu.edu.cn

RECEIVED 27 April 2024
ACCEPTED 22 May 2024
PUBLISHED 17 June 2024

CITATION
Wu Q, Li Y, Wang Q, Zhao X, Sun D and Liu B (2024), Identification of DNA motif pairs on paired sequences based on composite heterogeneous graph.
Front. Genet. 15:1424085.
doi: 10.3389/fgene.2024.1424085

Identification of DNA motif pairs on paired sequences based on composite heterogeneous graph

Qiuqin Wu[1], Yang Li[2], Qi Wang[1], Xiaoyu Zhao[1], Duanchen Sun[1]* and Bingqiang Liu[1]*

[1]School of Mathematics, Shandong University, Jinan, China, [2]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, United States

Motivation: The interaction between DNA motifs (DNA motif pairs) influences gene expression through partnership or competition in the process of gene regulation. Potential chromatin interactions between different DNA motifs have been implicated in various diseases. However, current methods for identifying DNA motif pairs rely on the recognition of single DNA motifs or probabilities, which may result in local optimal solutions and can be sensitive to the choice of initial values. A method for precisely identifying DNA motif pairs is still lacking.

Results: Here, we propose a novel computational method for predicting DNA Motif Pairs based on Composite Heterogeneous Graph (MPCHG). This approach leverages a composite heterogeneous graph model to identify DNA motif pairs on paired sequences. Compared with the existing methods, MPCHG has greatly improved the accuracy of motifs prediction. Furthermore, the predicted DNA motifs demonstrate heightened DNase accessibility than the background sequences. Notably, the two DNA motifs forming a pair exhibit functional consistency. Importantly, the interacting TF pairs obtained by predicted DNA motif pairs were significantly enriched with known interacting TF pairs, suggesting their potential contribution to chromatin interactions. Collectively, we believe that these identified DNA motif pairs held substantial implications for revealing gene transcriptional regulation under long-range chromatin interactions.

KEYWORDS

DNA motifs, DNA motif pairs, chromatin interactions, TF pairs, gene transcriptional regulation

1 Introduction

The identification and recognition of DNA motifs binding to transcription factors (TFs) are pivotal for comprehending the regulatory mechanisms governing gene expression and cellular processes (Wong et al., 2013). A DNA motif denotes to a short, similarly repeated pattern of nucleotides that holds biological significance (Hashim et al., 2019). Deciphering these binding DNA motifs provides researchers with insights into the regulation and control of genes, fostering a deeper understanding of diverse biological phenomena (Liu et al., 2018; Yang et al., 2019; Li et al., 2024; Wang et al., 2024). With the development of high-throughput technology, several experimental techniques are available for determining TF binding DNA motifs, such as Chromatin Immunoprecipitation (ChIP) (Park, 2009), Electrophoretic Mobility Shift Assay (EMSA) (Hellman and Fried, 2007), DNA Affinity Purification Sequencing (DAP-seq) (Bartlett et al., 2017), and Systematic Evolution of

Ligands by Exponential Enrichment (SELEX) (Gold, 2015). Moreover, researchers can access relevant databases to query for associated DNA motifs. For instance, JASPAR (Castro-Mondragon et al., 2022) is a widely utilized DNA motif database for storing and analyzing transcription factor binding site. TRANSFAC (Wingender et al., 2000) is a classic database containing DNA motifs of transcription factors and regulatory elements, offering a wealth of DNA motif data and associated biological information. Other databases include UniProbe, Cis-BP, motifMap, ScerTF, TFcat, and FlyTF (Fulton et al., 2009; Pfreundt et al., 2010; Daily et al., 2011; Robasky and Bulyk, 2011; Spivak and Stormo, 2012; Weirauch et al., 2014). However, the action of a single DNA motif is limited, and actual gene regulation often involves intricate interactions among multiple DNA motifs, giving rise to DNA motif pairs (Pilpel et al., 2001). These pairs of DNA motifs play a pivotal role in maintaining the accuracy and flexibility of gene expression (Clauss and Lu, 2023).

When two DNA motifs coexist and interact in a specific manner during gene regulation, they can either cooperate or compete to influence gene expression. This is pivotal for unraveling the intricate mechanisms of gene regulation networks, cell signaling, and biological processes (Kim and Wysocka, 2023). Moreover, these predictions of the interaction between DNA motifs find broad applications in bioinformatics, facilitating genome annotation and the anticipation of protein-nucleic acid interactions, thereby equipping researchers with potent tools to decipher biological data (Khodabandelou et al., 2020; Wang et al., 2022). Lastly, the underlying chromatin interactions between different DNA motifs are associated with various diseases (Bhatia and Kleinjan, 2014). Consequently, predictions based on DNA motif pairs hold promise for discovering new drug targets and innovations in the field of biotechnology, deepening our understanding of gene regulation networks (Makolo and Suberu, 2016).

The essence of DNA motif pairs lies in discerning pattern pairs, specifically identifying statistically significant pattern pairs within two correlated sequences, derived from different sequences. Current methods for identifying DNA motif pairs can be broadly classified into two types. The first approach is direct, involving the independent identification of statistically significant DNA motifs from two correlated sequences. Subsequently, the threshold is calculated to combine the DNA motifs on both sides of sequences to select statistically significant DNA motif pairs. This method may result in the exclusion of DNA motifs capable of forming pairs but are underrepresented. The second approach is based on statistical significance and involves predicting DNA motif pairs through a global optimization model. This method requires constructing a well-designed model for predicting DNA motif pairs. The algorithm developed by Ka-Chun Wong's research group in 2016, referred to as Wong's 2016 (Wong et al., 2016), and EPmotifPair (Wang et al., 2022) both belong to the first category of methods in existing approaches on HI-C (van Berkum et al., 2010) data for predicting DNA motif pairs. Wong's 2016 is presently the first method for identifying DNA motif pairs on HI-C data. It can more flexibly learn sequence features in different directions, such that disturbances in predictions on one side may not affect predictions on the other side. EPmotifPair (Wang et al., 2022) predicts DNA motif pairs in a set of sequences integrated from enhancer sequences and promoter sequences. By comprehensively considering multiple co-occurring sequence patterns, it reduces the error rate compared to the separate prediction of DNA motifs. MotifHyades (Wong, 2017) belongs to the second category of methods for predicting DNA motif pairs. It adopts the probability model and utilizes two derived optimization algorithms to find DNA motif pairs with linear complexities. However, Wong's 2016 (Wong et al., 2016) not only overlooks underrepresented DNA motifs that could have formed pairs but is also time-consuming. EPmotifPair (Wang et al., 2022) not only fails to account for potential interactions between DNA motifs but also requires the specification of numerous parameters, such as the predetermined number of DNA motifs. MotifHyades (Wong, 2017) improves the computational speed and accuracy compared with Wong's 2016 (Wong et al., 2016), but it is sensitive to the choice of the initial value. Additionally, the probability model adopted by MotifHyades (Wong, 2017) assumes conditional independence within each sequence pair, disregarding potential interactions among DNA motifs.

To address the aforementioned challenges, we propose a graph theory-based approach named MPCHG. The methodology is elucidated in Figure 1 (This paper takes Enhancer-Promoter as an example). It helps capture multiple relationships between different *k*-mers, including both within-sequence and between-sequence relationships. Subsequently, a community detection algorithm is employed to obtain a dense subgraph, considering not only the topology of the network but also the practical significance of node connections. Importantly, we refrain from predefining the length of DNA motifs and the number of DNA motif pairs, avoiding the loss of some important DNA motifs or the presence of high noise. We apply MPCHG to analyze seven sets of HI-C data. The results reveal a higher proportion of DNA motifs matching existing databases for predicted DNA motif pairs. The identified paired DNA motifs demonstrate higher DNase accessibility than the background sequences, and the functional consistency of DNA motifs within pairs is evident. Particularly noteworthy is the acquisition of predicted TF pairs from the predicted DNA motif pairs, and we discover that the predicted TF pairs are enriched with the interacting TFs in the STRING database. It can be seen that predicting DNA motif pairs on HI-C data can help us understand the regulatory mechanisms of genes.

