Three-dimensional genomic data interpretation and transcriptional regulatory element identification-final

Qitong Huang



Propositions

1. Transcription regulatory element sequences possess linguistic characteristics.
(this thesis)

2. The 25 bp sequence of MAE-seq enables more precise identification of enhancers.
(this thesis)

3. The interpretability of a model is equally important as its accuracy.

4. Reporting "non-significant" as "no difference" is using statistical uncertainty to masquerade as certainty.

5. Both personal and family health are the foundation of a successful career.

6. Artificial intelligence makes it easy to generate data, but difficult to verify it.

Three-dimensional genomic data interpretation and transcriptional regulatory element identification

Qitong Huang

Thesis committee

Promotor

Prof. Dr M.A.M. Groenen

Professor in Animal Breeding and Genomics

Wageningen University & Research

Co-promotor

Dr O. Madsen

Associate Professor in Animal Breeding and Genomics

Wageningen University & Research

Prof. Dr Yubo Zhang

Animal Genomics Research Center

Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Other members

Prof. Dr D. de Ridder, Wageningen University & Research

Dr E. de Wit, Netherlands Cancer Institute, Amsterdam

Dr A. Bossers, Utrecht University

Dr P.J. Schaap, Wageningen University & Research

This research was conducted under the auspices of the Graduate School Wageningen Institute of Animal Sciences (WIAS).

Three-dimensional genomic data interpretation and transcriptional regulatory element identification

Qitong Huang

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr C. Kroeze,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 21 October 2026
at 10.30 a.m. in the Omnia Auditorium.

Qitong Huang
Three-dimensional genomic data interpretation and transcriptional regulatory element identification, 136 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2026)
With references, with summary in English

DOI https://doi.org/10.18174/681026

Three-dimensional genomic data interpretation and transcriptional regulatory element identification

Contents

Chapter 1 General Introduction **7**
Chapter 2 Accurate and fast loop calling from Hi-C data by TLoops **29**
Chapter 3 A single-cell genomic strategy for alternative transcript start sites identification **43**
Chapter 4 MAE-seq refines regulatory elements across the genome **60**
Chapter 5 Identification of enhancer core sequence from MAE-seq data by multiple kernel learning classifier **96**
Chapter 6 General Discussion **108**
Summary **122**
Appendices **127**

Chapter 1

General Introduction

1 General Introduction

1.1 Chromatin conformation and transcriptional regulatory elements

A basic challenge in the understanding of molecular biology is how a fertilized egg develops into an individual that consists of hundreds of different cell types. Decades of research have revealed that zygote development is a complex process that involves spatiotemporal specific gene expression as well as regulatory element modulation [1, 2]. In recent years researchers have discovered that the genome structure is not the one-dimensional structure as depicted in many biological text books, but instead has a complex three-dimensional organization in the nucleus [3, 4]. Meanwhile, as different types of (epi)genomic sequencing have advanced, along with a deeper understanding of molecular mechanisms and non-coding transcriptional regulatory sequences, it has been revealed that biological phenotypes are regulated by complex gene regulatory networks (GRNs). These networks are comprised of multiple genes, regulatory elements, transcription factors (TFs), and co-factors [5]. Thus, expression of a gene is not modulated only by one gene, TF, or regulatory element, instead it tends to be regulated by multiple components of regulatory elements and multiply complex interaction mechanisms. Moreover, it was shown that the chromatin architecture could profoundly affect the GRNs [6]. For example, the folding of chromatin enables a distal regulatory element to interact with its target gene(s) hereby regulating the expression of genes. Consequently, it has become essential to explore transcriptional regulatory elements and their networks by chromatin conformation analysis. 3D genomes and structural biology data have been integrated with transcriptomics and other omics approaches to significantly improve the interpretation of GRNs and their biological effects.

In eukaryotic cells, DNA is highly condensed and stored in the nucleus in a structure called chromatin [7]. DNA replication, repair, transcription and expression processes are precisely regulated by the conformation of the chromatin. Many studies have shown that specific chromatin conformations can selectively activate or silence gene expression, thereby controlling cell self-maintenance or direct differentiation and determining cell tissue specificity and fate [8]. Hi-C [9] and ChIA-PET [10] technologies can uncover the spatial chromatin conformation, from which researchers can directly infer the interactions among genes and regulatory elements (Figure 1, Table 1). In the data analysis of Hi-C and related technologies, researchers integrated the spatial conformation of chromatin into A/B compartment, TAD and chromatin loop conformations (Figure 1). While these conformations allow us to understand changes in gene expression, they also help us more precisely to identify transcriptional regulatory elements. At the same time, a large number of transcriptional regulatory element sequences obtained based on other technologies such as MAE-seq [11], enhancer-screening [12] and eRNA capture [13] can also be assisted in verification through three-dimensional conformation.

Recent advances in computational and experimental methods have facilitated the

identification of transcriptional regulatory elements. Meanwhile, there have been continuous improvements to chromatin conformation detection techniques. Effectively analyzing and integrating these various datasets is crucial. Many tools for analyzing Hi-C data and the identification of transcriptional regulatory elements have been developed. However, they all have advantages and disadvantages, and their accuracy can vary greatly. Furthermore, there is currently a need to develop uniform standards for these tools [14]. It has also been reported that the length of transcriptional regulatory elements obtained from e.g. chromatin immunoprecipitation sequencing (ChIP-seq) data may exceed the actual size of the DNA element where TFs binds [15]. This highlights the need for a better understanding of the relationship between chromatin conformation and transcriptional regulatory factors, which has become an increasingly intriguing area of research.

Table 1. Comparisons of different chromatin conformation capture techniques. This table summarizes commonly used 3D genome mapping methods and lists their core features, main advantages, and resolution ranges.

Technique	Features	Advantages	Resolution Range
4C (Chromosome Conformation Capture)	- Detection of long-range chromatin interactions	- Targeted analysis of specific genomic regions	Kilobase range
5C (Chromosome Conformation Capture Carbon Copy)	- Analysis of specific chromatin interactions	- Quantitative measurement of interaction frequencies	Kilobase range
Hi-C	- Genome-wide analysis of chromatin interactions	- High-resolution mapping of interactions across the genome	Kilobase range
ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)	- Identification of protein-mediated chromatin interactions	- Integration of protein-DNA interactions with chromatin structure	Kilobase range
In situ Hi-C	- Chromatin conformation analysis in native nuclear context	- Preserves spatial organization of chromatin within the nucleus	Kilobase range
High-efficiency Enzyme digestion DNase Hi-C	- Sensitive detection of chromatin interactions	- Enhanced detection of long-range interactions in low-input samples	Kilobase range
Capture-Hi-C	- Targeted analysis of specific chromatin interactions	- Enables focused investigation of interactions in specific regions	Kilobase range
Single-cell Hi-C	- Chromatin conformation analysis at single-cell resolution	- Provides insights into cell-to-cell variability in 3D genome	Kilobase to megabase range
DLO Hi-C	- Analysis of chromatin loops using DNA-ligation-only Hi-C	- Simpler protocol without restriction enzyme digestion	Kilobase range

1.2 Chromatin conformation

1.2.1 One-dimensional chromatin conformation

Nucleosomes are the basic unit of chromatin, first discovered in 1974 through MNase digestion and electron microscope experiments by Kornberg [7]. Under the electron microscope view, nucleosomes are connected to other nucleosomes, presenting a classic "beads-on-a-string" structure [16] (Figure 1). It was not until 1997 that the first high-resolution (2.8 Å) crystal structure of nucleosomes was successfully obtained [17]. Structural analysis showed that the nucleosome was formed by an approximately 147 bp long DNA strand wound around a histone octamer structure (2 molecules of each histone H2A, H2B, H3 and H4 form a histone octamer). The DNA passed through 14 Å minor grooves and interacts with histone octamers to form a tight structure.

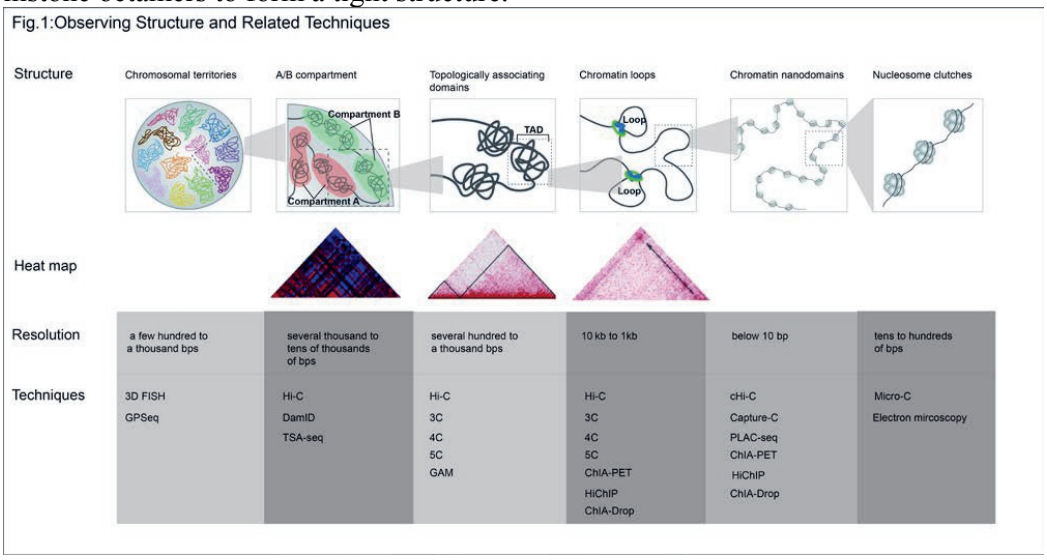


Figure 1. A schematic depicting of the structural conformation of DNA within the cell (top), representative heatmaps generated by Hi-C (accession number: GM12878) for resolving DNA architecture (middle), and a comparative summary of the techniques and their resolution capabilities for determining DNA conformation (bottom). See details of some techniques in table 1.

The discovery and determination of nucleosome structure was a milestone event in studying chromatin conformation. It also successfully opened the prelude of interpreting genomic information from chromatin structure to function. Studies have shown that the distribution of nucleosomes across the genome have particular preference: nucleosomes are sparsely distributed in regulatory regions, such as promoters, terminators and transcription factor binding sites (TFBS), and densely distributed in the rest of the genome like the transcribed gene-body [18]. However, the experiments of Saragosti et al. [19] on the nucleosome-free region and DNase hypersensitive region of the SV40 minichromosome showed that the nucleosome depleted region was highly consistent with the active gene regulatory region - the so-called DNase hypersensitive region. These results suggested that

nucleosome positioning affects gene expression and had important biological implications. At the same time, many studies have also shown that the positioning of nucleosomes is closely related to processes such as DNA replication, repair, transcription and alternative splicing [20]. For example, by observing the behaviour of single RNA type II polymerase (RNA Pol II) in the transcription elongation stage through optical tweezers experiments, Hodges et al. [18] found that the nucleosomes in the transcribed region would cause RNA Pol II to pause multiple times, resulting in a decrease in the transcription rate. Furthermore, results of nucleosomes in the promoter region of yeast showed that the number of nucleosomes in the promoter region was negatively correlated with the transcription rate of genes downstream of the promoter [21].

Currently, the main modes of action of nucleosome positioning affecting gene expression can be summarized into two categories: one is the identification of nucleosomes as markers for transcription start site (TSS), and the other is the inhibitory effect of nucleosomes. By studying the relationship between nucleosome positioning and TSS, a stable and open nucleosome-free region (NFR) was found in the promoter region, where RNA Pol II docks upon gene transcription initiation. At the downstream +1 nucleosome of NFR, RNA Pol II plays a role in regulating gene transcription and expression [22]. For genes lacking a core promoter element, the transcription complex can still correctly identify the transcription start site, but replacing the NFR without changing the core promoter sequence will shift the TSS's position [23]. This implies that nucleosome positioning is likely an important marker for the accurate recognition of TSS by the transcriptional machinery. The inhibitory effect of nucleosomes is primarily associated with the histone components of the nucleosome itself. In vitro transcription experiments with reconstituted chromatin have shown that when naked DNA forms nucleosomes with the core histones (H2A, H2B, and H4) at a density of one nucleosome per 200 bp of DNA, 75% of transcription is inhibited. The degree of inhibition is further increased when nucleosomes are cross-linked with histone H1 [24].

1.2.2 Two-dimensional chromatin conformation

Through the formation of nucleosomes, the length of DNA is compressed by a factor of 6–7, but this is far from being sufficient for storing the genome in the nucleus, which is only 10 μm in diameter. This means that chromatin has a higher-order conformational type than nucleosomes. In 1979, Finch et al. [25] first observed chromatin fibers with a diameter of about 30 nm by electron microscopy. The electron microscopy results suggested that the structure comprised 11 nm nucleosome beaded structures. The discovery of the 30 nm chromatin fibers formally advances the study of chromatin conformation to a two-dimensional level focusing on nucleosome interaction. For the 30 nm chromatin fibers many different structural models have been suggested, such as the Solenoid model [26], Helical Ribbon model [27], and two-start Zigzag model [28]. In 2014, cryo-electron microscopy single-particle three-dimensional reconstruction technology

successfully obtained the high-definition three-dimensional structure of the 30 nm chromatin fibers [29], showing that the 30 nm chromatin fibers had four nucleosomes as structural units; each unit twisted and folded to form a left-handed double helical advanced structure (termed tetranucleosome). The result of the tetranucleosome structural unit was confirmed in a study applying chromatin single-molecule mechanics [30], showing that the structure is a stable intermediate state in chromatin folding/unfolding. In this study it was suggested that the tetranucleosome might be a structure with critical gene regulatory functions by regulating gene transcription through the tightness of the chromatin fiber structure. One limitation of studies on the 30 nm chromatin fibers is that they all are based on *in vitro* experiments, and scientists have still not demonstrated the existence of this structure in cells.

1.2.3 Three-dimensional chromatin conformation

Research on the three-dimensional conformation of chromatin dates back to as early as 1885 where Rabl [31] observed different chromosomal regions within the cell nucleus. Subsequently, through techniques such as fluorescence staining and microscopy, many experiments have confirmed the presence of distinct three-dimensional structures in the cell nucleus. However, due to a lack of microscopic evidence, the study of chromatin's three-dimensional conformation did not receive much attention [32]. With the development of new sequencing technology and the completion of the Human Genome Project (HGP) [33] and the Human Genome Encyclopedia Project (Encyclopedia of DNA elements, ENCODE) [34], scientists gradually realized that the spatial structure of a genome is not linearly arranged in one dimension. Therefore, the linear molecular model had to be revised to explain these newly discovered discrete regulatory elements and the relationship between structural variation and gene function. So, since 2002, there has been a growing body of research focusing on the field of three-dimensional genomics. In 2002, Dekker et al. [35] developed a new technique called Chromosome Conformation Capture (3C) to measure specific point-to-point chromatin interactions. This technology shifted the understanding of the one-dimensional DNA sequences to a three-dimensional level. 3C became the basis for the development of subsequent variety of three-dimensional genome sequencing technology. Currently, conformational capture technologies derived from 3C technology include, among others, 4C [36], 5C [37], Hi-C [9], ChIA-PET [10], in situ Hi-C [8], high-efficiency Enzyme digestion DNase Hi-C [38], hybridization probe Capture-Hi-C [39], single-cell Hi-C [40], DLO Hi-C [41] (Table 1). These assays have provided the essential basis of long-distance interaction between genes, the interpretation of the interaction between transcription factors and chromatin, as well as the spatial conformation of interacting chromatin in the nucleus. In addition, with the further development of the electron microscopy technology, the ultra-high resolution electron microscopy technology has also provided an additional perspective for the study of the three-dimensional conformation of chromatin [42].

Based on the different conformational capture technologies and electron microscopy technology a three-dimensional hierarchical hypothetical model of eukaryotic chromatin has been proposed. On different spatial scales, these hierarchical structures are: 1) chromosome territory (CT), 2) chromatin compartment A/B [10], 3) topologically associating domain (TAD) [43] and 4) chromatin loops [8] (Figure 1).

Chromosome territory

Many studies have shown that the radial positioning of CT has a well-defined preference in the nucleus and is highly conserved in evolution. Each chromosome occupies a specific, non-overlapping area with limit contact between chromosomes. These areas, that act as a barrier, are the CTs. In general, small, gene-rich chromosomes are located near the center of the nucleus and gene-poor chromosomes at the nuclear periphery [32]. In a study of bovine embryos Koehler et al. [44] found that the first appearance of CT radial positioning preference in the nucleus was related to the activation of the embryonic genome, and was fully established at the blastocyst stage. In addition, other studies have shown [45] that the localization of CT is related to cell replication time and transcriptional activity. Early-replicating regions and active genes tend to be located in the center of the nucleus. In contrast, late-replicating regions and repressed genes tend to be found at the nuclear periphery. This phenomenon is consistent with the generally accepted notion that the nuclear lamina represses gene expression. However, there are exceptions like the CT localization in retinal rod cells of nocturnal mammals which is opposite to that of most diurnal eukaryotes, with heterochromatin located in the nucleus center and euchromatin located in the nucleus periphery [46].

Compartment A/B

In 2009, Lieberman-Aiden et al. [10] applied Hi-C to uncover the three-dimensional configuration of the chromatin of human lymphoblastoid cells. They revealed an important feature of the spatial arrangement of chromatin: chromatin is composed of two alternating compartment called A and B. Compartment A is characterized by high gene density, active transcription, and higher accessibility (such as being highly sensitive to DNase I). In contrast, compartment B exhibits dense folding, low gene density, and enrichment of histones with inhibitory gene expression features, such as H3K27me3. By analyzing associations, the researchers correlated Hi-C model data for each chromosome with replication timing for other epigenetic marks and discovered that compartment A and B corresponded respectively to cytological maps of euchromatin and heterochromatin. Besides gene expression levels, the distribution of compartment B significantly overlaps with nuclear lamina-associated domains (LADs) primarily located in the nuclear fiber layer nearby [47]. These findings suggest that compartment A/B is essentially a rediscovery of the distinct subnuclear localizations of euchromatin and heterochromatin.

Topologically associating domain

TADs are units of the genome that range in length from hundreds of kilobases to several megabases and contain one or more genes and associated regulatory elements. They can be considered as highly self-interacting genome units. Research has shown that insulator-binding protein CTCF, housekeeping genes, tRNA, short interspersed elements (SINEs), histones H3K4me3 and H3K36me3, and the cohesin complex are enriched at TAD boundaries, suggesting that these factors play a role in TAD formation [43]. TADs are highly conserved across different species, cell types, and evolutionary stages [48]. For example, 50%-70% of TADs are conserved in mouse and human embryonic stem cells [49], and their positions hardly change during differentiation [50]. Thus, TADs are considered the fundamental structural unit of chromatin folding.

The relationship between TADs and gene expression has long been the focus of chromatin research. Various studies have demonstrated a close association between TADs and various biological phenomena. For instance, TAD boundaries are highly consistent with the start of replication domains [51] during embryonic stem cell differentiation where TADs are established [52] and in mouse cells the silencing of the female X chromosome coincides with the reactivation of TAD structure [53]. Furthermore, the movement or disappearance of the TAD boundary has been linked to certain diseases and stress responses [54]. These findings suggest that TADs represent a fundamental functional unit of the genome, with internal coordinated gene regulation and border barrier function between TADs which likely serves as the primary mode of action for their biological functions.

Chromatin loop

Based on how chromatin fibers aggregate, infrequent contact between two distant genomic regions will happen due to random collisions [55]. However, some sites have a significantly higher frequency of incidental contact than expected by chance, which leads to the formation of a circular chromatin structure, called a loop, that connects these high-frequency interaction sites. Loop formation is mediated by factors such as CTCF and cohesin. Usually, loop formation occurs within the same TAD or sub-TAD (a smaller structural unit within the TAD, defining smaller chromosomal subregions). Depending on the connecting elements at both ends, the loop can be further classified as a promoter-enhancer loop, promoter-promoter loop, enhancer-enhancer loop, and so on. These different loops interact to form the complex regulatory network needed for gene expression, as described in various studies [56].

A typical example is the interaction between the promoter of the β-globin gene cluster and the locus control regions (LCR) located 30–60 kb upstream of the gene-cluster. When these regions form chromatin loops, the expression of the β-globin gene cluster is greatly enhanced, whereas a lack of observed loops results in low levels of gene expression [57]. Recently, Debruyne et al. [52] have shown that changes in the loop structure regulated by the CTCF homologue BORIS can lead to the development of super-enhancers in cancer cells that are resistant to treatment. These super-enhancers comprise multiple enhancers and have a high

transcription factor density, which promotes the formation of genomic regions with gene expression, ultimately leading to the expression of a class of pro-neural transcription factors and the development of a phenotype resistant to ALK inhibition [58]. Deng et al. [59] utilized the fusion of a multi-adaptor protein (LBD1) to a zinc finger protein, which facilitated the recruitment of this complex to the target site at the β-globin promoter. This induced the formation of a chromatin loop between the enhancer and the promoter, resulting in a significant enhancement of β-globin gene expression that was originally repressed. Furthermore, using the CRISPR/Cas9 technology to delete a prostate cancer risk region, hereby destroying the loop structure formed by the interaction between this region and *HOXA13*, significantly up-regulated the expression of the otherwise suppressed *HOXA13* gene, and successfully activated the expression of the oncogene *GATA2* induced by *HOXA13* [60]. In conclusion, this and other observations show that chromatin loops play a crucial role in gene activation and silencing and are essential in the basic process of transcribing DNA-encoded information to RNA.

1.3 Identification of chromatin loop

With the reduction of sequencing costs, that has been ongoing for the last couple of decades, the amount of Hi-C data is constantly increasing, and the resolution is getting better and higher. The existing Hi-C experiment utilizes 4.9×10^9 interaction counts to attain a maximum resolution of 1 kb for chromatin interactions in the human genome atlas [8]. With higher resolution the amount of data also increases significantly. Thus, the improvement in resolution has brought opportunities to study the fine structure of the three-dimensional genome but has also brought challenges to computing resources and analysis methods. An interaction matrix is usually used to reflect the resolution of Hi-C data. The rows and columns of the Hi-C contact matrix represent different regions of the genome, typically defined as discrete regions on chromosomes. Each element of the matrix represents the interaction frequency or contact strength between the corresponding regions. Because DNA molecules have the characteristics of multi-molecular polymers, two positions at farther linear distance have a weaker random interaction. Therefore, the intrachromosomal interaction matrix shows the tendency that the farther it deviates from the diagonal, the smaller the intensity. In Hi-C experiments, interactions on the diagonal occur more frequently because they typically involve shorter distances between regions on the chromosome. However, the cross-linking and sequencing steps in the Hi-C experiment can introduce noise, such as uneven cross-linking efficiency, biases in DNA fragment ligation, and sequencing errors. These noise sources manifest as higher values along the diagonal in the Hi-C matrix. This uncoordinated interaction mode is not conducive to the feature extraction of the chromatin hierarchy. In addition, due to the influenced of the nature of the genome and analytical methods, the interaction matrix obtained by the Hi-C experiment is affected by factors such as uneven lengths of enzyme fragments and inconsistent GC content of chimeric molecules. The strength of

interactions in different regions is therefore not suitable for direct comparison. Hence, precisely identifying chromatin loops is the primary goal of analyzing Hi-C data. The loop calling process is divided into preprocessing and significant interaction analysis. Preprocessing includes sequence alignment, filtering, binning and normalization. The normalization part will be explained in detail below.

1.3.1 Normalization

The original interaction matrix obtained after read positioning, filtering, and binning is affected by systematic deviations such as the length of restriction enzyme digestion DNA fragments and GC content. These deviations influence the comparability of interaction strengths between different genomic regions, affecting subsequent feature extraction such as the neglection of long-range interaction and false positive fragments enrichment. To reduce this bias, a standardized correction is required. Commonly used standardization methods can be divided into deterministic factor correction and matrix balancing.

The basic idea of determining factor correction is to determine the sources of systematic bias and then constructing a probability model to calculate the expected interaction to correct the original interaction matrix. Yaffe et al. [61] constructed a joint probability correction model based on the three systematic error sources, including enzyme digestion site distance, chimeric molecule GC content and sequence specificity, learning the maximum likelihood parameters through the interaction data and then estimating the expected interaction. This method needs to evaluate many parameters, and the maximum likelihood is calculated at the resolution of the restriction enzyme digestion DNA fragments which leads to high computational costs. A parametric model, HiCNorm, has been developed to improve this method [56]. HiCNorm uses a method based on Poisson regression to calculate the correction model which reduces model parameters and estimates parameters on lower-resolution interactions, hereby improving the correction effect and computational efficiency [62].

The matrix balance method assumes that any two regions on the genome should have equal visibility, without taking into account any specific sources of systematic bias [63]. This implies that the sum of valid interaction numbers is assumed to be similar. Based on this assumption, the hiclib approach aims to reach a point where the matrix estimation and bias estimation no longer significantly change through iterative optimization. To enhance convergence speed, Rao et al. [8] incorporated the Knight-Ruiz method into the matrix balance approach by constructing a weight matrix depends on regional activity. The HiCorrector approach, on the other hand, addresses the issue of large interaction matrices at high resolution by dividing them into blocks and processing them in parallel, hereby achieving the objective of matrix balance [64].

Despite being widely used, the two types of normalization methods are based on different assumptions and should therefore be used based on specific needs in each application. The deterministic factor correction method allows for the exclusion of other systematic deviations in the data, such as those found in cancer genomes,

while the matrix balance algorithm will eliminate such effect. The deterministic factor correction method, therefore, is more appropriate in this analysis process, though if the bias needs to be preserved, the deterministic factor correction method may be more appropriate. It is worth noting that the matrix balance method does not necessarily offer equal "visibility" to all genomic regions, such as the super-enhancer area which tends to have more long-range interactions. In this case, the deterministic factor correction method may be more suitable in preserving this feature [65]. It is therefore important to select the appropriate standardized method for the type of data to be analyzed.

1.3.2 Significant interaction analysis

In the interaction matrix obtained from Hi-C experiments, the majority of interactions are not chromatin loop signals. It is therefore crucial to identify significant interactions in the matrix to gain insight into chromatin structure and gene regulation. The fundamental concept for identifying significant interactions is to statistically model the elements in the matrix and establish a background model to differentiate significant interactions. This process includes identifying both significant interactions and significant differences in interactions.

The discrimination of significant interactions can be divided into two types. One type considers the influence of systematic biases or linear genomic distance and directly models individual interaction elements [66, 67]. For example, the Fit-HiC approach estimates the effect of distance on random polymer interactions by modeling the interaction-distance relationship twice and thereafter models interactions using a binomial distribution to determine significance [67]. The other type considers the influence of neighboring interactions or other hierarchical structures like TADs [68, 69] and is generally applicable to higher-resolution interactions and can, to some extent, improve the accuracy of the results. For instance, the PSYCHIC method constructs a background model based on TAD structural domains to identify significant interactions between enhancers and promoters. Additionally, for Capture Hi-C data, the CHiCAGO method [70] models interactions using a combination of negative binomial and Poisson random variables, where the expected value of the negative binomial random variable is regarded as a function of distance and is used to characterize Brownian collision. The expected value of the Poisson random variable is regarded to be a function related to fragment properties and is used to characterize technical errors. In terms of identifying significantly different interactions, the initial approach uses the binomial distribution to model the interaction at the same distance and to determine the P value of the different interactions, and then established the background by combining biological replicates from different samples to determine the difference interaction and calculating the false discovery rate (FDR) [43]. Subsequently, HiBrowse [71] and DiffHiC [72] methods use the gene differential expression identification software edgeR to interact with biological replicates to identify significant differences. The FIND method [73] considers the characteristics of adjacent interactions in the high-resolution interaction matrix and

uses the spatial Poisson process to identify differential interactions, which improves the signal-to-noise ratio of the identification results. Significant differential interactions identification is an extension of the identification of significant interactions.
It is important to note that the methods for calling chromatin loops are continuously developing, and different approaches may produce distinct outcomes. The selection of a specific method should depend on the research question and the characteristics of the data being analyzed. Additionally, experimental factors, such as Hi-C data quality and the selection of restriction enzymes, can influence the accuracy of chromatin loop calling [8, 67]. The methods for calling chromatin loops typically rely on the concept of "fractal globules" and utilize the power law method to approximate and estimate the expected Hi-C interaction frequency [10, 43]. The fractal globule theory proposes that chromatin forms a continuous fractal path that densely occupies the three-dimensional space without crossing itself [10]. However, recent research has shown that chromosomes may contain internal knots [74], and it is therefore critical to account for the effects of knotted structures when considering the power-law dependence between fractal globules and equilibrium structures. Recently, with the cLoops approach, good results were achieved by applying the DBSACN clustering method to call loops [75]. This inspired me to use affinity as a clustering condition and use dynamic windows instead of binning to identify chromatin loops more accurately (this thesis, Chapter 2).

1.4 Approaches for capturing chromatin conformation

In general, at present, there are two main types of technique used to capture chromatin conformation, which are microscopy-based techniques and sequencing data-based techniques. Hi-C and its derived technologies are essential sequence-based tools for the identification and analysis of chromatin conformation. In addition to Hi-C, there are several other key techniques available, including fluorescence in situ hybridization (FISH) [76], ChIA-PET and its derivatives (Table 1, Figure 1), and live microscopy methods. The diverse range of techniques allows for a more comprehensive exploration of chromatin interactions, hereby providing a more accurate analysis of biological functions.

1.4.1 Microscopy-based techniques

Major technological advancements in light microscopy have overcome the limitations of studying chromosome conformation and individual chromatin contacts, which were previously restricted by the small number of loci that could be probed simultaneously. There are currently three main approaches to improve resolution and overcome diffraction limitations. The first method, structured illumination microscopy (SIM) [77], doubles the resolution rate by using non-uniform illumination patterns to analyze resolution in Fourier mathematical space. SIM has been used to subdivide mammalian topologically associated domains (TADs) into smaller chromatin nanodomains (CNDs), a substructure that cannot be directly identified in Hi-C experiments. The second approach is

single-molecule localization microscopy (SMLM) [78], a super-resolution microscopy method that utilizes fluorophores which can switch between a fluorescent and dark state. SMLM relies on random excitation and detection of individual fluorophores, followed by precise localization of emission centers to produce super-resolution images with lateral resolution as low as 20 nm in biological samples. The third method is stimulated emission depletion (STED) [79], which uses a confocal microscope-like setup with a central zero-intensity ring-shaped laser to deplete emission at the periphery, allowing emission only at the center and resulting in a subdiffractional point spread function. While these three methods greatly increase resolution, their throughput is limited, and the number of loci that can be analyzed simultaneously is limited. More recently, investigators have combined oligonucleotide-based FISH protocols with microfluidics to allow for multiple cycles of hybridization [80], enabling sequential probe hybridization and assaying of multiple distinct loci in 3D nuclear space with high precision, an approach to overcome some of the throughput restrictions.

1.4.2 Sequencing data based techniques

Methods for analyzing chromatin structure using sequencing involves crosslinking fragments of chromatin that are in close proximity. The crosslinked fragments are then isolated and sequenced followed by bioinformatic analyses to determine the chromatin structure based on counting the number of interacting fragments. There are currently two main sequencing technologies: one is based on formaldehyde cross-linked chromatin fragment ligation such as 3C and its derived technologies, while the other does not involve cross-linked fragment ligation such as GAM [81], TSA-seq [82] and DamID [83].

As introduced in section1.2.3, Hi-C and its derived technologies are the main methods based on the connection of formaldehyde cross-linked chromatin fragments. Single-cell Hi-C can identify the cell-to-cell variability of 3D chromatin organization, while traditional Hi-C provides an average of chromatin organization in the cell population present in a sample [40]. Micro-C preserves nucleosome positioning information that can be analyzed jointly with chromatin contact information from a single dataset [84]. In contrast, ChIA-PET is a technique that captures specific chromatin interactions by using an antibody to enrich for chromatin regions bound by a specific protein of interest. It provides insights into the regulatory elements of the targeted protein or histone mark and offers higher resolution (hundreds of base pairs) than Hi-C for specific protein-mediated interactions [10]. HiChIP [85] and PLAC-seq [86] approaches create an optimal environment for proximity junctions, allowing the identification of subsets of interactions associated with specific protein binding, hereby improving data accuracy compared to ChIA-PET. It is important to note that both Hi-C and ChIA-PET are incapable of detecting multiple interactions. To depict more precise transcriptional regulatory networks, we need more precise technologies to decipher chromatin architecture and identify transcriptional regulatory elements. Therefore, we have developed single-cell TSS-seq (scTSS-seq) technology and MAE-seq to

achieve the identification and annotation of promoters and enhancers (Chapters 3 and 4 this thesis).

1.5 Three-dimensional genomic data and transcriptional regulatory element identification

Recent advances in high-throughput sequencing technologies and computational methods have made it possible to analyze 3D genomic data and identify transcriptional regulatory element (TREs). To identify TREs in 3D genomic data, researchers can use a variety of computational methods, including clustering, network analysis, and machine learning algorithms. These methods can help identify transcription regulatory networks in the genome and predict the functions of newly discovered regulatory elements. Overall, the integration of 3D genomic data with the identification of TRE is highly relevant as it can yield important insights into their connection with development and variations in traits or phenotypes.

1.5.1 DNA motifs calling

The identification of TREs requires the pivotal process of DNA motif discovery. TFBS, essential components of cis-regulatory elements, precisely regulate gene expression. TFBS motifs are short DNA sequences that exhibit specific sequence consensus within the TFBS [87]. ChIP-seq data and mRNA profiling are sources for collecting regions for motif discovery. Regular expressions and position weight matrices (PWM) are common tools for characterizing consensus sequence patterns [88]. Regular expressions consist of a string of DNA sequences, whereas PWM is a statistical model that describes the relative occurrence frequencies of bases at specific positions. Relative entropy is used to estimate the significance of PWM, which is a weighted average of the binding energies of each site [89]. Therefore, PWM is a powerful tool for characterizing sequence patterns and their significance in TFBS motif discovery [90].

1.5.2 Cis-Regulatory Elements and cis-regulatory modules predictions

In order to drive a specific expression pattern of its target gene, a regulatory element typically contains a local concentration of multiply TFBSs (known as a cis-regulatory module or CRM) rather than single isolated and scattered binding sites for transcription factors [91]. The distance between each binding site in a CRM is not fixed, and there are no restrictions on the order in which binding sites are arranged [92]. Due to this variability, researchers often focus on clusters of CRMs. One common tool for identifying these clusters is a sliding window that scans the data and requires a certain number of predicted matches to a given PWM to reach a threshold value within the window [93]. However, the choice of a threshold is subjective and can therefore lead to variable results. To address this issue, researchers have developed a clustering method using a hidden Markov model (HMM). The HMM sets background states within and between clusters and compares the likelihood of the CRM sequence to the background model [94]. By

applying microarrays or RNA-seq data, researchers are able to identify the functional targets of CRMs.