2 Methods

2.1 Data collection

The input data comprises of Hi-C data from seven sets derived from six distinct cell lines, namely,: K562, GM12878, HeLa-S3, HUVEC, IMR90, and NHEK. Two sets of Hi-C data (referred to as K562_1 and K562_2, respectively) are obtained from the K562 cell line, featuring variations in data preprocessing and annotation approaches. A set of protein-protein interaction data retrieved from the STRING database (Mering et al., 2003) serves as benchmark data to assess the performance of predicted DNA motif pairs. The first set of processed Hi-C data from the K562 cell line (K562_1) is acquired from the article published by

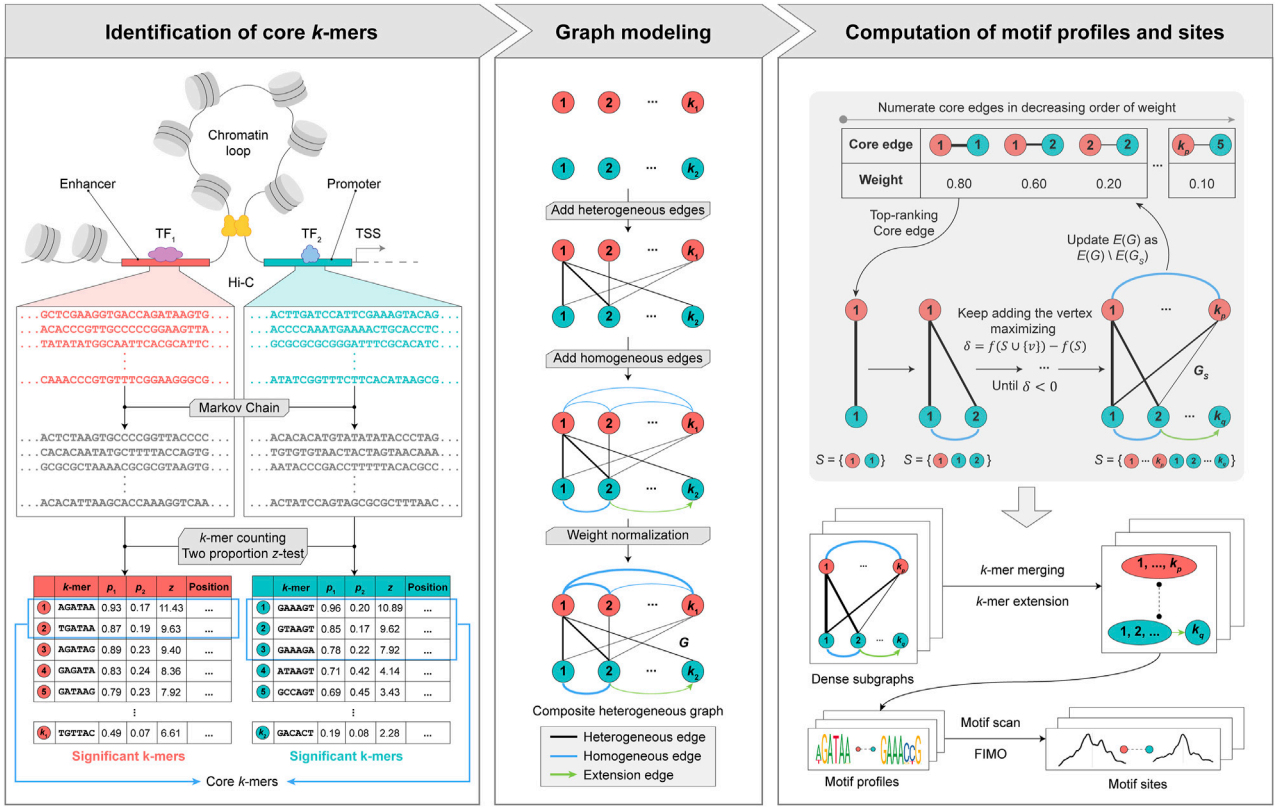

FIGURE 1
Overview of MPCHG. Enhancer (red)—Promoter (blue) interaction is used as an example for DNA motif pairs identification. p_1 denotes the frequencies of *k*-mers in the real sequence sets and p_2 denotes the frequencies of *k*-mers in the background sequence sets. *z* denotes *z*-score, which is used to measure the significance of *k*-mers. The *k*-mers are arranged in descending order by the size of the *z*-score. The *k*-mers framed by the blue square indicates core *k*-mers. In section Graph modeling, the black lines represent the heterogeneous edges, which connect different types of *k*-mers, and the thickness of the lines indicates the weight of the connected edges, the greater the weight, the thicker the lines. The blue line represents the homogeneous edge, and they will connect the overlapping *k*-mers, and the thickness of the lines indicates the weight of the connected edges, the thicker the line, the greater the weight of the connected edge. The green arrow is called extension edge, indicating the overlap between the two *k*-mers, which can be used to merge and extend the two *k*-mers in subsequent steps.

Ka-Chun Wong in 2016 (Wong et al., 2016). In this study, chromatin fragments are classified into four categories: E (Enhancer), TSS (Promoter), WE (Weak Enhancer), and PF (Promoter-Flanking Region). These categories collectively form 10 interacting pairs, resulting in a total of 74,552 long-range regulatory region pairs. The number of each interaction type is detailed in Supplementary Figure S1A. The remaining six sets of processed HI-C data are sourced from the article published by Wang in 2022 (Wang et al., 2022). Which are normalized using the Knight and Ruiz normalization vectors (Lyu et al., 2020) by Rao *et al.* (Wang et al., 2022). Notably, their chromatin interaction type is exclusively Promoter-Enhancer, in contrast to the first set of data. The long-range regulatory region pairs are summarized in Supplementary Figure S1B. In pursuit of elucidating the mechanism of DNA motif interactions, protein-protein interaction data are obtained from the STRING database, resulting in the extraction of 4,950,896 pairs of experimentally validated data. By comparing the protein names in the STRING database with transcription factors (TFs) in the JASPAR database, experimentally verified TF-TF interactions are identified, encompassing a total of 65,290 TF-TF interactions, involving 583 TFs.

2.2 Generation of background sequences

We utilize a third-order Markov model (Eddy, 2004) to create background sequences corresponding to each sequence (referred to as the real sequence) within the input sequence pairs. The generated background sequences are designed to align with the number and length of the given chromatin sequences, and their composition is determined by the nucleotide frequencies observed in the dataset.

2.3 Identification of significant **k**-mers

We enumerate all possible *k*-mers (with $k = 6$ by default) employing a sliding window approach in both the real and background sequence sets concurrently. Let $n_F(k_i)$ and $n_B(k_i)$ represent the counts of occurrences of a *k*-mer k_i in the real and background sequence sets, respectively. Similarly, let $p_F(k_i)$ and $p_B(k_i)$ denote the frequency of each *k*-mer k_i in the real and background sequence sets, respectively. Recognizing the reverse complementary nature of DNA, we define the frequency of a *k*-mer as the sum of the frequencies of the *k*-mer and its reverse

complementary counterpart. Additionally, we exclude *k*-mers such as AAAAAA due to insufficient variation and discriminative power. Including them in the statistics could introduce noise and compromise the performance of the model. Assuming that the frequency distribution of *k*-mers follows a normal distribution, we retain *k*-mers with frequencies exceeding one standard deviation in the real sequences, deeming these *k*-mers as significant. Subsequently, we maintain the same selection *k*-mers in the background sequences. Following this, we use a two-proportion z-test with the null hypothesis that the frequencies $p_F(k_i)$ in the real sequence sets and $p_B(k_i)$ in the background sequence sets are the same to evaluate the significance of *k*-mers occurrences (Eqs 1-4):

$$H_0:\ p_F(k_i) = p_B(k_i), \tag{1}$$

$$H_1:\ p_F(k_i) > p_B(k_i), \tag{2}$$

$$z_i = \frac{p_F(k_i) - p_B(k_i)}{\sqrt{p_i(1-p_i)\left(\frac{1}{\sum n_F(k_i)} + \frac{1}{\sum n_B(k_i)}\right)}}, \tag{3}$$

where,

$$p_F(k_i) = \frac{n_F(k_i)}{\sum_j n_F(k_i)},\ p_B(k_i) = \frac{n_B(k_i)}{\sum_j n_B(k_i)},\ p_i = \frac{n_F(k_i) + n_B(k_i)}{\sum n_F(k_i) + \sum n_B(k_i)}. \tag{4}$$

Where the *k*-mer k_i is considered a core *k*-mer if it corresponds to a z-score greater than 1.96.