1.5.3 Different transcriptional regulatory elements identifications by machine learning methods

With the increasing use of next-generation sequencing, the integrated analysis of multi-omics data has become powerful for identifying transcriptional regulatory elements in addition to the basic method of motif calling. For different types of regulatory elements, appropriated tools have been developed to identify and predict them. Enhancer identification is an application field of multi-omics analysis. Enhancers are identified by examining whether a DNA sequence has enhancer activity, which can be profiled using multiple data types, such as chromatin accessibility, histone marks, TFBSs, STARR-seq, and enhancer RNA (eRNA) data. However, different data types and analysis strategies require different approaches that have different strengths and weaknesses. Evolutionary conservation analysis is a simple method to calculate conserved motifs across species, but it is less sensitive to detect species (or tissue) specific enhancers [95]. Sequence characteristic analysis (k-mer analysis) is easier to compute, but the prediction power for new enhancer is limited [96]. Using ChIP-seq technology, histone marks and TFBS information can be mined, but the conservation among different cell types requires further examination and data to reliable predict enhancers [97]. Combining TFBS with chromatin accessibility data can efficiently identify transcriptional regulatory elements, but determining the specific type of element is still a challenge [98]. STARR-seq data is an effective method for testing enhancer activity but this method has some limitations as it cannot discover novel enhancers, because of its dependence of annotated reference, and has limited sensitivity of weak enhancers and deficiency of specificity [99]. Furthermore, eRNA data analysis is also an efficient method to identify enhancers, but not all enhancers are known to produce eRNAs [13]. To address the challenges posed by these different types of data, integrative applications and algorithms have been developed. Clustering of epigenetic profiles, chromatin accessibility data, and TFBSs is a common method for generating global epigenetic maps for different cell lines and tissues. After identifying the target DNA regions, unsupervised learning techniques are applied to classify the regulatory class of each region. Clustering of chromatin profiles is used to compute the similarity of chromatin profiles with existing annotated segments and classify the elements. In addition, machine learning approaches are also applied to predict enhancers. Support vector machines (SVMs), artificial neural networks (ANNs), decision trees (DTs), random forests (RFs), probabilistic graphical models (PGMs), and ensemble techniques have all been successfully applied for the identification of enhancers. ChromHMM [100] is a computational method for modeling chromatin states by integrating multiple genomic datasets, such as ChIP-seq and DNA methylation sequencing. It builds a probabilistic model to predict genome-wide chromatin state patterns and identify regulatory elements - including enhancers - thereby systematically

revealing functional regulatory mechanisms of the genome. CSI-ANN uses chromatin signatures to classify regulatory elements by artificial neural networks (ANN). CSI-ANN employs ANN to learn the relationship between chromatin states and gene regulatory elements. ANN consists of multiple layers and nodes, where each node performs weighted summation and activation functions. By constructing an appropriate network structure, the ANN can learn and capture the nonlinear relationship between chromatin states and regulatory elements [101]. EnhancerFinder applies multiple kernel learning (MKL) to classify elements based on SVMs. By leveraging the strengths of multiple kernel functions, MKL can improve the performance of enhancer prediction models. The combination of different kernels helps overcome limitations or biases associated with individual kernels, reducing the risk of underfitting and/or overfitting. MKL's ability to adaptively learn the optimal combination of kernels allows it to capture the most relevant information for enhancer identification, leading to better prediction accuracy [102]. Random forest algorithm reduces the effects of class-imbalance between enhancer/non-enhancer samples and eliminates limitations coming from the small size of the training data. Random Forest provides feature importance evaluation, enabling the understanding of which chromatin features have the most significant impact on enhancer identification. Additionally, model evaluation and validation, such as using cross-validation techniques, ensure the performance and reliability of the Random Forest model in enhancer prediction tasks [103].

1.5.4 Identifying Regulatory Elements via Deep Learning

Deep learning is a type of neural network that consists of at least two layers. For protein-DNA binding prediction, DeepBind was the first to use deep learning [89]. Its network architecture comprises a convolutional layer, pooling layer, and fully connected layer, which can be applied to various data types such as PBMs, ChIP-seq, HT-SELEX, RNAcompete, and CLIP-seq (Table 2). DanQ combines the strengths of Convolutional Neural Networks and Recurrent Neural Networks to predict DNA sequence functions [104]. Its architecture comprises a convolutional layer followed by a max pooling layer and a bidirectional recurrent layer that processes the sequence from both directions, allowing for learning of both upstream and downstream biological contexts. DeepSEA is an algorithmic framework for noncoding-variant effect prediction, based on a fully sequence-based approach [105]. It simultaneously predicts large-scale chromatin-profiling data such as TF binding, DNase I hypersensitive sites, and histone mark profiles, learning regulatory sequence code from genomic sequences]. BiRen uses a hybrid deep learning architecture to predict enhancers based on DNA sequence alone. Compared to other enhancer predictors based on sequence characteristics, BiRen exhibits superior accuracy, robustness, and generalizability in enhancer prediction [106].

Table 2. Comparison of techniques for studying protein–nucleic acid interactions. This table summarizes commonly used methods for analyzing protein–DNA or protein–RNA binding, listing their key features and source data types.

Technique	Features	Source Data
PBMs	- Global analysis of protein-DNA binding preferences	- Large-scale DNA arrays with known sequences
	High-throughput screening of protein-DNA interactions	
ChIP-seq	- Genome-wide mapping of protein-DNA binding sites	- Crosslinked chromatin samples
	- Identification of histone modifications and other DNA-binding proteins	- DNA sequencing data
HT-SELEX	- High-throughput identification of RNA binding sequences	- In vitro synthesized diverse RNA pools
	- Characterization of protein-RNA interaction preferences	
RNAcompete	- Determination of RNA binding preferences of RNA-binding proteins	- In vitro synthesized RNA array
	- Characterization of RNA-protein interaction motifs	
CLIP-seq	- Genome-wide mapping of protein-RNA interaction sites	- Crosslinked protein-RNA complexes
	- Identification of RNA binding sites for specific RNA-binding proteins	- RNA sequencing data

1.6 Aim and outline of this thesis

In this thesis I enhanced the identification method for chromatin loops using Hi-C data and assessed the impact of chromosome knots on the interaction signal's expected value and the effectiveness of AP-Cluster in detecting significant interactions (Chapter 2). Next, I utilized scTss-seq and bulk RNA-seq technologies to identify alternative transcription start site. Through machine learning methods, I conducted sequence comparisons between novel TSS-promoter and known TSS-promoter, developing the bioinformatics validation in Chapter 3. Subsequently, I designed computational models and wrote a software for analyzing MAE-seq data, which is a technique for identifying the shortest core sequence of enhancers (Chapter 4). In Chapter 5, I utilized the multiple kernels learning (MKL) approach, to identify the sequence features of existing transcriptional regulatory elements and created an efficient sequence feature classifier for recognizing transcriptional regulatory elements using MAE-seq data. Multiple kernel learning refers to a set of machine learning methods that use a predefined set of kernels and learn an optimal linear or non-linear combination of kernels as part of the algorithm. I used this algorithm to develop a classifier to identify the sequence feature of transcriptional regulatory elements, especially for enhancers. I furthermore compared the result among different machine learning algorithms, developed an innovation sequencing coding and solved the problem of inconsistent scoring of forward and reverse chain sequences.

1.7 Reference

1. Dekker, J. and L. Mirny, *The 3D Genome as Moderator of Chromosomal Communication.* Cell, 2016. **164**(6): p. 1110-1121.
2. Preissl, S., K.J. Gaulton, and B. Ren, *Characterizing cis-regulatory elements using single-cell epigenomics.* Nat Rev Genet, 2023. **24**(1): p. 21-43.
3. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.
4. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annu Rev Biochem, 2012. **81**: p. 145-66.
5. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
6. Sun, D.Q., L. Tian, and B.G. Ma, *Spatial organization of the transcriptional regulatory network of Saccharomyces cerevisiae.* FEBS Lett, 2019. **593**(8): p. 876-884.
7. Kornberg, R.D., *Chromatin structure: a repeating unit of histones and DNA.* Science, 1974. **184**(4139): p. 868-71.
8. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
9. Fullwood, M.J., et al., *An oestrogen-receptor-alpha-bound human chromatin interactome.* Nature, 2009. **462**(7269): p. 58-64.
10. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.
11. Zhu, X., et al., *MAE-seq refines regulatory elements across the genome.* Nucleic Acids Res, 2024. **52**(2): p. e9.
12. Kheradpour, P., et al., *Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.* Genome Res, 2013. **23**(5): p. 800-11.
13. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.
14. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
15. Klein, J.C., et al., *A systematic evaluation of the design and context dependencies of massively parallel reporter assays.* Nat Methods, 2020. **17**(11): p. 1083-1091.
16. Oudet, P., M. Gross-Bellard, and P. Chambon, *Electron microscopic and biochemical evidence that chromatin structure is a repeating unit.* Cell, 1975. **4**(4): p. 281-300.
17. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.
18. Sekinger, E.A., Z. Moqtaderi, and K. Struhl, *Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast.* Mol Cell, 2005. **18**(6): p. 735-48.
19. Saragosti, S., G. Moyne, and M. Yaniv, *Absence of nucleosomes in a fraction of SV40 chromatin between the origin of replication and the region coding for the late leader RNA.* Cell, 1980. **20**(1): p. 65-73.
20. Hodges, C., et al., *Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II.* Science, 2009. **325**(5940): p. 626-8.
21. Bernstein, B.E., et al., *Global nucleosome occupancy in yeast.* Genome Biol, 2004. **5**(9): p. R62.
22. Mavrich, T.N., et al., *Nucleosome organization in the Drosophila genome.* Nature, 2008. **453**(7193): p. 358-62.
23. Leimgruber, E., et al., *Nucleosome eviction from MHC class II promoters controls positioning of the transcription start site.* Nucleic Acids Res, 2009. **37**(8): p. 2514-28.
24. Laybourn, P.J. and J.T. Kadonaga, *Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II.* Science, 1991. **254**(5029): p. 238-45.
25. Finch, J.T. and A. Klug, *Solenoidal model for superstructure in chromatin.* Proc Natl Acad Sci U S A, 1976. **73**(6): p. 1897-901.
26. Widom, J. and A. Klug, *Structure of the 300A chromatin filament: X-ray diffraction from oriented samples.* Cell, 1985. **43**(1): p. 207-13.
27. Woodcock, C.L., L.L. Frado, and J.B. Rattner, *The higher-order structure of chromatin: evidence for a helical ribbon arrangement.* J Cell Biol, 1984. **99**(1 Pt 1): p. 42-52.
28. Dorigo, B., et al., *Nucleosome arrays reveal the two-start organization of the chromatin fiber.* Science, 2004. **306**(5701): p. 1571-3.
29. Song, F., et al., *Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units.*

Science, 2014. **344**(6182): p. 376-80.
30. Li, W., et al., *FACT Remodels the Tetranucleosomal Unit of Chromatin Fibers for Gene Transcription.* Mol Cell, 2016. **64**(1): p. 120-133.
31. Rabl, C., *Uber Zelltheilung*. 1887: W. Engelmann.
32. Cremer, T. and M. Cremer, *Chromosome territories.* Cold Spring Harb Perspect Biol, 2010. **2**(3): p. a003889.
33. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
34. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
35. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.
36. Simonis, M., et al., *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).* Nat Genet, 2006. **38**(11): p. 1348-54.
37. Dostie, J., et al., *Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.* Genome Res, 2006. **16**(10): p. 1299-309.
38. Ma, W., et al., *Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes.* Nat Methods, 2015. **12**(1): p. 71-8.
39. Schoenfelder, S., et al., *Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells.* Nat Genet, 2010. **42**(1): p. 53-61.
40. Nagano, T., et al., *Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.* Nature, 2013. **502**(7469): p. 59-64.
41. Lin, D., et al., *Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture.* Nat Genet, 2018. **50**(5): p. 754-763.
42. Chen, X., et al., *Structural insights into preinitiation complex assembly on core promoters.* Science, 2021. **372**(6541).
43. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
44. Koehler, D., et al., *Changes of higher order chromatin arrangements during major genome activation in bovine preimplantation embryos.* Exp Cell Res, 2009. **315**(12): p. 2053-63.
45. Cremer, M., et al., *Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes.* Methods Mol Biol, 2008. **463**: p. 205-39.
46. Solovei, I., et al., *Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution.* Cell, 2009. **137**(2): p. 356-68.
47. Ryba, T., et al., *Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types.* Genome Res, 2010. **20**(6): p. 761-70.
48. Dekker, J. and E. Heard, *Structural and functional diversity of Topologically Associating Domains.* FEBS Lett, 2015. **589**(20 Pt A): p. 2877-84.
49. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation.* Nature, 2015. **518**(7539): p. 331-6.
50. Battulin, N., et al., *Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach.* Genome Biol, 2015. **16**(1): p. 77.
51. Pope, B.D., et al., *Topologically associating domains are stable units of replication-timing regulation.* Nature, 2014. **515**(7527): p. 402-5.
52. Ke, Y., et al., *3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis.* Cell, 2017. **170**(2): p. 367-381 e20.
53. Wang, C.Y., et al., *SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X.* Cell, 2018. **174**(2): p. 406-421 e25.
54. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-1025.
55. Bornfleth, H., et al., *Quantitative motion analysis of subchromosomal foci in living cells using four-dimensional microscopy.* Biophys J, 1999. **77**(5): p. 2871-86.
56. Zhang, Y., et al., *Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations.* Nature, 2013. **504**(7479): p. 306-310.
57. Deng, W., et al., *Reactivation of developmentally silenced globin genes by forced chromatin looping.* Cell, 2014. **158**(4): p. 849-860.
58. Debruyne, D.N., et al., *BORIS promotes chromatin regulatory interactions in treatment-resistant cancer cells.* Nature, 2019. **572**(7771): p. 676-680.
59. Deng, W., et al., *Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor.* Cell, 2012. **149**(6): p. 1233-44.
60. Luo, Z., et al., *A Prostate Cancer Risk Element Functions as a Repressive Loop that Regulates HOXA13.* Cell Rep, 2017. **21**(6): p. 1411-1417.

61. Yaffe, E. and A. Tanay, *Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.* Nat Genet, 2011. **43**(11): p. 1059-65.
62. Hu, M., et al., *HiCNorm: removing biases in Hi-C data via Poisson regression.* Bioinformatics, 2012. **28**(23): p. 3131-3.
63. Imakaev, M., et al., *Iterative correction of Hi-C data reveals hallmarks of chromosome organization.* Nat Methods, 2012. **9**(10): p. 999-1003.
64. Li, W., et al., *Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data.* Bioinformatics, 2015. **31**(6): p. 960-2.
65. Schmitt, A.D., et al., *A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.* Cell Rep, 2016. **17**(8): p. 2042-2059.
66. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.
67. Ay, F., T.L. Bailey, and W.S. Noble, *Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.* Genome Res, 2014. **24**(6): p. 999-1011.
68. Xu, Z., et al., *A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data.* Bioinformatics, 2016. **32**(5): p. 650-6.
69. Ron, G., et al., *Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains.* Nat Commun, 2017. **8**(1): p. 2237.
70. Cairns, J., et al., *CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data.* Genome Biol, 2016. **17**(1): p. 127.
71. Paulsen, J., et al., *HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization.* Bioinformatics, 2014. **30**(11): p. 1620-2.
72. Lun, A.T. and G.K. Smyth, *diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data.* BMC Bioinformatics, 2015. **16**: p. 258.
73. Djekidel, M.N., Y. Chen, and M.Q. Zhang, *FIND: difFerential chromatin INteractions Detection using a spatial Poisson process.* Genome Res, 2018. **28**(3): p. 412-22.
74. Siebert, J.T., et al., *Are There Knots in Chromosomes?* Polymers (Basel), 2017. **9**(8).
75. Cao, Y., et al., *Accurate loop calling for 3D genomic data with cLoops.* Bioinformatics, 2020. **36**(3): p. 666-675.
76. Bayani, J. and J.A. Squire, *Fluorescence in situ Hybridization (FISH).* Curr Protoc Cell Biol, 2004. **Chapter 22**: p. Unit 22 4.
77. Gustafsson, M.G., et al., *Three-dimensional resolution doubling in wide-field fluorescence microscopy by structured illumination.* Biophys J, 2008. **94**(12): p. 4957-70.
78. Huang, B., H. Babcock, and X. Zhuang, *Breaking the diffraction barrier: super-resolution imaging of cells.* Cell, 2010. **143**(7): p. 1047-58.
79. Hell, S.W. and J. Wichmann, *Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy.* Opt Lett, 1994. **19**(11): p. 780-2.
80. Beliveau, B.J., et al., *Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes.* Proc Natl Acad Sci U S A, 2012. **109**(52): p. 21301-6.
81. Quinodoz, S.A., et al., *Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus.* Cell, 2018. **174**(3): p. 744-757 e24.
82. Zhang, L., et al., *TSA-seq reveals a largely conserved genome organization relative to nuclear speckles with small position changes tightly correlated with gene expression changes.* Genome Res, 2020. **31**(2): p. 251-64.
83. Guelen, L., et al., *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.* Nature, 2008. **453**(7197): p. 948-51.
84. Hsieh, T.H., et al., *Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C.* Cell, 2015. **162**(1): p. 108-19.
85. Mumbach, M.R., et al., *HiChIP: efficient and sensitive analysis of protein-directed genome architecture.* Nat Methods, 2016. **13**(11): p. 919-922.
86. Li, X., et al., *Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions.* Nat Protoc, 2017. **12**(5): p. 899-915.
87. GuhaThakurta, D., *Computational identification of transcriptional regulatory elements in DNA sequence.* Nucleic Acids Res, 2006. **34**(12): p. 3585-98.
88. Liu, X.S., D.L. Brutlag, and J.S. Liu, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.* Nat Biotechnol, 2002. **20**(8): p. 835-9.
89. Alipanahi, B., et al., *Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.* Nat Biotechnol, 2015. **33**(8): p. 831-8.
90. Berg, O.G. and P.H. von Hippel, *Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.* J Mol Biol, 1987. **193**(4): p. 723-50.
91. Aerts, S., *Computational strategies for the genome-wide identification of cis-regulatory elements and*

transcriptional targets. Curr Top Dev Biol, 2012. **98**: p. 121-45.
92. Kulkarni, M.M. and D.N. Arnosti, *Information display by transcriptional enhancers.* Development, 2003. **130**(26): p. 6569-75.
93. Markstein, M., et al., *Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.* Proc Natl Acad Sci U S A, 2002. **99**(2): p. 763-8.
94. Crowley, E.M., K. Roeder, and M. Bina, *A statistical model for locating regulatory regions in genomic DNA.* J Mol Biol, 1997. **268**(1): p. 8-14.
95. Visel, A., J. Bristow, and L.A. Pennacchio, *Enhancer identification through comparative genomics.* Semin Cell Dev Biol, 2007. **18**(1): p. 140-52.
96. Fletez-Brant, C., et al., *kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W544-56.
97. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.
98. Boyle, A.P., et al., *High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells.* Genome Res, 2011. **21**(3): p. 456-64.
99. Kwasnieski, J.C., et al., *High-throughput functional testing of ENCODE segmentation predictions.* Genome Res, 2014. **24**(10): p. 1595-602.
100. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization.* Nat Methods, 2012. **9**(3): p. 215-6.
101. Firpi, H.A., D. Ucar, and K. Tan, *Discover regulatory DNA elements using chromatin signatures and artificial neural network.* Bioinformatics, 2010. **26**(13): p. 1579-86.
102. Erwin, G.D., et al., *Integrating diverse datasets improves developmental enhancer prediction.* PLoS Comput Biol, 2014. **10**(6): p. e1003677.
103. Rajagopal, N., et al., *RFECS: a random-forest based algorithm for enhancer identification from chromatin state.* PLoS Comput Biol, 2013. **9**(3): p. e1002968.
104. Quang, D. and X. Xie, *DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences.* Nucleic Acids Res, 2016. **44**(11): p. e107.
105. Zhou, J. and O.G. Troyanskaya, *Predicting effects of noncoding variants with deep learning-based sequence model.* Nat Methods, 2015. **12**(10): p. 931-4.
106. Yang, B., et al., *BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone.* Bioinformatics, 2017. **33**(13): p. 1930-1936.

Chapter 2

Accurate and fast loop calling from Hi-C data by TLoops

Unpublished

Qitong Huang[†], Siyuan Kong[†], Jing Luo[†], Martien A.M. Groenen, Ole Madsen and Yubo Zhang. Accurate and fast loop calling from Hi-C data by TLoops. [†]These authors contributed equally to this work.

Abstract
Chromatin loops are crucial for gene regulation and consequent cellular processes. The Hi-C technique is commonly used to study the three-dimensional structure of the genome and to identify chromatin loops. Several computational methods have been developed to identify chromatin loops, but these methods may yield different results due to the continuous development of computational methods and variation in experimental setups. The fractal globule theory is commonly used to approximate the expected Hi-C interaction frequency. It is generally accepted that the decay of contact probability with increasing genomic distance observed in interphase chromosomes resembles that of a "crumpled globule" configuration, as predicted by models of condensed long polymers without topological entanglements. However, recent studies suggest that chromosomal DNA may also contain topological knots. Moreover, the chromatin loop extrusion mechanism has been implicated not only in the dynamic formation of chromatin loops, but also in facilitating the resolution of such knots, thereby maintaining genomic organization and accessibility. We developed a new computational tool, TLoops, designed to enhance the accuracy of chromatin loop identification while minimizing computational resource demands. In addition, TLoops enables the modelling and detection of chromatin loops that arise through topological untangling processes. TLoops utilizes AP-cluster and knot simulation to perform loop calling, and its effectiveness and efficiency have been demonstrated through comparisons with existing tools. TLoops represents a promising direction for improving the accuracy and efficiency of chromatin loop identification.

Keywords
Hi-C, chromatin loop, DNA knots, extrusion

Availability and implementation
The tools of TLoops are available at: https://gitee.com/agiswurhqt/tloops

1 Introduction
Chromatin loops refer to the physical organization of DNA within the eukaryotic nucleus. They play essential roles in gene regulation and other cellular processes. By bringing distant genome regions closer, chromatin loops can facilitate interactions between regulatory elements that control gene expression, such as enhancers and promoters, to influence development and cell fate decisions [1, 2].
Hi-C is a technique for studying the three-dimensional structure of the genome and identifying chromatin loops. It is based on crosslinking cellular chromatin with formaldehyde, digesting the DNA with restriction enzymes, and ligating the fragments under conditions that favour proximity ligation [3]. The resulting chimeric DNA molecules are sequenced and the frequency of junction events between different regions of the genome is used to infer their physical proximity in the nucleus. To identify chromatin loops from Hi-C data, several computational methods have been developed. These methods use various algorithms to analyze

the frequency of interactions between pairs of genomic regions and identify regions with higher interaction frequencies than expected by chance. Some commonly used chromatin loop calling methods include HiCCUPS [4], Fit-Hi-C [5], GOTHiC [6], HOMER [7], FastHiC [8] and cLoops [9]. Among them, Fit-Hi-C, GOTHiC, and HOMER all model the number of interaction fragments using a binomial distribution. HiCCUPS is part of the juicer-tools suite and applies a chromatin loop calling algorithm that utilizes a doughnut-shaped sliding window approach. HiCCUPS considers all elements in a Hi-C interaction matrix as potential peaks and tests if the surrounding region significantly differs from neighbouring interactions. cLoops, another chromatin loop calling algorithm, is based on the clustering algorithm cDBSCAN and directly analyzes the paired-end tags (PETs) in Hi-C data to identify candidate loops. Additionally, FastHiC employs a hidden Markov random field based Bayesian (HMRFBayes) method, explicitly modelling spatial dependency among adjacent loci to improve the robustness and statistical power of the technique.

However, methods for chromatin loop calling are still evolving, and different methods have been shown to yield different results. The choice of method may depend on the specific research question and the characteristics of the data analyzed. In addition, experimental factors such as the quality of Hi-C data and the choice of restriction enzymes can also affect the accuracy of chromatin loop calling [4, 5]. The described methods are often based on the concept of "fractal globules", with the power law method approximating the expected Hi-C interaction frequency [1, 10]. The fractal globule theory postulates that chromatin forms a continuous fractal path that densely fills the three-dimensional space without crossing itself [1]. Studies, however, have shown that chromosomes might have knots [11, 12]. The inadvertent consequence of DNA-DNA interactions occurring within highly crowded DNA molecules in living cells is the formation of tangled DNA [13], where the tangled chromatin fibers may exhibit the scaling properties of a fractal globule [14]. The power-law dependence between contact probability and sequence separation lies between fractal globule and equilibrium structures. While taking into account the power-law relationship between fractal globules and equilibrium structures, one must also acknowledge the impact of knotted structures. Oligopaint is a Fluorescence in situ Hybridization technique that precisely localizes genomic regions in numerous fixed cells by progressively labelling adjacent 30 kb chromatin segments, which assemble into ~2 Mb segments [15]. Each segment is targeted with approximately 300 fluorescent probes, and its three-dimensional diffraction-limited image is analyzed to obtain centroid coordinates. To construct closed curves, polygonal lines connecting these centroids are completed 100 times. Two long, parallel segments extend from the endpoints and are then joined, with their spatial orientation uniformly varied each time. The knot type of each closed curve is identified by computing the Jones polynomial [15]. Oligopaint analysis revealed that approximately 50% of chromosomes were classified as knotless, while the remainder exhibited distinct knotted topologies. Topological modelling further indicated a strong association between chromatin loop formation and the untangling of knots [15]. Specifically, the loop extrusion

mechanism was found to facilitate not only topological untangling and decatenation, but also the spatial segregation of chromatin loops, thereby contributing to the establishment of discrete chromosome territories. Chromatin loops generated via such untangling processes manifest distinct features in Hi-C data, typically characterized by attenuated upstream and downstream interaction signals and the absence of clearly defined topologically associating domains (TADs). In contrast to canonical loops with prominent TAD structures and enriched local contacts, untangling-driven loops are more challenging to detect due to their longer interaction ranges and the lack of adjacent signal reinforcement. Notably, the disappearance of a well-defined TAD structure does not imply a complete loss of local interaction signals associated with the knot class of chromatin loops. Instead, the post-untangling chromatin configuration often retains features of a residual or weakened TAD-like structure. The corresponding topological organization is illustrated in Fig. 1.

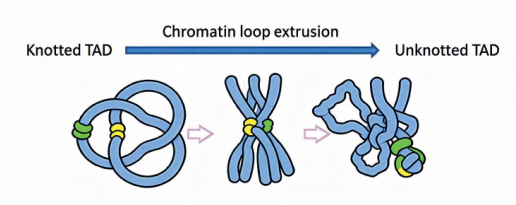


Figure 1. Schematic overview of chromatin loop extrusion. In the first stage: In the untangling process of knotted TAD the green and yellow sites are segments on the trefoil that do not interact with each other. In the second stage: After the trefoil unknotting, the yellow sites gradually approach the green sites, and the TAD structure begins to disappear. In the third stage: The yellow sites and the green sites form a chromatin loop, adjacent interactions are reduced, and TAD disappears.

To address the challenges to accurately detect chromatin loops, particularly those associated with topological unwinding, we developed a new computational tool, TLoops, which improves loop identification accuracy while significantly reducing computational resource requirements. Moreover, TLoops is specifically designed to detect chromatin loops arising from unwinding-mediated structural reorganization. TLoops performs loop calling in three steps: determining the scan window, applying AP-clustering [16], and statistically testing the mathematical expectation estimate based on weakened TAD-like structure simulation. By comparing TLoops with existing tools like HiCCUPS, Fit-Hi-C, and HOMER, the effectiveness and efficiency of TLoops have been demonstrated. At the same time, when using the software, we can apply the sub-function UKLoop in TLoops separately to directly identify the chromatin loops generated by the unknotted effect to supplement the results of other software.

2 Materials and methods

2.1 Public data used

Mouse Hi-C datasets used in this study were retrieved from the GEO repository with accession numbers GSM862721, GSM1718027 and GSM862720.

2.2 Data pre-processing

For the Hi-C data, the raw fastq data was analyzed by fastqc for quality control, and the adapters were trimmed by cutadapt (1.8.dev0) using default settings. We applied HiC-Pro (v2.11.0) [16] to align against the mouse reference genome (mm9) and default parameters were used to filter for noise.

2.3 Input and output of TLoops and visualization

TLoops utilizes the validedpair file from HiC-Pro as the input data. TLoops generates a table with loop id, position, interactive strength and significance. Juice-box [17] was applied to visualize the heatmap.

2.4 Normalization

We perform data normalization based on local GC content, mappability of reads, and fragment length within sliding windows instead of global normalization. Therefore, we cannot assume equal visibility of all loci. We used generalized linear models to calibrate the obtained windows.

2.5 Dynamic sliding window

Our approach differs from the current binning method as we developed a minimum sliding window technique to analyze Hi-C interaction signals. The sliding window is composed of two components: the inner window and the outer window. The inner window's size can be manually set, and it slides horizontally over a constant distance while tracking the number of interactive signals within its range. The outer window is a concentric circle surrounding the inner window, which expands outward at a fixed interval. We utilized the Poisson test to compare the statistical significance of the two windows. If a significant difference is detected, the outer window expansion stops, and the signal inside the outer window is retained as the candidate for further testing.

2.6 AP-Cluster analysis

The affinity propagation clustering (AP-Cluster) algorithm [18] is employed to evaluate the degree of similarity among signal distributions in the dynamic sliding window step. If only one cluster is identified, the inspection moves on to the next step. However, if multiple clusters are found, the current candidate is divided into subsets based on the number of clusters, and each subset is examined independently. The cluster calculation formulas are as follows:

$$\mathrm{r}(i,k) \leftarrow s(i,k) - \max_{k' \text{s.t.} k' \neq k} \{a(i,k') + s(i,k')\}$$

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' \text{s.t.} i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}$$

where i represents the interaction signal at the centre of the sliding window, and k represents the interaction signal of other candidate samples within the sliding window.

2.7 Unknotted TAD simulation and statistical model

After all candidates have been identified, the expected interaction of the window is calculated by structural modelling, and statistical test is performed on them one by one. The structural modelling calculation is performed using an adjusted power-law decay exponent λ. When candidates are located within unknotted TADs, they have a higher probability as determined by the chromatin unknotting effect. When the candidate cluster corresponds to unknotted TADs, the surrounding interaction signals typically exhibit a sharp decline. This contrasts with canonical TAD-associated loops, which maintain continuous interaction patterns in both horizontal and vertical dimensions of the contact matrix. As a result, the post-clustering affinity for unknotted regions may not differ substantially from the background interaction level. Upon completion of the unknotting process, the interaction signals within the original TAD region are expected to become more uniformly distributed. To account for this, we redefine the expected interaction baseline using the local mean signal intensity within a defined genomic window (default: 10 kb), rather than relying solely on genomic distance-dependent decay. This adjusted baseline is then used to assess the statistical significance of affinity deviations. Finally, we use the cumulative distribution function of the negative binomial distribution to calculate the probability of the observed contact count being lower than or equal to the expected contact count and correct the significance using the Benjamini-Hochberg method.

3 Results

3.1 Dynamic sliding window

The dynamic sliding window was designed to mitigate resolution interference from set bins. To determine the initial size of the sliding window and step size, we compared candidate generation, calculation speed, and window average density. We found that an excess of candidates would not affect the accuracy of results due to the clustering and statistical verification of subsequent steps. Window density reflects the enrichment of chromatin loop interaction signals, with higher densities indicating a greater likelihood of a signal being a chromatin loop. We choose 10 kb as the default starting value for the internal window, 20 kb as the default starting value for the external window, and 10 kb as the default value for the step size. The double sliding window follows these rules:

a. The internal and external windows slide horizontally along the same interaction distance.
b. When there is a significant density difference in the region, record all signals within the external window (including the internal window), and continue

scanning from the last signal in the window.

c. When there is no significant density difference in the region, the external window gradually expands to the maximum setting value. If there is a significant density difference perform b, otherwise, slide the internal window with a fixed step size and perform a.

It is important to point out that the internal window size and sliding step employed here are not equivalent to the typical currently used bin size. The contents within each window are clustered to obtain the optimal affinity region. The first step involves screening for density differences to determine the candidates, thereby reducing the number of calculations in later steps. Additionally, the default initial sliding interaction distance is set to 20kb, which eliminates the diagonal signal interference [19].

After determining the windows, we normalized and corrected all candidate windows. Currently, there are two main normalization methods: normalization factor correction and matrix balancing correction. The basic idea of normalization factor correction is to identify the source of systematic bias and construct a probability model to calculate the expected interaction for correcting the original interaction matrix. These normalization methods mainly consider three sources of systematic error: the fragment lengths cut by the restriction enzyme, the GC content of the chimeric molecule, and the sequence specificity (highly repetitive sequences) to construct a joint probability correction model [20]. The matrix balancing method does not consider specific sources of systematic bias and assumes that any two regions on the genome should have the same visibility [21], that is, the sum of effective interaction numbers should be equal. We use the normalization factor correction method for data normalization in our sliding window. Firstly, local normalization is more suitable for combining local features. At the same time, some genomic regions may not have the same "visibility" due to their inherent properties. For example, super enhancer regions have more long-range interactions, and using the method of determining correction factors is more advantageous for preserving this feature [22].

3.2 Affinity propagation clustering algorithm

Affinity Propagation clustering is a graph-based clustering algorithm that can automatically cluster data points without pre-specifying the number of clusters. The algorithm is based on a similarity measure between each pair of data points, which can be any computable metric such as Euclidean distance, Pearson correlation coefficient, etc. This feature is consistent with the characteristics of chromatin loop interaction signals. In cLoops [9], the cDBSCAN algorithm based on KD-Tree implementation in the C programming language was used for neighbour cluster search. However, this method still requires predefining density [23]. We use AP-Clustering to avoid this drawback as AP-Clustering does not require predefined centre points.

Before the calculation, it is necessary to determine the initial values of damping and preference. In AP-Clustering, preference determines the likelihood of each sample being chosen as a cluster center - the higher the value, the more clusters are

formed; the lower the value, the fewer clusters. It is typically set to the median of the similarity matrix. Damping, on the other hand, is a smoothing factor used during the message-passing process to prevent oscillation and non-convergence, usually ranging from 0.5 to 0.9, with a default of 0.5. Together, these two parameters influence both the number and stability of the resulting clusters. We compared the relationship between damping factor and the number of clusters (Fig 2A) and the relationship between preference and the number of clusters (Fig 2B). It can be seen that the number of clusters is less affected by damping, and a value of 0.5 balances the calculation speed and clustering effect. The preference parameter is initialized with the median similarity between data points. The clustering effect within the sliding window is shown in Fig 2C and Supplementary Fig 1 A-C. The Euclidean distance was used as the measurement for calculation, and a unified damping and preference were applied to all windows. On average, each window contained 1.3 clusters, which is consistent with the principle of local enrichment of chromatin domain interactions.