2.4 Construction of composite heterogeneous graph

We treat each *k*-mer as a node and construct a composite heterogeneous graph by establishing edges between them. Based on the positional information of each type of *k*-mer in the real sequence pairs, if two distinct types of *k*-mers are situated in different sequences within a sequence pair, we establish a connection between these two *k*-mers, referring to this connection as pair edges. The weights for pair edges are computed using Eq. 5. The first term in Eq. 5 assesses the practical significance of the edge connection between nodes v_i and u_j based on the number of sequence pairs they co-occur in. If they appear frequently together, the edge weight will be higher. The second and third terms in Eq. 5 consider the topological structure of the graph. They incorporate the number of neighborhoods for nodes v_i and u_j, respectively, relative to the total number of k-mers belonging to enhancers and promoters. This helps balance the importance of the nodes in the graph. Next, we introduce the concept of a neighborhood: for *k*-mers of the same type (promoter or enhancer), if one *k*-mer differs from another *k*-mer by only one mismatched base or has at least four consecutive identical bases, we consider the two *k*-mers as neighbors and establish a connection between them, denotes as neighborhood edges. The weights for neighborhood edges are determined using Eq. 6. It considers the proportion of common k-mers between nodes v_p and v_q relative to the total number of k-mers in each node. Higher weights indicate a higher similarity or overlap between the k-mers, which signifies a stronger relationship in the graph. Finally, we normalize the weights for the edges of the graph *G* using Eq. 7 for ensuring that the weights are scaled appropriately relative to each other. In this framework, *k*-mers, treated as nodes, and the interconnected edges between *k*-mers collectively form the weighted heterogeneous graph *G*.

$$\omega(v_i, u_j) = \frac{N(v_i, u_j) - N_{min}(v_i, u_j)}{N_{max}(v_i, u_j) - N_{min}(v_i, u_j)} + \frac{L(v_i)}{n(E)} + \frac{L(u_j)}{m(TSS)}, \tag{5}$$

$$\omega(v_p, v_q) = \frac{L(v_p \cap v_q)}{L(v_p)} + \frac{L(v_p \cap v_q)}{L(v_q)}, \tag{6}$$

$$\omega' = \frac{\omega - \omega_{min}}{\omega_{max} - \omega_{min}}, \tag{7}$$

where, v_i is a *k*-mer belonging to enhancer sequences, u_j is a *k*-mer belonging to promoter sequences, $N(v_i, u_j)$ represents the number of sequence pairs in which v_i and u_j belong, $L(v_i)$ and $L(u_j)$ represent the number of neighborhoods for v_i and u_j separately, $n(E)$ and $m(TSS)$ represent the num of *k*-mers belonging to enhancers and promoters, respectively. $L(v_p \cap v_q)$ represents the num of union of v_p and v_q, $L(v_p)$ and $L(v_q)$ represent the num of v_p and v_q, separately. ω denotes the weights of edges in graph *G*, ω_{max} and ω_{min} represent the maximum and minimum weights of edges in graph *G*, respectively.

2.5 The acquisition of dense subgraphs

We apply a community discover detection algorithm to identify dense subgraphs. Firstly, we define the fitness function for evaluating the density of a subgraph. Let *S* be a connected subgraph of graph *G*, where *V* represents the vertex set of subgraphs *S*, and E_S represents the edge set of S. Let $n_S = |V_S|$ and $m_S = |E_S|$. By adding $\frac{n_S(n_S-1)}{2} - m_S$ edges to *S*, we form a complete graph S', with the newly added weights are set to the average weight of graph *G*. The density of subgraph *S* is assessed by considering the difference in weights between the existing edges in *S* and the newly added edges. Eq. 8 outlines the community evaluation function $f(S)$ for subgraph *S*:

$$f(S) = \sum_{v_i u_j \in E(S)} \omega(v_i u_j) - \frac{1}{2|E(G)|}(n_S(n_S-1) - 2m_S) \cdot \sum_{v_i u_j \in E(G)} \omega(v_i u_j). \tag{8}$$

Obviously, the larger $f(S)$, the denser the subgraph *S* in the given sense. For a node $v \notin V_S$, the fitness function $\delta_S(v) = f(S \cup \{v\}) - f(S)$ for v in *S* is defined, and the node that the maximizes fitness function, i.e., $\delta_S(v) > 0$, is added to the existing subgraph *S*.

It is worth noting that when identifying dense subgraphs, we select the point with the highest number of neighborhoods in the core pairs, possessing the highest weight, as the initial point. An iterative process ensues, continuing until no node is found that satisfies the condition, resulting in the formation of the current dense subgraph. Subsequently, we select the two nodes from the pair with the highest weight among the remaining core pairs as the initial nodes for the iterative process. Nodes that do not belong to any dense subgraph are considered isolated points and are excluded from the analysis. For the resulting dense subgraph $C = \{C_1, C_2, \cdots, C_t\}$, where t denotes the number of obtained

dense subgraphs, we define the overlap degree of nodes of subgraph C_i and subgraph $C_j\,(1 \le i, j \le t)$ as $|C_i \cap C_j| / \min(|C_i|, |C_j|)$. Simply put, it is the count of shared nodes between both C_i and C_j divided by the smaller of the two sets' node counts. If the overlap degree of nodes is greater than 0.5, we merge the two subgraphs C_i and C_j. Additionally, during the process of obtaining a dense subgraph, we record the type to which each *k*-mer belongs in the subgraph, as well as the weight of a *k*-mer pair formed from two types of *k*-mer.

2.6 Merger and extension of **k**-mers

We extend the *k*-mers identified in the dense subgraphs obtained in the previous step. First, considering that we have recorded the type to which each *k*-mer belongs in the subgraph, we categorize all *k*-mers in each dense subgraph into two groups: enhancer *k*-mers and promoter *k*-mers. The two *k*-mers corresponding to the most weighted *k*-mer pair in each dense subgraph serve as the centers for the two types of *k*-mers. Next, we compare each *k*-mer in each type to the central *k*-mer, determining the position of each *k*-mer by assessing whether the relationship is a mismatch or an overlap. During the construction the position weight matrix (PWM), the frequency of each base corresponds to the frequency of its *k*-mer in the sequence. The two PWMs obtained from the dense subgraph constitute the initial DNA motif pairs. Subsequently, we use FIMO to scan the positions of the two PWMs in the real sequences. If the two PWMs appear in the sequence pair respectively, we consider them to be the final DNA motif pairs.

2.7 Evaluation methods for predicted DNA motif pairs

Three evaluation methods are introduced to assess the performance of the predicted DNA motif pairs. Two of these methods are utilized to evaluate the accuracy of the predicted DNA motif pairs, while the third method is employed to assess the enrichment of the predicted DNA motif pairs.

The first evaluation method is DNA motif pair distance (*MPD*), which is defined by MotifHyades and computed using Eq. 9 (Wong, 2017). The metric *MPD* is employed to assess how well the predicted DNA motif pairs $M = \{(M_P^i, M_E^i) \mid i \in N, i \le K\}$ can be matched to the known DNA motif pairs $m = \{(m_P^i, m_E^i) \mid i \in N, i \le K\}$ inserted into simulated sequence pairs:

$$MPD = \frac{1}{K}\sum_{i=1}^{K} min_x \left(D\left(m_P^i, M_P^x\right) + D\left(m_E^i, M_E^x\right)\right), \tag{9}$$

where $D(H_1, H_2)$ denoted the standard DNA motif distance between DNA motif H_1 and H_2 (Wong et al., 2013).

The second evaluation metric is DNA motif pair found ratio (*MPFR*), which is computed by Eq. 10 and used to estimate how many statistically significant DNA motif pairs are found correctly. A DNA motif H_1 is deemed a statistically significant ($p < 0.005$) match to another DNA motif H_2 when the standard DNA motif distance $D(H_1, H_2)$ is less than 0.5 according to the empirical distribution of random DNA motif patterns (Wong et al., 2013).

$$\text{MPFR} = \frac{1}{K}\sum_{i=1}^{K} I\left[D\left(m_P^i, M_P^{x'}\right) < 0.5 \wedge \left(m_E^i, M_E^{x'}\right) < 0.5\right], \tag{10}$$

where $x' = \arg min_x \left(D\left(m_P^i, M_P^x\right) + D\left(m_E^i, M_E^x\right)\right)$ and I[*condition*] is the Iverson bracket used in mathematical notation and represents logical true-or-false conditions.

The third evaluation metric involves assessing the statistical significance of the enrichment of the predicted TF pairs with known TF pairs through hypergeometric testing, as computed using Eqs 11, 12:

$$p_{value} = phyper\left(m, \frac{n(n-1)}{2}, M, \frac{N(N-1)}{2}\right), \tag{11}$$

where,

$$phyper(x_1, y_1, x_2, y_2) = \sum_{k=x_1}^{\min(y_1, x_2)} \frac{y_1!(y_2 - y_1)!x_2!(y_2 - x_2)!}{y_2!k!(y_1 - k)!(y_2 - x_2 - y_1 + k)!}, \tag{12}$$

x_1, y_1, x_2 and y_2 are any non-negative integers. N corresponds to the number of TFs in the STRING database, while M represents the number of TF pairs. Similarly, n and m denote the number of TFs in the predicted TF pairs and the number of predicted TF pairs, respectively.

3 Results

3.1 Benchmarking MPCHG on simulation datasets and real datasets

To assess the accuracy of predicted motif pairs, we generated a total of 9000 sets of simulated data for different parameters and computed both the DNA motif pair distance and DNA motif pair found ratio for these 9000 sets of simulated data. Additionally, to explore the biological significance of predicted motif pairs, we identified that they may contribute to chromatin interactions based on the transcription factors they bind. Finally, we compared the accuracy of predicted motifs with existing software and found MPCHG to exhibit higher accuracy.