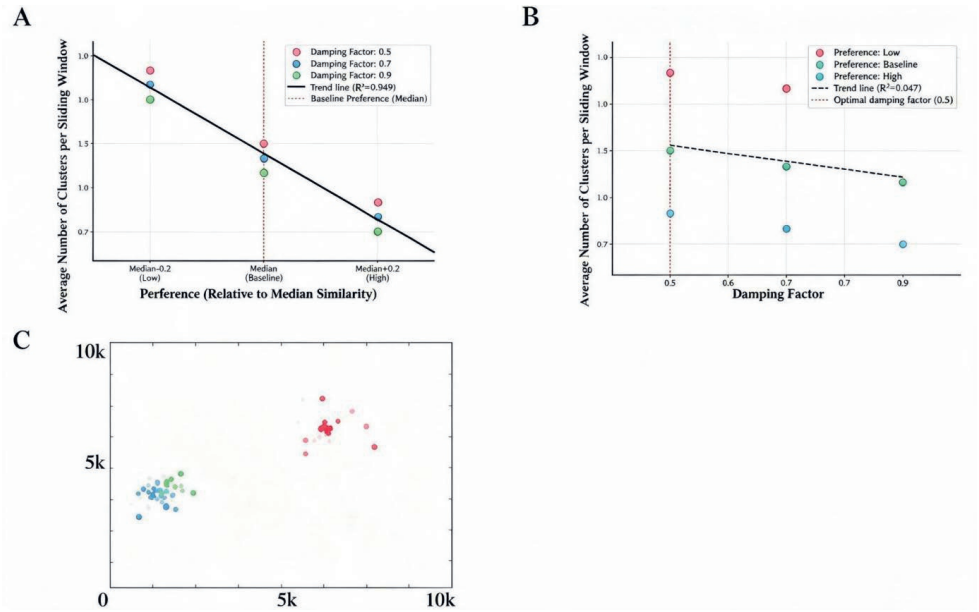


Figure 2 . (A) Relationship between preference and cluster size. Baseline indicates the number of clusters with the default preference, and different color dots represent different damping values under that preference. (B) Relationship between damping factor and cluster size. The red baseline indicates the number of clusters when the damping factor is 0.5. The red, green, and blue dots represent the number of clusters when preference is set to high, medium, and low, respectively. (C) Clustering effect within sliding windows. The blue-green point set represents cluster 1, and the red-pink point set represents cluster 2.

3.3 Mathematical analysis of unknotted TAD simulation

After obtaining the target clusters, a significant interaction analysis is performed on these clusters. When calculating the expected interaction frequency, researchers usually estimate it based on the fractal globule model [1]. In this theory, the

chromatin spatial state is a knot-free state. However, single-cell Hi-C methods have opened up the possibility of reconstructing individual chromosome chromatin paths when cells are fixed, and recent studies using numerical simulations to analyze single-cell Hi-C data have shown that chromatin fibers within a single chromosome can knot [12, 15, 24]. Dusan et al. [25] inferred that chromatin loop extrusion can promote even highly complex knot untying by modelling a standard bead-rod self-avoiding chain. The presence of knot structures and the phenomenon of knot untying indicate the need for correction of existing calculations of interaction expected values. Chromosomes clearly belong to the category of non-equilibrium systems, lying between equilibrium systems and fractal sphere models [26]. Therefore, we assume that a portion of the chromatin loops may be affected by the phenomenon of loop extrusion, and this type of loop is likely to be located out of the boundaries of TAD structures or TAD-like structures [27] while the signal of the original TAD structure is greatly weakened due to the unbinding effect (Fig 3A).

Therefore, we adjusted the λ of these loops when calculating the expected interaction. We used the signals within the class TAD where the loops are located to refit and evaluate λ. We then used a negative binomial random variable to model significance after obtaining the expected values. We identified approximately 83% TAD-like loops and 17% independent loops (Fig 3B). The affinity of unknotted loops is smaller than the original expectation based on the power law distribution of the total number of interactions (Fig 3C). This suggests that the decrease in expected values may be because of loop untangling.

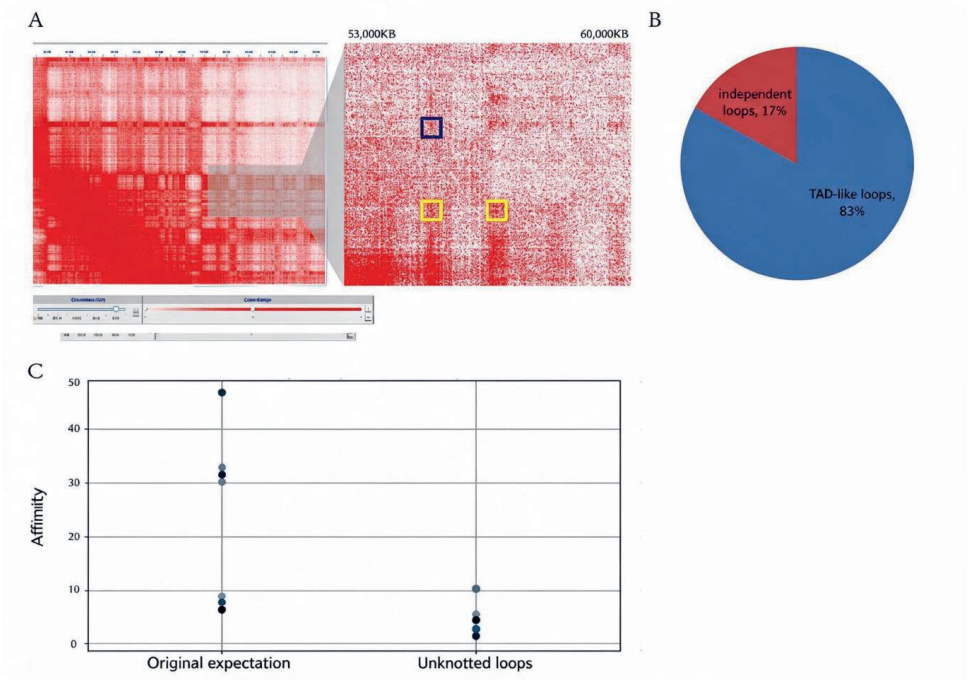


Figure 3. (A) TAD loop and unknotted TAD loop A in chromosome 14. The yellow boxes represent TAD loops, and the blue box represent an unknotted TAD loop. The data is from GSM862721 (B) The pie chart of the ratio of independent loops to TAD-like loops in the test data. (C) Numerical comparison of affinity of unknotted loops with original expectation.

3.4 Performance comparisons

3.4.1 Effect of sequencing depth

To test the sensitivity of TLoops to data volume, we performed loop calling on mouse chromosome 14 using varying amounts of data, ranging from 20% to 100% of the three Hi-C library data (Supplementary Table 1). The semaphore and the total number of loops showed a linear relationship ($R^2 > 0.961$). This indicates that TLoops are effective not only for large but also for small datasets.

3.4.2 Comparison of Different Methods

We compared TLoops with three commonly used Hi-C loop calling tools, namely HiCCUPS, Fit-Hi-C, and HOMER. Additionally, we compared the performance of these four tools with eHi-C data (GSM1718027) [28]. Regarding the total number of loops, ,Fit-Hi-C identified the most, while Tloops, HiCCUPS, and HOMER performed approximately equal (Fig 4.A). HOMER and Fit-Hi-C identified more near-range interactions, while Tloops and HiCCUPS were more balanced (Fig 4.A). We then compared the performance of different tools using Juicer Aggregated Peak Analysis (APA), and the results showed that Tloops and HiCCUPS outperformed HOMER and Fit-Hi-C (Fig4.B).

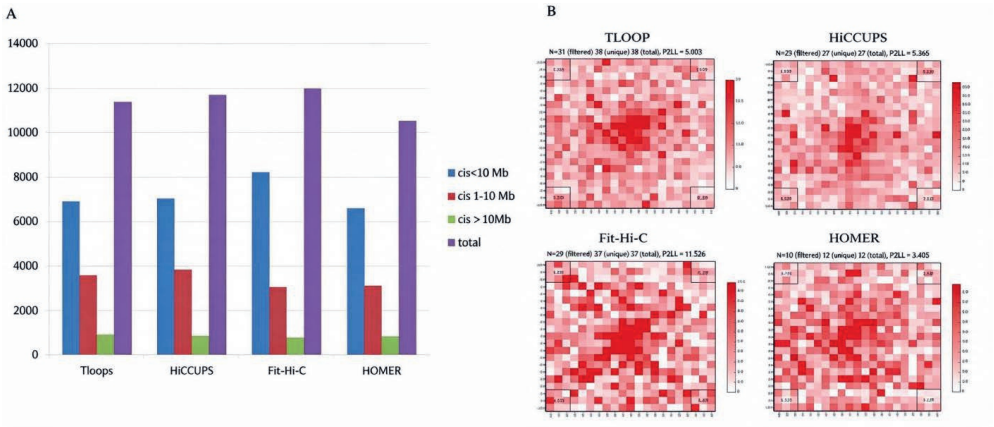


Figure 4. (A) Comparison of cis loop counts for different loop callers. (B) The APA graphs of different loop callers. The higher the P2LL value, the more prominent the peak signal of the ring is relative to its immediate background, and the stronger the ring is considered to be. The Fit-Hi-C value is the highest, which is related to the number of loops. The results for Tloop and HICCUPS were similar, while the lowest values were observed with HOMER, which was in line with expectations.

4 Discussion

The close association observed in our study between resolved knot structures and the boundaries of topologically associating domains (TADs) suggests a mechanistic relationship between the knot and TAD structures. Notably, the chromatin loops identified through our TLoops analysis frequently emerged from or anchored to these untangling sites. This aligns with the model proposed by Racko et al. [25], who suggested that loop extrusion can be instrumental in resolving topological constraints. Our results extend this concept by providing empirical evidence that the resolution of specific knotted configurations is directly coupled to the formation of stable loops, which in turn appear to participate in defining or reinforcing a subset of TAD boundaries. Consequently, the untangling process may not merely remove entanglement but actively reorganize local chromatin architecture, giving rise to new, persistent loops as a direct topological outcome.

Furthermore, the emergence of these untangling-associated loops has significant implications for interpreting genomic interaction data. We inferred that loops embedded within previously knotted regions were often masked in standard Hi-C contact maps due to locally enriched, diffuse interaction signals - a compactness effect also noted in studies of highly transcribed chromatin [29]. This highlights a limitation of conventional Hi-C analytical pipelines, which can overlook complex topological features [30]. Our integration of mathematical modeling with TLoops detection helped circumvent this issue, enabling the characterization of otherwise obscured structures. Ultimately, reconciling such topological dynamics with population-averaged Hi-C data remains a critical challenge for fully elucidating the spatiotemporal plasticity of the genome's three-dimensional structure.

5 Conclusion

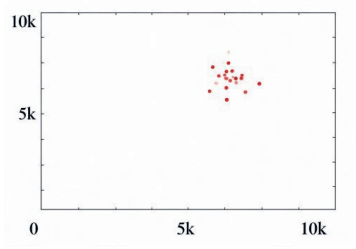
TLoops is a chromatin loop identification tool based on dynamic sliding window, AP-Clustering, and statistical testing. AP-Clustering overcomes the disadvantages of other clustering methods, and its algorithm better represents the enrichment effect of chromatin loop signals. Compared to other Hi-C loop calling tools, TLoops performs accurately and quickly in identifying chromatin loops.

Meanwhile, the effect of chromatin structure on loop classification is a topic worth further investigating. In this study, we have not yet been able to accurately classify loops based on the effect of chromatin structure, such as whether the interaction matrix can form a fixed mapping with different types of loop. This is the direction for Tloops to improve in the future.

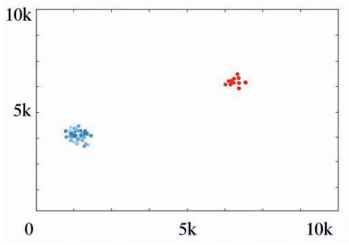
Supplementary

Supplementary figure 1

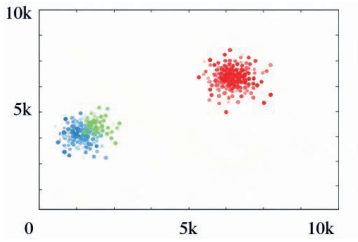
A



B



C



Supplementary figures A,B and C: Clustering is performed by randomly sampling different numbers of signals from the interacting signals. And filter out signals outside the window. The random sampling outcomes show instances of one(A), two(B), and three clusters(C) appearing in the same window size. In particular, when the centers of two clusters lie very close together (C), they are combined into a single cluster.

Supplementary Table 1

The three mouse Hi-C datasets used in this study were retrieved from the GEO repository: accession numbers GSM862721, GSM1718027 and GSM862720.

Number of loops called by different amount of reads

Subsample (% of total reads)	GSM862721	GSM1718027	GSM862720
1	553	492	505
0.8	461	392	447
0.5	231	158	181
0.2	92	45	69

6 References

1. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.
2. Zheng, H. and W. Xie, *The role of 3D genome organization in development and cell differentiation.* Nat Rev Mol Cell Biol, 2019. **20**(9): p. 535-550.
3. Sati, S. and G. Cavalli, *Chromosome conformation capture technologies and their impact in understanding genome function.* Chromosoma, 2017. **126**(1): p. 33-44.
4. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
5. Ay, F., T.L. Bailey, and W.S. Noble, *Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.* Genome Res, 2014. **24**(6): p. 999-1011.
6. Mifsud, B., et al., *GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data.* PLoS One, 2017. **12**(4): p. e0174744.
7. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.
8. Xu, Z., et al., *FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data.* Bioinformatics, 2016. **32**(17): p. 2692-5.
9. Cao, Y., et al., *Accurate loop calling for 3D genomic data with cLoops.* Bioinformatics, 2020. **36**(3): p. 666-675.
10. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
11. Siebert, J.T., et al., *Are There Knots in Chromosomes?* Polymers (Basel), 2017. **9**(8).
12. Sulkowska, J.I., et al., *KnotGenome: a server to analyze entanglements of chromosomes.* Nucleic Acids Res, 2018. **46**(W1): p. W17-W24.
13. Valdes, A., et al., *DNA knots occur in intracellular chromatin.* Nucleic Acids Res, 2018. **46**(2): p. 650-660.
14. Arsuaga, J., et al., *Current theoretical models fail to predict the topological complexity of the human genome.* Front Mol Biosci, 2015. **2**: p. 48.
15. Goundaroulis, D., E. Lieberman Aiden, and A. Stasiak, *Chromatin Is Frequently Unknotted at the Megabase Scale.* Biophys J, 2020. **118**(9): p. 2268-2279.
16. Servant, N., et al., *HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.* Genome Biol, 2015. **16**: p. 259.
17. Durand, N.C., et al., *Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.* Cell Syst, 2016. **3**(1): p. 95-8.
18. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-6.
19. Hu, M., et al., *HiCNorm: removing biases in Hi-C data via Poisson regression.* Bioinformatics, 2012. **28**(23): p. 3131-3.
20. Yaffe, E. and A. Tanay, *Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.* Nat Genet, 2011. **43**(11): p. 1059-65.
21. Imakaev, M., et al., *Iterative correction of Hi-C data reveals hallmarks of chromosome organization.* Nat Methods, 2012. **9**(10): p. 999-1003.
22. Schmitt, A.D., et al., *A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.* Cell Rep, 2016. **17**(8): p. 2042-2059.
23. Dudik, J.M., et al., *A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals.* Comput Biol Med, 2015. **59**: p. 10-18.
24. Stevens, T.J., et al., *3D structures of individual mammalian genomes studied by single-cell Hi-C.* Nature, 2017. **544**(7648): p. 59-64.
25. Racko, D., et al., *Chromatin Loop Extrusion and Chromatin Unknotting.* Polymers (Basel), 2018. **10**(10).
26. Halverson, J.D., et al., *From a melt of rings to chromosome territories: the role of topological constraints in genome folding.* Rep Prog Phys, 2014. **77**(2): p. 022601.
27. Uuskula-Reimand, L., et al., *Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders.* Genome Biol, 2016. **17**(1): p. 182.
28. Kong, S., et al., *Exonuclease combinations reduce noises in 3D genomics technologies.* Nucleic Acids Res, 2020. **48**(8): p. e44.
29. Bonev, B., et al., *Multiscale 3D Genome Rewiring during Mouse Neural Development.* Cell, 2017. **171**(3): p. 557-572 e24.
30. Finn, E.H. and T. Misteli, *Molecular basis and biological function of variability in spatial genome organization.* Science, 2019. **365**(6457).

Chapter 3

A single-cell genomic strategy for alternative transcript start sites identification

Published

Yanling Peng, Qitong Huang, Danli Liu, Siyuan Kong, Rui Kamada, Keiko Ozato, Yubo Zhang, Jun Zhu. (2024) A single-cell genomic strategy for alternative transcript start sites identification. *Biotechnol J* DOI: 10.1002/biot.202300516

Abstract
Alternative transcription start sites (TSSs) usage plays a critical role in gene transcription regulation in mammals. However, precisely identifying alternative TSSs remains challenging at the genome-wide level. We report a single-cell genomic technology for alternative TSSs annotation and cell heterogeneity detection. In the method, we utilize Fluidigm C1 system to capture individual cells of interest, SMARTer cDNA synthesis kit to recover full-length cDNAs, and dual priming oligonucleotide system to specifically enrich TSSs for genomic analysis. We apply this method to a genome-wide study of alternative TSSs identification in two different IFN-β stimulated mouse embryonic fibroblasts (MEFs). The data clearly discriminate two IFN-β stimulated MEFs. Moreover, our results indicate 81% expressed genes in these two cell types containing multiple TSSs, which is much higher than previous predictions based on Cap-Analysis Gene Expression (CAGE) (58%) or empirical determination (54%) in various cell types. This indicates that alternative TSSs are more pervasive than expected and implies our strategy could position them at an unprecedented sensitivity. It would be helpful for elucidating their biological insights in the future.

KEYWORDS
analytical biotechnology, bioinformatics

Abbreviations: ALK, anaplastic lymphoma kinase; CAGE, Cap-Analysis Gene Expression; DPO, dual priming oligonucleotides; ISG, IFN-stimulated gene; MEF, mouse embryonic fibroblast; ScISOr-Seq, single-cell isoform RNA-Seq; scTSS-seq, single cell TSSs sequencing; SUPeR-seq, single-cell universal poly(A)-independent RNA sequencing; TP, tumor protein; t-SNE, t-distributed Stochastic Neighbor Embedding; TSS, transcription start site.
Yanling Peng, Qitong Huang, and Danli Liu contributed equally to this study.

1 INTRODUCTION

Alternative transcription start sites (TSSs) usage is a common phenomenon. More than 50% of genes contain multiple TSSs, and it plays an important role in gene transcription regulation in mammals [1–3]. For example, a novel transcript of the anaplastic lymphoma kinase (*ALK*) initiates from a de novo alternative transcription initiation site in *ALK* intron 19, which leads to the expression of a novel ALK isoform and results in melanomas [1]. Similarly, alternative TSSs usage of tumor protein (TP) p53 family gene (p73) could result in full-length or shorter isoform. Elevated expression of the shorter isoform of p73 (deltaNp73), which inhibits apoptosis, has been associated with tumor progression and poor prognosis in several human cancers, including neuroblastoma, lung and ovarian carcinomas

[2]. Therefore, profiling of genome-wide alternative TSSs is expected to facilitate a better understanding of transcriptome complexity.

Cap-Analysis Gene Expression (CAGE) had the high performance of the 5′RNA-seq methods [4]. Based on the CAGE method, a series of 5′end-capture techniques have been derived. For example, nAnT-iCAGE [5] the less biased method for genome-wide identification of TSSs, nanocage [6,7] an approach for with limited amounts of RNA, and SLIC-CAGE [8], a super-low input carrier CAGE approach. As technology continues to evolve and integrate, the accuracy and comprehensiveness of single cell TSS identification continues to improve, for example scCAT-seq [9], a method for sequencing the 5′end of a single cell to infer the location of the TSS, and CamoTSS [10], a computational method that effectively detects alternative TSS. During the past decades, single-cell sequencing methods have made rapid advances on profiling genomics at unprecedented resolution [11]. Traditionally, bulk cell sequencing methods can only reveal the average expression signal from an ensemble of cells assumed to be one homogeneous population/tissue status [11]. Therefore, signals from small cell numbers or low densities could hardly be detected. Using single-cell technologies, numerous novel transcripts/cell types thereby have been identified [12]. For examples, based on single- cell universal poly(A)-independent RNA sequencing (SUPeR-seq), 913 novel linear transcripts and 2891 circRNAs in mouse preimplantation embryos have been found [12]. Based on single-cell isoform RNA-seq (ScISOr-seq), 16,872 novel isoforms and 18,173 known isoforms in thousands of cerebellar cells have been detected [13]. Thus, single cell sequencing is able to detect cell heterogeneity and lowly expressed alternative TSSs which may be un-achievable by bulk cell technologies.

Here, to enable genome-wide alternative TSSs annotation and cell heterogeneity detection, we have developed single cell TSSs sequencing (scTSS-seq) method (Figure 1). In this method, we isolate single cells in Fluidigm C1 system, capture full-length cDNAs using SMARTer cDNA synthesis kit, and enrich 5′end cDNA tags specifically with customized dual priming oligonucleotides (DPO) [14] in Nextera XT DNA Library Prep Kit. DPO contains a longer 5′- segment, a shorter 3′-segment, a poly(I) linker between these two segments [14]. This enables enrichment of 5′end cDNA tags precisely even under sub-optimal PCR conditions [14]. We then apply this method to naïve 1 h and IFN-β pretreated 1 h mouse embryonic fibroblasts (MEFs) (Materials and Methods) (Figure 2). Furthermore, in conjunction with RNA-seq, we examine whether pervasive alternative TSSs contribute to transcription memory following IFN-β re-stimulation. This study detects a larger proportion of genes containing alternative TSSs than empirical estimations and shows that alternative TSSs are cell-type-specific and may be involved in transcriptional memory.

2 MATERIALS AND METHODS

2.1 Cell preparation

The preparation of MEFs was similar to the protocol of Kamada et al [15]. In brief, MEFs were separated from day 13.5 mouse embryos, and then cultured in Dulbecco's minimal essential medium (DMEM) with 10% fetal bovine serum (FBS) at 37◦C with 5% CO2. Later, they were treated with 100 units/mL of murine recombinant IFN-β (PBL Interferon Source) with two different ways for indicated periods (Figure 2). For naïve 0 h MEFs, MEFs were treated without IFN-β for 6 h, washed, and left without IFN-β for 24 h, and then harvested. For pretreated 0 h MEFs, MEFs were treated with IFN-β for 6 h, washed, and then left without IFN-β for 24 h, and then harvested. In this study, both naïve and pretreated MEFs were restimulated with IFN-β for 1 h and then harvested for scTSS-seq, which were termed as naïve 1 h, and pretreated 1 h MEFs, respectively (Figure 2). For bulk cell RNA-seq, besides naïve 1 h and pretreated 1 h MEFs, naïve 0 h and pretreated 0 h MEFs were used.

2.2 scTSS-seq experiment procedure

A total of 96 naïve 1 h and 96 pretreated 1 h MEFs were captured using the C1 Fluidigm system according to the manufacturer's instructions. Upon capture, reverse transcription and cDNA preamplification were performed in the C1 Fluidigm system using the SMARTer PCR cDNA Synthesis Kit (Clontech 634926) according to manufacturer's recommendations. Full-length cDNAs were harvested and diluted to a range of 0.2 ng/µl, and Nextera libraries were prepared using the Nextera DNA Sample Preparation Kit (Catalog No. FC-131-1096). Nex- tera P5 (Illumina) and custom-made P7 index primers (Table S1) were used to amplify the tagmented fragments. The P7 index primers were designed as DPO. The details of the scTSS-seq library construction procedure were given in Supplementary File 1. Libraries were pooled and single-end sequencing was performed on HiSeq 2000 (Illumina).

2.3 scTSS data processing

Reads were subjected to adapter and quality trimming with Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_ galore/; v0.4.4; options: default parameters). The trimmed reads were then aligned to the mouse genome (mm9) using bowtie2[16] (v2.29, options: default parameters). HOMER[17] tag

directory was created with “makeTagDirectory” function (options: default parameters) for each cell. For each tag directory, peaks were called with HOMER “findpeaks” function [17] (options: -style tss -ntagThreshold 200). For each cell, peaks were then annotated with HOMER “annotatePeaks.pl” function [17], three different mm9 transcript models were used for annotations, including Refseq (HOMER default), UCSC (https://genome.ucsc.edu/cgi-bin/hgTables), and Genecode V1 (https://www.gencodegenes.org/mouse/release_M1.html). In these models, annotated promoters were defined as regions between downstream 1000 bp and upstream 100 bp of TSSs. For each cell, peaks overlapped with annotated promoters in any of three models were termed as known TSSs, while the remaining were novel TSSs (Figure 2). Furthermore, we downloaded mouse mm10 G-quadruplex (G4) data from EndoQuad (https://endoquad.chenzxlab.cn/). Since we utilized the mm9 version of the reference sequence, we employed liftover to convert the database to the mm9 version. Subsequently, we performed intersect analysis on our TSCs using bedtools (v2.17.0). We removed the novel TSCs overlapped with G4 database.

Low quality cells would have lost transcripts prior to cell lysis and their exon numbers would be larger in broken cells [18]. Thus, we used the ratios between known TSSs numbers and identified TSSs numbers in total for each cell (known TSS ratios) for filtering low quality cells. In this study, we used threshold as z-score of known TSSs ratios of 1 for removing cells. The remaining cells were considered as high quality cells and used for downstream analysis.

To produce the matrix of TSSs tags across cells, we merged TSSs when their 5′end were located around ±2 bp across all high quality cells. For further analysis, the expression levels of TSSs were normalized to counts per million, with calculation by counts per mapped reads per 10 million. The expression levels of genes were calculated as the expression levels of TSSs tags within Refseq genes on aver- age. t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis and differential analysis for TSSs were performed by using R pack- ages Seurat [19]. To know how many novel TSSs were overlapped in other TSSs datasets across difference cell types, we downloaded TSSs coordinates (bed format) from different cell types or tissues such as 10Thalf, 3T3, ATDC5, and embryos (ftp://ftp.hgc.jp/pub/hgc/db/dbtss/ dbtss_ver8/mm9/TSSseq/bed/).

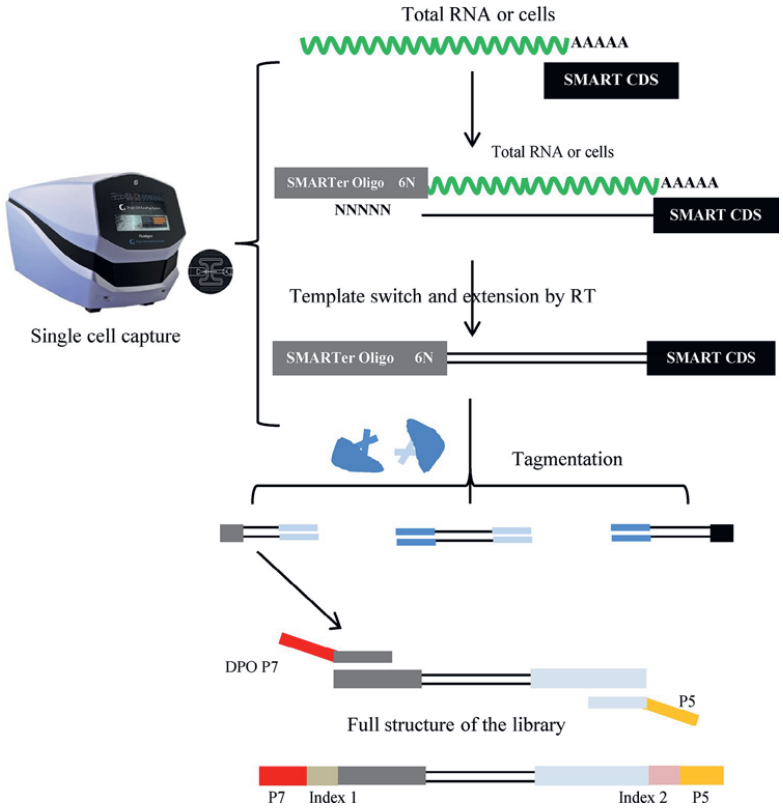


Figure 1. Schematic of the scTSS-seq method. P5, Nextera XT illumina P5 primers; DPO P7, custom-made DPO P7 primers. DPO, dual priming oligonucleotides; scTSS-seq, single cell TSSs sequencing.

2.4 RNA-seq procedure and analysis

The library construction of RNA-seq was similar to Kamada et al. [15]. In contrast, single-end sequencing on a HiSeq2000 (Illumina) was used in our study. The obtained sequences were aligned to the mouse genome mm9 using Tophat [20] (v2.29, options: default parameters). Transcript abundance was quantified using Cufflinks [21] (v1.2.1, options: default parameters). Differential expressed genes (DEGs) are defined as the Fold change (FC) ≥ 2 of gene expression levels and the p values ≤ 0.05 using the program Cuffdiff [22]. To determine IFN-stimulated genes (ISGs), we first identified genes showing significantly ($p < 0.01$, FC ≥ 2) higher transcript expression (RPKM) in treated cells following IFN treatment, that is, treated 1 h versus treated 0 h, which detected 472 upregulated genes. Then, we identified genes showing significantly ($p < 0.01$, FC ≥ 2) higher transcript expression (RPKM) in untreated cells following IFN treatment, that is, untreated 1

h and untreated 0 h, which included 346 upregulated genes, while 88 genes were defined as ISGs as their gene expression levels higher upregulated in treated cells than untreated cells (expression of treated upregulated genes/untreated upregulated genes > 1). We also identified 188 genes downregulated in treated cells and 871 genes upregulated in untreated cells during IFN treatment. A total of 24 was defined as non-ISGs as their gene expression levels higher downregulated in treated cells than untreated cells (expression of treated downregulated/untreated downregulated genes < 1). Gene Ontology (GO) analysis of different categories of ISGs were performed by using R packages ClusterProfiler [23]. Terms with $p < 0.05$ were regarded as significantly enriched.

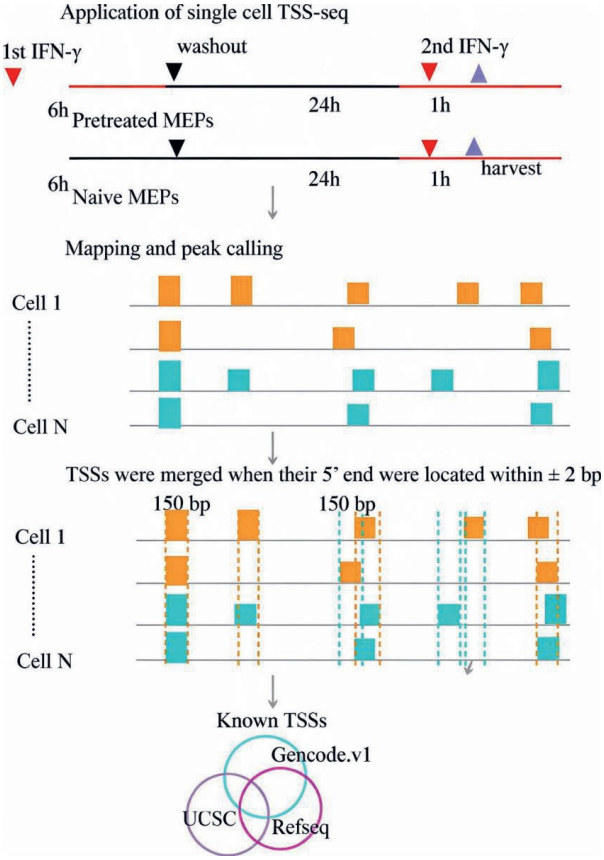


Figure 2. scTSS-seq applications and TSSs identification schematic in naïve 1 h and pretreated 1 h MEFs. MEF, mouse embryonic fibroblast; scTSS-seq, single cell TSSs sequencing; TSS, transcription start site.

3 Result

3.1 The performance of scTSS-seq

In this study we filtered the possible false positive TSSs based on known G4 database and removed six naïve and three pretreated cells based on quality metrics (Section 2). At the single cell level we detected 5002 TSSs in median and 2713 known TSSs (54.2% in total) per cell (Figure 3A). When all single cells were pooled together our study detected 454,719 TSSs in total and 143,923 known TSSs (34% in total) (Figure 3B). In comparison, C1 CAGE, an alternative TSS identification method, detected a median of 2788 CAGE clusters and 948 known TSSs (34% in total) per cell by applying C1 CAGE to 151 A514 cells following TGF-β stimulation (40 cells for 0 h, 41 cells for 6 h, and 70 cells for 24 h) [26]. Moreover, our results identified a median of 1150 novel TSSs (23% in total of TSSs) per cell (Figure 3A) and a total of 155,485 novel TSSs (34%) across all single cells overlapped with other TSSs in dbTSSs (Figure 3B), indicating a large part of TSSs that were not overlapped with annotated promoters were identified in other cell types. In addition, to verify the stability of the identified signals, we used a logistic regression-based approach to train the parameters peak score, focus ratio, distance to nearest RefSeq TSS, and cluster volume (Supplementary 2). All the results suggested that our method was efficient at determining TSSs.

Later, we investigated the correlation of gene expression between scTSS-seq and bulk RNA-seq dataset for each cell type. For scTSS-seq data, gene expression levels were calculated as the sum of genic TSSs tags within the genes (CPM). The expression profiles of scTSS-seq data and bulk RNA-seq data were well correlated (rho = 0.65 for naïve 1 h MEFs, rho = 0.71 for pretreated 1 h MEFs, $p < 0.001$, Spearman correlation, Figure 3C, D). This indicated that TSSs data were of good quality. To test whether scTSS-seq was able to differentiate cell types as well as other single-cell technologies, we employed the t-SNE method for non- linear dimension reduction and data visualization method. The result suggested that two prominent clusters consistent with the two well-predefined cell types were classified (Figure 3E), suggesting that scTSS-seq could detect cell heterogeneity.

Furthermore, we characterized the distribution of TSSs across 183 high quality MEFs. About 363,384 TSSs were present in a single cell (cell-unique TSSs) while 91,407 occurred in more than one cell (cell-shared TSSs) (Figure 3F). About 24% of cell-unique TSSs were known TSSs, while more than 62% cell-shared TSSs were known TSSs (Figure 3F). These results further supported that single cell technologies were able to detect numerous novel TSSs or transcripts [12].

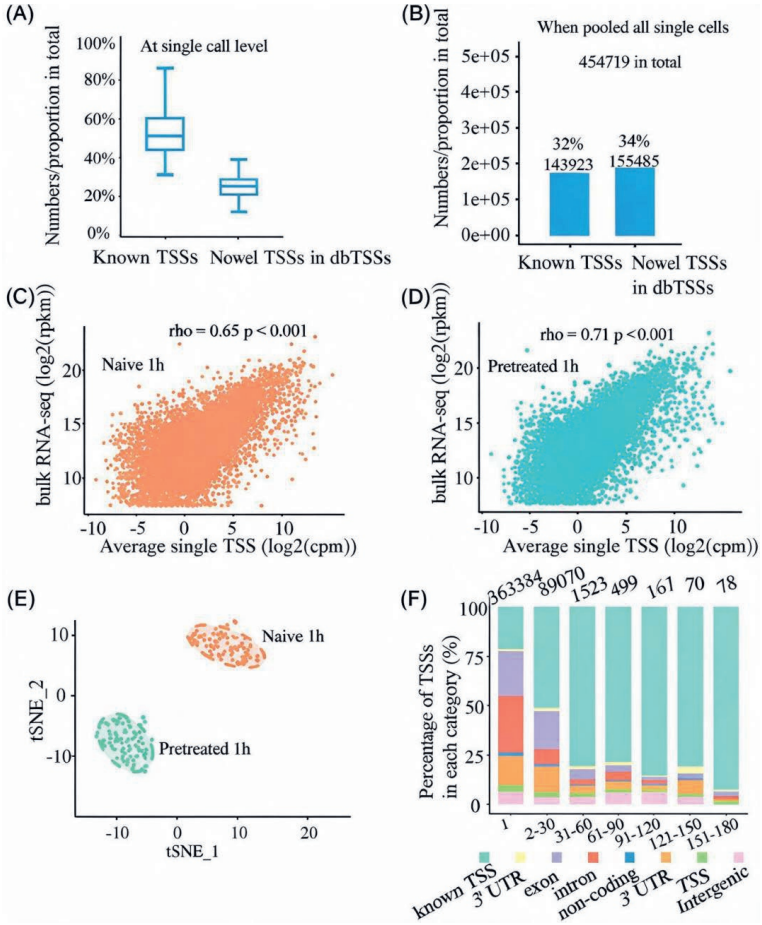


Figure 3. The performance of scTSS-seq. (A, B) TSSs identification at single cell level and when all single cells were pooled together. (C, D) The Spearman correlation of gene expression profile based on scTSS-seq data and bulk RNA-seq data for naïve 1 h and pretreated 1 h MEFs. (E) t-SNE analysis based on TSSs expression levels. (F) Distribution of genomic annotation for TSSs. Note: 1, TSSs occurring in one cell; 2–30 TSSs occurring between 2 and 30 cells; 31–60, TSSs occurring between 31 and 60 cells and the remaining followed the similar rule to name. MEF, mouse embryonic fibroblast; scTSS-seq, single cell TSSs sequencing; t-SNE, t-distributed Stochastic Neighbor Embedding; TSS, transcription start site.