3.1.1 The DNA motif pairs predicted by MPCHG obtained high quality DNA motif pair distance and DNA motif pair found ratio on different simulation data

We generate 9000 sets of simulation data to evaluate the performance of MPCHG. The simulated sequences follow a Gaussian distribution with a mean of 500 nucleotides and standard deviation of 20 nucleotides for basic benchmarking. The number of DNA sequence pairs T) is varied from 100 to 1000. Subsequently, we randomly select H DNA motif profile matrices from the JASPAR database, and the number of DNA motif pairs is varied from 3 to 100 through random combinations. We select the base-generating string corresponding to the number with the highest probability based on the distribution of bases at each position within each profile matrix. Afterward, we randomly replace the selected strings in the sequence pairs with the generated string pairs. The complete performance spectrum is visualized in Figure 2A and Figure 2B. The DNA motif pairs identified by MPCHG consistently exhibited high-quality DNA

10.3389/fgene.2024.1424085

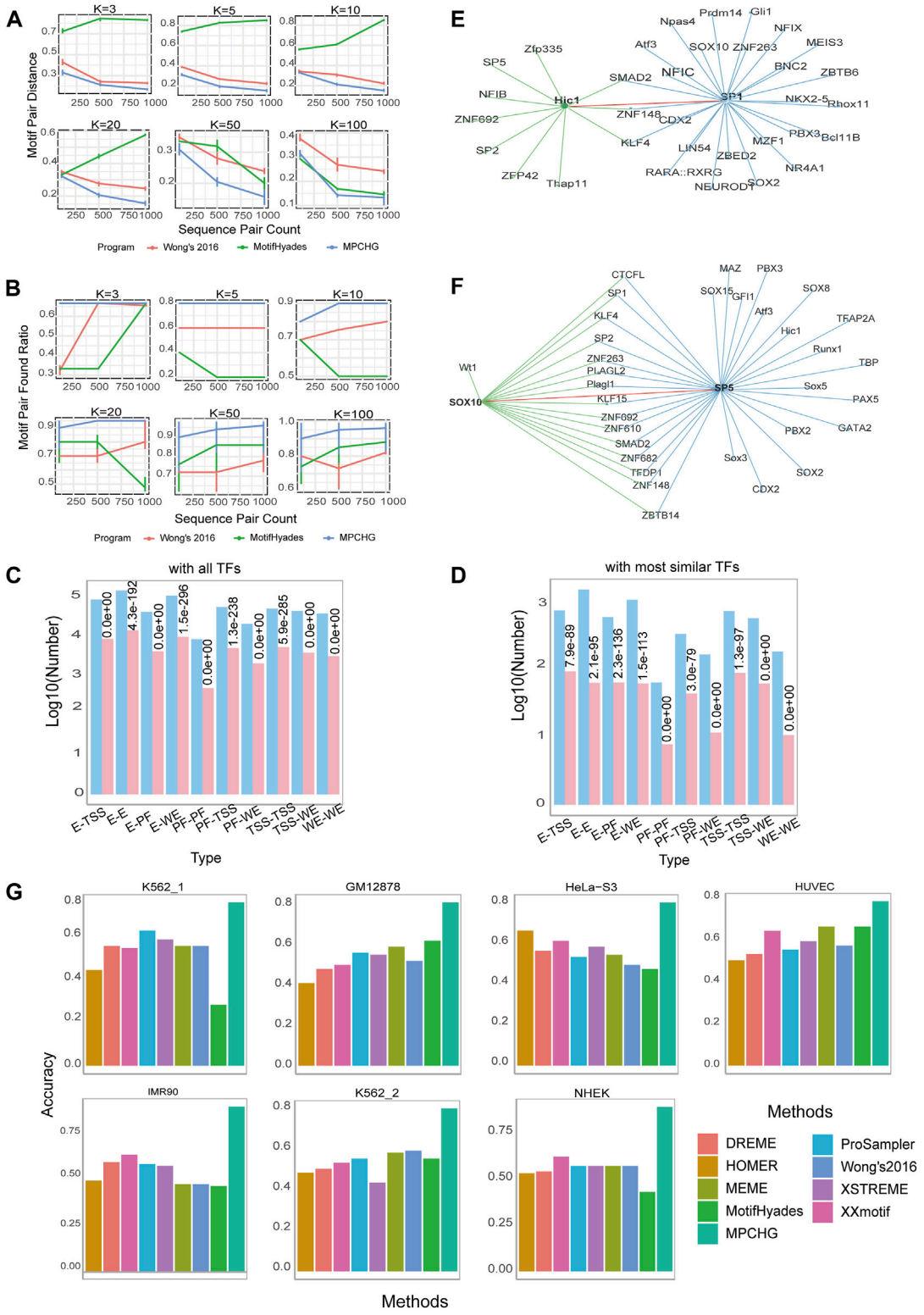

FIGURE 2
Performance of MPCHG on simulation datasets and real datasets. **(A)** Line chart for Motif Pair Distances (i.e., MPD). on known DNA motifs from JASPAR. **(B)** Line chart for Motif Pair Found Ratio (i.e., MPFR). on known DNA motifs from JASPAR. **(C,D)** Histogram on the predicted DNA motif pairs enriched with known interacting TF pairs. The red columns indicate the predicted log10 (TF pairs num) and the blue columns indicate the predicted log10 (TF pairs num supported by STRING database, which is experimentally proven). **(E)** The TF pair Hic1-SP1 in the network of the TFs corresponding to the predicted DNA motifs in K562_1 cell line. The green line represents the TFs interacting with TF Hic1 and the blue line represents the TFs interacting with TF SP1. The red line indicates the interaction between Hic1 and SP1, leaving out some of the lines between the interacting TFs. **(F)** The TF pair SOX10-

(*Continued*)

FIGURE 2 (Continued)
SP5 in the network of the TFs corresponding to the predicted DNA motifs in HUVEC cell line. The green line represents the TFs interacting with TF SOX10 and the blue line represents the TFs interacting with TF SP5. The red line indicates the interaction between SOX10 and SP5, leaving out some of the lines between the interacting TFs. **(G)** Histogram of the accuracy comparison between MPCHG and other six methods on seven datasets. The horizontal axis represents different methods, while the vertical axis indicates the accuracy of predicted DNA motifs.

motif pair distances and DNA motif pair found ratios across diverse simulation datasets. In addition, we can see from Figure 2A and Figure 2B that with more sequence pairs, MPD decreases while MPFR increases, suggesting MPCHG's better generalization on larger datasets and its potential for enhanced robustness, leading to more reliable and accurate predictions.

3.1.2 The interacting TF pairs obtained by predicted DNA motif pairs were significantly enriched with known interacting TF pairs and the TF pairs obtained by predicted DNA motif pairs may contribute to chromatin interactions

It is widely recognized that the interaction between DNA motifs is facilitated by transcription factors (TFs). Thus, we predict TF interactions based on the interactions between DNA motifs (Yu et al., 2006). Utilizing the JASPAR database, we can retrieve information about which TFs bind to each DNA motif. Subsequently, we compare the predicted DNA motifs with DNA motifs in the JASPAR (NON-REDUNDANT) DNA-JASPAR CORE (2022) vertebrates database to identify the TFs associated with the predicted DNA motifs. Based on the interactions between DNA motifs and the TFs bound by each DNA motif, we derive TF pairs. During the process of obtaining TF pairs from DNA motif pairs, it is noteworthy that a DNA motif may bind to multiple TFs. Therefore, we consider two approaches for the TFs associated with a predicted DNA motif: one involves including all TFs for the predicted DNA motif, while the other involves considering only 1 TF. For a given DNA motif, we first identify the most similar DNA motif in the database, i.e., the DNA motif corresponding to the lowest *p_value*. Subsequently, we designate the TF of this most similar DNA motif as the TF of our predicted DNA motif. This yields two types of TF pairs corresponding to predicted DNA motif pairs.

To assess whether the predicted TF pairs are enriched with known TF pairs, we collect experimentally validated interacting TFs in STRING database (Szklarczyk et al., 2023). Then, we use hypergeometric testing to calculate the p_{value}, evaluating the statistical significance of the enrichment of the predicted TF pairs with known TF pairs. The findings for the K562_1 cell line are illustrated in Figure 2C and Figure 2D, while the results for the remaining 6 cell lines are presented in Supplementary Figures S2A, S2B. These figures unveil a notable and statistically significant enrichment of the predicted interacting transcription factor (TF) pairs with the established TF interactions in the STRING database.

The TF pairs we have predicted are likely to play a role in chromatin interactions. To illustrate, by comparing our predicted TF pairs with experimentally validated TF pairs in the STRING database, we identify a novel predicted TF pair, HIC1-SP1, as depicted in Figure 2E. HIC1 is a transcription factor (TF) classified as a member of the BTB/POZ (Broad complex, Tramtrack, Bric à brac or poxvirus and zinc finger) zinc finger family. These TFs are characterized by the presence of an N-terminal POZ domain involved in protein-protein interactions and a C-terminal zinc-finger binding domain for direct DNA interaction. A recent report reveals that HIC1 can act as both a transcriptional repressor and an activator during induction of human regulatory T cells (Ray and Chang, 2020). SP1, also known as specificity protein 1*, is a protein that in humans is encoded by the SP1 gene. The protein encoded by this gene is a zinc finger transcription factor that binds to GC-rich DNA motifs of many promoters (Al-Sarraj et al., 2005). Notably, Hypoxia repressed SIRT1 transcription through promoting the competition between Sp1 and HIC1 on the SIRT1 proximal promoter in a SUMOylation-dependent manner (Sun et al., 2013). Based on this, the competitive relationship between SP1 and HIC1 may regulate gene transcription by influencing chromatin structure and status. Furthermore, another novel DNA motif pair, SOX10-SP5, as illustrated in Figure 2F, is predicted in HUVEC cell line. Sox10 is present in all neural crest cells and plays a particularly vital role in determining the fate, viability, and maturation of Schwann cells originating from neural crest stem cells (Mao et al., 2014). SP5 binds to the GC box, a DNA motif present in the promoter of a very large number of genes (Harrison et al., 2000), and is an essential early regulator of neural crest specification in *xenopus* (Park et al., 2013). Furthermore, experimentally validated by Choi et al. demonstrated that knocking down Sp5 on the initial steps of neural crest development could result in complete loss or reduction of the expression of NC markers Sox10 (Park et al., 2013). Thus, it is likely that the interaction of SOX10-SP5 contributes to chromatin interactions, allowing their transcripts to co-localize in the neural crest region (Park et al., 2013).

3.1.3 MPCHG achieved a higher accuracy than existing methods in identifying DNA motifs

We finally assess the accuracy of the DNA motifs obtained in the intermediate process to understand the degree of overlap with existing DNA motifs. We conduct a comparative analysis of MPCHG against six state-of-the-art DNA motif-finding tools, namely, DREME (Bailey, 2011), HOMER (Heinz et al., 2010), MEME (Bailey et al., 2006), ProSampler (Li et al., 2019), XSTREME (Grant and Bailey, 2021), and XXmotif (Hartmann et al., 2013). All these tools utilize the JASPAR database as a reference and employed the TomTom software (Gupta et al., 2007) with default parameters for assessment. In particular, MEME requires user to specify the number of DNA motifs, and after systematic testing at output settings of 50, 100, 150, and 200 DNA motifs, the optimal parameter of 50 is determined (yielding the highest accuracy in comparison with the JASPAR database). The remaining parameters of the above-mentioned algorithm are set to their default values, the accuracy of each method can be observed in Figure 2G. The results show that the

10.3389/fgene.2024.1424085

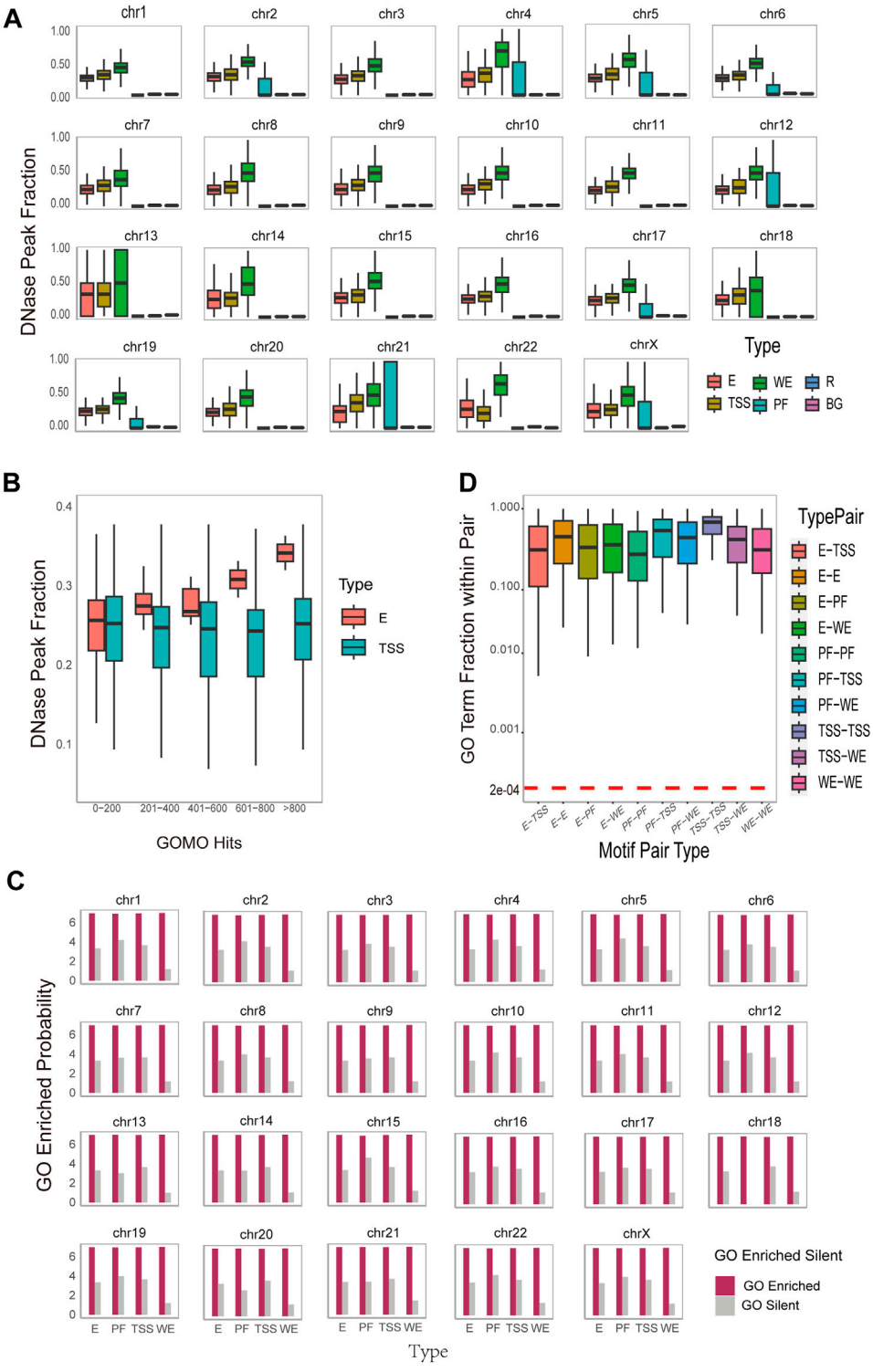

FIGURE 3
Functional and Spatial-level analysis of identified motif pairs. **(A)**. Box plots on the DNase hypersensitivity peak fraction of the DNA motifs found on different region types (i.e., WE (Weak Enhancer), E (Enhancer), TSS (Promoter), PF (Promoter-Flanking Region), R (Regulatory Region Background), BG (Background)) on different chromosomes. The horizontal axis represents different type of DNA motifs, while the vertical axis, DNase Peak Fraction, represents the ratio of the number of DNA motifs that overlap with DNase hypersensitive sites to the total DNA motifs. **(B)**. Box plots on the DNase hypersensitivity peak fraction of the DNA motifs found on different region (i.e., E (Enhancer) and TSS (Promoter)) with varying numbers of enriched Gene Ontology (GO) terms. **(C)**. Histogram of GOMO gene ontology enrichment results, with DNA motifs identified and sorted by type (horizontal axis), and the vertical axis is converted to 7+log(probability the proportion of DNA motifs with at least one GO term in each type). For each DNA motif, the term "GO Enriched" indicates that it has at least one statistically significant GO term identified by GOMO, while the term "Silent" indicates that there is no statistically significant GO term identified by GOMO. **(D)**. Boxplot on the overlap coefficients (Szymkiewicz-Simpson coefficients) between the enriched GO terms of
(*Continued*)

FIGURE 3 (Continued)
the first DNA motif and those of the second DNA motif within each DNA motif pair. The arrangement is sorted by type on the horizontal axis. A horizontal red dashed line serves as a reference for the expected overlap coefficient under the null hypothesis. This assumption posits that the overlap is entirely random, featuring a uniform hit distribution for all identified GO terms in this study conducted by GOMO.

accuracy of motifs predicted by MPCHG on 7 sets of data ranges from 75.0% to 88.7%. In contrast, the accuracy of the other six methods range from 28.0% to 66.7%. Where, MotifHyades exhibits the lowest accuracy at 28.0% on K562_1 cell line. Overall, MPCHG demonstrates an improvement of around 60% in accuracy compared to the other six methods. Notably, the accuracy of MPCHG averaged around 80% across various cell lines, indicating its high robustness.

3.2 DNA motif spatial accessibility and functional correlations provide insights into predicted DNA motif pairs

Exploring the spatial accessibility of motif pairs can unveil their mechanisms of action in gene regulation. By assessing the spatial accessibility of these motif pairs, we can determine which gene regions are more prone to transcription factor binding, thus gaining deeper insights into key nodes within the gene regulatory network. Furthermore, investigating the functional correlations between motif pairs can reveal their synergistic roles and functional regulations in biological processes, thereby understanding their functions and regulatory mechanisms in specific biological processes.