3.2 Alternative TSSs detection at single cell level

To characterize the distribution of alternative TSSs and to compare with other bulk cell 5′end cDNA technologies, we first excluded the TSSs occurring in one cell and then counted alternative TSSs across naïve and pretreated MEFs. In this study, we found 91,407 TSSs occurring in more than two cells, of which 89,159 genic TSSs belonged to 11,557 Refseq genes. Among these genes, 9,305 (81% in total) had more than one TSS, and most genes contained 2–10 TSSs (Figure 4A). With bulk

CAGE, 11,264 (58% in total) genes had alternative TSSs in mammals [24]. With empirical determination according to the UCSC database, 54% of these genes exhibited alternative TSSs [3]. This suggested our method was capable of detecting alternative TSSs at genome-wide scale sensitively.

To explore whether the alternative TSSs numbers changes were correlated with their gene expression. Based on the matrix of the average TSSs expression level per gene using Seurat [25] and differential gene expression analysis of bulk RNA-seq, we detected 67 significant differentially expressed genes between pretreated 1 h and naïve 1 h following IFN stimulation, including 52 genes upregulated (average log2FC > 1, p value adjust < 0.05) and 17 genes downregulated (average log2FC < −1, p value adjust < 0.05) in pretreated 1 h compared to naïve 1 h cells (Figure 4B). The alternative TSSs numbers within upregulated genes were significantly larger in pretreated 1 h than in naïve 1 h MEFs ($p < 0.001$, the paired Wilcoxon signed rank test with continuity correction). In contrast, the alternative TSSs numbers within downregulated genes were significantly smaller in pretreated 1 h compared to naïve 1 h MEFs ($p > 0.05$, the paired Wilcoxon signed-rank test with continuity correction). Later, we investigated the changes of alternative TSSs numbers for similar expressed genes between two cell types. A total of 1,178 genes exhibited similar expression levels (abs(log2(FC)) < 1, $p > 0.05$) between pretreated 1 h and naïve 1 h MEFs. The numbers of alternative TSSs within these genes were significantly larger in pretreated 1 h than naïve 1 h MEFs ($p < 0.001$, the paired Wilcoxon signed-rank test with continuity correction). These results indicated that alternative TSSs may be involved in transcription regulation using varying mechanisms, such as dependent or independent of changes in gene expression.

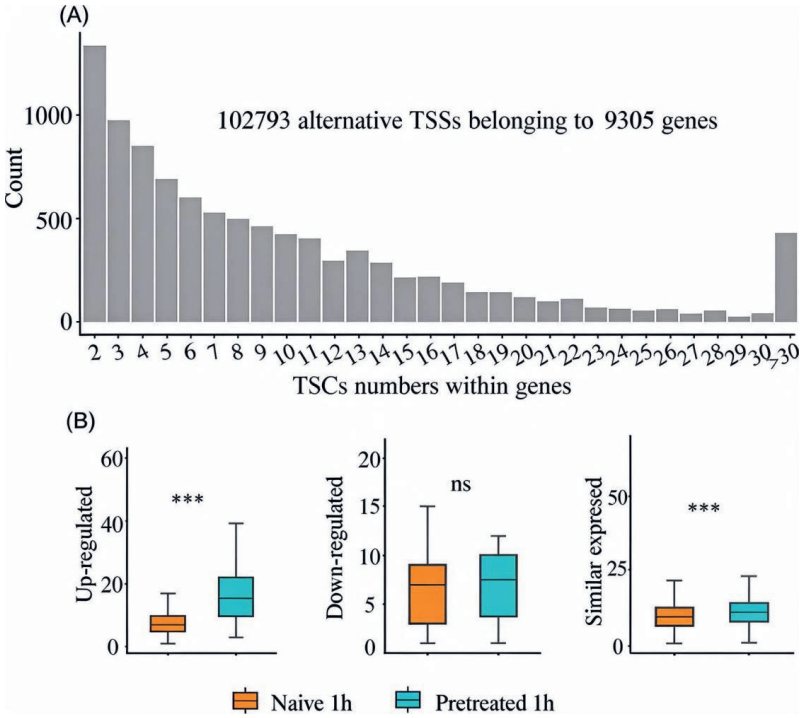


*Figure 4. Alternative TSSs usage at single cell resolution. (A) The distribution of alternative TSSs. (B) The comparison of TSSs numbers within genes between naïve 1 h and pretreated 1 h MEFs for unregulated, downregulated and similar expressed genes. Note: ***p < 0.001 the paired Wilcoxon signed-rank test are performed. MEF, mouse embryonic fibroblast; TSS, transcription start site.*

3.3 Alternative TSSs usage is correlated with transcriptional memory

With IFN-β restimulation in MEFs, it has been reported that the recruitment of RNA Pol II in memory ISGs is faster and greater than that in non-memory ISGs [15]. To test whether alternative TSSs usage affected transcription memory, we combined bulk RNA-seq data to analyze the changes of alternative TSSs numbers for ISGs across naïve and pretreated 1 h MEFs. Taking similar strategies with Kamada et al. [15], we identified 112 upregulated ISGs based on bulk RNA- seq data. Of these, 88 (79%) ISGs mRNA levels were >1-fold higher in pretreated 1 h than naïve 1 h MEFs (>1 in RPKM), which was defined as memory ISGs; while 24 (21% in total) ISGs mRNA levels were ≤1-fold higher (≤1 in RPKM), which was defined as nonmemory ISGs (Figure 5A). GO analyses showed that memory and non-memory ISGs shared related categories, such as innate immune and defense responses. The GO results were similar in naïve and pre- treated MEFs restimulated by IFN-β in previous study [15]. For these 88 memory ISGs, the alternative TSSs numbers within memory ISGs were significantly larger in pretreated 1 h than in naïve 1 h MEFs (Figure 5B) ($p < 0.001$, the paired Wilcoxon

signed-rank test with continuity correction). For example, memory IGS H2-K1 in pretreated 1 h have more prevalent alternative TSSs usage than in naïve 1 h MEFs (Figure 5C). For 37 non-memory ISGs (Figure 5B), alternative TSSs numbers within non-memory ISGs showed no significant differences between pretreated 1 h and naïve 1 h MEFs ($p > 0.01$, the paired Wilcoxon test with continuity correction). These results supported that TSSs provided a platform for RNA Pol II binding, while the rapid recruitments of RNA Pol II binding were necessary for transcription memory [15].

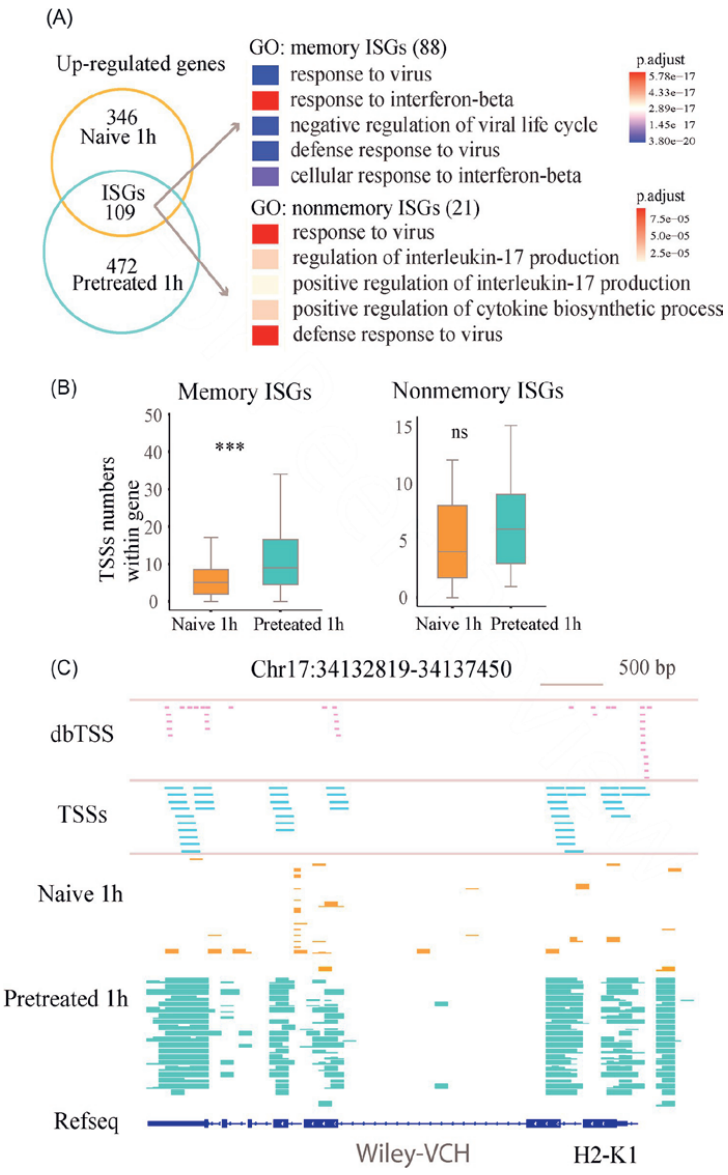


Figure 5. Memory ISGs related-TSSs biological process. (A) The identification of memory and non-memory ISGs and GO analysis for the identified memory and nonmemory ISGs. (B) Comparison of alternative TSSs numbers per gene between naïve 1 h and pretreated 1 h MEFs, the paired Wilcoxon signed-rank test is performed, (C) example of TSSs distribution for ISGs, H2-K1. GO, Gene Ontology; ISG, IFN-stimulated gene; MEF, mouse embryonic fibroblast; TSS, transcription start site.

4 Discussion

In this study, we have developed scTSS-seq for genome-wide alternative TSSs detection at single cell level. This method utilizes Fluidigm C1 system, SMART technology, and DPO system to increase detection sensitivity of transcripts and genes, full-length rate of cDNAs, and enrichment specificity [14,26,27] respectively. The results indicate scTSS-seq to be as accurate as C1 CAGE [28]. The previous study indicated the six 5′RNA-seq methods included CAGE, RAMPAGE, STRT, NanoCAGE, oligo-capping, and GRO-cap, the study indicated that CAGE performed best overall by comparing RNA input, per-sample cost, the amount of time and number of steps per library for each method, and the rate of false positives [4]. C1-CAGE method allows 5′-end transcript profiling with strand information in a single cell [28]. In addition, this method is able to detect cell heterogeneity. Further- more, higher rates (82%) of alternative TSSs are detected compared to predictions based on CAGE (58%) or empirical determination (54%) in various cells. Thus, scTSS-seq is an efficient method for positioning alternative TSSs at the genome-wide scale.

At the same time, we have detected that memory ISGs have larger numbers of alternative TSSs in pretreated 1 h than in naïve 1 h MEFs, implying that alternative TSSs usage is cell-type-specific and may be contributed to transcriptional memory [15]. However, it still remains challenging to validate the role of alternative TSSs on memory ISGs in single cells so far. Firstly, cell viability is required when isolating single-cells for the purpose of production of monoclonal cell culture, which would be affected during single cell separation [29]. Secondly, the amount of protein and RNA of single cells may be not enough for histochemical experiments.[30]. Thirdly, the quality and amount of antibodies would be limited for validation [31]. Despite these factors, we have developed a reliable method for genome-wide alternative TSSs detection and to dissect a more pervasive alternative TSSs phenomenon in cells. Meanwhile, we have provided useful alternative TSSs resources for future biological functional research.

AUTHOR CONTRIBUTIONS

Yubo Zhang and Jun Zhu conceived the project, supervised, and designed the study. Yubo Zhang and Rui Kamada performed experiments. Keiko Ozato supervised the

culture study. Yanling Peng and Qitong Huang carried out computational analyses. Yanling Peng, Danli Liu, Siyuan Kong, Yubo Zhang, and Jun Zhu wrote the manuscript.

ACKNOWLEDGMENTS
This work was supported by National Key Research and Development Program of China [2018YFA0903201], National Natural Science Foundation of China [2017M620977], the Science, Technology and Innovation Commission of Shenzhen Municipality [JCYJ20180306173714935], NIH Intramural research program at NICHD [ZIA HD001310-35], Funding for open access charge: National Key Research and Development Program of China, the National Natural Science Foundation of China (No.32202653), and the Project Funded by the China Postdoctoral Science Foundation (No. BX2021367 and 2021M703543)

CONFLICT OF INTEREST STATEMENT
The authors declare no competing interests.

DATA AVAILABILITY STATEMENT
All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https:// www.ncbi.nlm.nih.gov/geo/) under accession number GSE174480.

SUPPLEMENTARY DATA
Supplementary Data are available Online**.**
(https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/biot.202300516)

5 REFERENCES

1. Wiesner, T., Lee, W., Obenauf, A. C., Ran, L., Murali, R., Zhang, Q. F., Wong, E. W. P., Hu, W., Scott, S. N., Shah, R. H., Landa, I., Button,
J., Lailler, N., Sboner, A., Gao, D., Murphy, D. A., Cao, Z., Shukla, S., Hollmann, T. J., Chi, P. (2015). Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature, 526, 453–457.
2. Murray-Zmijewski, F., Lane, D. P., & Bourdon, J. C. (2006). p53/p63/p73 isoforms: An orchestra of isoforms to harmonise cell differentiation and response to stress. Cell Death and Differentiation, 13, 962–972.
3. Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C. M., Gibson, D., Gonzalez, J. N., Guruvadoo, L., Haeussler, M., Heitner, S., Hinrichs, A. S., Karolchik, D., Lee, B. T., Lee, C. M., Nejad, P., Villarrea, C., Vivian, J. (2017). The UCSC Genome Browser database: 2017 Update. Nucleic Acids Research, 45, D626–D634.
4. Adiconis, X., Haber, A. L., Simmons, S. K., Levy, A. L, Moonshine, A., Ji, Z., Busby, M. A., Shi, X., Jacques, J., Lancaster, M. A., Pan, J. Q., Regev, A., & Levin, J. Z. (2018). Comprehensive comparative analysis of 5′ -end RNA-sequencing methods. Nature Methods, 15, 505–511.
5. Murata, M., Nishiyori-Sueki, H., Kojima-Ishiyama, M., Carninci, P., Hayashizaki, Y., & Itoh, M. (2014). Detecting expressed genes using CAGE. Methods in Molecular Biology, 1164, 67–85.
6. Salimullah, M., Mizuho, S., Plessy, C., & Carninci, P. (2011). NanoCAGE: A high-resolution technique to

discover and interrogate cells transcriptome. Cold Spring Harbor Protocols, 2011(1), pdb.prot5559.
7. Poulain, S., Kato, S., Arnaud, O., Morlighem, J. É., Suzuki, M., Plessy, C., & Harbers, M. (2017). NanoCAGE: A method for the analysis of coding and noncoding 5'-capped transcriptomes. Promoter Associated RNA: Methods and Protocols, 1543, 57 – 109.
8. Cvetesic, N., Leitch, H. G., Borkowska, M., Müller, F., Carninci, P., Hajkova, P., & Lenhard, B. (2018). SLIC-CAGE: High-resolution tran- scription start site mapping using nanogram-levels of total RNA. Genome Research, 28(12), 1943–1956.
9. Liu, L., Liu, C., Quintero, A., Wu, L., Yuan, Y., Wang, M., Cheng, M., Leng, L., Xu, L., Dong, G., Li, R., Liu, Y., Wei, X., Xu, J., Chen, X., Lu, H., Chen, D., Wang, Q., Zhou, Q.,Xu, X. (2019). Deconvolution of single-cell multi- omics layers reveals regulatory heterogeneity. Nature Communications, 10(1), 470.
10. Hou, R., Hon, C. C., & Huang, Y. (2023). CamoTSS: Analysis of alternative transcription start sites for cellular phenotypes and regula tory patterns from 5' scRNA-seq data. Nature Communications, 14(1), 7240.
11. Tang, X. M., Huang, Y. M., Lei, J. L., Luo, H., & Zhu, X. (2019). The single- cell sequencing: New developments and medical applications. Cell & Bioscience, 9, 53.
12. Fan, X. Y., Zhang, X. N., Wu, X. L., Guo, H. S., Hu, Y., Tang, F., & Huang, Y. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biology, 16, 148.
13. Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E., & Tilgner, H. U. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nature Biotechnology, 36, 1197–1202.
14. Chun, J. Y., Kim, K. J., Hwang, I. T., Kim, Y. J., Lee, D.-H., Lee, I.-K., & Kim, J.-K. (2007). Dual priming oligonucleotide system for the multiplex detection of respiratory viruses and SNP genotyping of CYP2C19 gene. Nucleic Acids Research, 35, e40.
15. Kamada, R., Yang, W., Zhang, Y., Patel, M. C., Yang, Y., Ouda, R., Dey, A., Wakabayashi, Y., Sakaguchi, K., Fujita, T., Tamura, T., Zhu, J., & Ozato, K. (2018). Interferon stimulation creates chromatin marks and establishes transcriptional memory. Proceedings of the National Academy of Sciences of the United States of America, 115, E9162–E9171.
16. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9, 357–359.
17. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell, 38, 576–589.
18. Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., & Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. Genome Biology, 17, 29.
19. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of single-cell data. Cell, 177, 1888– 1902.e21.
20. Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111.
21. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Tran- script assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28, 511–515.
22. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28(5), 511–515.
23. Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). ClusterProfiler: An R Package for comparing biological themes among gene clusters. Omics, 16, 284–287.
24. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M.,Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nature Genetics, 38, 626–635.
25. Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology, 42(2), 293–304.
26. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature Methods, 10, 1096–1098.