3.2.1 DNA motifs predicted by MPCHG are spatial accessible and DNase peak fractions of different type of DNA motifs have different correlation to the number of enriched GO terms

Exploring the accessibility of DNA motifs is instrumental in identifying gene regions prone to transcription factor binding, offering insights into the underlying mechanisms of gene regulation. To investigate DNA motifs accessibility, we download the DNase Chip-seq peak-calling data (Supplementary Table S1) from the ENCODE consortium (Dunham et al., 2012) across 6 cell lines. We calculate how many DNA motifs overlap with DNase hypersensitive sites on the reference hg19 human genome. To measure the significance of DNase Peak Fraction, we adopt the approach used by Wong in 2016 (Wong et al., 2016). For each DNA motif instance, we randomly sample 100 sites of the same width from both the regulatory region and the entire region of the same chromosome. This process yields regulatory region background DNase peak fractions (denoted as R) and overall background DNase peak fractions (denoted as BG) for each chromosome, respectively. The result (Figure 3A) for K562_1 cell line and the result (Supplementary Figures S3–S8) for other 6 cell line illustrate the DNase Peak Fraction for different types of DNA motifs on each chromosome individually. As depicted in the Figure 3A, the DNase Peak Fraction consistently follows a pattern across various DNA motif types: WE motifs exhibit the highest DNase Peak Fraction, followed by TSS motifs, and the lowest PF, except for the 22nd chromosome. E motifs and TSS motifs have relatively almost the same size DNase Peak Fraction, but both are higher than R and BG motifs. This suggests that WE motifs are more inclined to be open, followed by TSS motifs, and this overlapping fraction is statistically significant. We conduct t-tests and Mann-Whitney tests to measure the statistical significance of the difference between the identified DNA motifs and those in the background region. The result indicates that all $p-values$ are less than 0.01, signifying a significant overlap between DNA motifs predicted by our method and DNase hypersensitive sites.

Furthermore, DNase peak fractions exhibit distinct correlations with the number of enriched GO terms for different type of DNA motifs. As illustrated in Figure 3B, for enhancer motifs, the DNase peak fraction displays a positively correlation with the number of enriched GO terms, while for TSS motifs, it remains almost unchanged. This observation may be attributed to the fact that enhancers, responsible for gene expression regulation, are typically located in open chromatin regions known as DNase hypersensitive sites. These sites, susceptible to nucleases like DNase I, represent chromatin regions that are not tightly bound in the nucleus, allowing easier access to DNA structures by regulatory elements such as transcription factors. Consequently, the increase in the number of GO terms associated with enhancer motif enrichment and their overlap ratio with DNase hypersensitive sites may be attributed to the likelihood of these enhancers being situated in open chromatin regions. This accessibility facilitates interactions with regulators, influencing the enrichment of GO terms. On the other hand, TSS motifs are commonly found in the promoter region of a gene, associated with the transcription start site. While these motifs play a crucial role in gene initiation, an increase in their number does not lead to a significant change in the overlap ratio with TSS. This is because the location of TSS motifs in the promoter region is relatively fixed, and there is no direct correlation with an increase in the number of GO terms. Despite the increase in enriched GO terms for enhancer motifs, these terms do not directly impact the distribution of TSS motifs. Therefore, the overlap ratio with TSS remains largely unchanged. This phenomenon underscores the importance of distinguishing between various regulatory elements and factors in the study of gene regulation. It emphasizes the necessity of considering their intricate interactions within the gene expression regulatory network.

3.2.2 DNA motifs predicted by MPCHG are enriched with GO terms and the two DNA motifs coupled within one DNA motif pair are functional consistency

Ontology enrichment analysis serves as a crucial bioinformatics tool, facilitating the identification of significant enrichment in a group of genes or gene-associated entities in biological functions and processes (Peng et al., 2019). This analysis provides comprehensive insights into the functional characteristics of the study subject, shedding light on its significant roles in biology. To conduct this analysis, we use GOMO software for Gene Ontology enrichment on

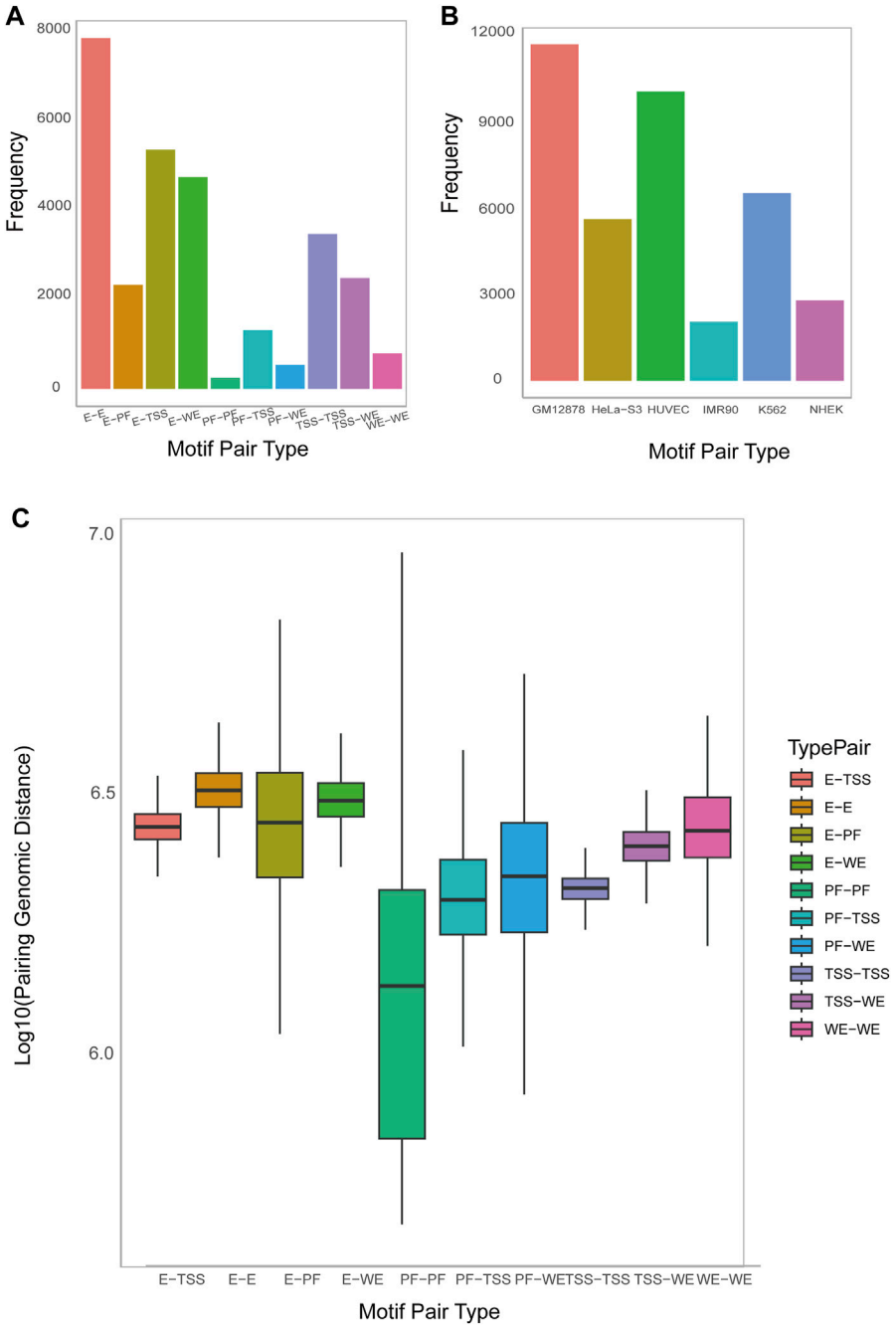

FIGURE 4
The num and spatial distribution relationship of identified DNA motif pairs. **(A)**. Histogram on the number of predicted motif pairs annotated to different types by ChromHMM and Segway (i.e., E (Enhancer), WE (Weak Enhancer), TSS (Promoter), and PF (Promoter-Flanking Region)). **(B)**. Histogram on the number of predicted motif pairs annotated by six cell lines. Their chromatin interaction type is exclusively Promoter-Enhancer. **(C)**. Boxplot on the average genomic distance between the motif instances of the first DNA motif and those of the second DNA motif within each DNA motif pair, sorted by type (horizontal axis).

each DNA motif obtained (Buske et al., 2010). In short, GOMO scans all promoters using the provided DNA motifs to determine if any DNA motif is significantly associated with genes linked to one or more Gene Ontology (GO) terms. This process is significant for understanding the biological roles of the DNA motifs. The results are depicted in Figure 3C and Supplementary Figures S9–S14. Notably, on average, more than 97% of DNA motifs exhibit enrichment for at least one GO term. This observation suggests that the predicted DNA motifs play a discernible role in gene regulation, cellular processes, or other biological functions, and their functions may be relatively extensive and universal. Among the top frequent terms, we observe the DNA motifs-related GO terms such as (GO:0048731 system development) (GO: 0048513 animal organ development), (GO:0030154 cell differentiation), and (GO:0003700 DNA-binding transcription factor activity).

Furthermore, our interest extends to the functional roles between the two DNA motifs within each DNA motif pair. To explore this, we calculate the overlap coefficient (Szymkiewicz-Simpson coefficient) between the enriched GO terms of the first DNA motif and those of the second DNA motif within each DNA motif pair. The results of the overlap coefficient are illustrated in Figure 3D and Supplementary Figure S15. The observed overlap coefficients are higher than expected, indicating a substantial overlap between the two DNA motif-related GO term set. This suggests a potential functional or biological correlation between the two motifs. Notably, the overlap coefficient for TSS-TSS interaction is the highest, implying that interactions between promoters may be functionally more closely related, involved in more common biological processes, and exhibit stronger functional correlations. These findings provide valuable insights for a deeper understanding of promoter interaction in gene regulatory network and biological processes. Additionally, they offer guidance for further functional annotation and research into regulatory mechanisms.

3.3 DNA motif pairs predicted by MPCHG unveiled genomic distance characteristics in human cell lines

To analyze genomic distance signatures within chromatin structures, we first counted the motif pairs predicted by MPCHG on seven cell lines. Through genomic distance analysis of the predicted motif pairs, MPCHG reveals the spatial relationships and interactions between the regulatory elements. Notably, these findings highlight the universality of long-distance regulatory mechanisms, and in particular enhancers play a key role in facilitating precise gene regulation.

3.3.1 DNA motif pairs were discovered by MPCHG on seven human cell lines

MPCHG has run on the seven cell lines (K562_1, GM12878, HeLa-S3, HUVEC, IMR90, K562_2, NHEK) to obtain ten thousand of DNA motif pairs. We counted the number of DNA motif pairs of 10 chromatin interaction types on the K562_1 cell line and the number of promoter-enhancer-pairs on the remaining six cell lines. The discovered DNA motif pairs are visualized Figure 4A and Figure 4B.

3.3.2 The genomic distance between DNA motifs pairs predicted by MPCHG revealed the interaction and relative position between the regulatory elements

Analyzing the distances between DNA motifs provides insights into the relative positioning and interactions of gene regulatory elements, indicating whether they are in close proximity or distantly located within the three-dimensional chromatin structure (Dekker and Misteli, 2015). This analysis enhances our understanding of the organization and spatial regulation of gene expression at the chromatin level. Therefore, Accordingly, we have computed the distance between DNA motifs of different interaction types. As depicted in Figure 4C, the interaction distance between E-E is the greatest, followed by E-WE, E-TSS, and E-PF and Supplementary Figure S16 also indicate that enhancers are far away from Promoters. This observation aligns with the widely accepted notion that enhancers are typically situated in regions far away from the genes they regulate, sometimes spanning millions of base pairs (bp). This long-distance regulatory action is facilitated through the establishment of chromatin loops, enabling effective and precise regulatory interactions.

4 Discussion

Identifying DNA motifs is of paramount importance in biology and computational biology. DNA motifs are short sequence patterns in protein or nucleic acid sequences that are functionally relevant. They are crucial for functional annotation, structure prediction, evolutionary relationships, and regulatory element recognition. Furthermore, the identification of DNA motif pairs in interacting sequences is also significant as it aids in predicting protein-protein interactions, drug design, and disease research. In conclusion, DNA motifs and their pairs play pivotal roles in biological research and medical applications.

Hence, we propose the MPCHG algorithm to identify tens of thousands of DNA motif pairs in the long-range chromatin interaction sequences. First, we use a 3-order Markov model to generate background sequences that matches the length and composition of the original sequence, ensuring statistical significance and rationality for *k*-mer seeds. In contrast to many algorithms that exhaustively determine DNA motif length within a specific range, our method extends the core DNA motif to both ends using a double-sample z-test. This approach aligns the predicted DNA motif more closely with real scenarios. At the same time, algorithms that set the DNA motif length in advance may miss some important DNA motifs or introduce high noise. Furthermore, we construct a composite heterogeneous graph for different types of *k*-mers (enhancer *k*-mers from enhancer sequences and promoter *k*-mers from promoter sequences). This graph connects *k*-mers of different types present in the same sequence pair. Simultaneously, it captures complex relationships among *k*-mers of the same type with mismatched or overlapping connections. To obtain a dense subgraph related to *k*-mers, we define the fitness function of the subgraph to assess its density. Nodes meeting specific conditions are extended to the current seed. Finally, we merge and extend the obtained subgraph to extract DNA motifs. Subsequently DNA motifs scanning enables the identification of DNA motif pairs.

Regarding the predicted DNA motif pairs, we conducted a thorough analysis covering various aspects. The accuracy rate, measured by comparing predicted DNA motifs with the JASPAR database using TOMTOM software, the accuracy of predicted DNA motifs ranged from 75.0% to 88.7%. Additionally, we employed GOMO software to explore the gene ontology enrichment of these predicted DNA motifs. The findings revealed that, on average, over 97% of DNA motifs are enriched for at least one GO term. This indicated that the predicted DNA motif play essential roles in gene regulation, cellular processes, or other biological functions, and their functions may be relatively extensive and universal. To further validate our predictions, we compare the predicted DNA motifs with DNase Chip-seq peak-calling data. The analysis demonstrated a significant overlap between DNA motifs predicted by our model and DNase hypersensitive sites. Notably, DNA motif pairs involving

enhancer or weak enhancer regions exhibited greater distance, aligning with the common understanding that regulatory components in enhancer regions are typically located far from their interacting partners, often spanning a large genomic distance. We extended our analysis to predict TF interactions based on the predicted DNA motif pairs. The result indicated that the predicted interacting TF pairs are significantly enriched with the known interacting TF pairs in STRING, as determined by hypergeometric testing. Moreover, we unveiled new TF interaction information, such as the interaction between HIC1 and SP1, suggesting a potential role in facilitating chromatin interactions and promoting gene transcription. Finally, to evaluate the generalization performance of our model, we tested it on six additional E-TSS datasets representing different cell lines (GM12878, HeLa-S3, HUVEC, IMR90, K562, and NHEK). The results demonstrated consistently good performance across these diverse datasets.

The prediction of DNA motif pairs stands as a critical challenge in bioinformatics, offering valuable insights into various biological processes, including gene regulation, protein-protein interactions, and RNA structures. While significant strides have been made in this field, the future holds immense potential for further advancements. Firstly, the continuous evolution of deep learning and artificial intelligence techniques, including innovative algorithms and graph neural networks, is expected to elevate the accuracy and reliability of DNA motif pair predictions. Secondly, the exploration of cross-species DNA motif pair prediction presents an intriguing challenge, offering opportunities to uncover conserved sequence patterns and explore evolutionary variations. Thirdly, the integration of diverse data sources, such as epigenetic data and protein interaction information, will contribute to more comprehensive annotations for predicted results, enhancing our understanding of the intricacies of biological systems. Additionally, applying DNA motif pair predictions in disease research and precision medicine holds promise for identifying potential disease markers or therapeutic targets. Lastly, the combination of DNA motif pair predictions with network interactions and systems biology approaches will enable the construction of comprehensive biological regulatory network models. This integrative approach has the potential to deepen our understanding of the fundamental principles of biology. In conclusion, ongoing research in predicting DNA motif pairs has significant potential to drive breakthroughs in biotechnology and medical advancements, fostering progress in the fields of biology and medicine.

Data availability statement

Original datasets are available in a publicly accessible repository: The original contributions presented in the study are publicly available. The input data of Hi-C data for six cell lines (K562, GM12878, HeLa-S3, HUVEC, IMR90, and NHEK) can be found here: https://doi.org/10.6084/m9.figshare.14192000. The protein-protein interaction data from the STRING database can be found here: https://cn.string-db.org/cgi/download?sessionId=blEDX7XXpqWC.

Author contributions

QWu: Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft. YL: Conceptualization, Project administration, Supervision, Writing–review and editing. QWa: Investigation, Methodology, Visualization, Writing–review and editing. XZ: Investigation, Methodology, Visualization, Writing–review and editing. DS: Conceptualization, Project administration, Resources, Supervision, Writing–review and editing. BL: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. National Key R&D Program of China (2020YFA0712400); National Nature Science Foundation of China (62272270 and 11931008); Shandong University Multidisciplinary Research and Innovation Team of Young Scholars (2020QNQT017). National Natural Science Foundation of China [No. 62202269]; Science Foundation Program of the Shandong Province (2023HWYQ-012); the Program of Qilu Young Scholars of Shandong University.

Acknowledgments

The authors would like to thank Yihua Wang for his assistance in downloading the data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1424085/full#supplementary-material

References

Al-Sarraj, A., Day, R. M., and Thiel, G. (2005). Specificity of transcriptional regulation by the zinc finger transcription factors Sp1, Sp3, and Egr-1. *J. Cell Biochem.* 94 (1), 153–167. doi:10.1002/jcb.20305

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27 (12), 1653–1659. doi:10.1093/bioinformatics/btr261

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi:10.1093/nar/gkl198

Bartlett, A., O'Malley, R. C., Huang, S. S. C., Galli, M., Nery, J. R., Gallavotti, A., et al. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* 12 (8), 1659–1672. doi:10.1038/nprot.2017.055

Bhatia, S., and Kleinjan, D. A. (2014). Disruption of long-range gene regulation in human genetic disease: a kaleidoscope of general principles, diverse mechanisms and unique phenotypic consequences. *Hum. Genet.* 133 (7), 815–845. doi:10.1007/s00439-014-1424-6

Buske, F. A., Boden, M., Bauer, D. C., and Bailey, T. L. (2010). Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* 26 (7), 860–866. doi:10.1093/bioinformatics/btq049

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 50 (D1), D165–D173. doi:10.1093/nar/gkab1113

Clauss, B., and Lu, M. (2023). A quantitative evaluation of topological motifs and their coupling in gene circuit state distributions. *iScience* 26 (2), 106029. doi:10.1016/j.isci.2023.106029

Daily, K., Patel, V. R., Rigor, P., Xie, X., and Baldi, P. (2011). MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *Bmc Bioinforma.* 12, 495. doi:10.1186/1471-2105-12-495

Dekker, J., and Misteli, T. (2015). Long-range chromatin interactions. *Csh Perspect. Biol.* 7 (10), a019356. doi:10.1101/cshperspect.a019356

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414), 57–74. doi:10.1038/nature11247

Eddy, S. R. (2004). What is a hidden Markov model? *Nat. Biotechnol.* 22 (10), 1315–1316. doi:10.1038/nbt1004-1315

Fulton, D. L., Sundararajan, S., Badis, G., Hughes, T. R., Wasserman, W. W., Roach, J. C., et al. (2009). TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.* 10 (3), R29. doi:10.1186/gb-2009-10-3-r29

Gold, L. (2015). SELEX: how it happened and where it will go. *J. Mol. Evol.* 81 (5-6), 140–143. doi:10.1007/s00239-015-9705-9

Grant, C. E., and Bailey, T. L. (2021). XSTREME: comprehensive motif analysis of biological sequence datasets. *BioRxiv* 2021-09. doi:10.1101/2021.09.02.458722

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8 (2), R24. doi:10.1186/gb-2007-8-2-r24

Harrison, S. M., Houzelstein, D., Dunwoodie, S. L., and Beddington, R. S. (2000). Sp5, a new member of the Sp1 family, is dynamically expressed during development and genetically interacts with Brachyury. *Dev. Biol.* 227 (2), 358–372. doi:10.1006/dbio.2000.9878

Hartmann, H., Guthohrlein, E. W., Siebert, M., Luehr, S., and Soding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.* 23 (1), 181–194. doi:10.1101/gr.139881.112

Hashim, F. A., Mabrouk, M. S., and Al-Atabany, W. (2019). Review of different sequence motif finding algorithms. *Avicenna J. Med. Biotechnol.* 11 (2), 130–148.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38 (4), 576–589. doi:10.1016/j.molcel.2010.05.004

Hellman, L. M., and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.* 2 (8), 1849–1861. doi:10.1038/nprot.2007.249

Khodabandelou, G., Routhier, E., and Mozziconacci, J. (2020). Genome annotation across species using deep convolutional neural networks. *PeerJ Comput. Sci.* 6, e278. doi:10.7717/peerj-cs.278

Kim, S., and Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* 83 (3), 373–392. doi:10.1016/j.molcel.2022.12.032

Li, Y., Ni, P., Zhang, S., Li, G., and Su, Z. (2019). ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. *Bioinformatics* 35 (22), 4632–4639. doi:10.1093/bioinformatics/btz290

Li, Y., Wang, Y., Wang, C., Ma, A., Ma, Q., and Liu, B. (2024). A weighted two-stage sequence alignment framework to identify motifs from ChIP-exo data. *Patterns* 5, 100927. doi:10.1016/j.patter.2024.100927

Liu, B., Yang, J., Li, Y., McDermaid, A., and Ma, Q. (2018). An algorithmic perspective of *de novo* cis-regulatory motif finding based on ChIP-seq data. *Brief. Bioinform* 19, 1069–1081. doi:10.1093/bib/bbx026

Lyu, H., Lium, E., and Wu, Z. (2020). Comparison of normalization methods for Hi-C data. *Biotechniques* 68 (2), 56–64. doi:10.2144/btn-2019-0105

Makolo, U. A., and Suberu, S. O. (2016). Gapped motif discovery with multi-objective genetic algorithm. *OALib* 03 (03), 1–6. doi:10.4236/oalib.1102293

Mao, Y., Reiprich, S., Wegner, M., and Fritzsch, B. (2014). Targeted deletion of Sox10 by Wnt1-cre defects neuronal migration and projection in the mouse inner ear. *Plos One* 9 (4), e94580. doi:10.1371/journal.pone.0094580

Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31 (1), 258–261. doi:10.1093/nar/gkg034

Park, D. S., Seo, J. H., Hong, M., Bang, W., Han, J. K., and Choi, S. C. (2013). Role of Sp5 as an essential early regulator of neural crest specification in xenopus. *Dev. Dyn.* 242 (12), 1382–1394. doi:10.1002/dvdy.24034

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10 (10), 669–680. doi:10.1038/nrg2641

Peng, J., Lu, G., Xue, H., Wang, T., and Shang, X. (2019). TS-GOEA: a web tool for tissue-specific gene set enrichment analysis based on gene ontology. *Bmc Bioinforma.* 20, 572. doi:10.1186/s12859-019-3125-6

Pfreundt, U., James, D. P., Tweedie, S., Wilson, D., Teichmann, S. A., and Adryan, B. (2010). FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res.* 38, D443–D447. doi:10.1093/nar/gkp910

Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29 (2), 153–159. doi:10.1038/ng724

Ray, H., and Chang, C. (2020). The transcription factor Hypermethylated in Cancer 1 (Hic1) regulates neural crest migration via interaction with Wnt signaling. *Dev. Biol.* 463 (2), 169–181. doi:10.1016/j.ydbio.2020.05.012

Robasky, K., and Bulyk, M. L. (2011). UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 39, D124–D128. doi:10.1093/nar/gkq992

Spivak, A. T., and Stormo, G. D. (2012). ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Res.* 40 (D1), D162–D168. doi:10.1093/nar/gkr1180

Sun, L., Li, H., Chen, J., Dehennaut, V., Zhao, Y., Yang, Y., et al. (2013). A SUMOylation-dependent pathway regulates SIRT1 transcription and lung cancer metastasis. *J. Natl. Cancer Inst.* 105 (12), 887–898. doi:10.1093/jnci/djt118

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 51 (D1), D638–D646. doi:10.1093/nar/gkac1000

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., et al. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* (39), e1869. doi:10.3791/1869

Wang, S., Hu, H., and Li, X. (2022). A systematic study of motif pairs that may facilitate enhancer-promoter interactions. *J. Integr. Bioinform.* 19 (1), 20210038. doi:10.1515/jib-2021-0038

Wang, Y., Li, Y., Wang, C., Lio, C.-W. J., Ma, Q., and Liu, B. (2024). CEMIG: prediction of the cis-regulatory motif using the de Bruijn graph from ATAC-seq. *Brief. Bioinform* 25, bbad505. doi:10.1093/bib/bbad505

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158 (6), 1431–1443. doi:10.1016/j.cell.2014.08.009

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., et al. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 28 (1), 316–319. doi:10.1093/nar/28.1.316

Wong, K. C. (2017). MotifHyades: expectation maximization for *de novo* DNA motif pair discovery on paired sequences. *Bioinformatics* 33 (19), 3028–3035. doi:10.1093/bioinformatics/btx381

Wong, K. C., Chan, T. M., Peng, C., Li, Y., and Zhang, Z. (2013). DNA motif elucidation using belief propagation. *Nucleic Acids Res.* 41 (16), e153. doi:10.1093/nar/gkt574

Wong, K. C., Li, Y., and Peng, C. (2016). Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells. *Bioinformatics* 32 (3), 321–324. doi:10.1093/bioinformatics/btv555

Yang, J., Ma, A., Hoppe, A. D., Wang, C., Liu, B., Ma, Q., et al. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Res.* 47, 7809–7824. doi:10.1093/nar/gkz672

Yu, X., Lin, J., Zack, D. J., and Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* 34 (17), 4925–4936. doi:10.1093/nar/gkl595