27. Hashimshony, T., Senderovich, N., Avital, G., & Klochendler, A. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. Genome Biology, 17, 77.
28. Kouno, T., Moody, J., Kwon, A. T. J., Shibayama, Y., Kato, S., Huang, Y., Böttcher, M., Motakis, E., Mendez, M., Severin, J., Luginbühl, J., Abugessaisa, I., Hasegawa, A., Takizawa, S., Arakawa, T., Furuno, M., Ramalingam, N., & West, J. (2019). C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nature Communications, 10, 360.
29. Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R., & Koltay, P. (2015). Technologies for single-cell isolation. International Journal of Molecular Sciences, 16, 16897–16919.
30. Romanov, R. A., Zeisel, A., Bakker, J., Girach, F., Hellysaz, A., Tomer, R., Alpár, A., Mulder, J., Clotman, F., Keimpema, E., Hsueh, B., Crow, A. K., Martens, H., Schwindling, C., Calvigioni, D., Bains, J. S., Máté, Z., Szabó, G., Yanagawa, Y. (2017). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. Nature Neuroscience, 20, 176–188.
31. Taussig, M. J., Fonseca, C., & Trimmer, J. S. (2018). Antibody validation: A view from the mountains. New Biotechnology, 45, 1–8.

Chapter 4

MAE-seq refines regulatory elements across the genome

Published

Xiusheng Zhu†, Qitong Huang†, Lei Huang†, Jing Luo†, Qing Li†, Dashuai Kong, Biao Deng, Yi Gu, Xueyan Wang, Chenying Li, Siyuan Kong*, and Yubo Zhang* *Nuclei Acids Res.* DOI: 10.1093/nar/gkad1129
†The first five authors should be regarded as Joint First Authors.

Abstract
Proper cell fate determination relies on precise spatial and temporal genome-wide cooperation between regulatory elements (REs) and their targeted genes. However, the lengths of REs defined using different methods vary, which indicates that there is sequence redundancy and that the context of the genome may be unintelligible. We developed a method called MAE-seq (Massive Active Enhancers by Sequencing) to experimentally identify functional REs at a 25-bp scale. In this study, MAE-seq was used to identify 626,879, 541,617 and 554,826 25-bp enhancers in mouse embryonic stem cells (mESCs), C2C12 and HEK 293T, respectively. Using ~1.6 trillion 25 bp DNA fragments and screening 12 billion cells, we identified 626,879 as active enhancers in mESCs as an example. Comparative analysis revealed that most of the histone modification datasets were annotated by MAE-seq loci. Furthermore, 33.85% (212,195) of the identified enhancers were identified as *de novo* ones with no epigenetic modification. Intriguingly, distinct chromatin states dictate the requirement for dissimilar cofactors in governing novel and known enhancers. Validation results show that these 25-bp sequences could act as a functional unit, which shows identical or similar expression patterns as the previously defined larger elements, Enhanced resolution facilitated the identification of numerous cell-specific enhancers and their accurate annotation as super enhancers. Moreover, we characterized novel elements capable of augmenting gene activity. By integrating with high-resolution Hi-C data, over 55.64% of novel elements may have a distal association with different targeted genes. For example, we found that the *Cdh1* gene interacts with one novel and two known REs in mESCs. The biological effects of these interactions were investigated using CRISPR-Cas9, revealing their role in coordinating *Cdh1* gene expression and mESC proliferation. Our study presents an experimental approach to refine the REs at 25-bp resolution, advancing the precision of genome annotation and unveiling the underlying genome context. This novel approach not only advances our understanding of gene regulation but also opens avenues for comprehensive exploration of the genomic landscape.

1 Introduction

Transcriptional regulatory elements (REs) are nucleotide sequences of genes involved in genetic transcription regulation. Accurate identification of REs is fundamental for annotating genomes and understanding gene expression patterns (1). According to the ENCODE (Encyclopedia of DNA Elements) project, 74.7% of the genome is transcribed, with 56.1% associated with modified histones, 15.2% in open-chromatin areas, 8.5% binding transcription factors (TFs), and 4.6% comprising methylated CpG dinucleotides (2). However, mounting evidence

suggests that many of these transcribed regions are nonfunctional (3–5). Whitfield et al. (6) found that less than 50% of predicted transcription factor binding sites (TFBSs) from ENCODE made a functional contribution to the promoter activity in four cell lines (K562, HCT116, HT1080 and HepG2). Non-coding loci without epigenomic signals can still be functional in the human genome (7). Despite advances in sequencing technology, genome annotation methods remain largely unchanged, leading to prevalent annotation errors (8). Therefore, there is a growing need for precise RE identification.

REs control gene expression temporally and spatially, playing pivotal roles in biological development, differentiation, and disease occurrence (9). The most common experimental method for genome-wide identification of transcriptional REs is chromatin immunoprecipitation, followed by high-throughput sequencing (ChIP-seq) (10) and chromatin-based methods. Unfortunately, due to the limitations of nucleosome structure or technical means, the lengths of the REs identified vary a lot. In eukaryotes, 147 bp of DNA are wrapped around an octamer of histones (two copies of each H2A, H2B, H3, and H4) to form the smallest unit of chromatin, the nucleosome (11). Therefore, it is tough for existing chromatin-based methods to refine the REs span to less than ~147 bp. For example, the RE length revealed by DHS-seq, FAIRE-seq and ChIP-seq is ~200–500 bp, ~200–1000 bp and ~200–700 bp, respectively (12–14). As a non-chromatin-based method, the length of the recently developed STARR-seq is 500–850 bp (15,16). These differences in length suggest that the sequences of annotated REs likely have redundancy. Meanwhile, the RE size predicted in the ENCODE project is much larger than its actual size (17). This implies that the information hidden in a large part of these discriminated nucleotides has not been properly interpreted, and that their potential functions might remain unknown. This indicates that accurate identification and characterization of REs need to be further clarified and their control of transcription in a cell-type specific manner might enlighten novel molecular mechanisms of gene regulation and diseases (18).

Massively parallel reporter assays (MPRAs) have been widely used for screening REs, including episomal and self-transcribed active regulatory region sequencing (STARR-seq) (19). hey sequence mRNA abundance in transfected cells to identify enhancer sequences. However, constructing output libraries for MPRAs is expensive and time-consuming. Additionally, bioinformatic methods have identified thousands of short core motifs spanning 6–14 nucleotides, but these motifs do not function as individual REs (20).

Cell type-specific enhancers, crucial in dictating cellular identity, frequently engage cell-specific transcription factors, including super enhancers (21, 22).

Existing enhancer investigations often involve multiple transcription factors, leading to complex analyses due to the variability in sequence lengths. Despite their large span, functional regions within super enhancers have been shown to be only modestly sized (23), necessitating experimental efforts for comprehensive annotation. Enhancer functionality hinges on cofactors, such as the COFs family (e.g. CDK9, BRD4) and chromatin-remodeling complexes (CRCs) like SMARCC1 and SMARCB1 (24–26). The interplay between enhancers, cofactors, and the epigenetic landscape merits further exploration.

In this study, we combined the concept of motifs with MPRAs to design short random nucleotides libraries with different lengths for high throughput discovery of functional ERs with minimum length, termed MAE-seq (Massively Active Enhancer for Sequencing). This approach allows high-throughput discovery of functional enhancer REs with minimum length. To screen functional enhancers, we used a minimal promoter to test the activity of the upstream enhancer of the fluorescent reporter gene, which remains silent in the absence of an enhancer or basal expression. Our results showed that ~25 bp is a suitable length for functional enhancer screening. We applied MAE-Seq in both human and mouse cells, identifying tens of thousands of 25-bp enhancer REs in these cell types. These data refine the sequence context of REs and uncover precise genetic information throughout the genome.

2 Materials and methods

2.1 Screening vector

The reporter vector, pMX-mP-mCherry based on the modified pMX-GFP retroviral vector (Cell Biolabs, USA), was used in the study. Minimal promoter and mCherry (miniP-mCherry) fragments were obtained from a modified pGL4.23 luciferase reporter vector (Promega, USA). Briefly, PGL4.23 was digested with XbaI/NcoI restriction enzymes to remove the original coding sequences of the *luc2* reporter gene and replace it with mCherry. The miniP-mCherry fragment was then amplified with primers containing NcoI or NotI restriction sites. Synthetic DNA fragments were cloned into NcoI and NotI digested pMX-GFP vectors. The resulting screening vector (see supplement file), pmx-mchery, was double digested with Hind III and xhoI to form a linearized vector that can be used for library generation.

2.2 Generation of random DNA libraries

A random DNA library was constructed following a two-step process (DNA fragment synthesis and seamless assembly cloning) (Supplementary Figure S1C). Random DNA fragments comprising different lengths of DNA fragments, homologous arms and Illumina adapters at both ends

(CTAACTGGCCGGTACCTGAGCTCGCTAGCCTCGAGTCGTCGGCAGCGT CAGATGTGTATAAGAGACAG-Randomsequence-CTGTCTCTTATACACATCT CCGAGCCCACGAGACAAGCTTAGACACTAGAGGGTATATAATGGAAGCT CG) were synthesized by GENEWIZ (Suzhou, China) and inserted into a reporter vector through homologous recombination methods following the Clone Express II One Step Cloning Kit protocol (Vazyme; Cat. No. C112). Sanger sequencing was conducted to check the quality of the library clones. We combined 80 homologous recombination reaction systems, purified them with 1.8 times AMPure XP Beads, transformed the purified product into electrocompetent cells (ThermoFisher Scientific; Cat. No. C6400) and transfered the transformed competent cells to 10L LB medium Amplify. When the OD value of the bacterial solution was 1.0, we extracted the plasmid according to the instructions of the kit (QIAGEN; Cat. No. 12362). These plasmid libraries were transfected into cells, and then cell genomic DNA was extracted. PCR amplification was performed with 100 ng genomic DNA as template (95°C 45s; followed by 15 cycles of 98°C for 15 s, 60°C for 30 s, 72°C for 3 min) and we combined 20 PCR reactions. The product was recovered by 2% agarose gel electrophoresis for 50 min (120v, 130mA) and sent for high-throughput sequencing. The data from the sequencing machine is the input library.

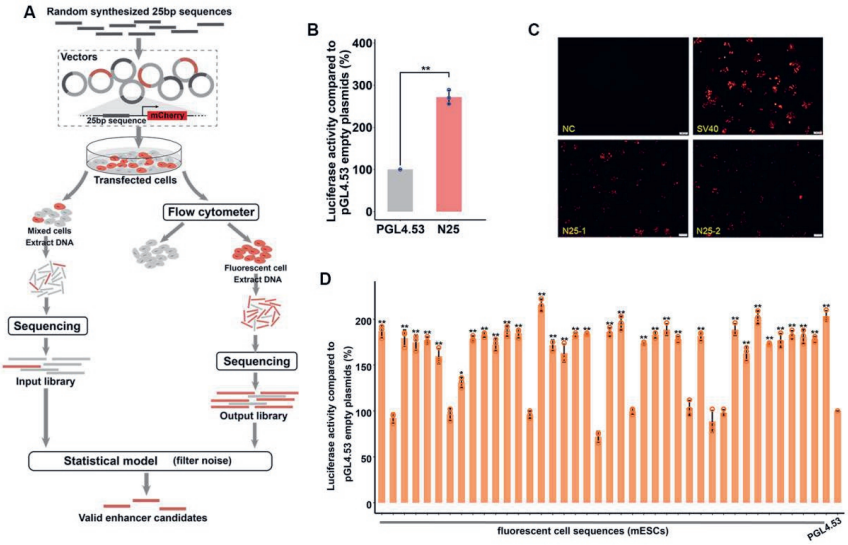


Figure 1. MAE-seq development. (A) The fragments to be screened were constructed into mCherry fluorescent reporter vector with mini promoter to form a plasmid library, and the cells after transfection of the plasmid library were divided into two groups, one group extracted all cell DNA specific amplification without fluorescence sorting and then high throughput sequencing (input), and one group extracted fluorescent cells by flow screening and then high throughput sequencing (output). The final enhancers are determined by joint analysis of input and output using the model we developed. (B) The dual luciferase activity of 25 bp DNA

*sequence was verified. The 25 bp sequence is selected from the reported long sequence enhancer. We selected 25 bp sequences of many different fragments, and only 25 bp with enhancer activity is displayed here to prove that the 25 bp DNA sequence has enhancer function. empty pGL4.53 plasmid was used as the control for baseline luciferase activity. The y axis represents the percentage of luciferase activity compared to pGL4.53 empty plasmids in the respective cells (n = 3 biological independent samples; bars show mean value ±s.e.m.; ns: non-significant, ** P < 0.01, calculated using t -test). (C) Intracellular activity validation of MAE-seq data. NC, empty reporter vector; SV40, SV40 enhancer; N25-1, N25-2, two randomly selected sequences of MAE-seq data. Constructed reporter vectors are transfected into mESCs to validate enhancer activity. NC is negative control and SV40 is a positive control. (D) The luciferase reporter system was used for activity validation of nine randomly selected MAE-seq loci in mESCs. empty pGL4.53 plasmid was used as the control for baseline luciferase activity. The y axis represents the percentage of luciferase activity compared to pGL4.53 empty plasmids in the respective cells (n = 3 biological independent samples; bars show mean value ±s.e.m.; ns: non-significant, ** P < 0.01, calculated using t -test).*

2.3 Cell culture and transfection

293T cell cultures were inoculated into 100 mM DMEM (ThermoFisher; Cat. No. 10566024) containing 10% FBS (Gibco; Cat. No.10437028) and 1% Penicillin Streptomycin P/S (Sigma; Cat. No. P0781) in a 37°C cell incubator. Transfection of plasmid libraries (1 ng DNA/1.2 × 10^7 cells) was performed using cells at 70–80% confluence using HD Transfection Reagent (FuGENE, Cat. No. E2311). Fluorescence cells were collected 48 hours after transfection.

E14 ESCs cultures were inoculated into a culture solution containing 100 mM ES medium (KO-DMEM (Gibco, 10829-018), 15%FBS (Gibco; Cat. No. 10099), 0.2 mM GlutaMAX™ Supplement (Gibco; Cat. No. 35050061), 0.1 mM MEM nonessential amino acids (NEAA) (Gibco; Cat. No. 11140–122), 50 U/ml penicillin/50 μg/mL streptomycin (Gibco; Cat. No. 15140-122), 1 U/ml LIF (Chemicon; Cat. No. ESG-1106), 0.1 mM 2-mercaptoethanol (BME) (Gibco; Cat. No. 21985023), 0.16 mM Vitamin C (Sigma; Cat. No. A4034), 1 μM PD0325901 and 3 μM CHIR-99021 (CT99021). Transfection of plasmid libraries (1μg DNA/1 × 10^6cells) was performed with cells at 70–80% confluence using Lipofectamine 3000 Transfection Reagent (Invitrogen, Cat. No. L3000075). Fluorescence cells were collected 48 h after transfection.

2.4 Generation of output libraries(one-step)

DNA was extracted (Vazyme; Cat. DC112) from the collected fluorescent cells. Components were added according to the TruePrep DNA Library Prep kit (Vazyme; Cat.TD202) and followed by one-step PCR (72°C for 3min, 98°C for 30s; followed by 5–15 cycles of 98°C for 15 s, 60°C for 30 s, 72°C for 3 min; 72°C for 5 min). The product was recovered by 2% agarose gel electrophoresis for

35 min (120v, 130mA) and sent for rubber recycling. The data obtained by high-throughput sequencing of the gum recovery products were output libraries. Sequence after one-step PCR amplification is 5′-AATGATACGGCGACCACCGAGATCTACACIIIIIIIITCGTCGGCAGCGTC AGATGTGTATAAGAGACAG-NNNNNN-CTGTCTCTTATACACATCTCCG AGCCCACGAGACIIIIIIIIATCTCGTATGCCGTCTTCTGCTTG-3′, IIIIIIII: Index 2 (i5), IIIIIIII: Index 1 (i7), -NNNNNN-: random sequence.

2.5 Luciferase assay

Sequences that required validation were sequenced (GENEWIZ, China). These fragments were then inserted in front of the PGK promoter of the luciferase plasmid pGL4.53 (Promega) using Gibson assembly. Cells were then co-transfected with the pRL-CMV Renilla reporter vector and the pGL4.53 vector with the silencer sequence inserted. The luciferase assay was performed using the Dual-Glo Luciferase Assay Kit from Promega according to the manufacturer's protocol. The original luciferase plasmid without any insertion was used as the control. All the luciferase assays were from three independent transfections performed on different days.

2.6 'LER' experimental assay

Five 25-bp loci of MAE-seq enhancers that overlapped peaks of ChIP-seq data were randomly selected for analysis. The activities of enhancers together with their 5′ and 3′ 200bp-flanking regions were validated, respectively. We termed MAE-seq enhancer, 5′ and 3′ flank sequences as 'E', 'L' and 'R' respectively. Six test groups were set up as follows: L, L-E, E, E-R, R, LR and L-E-R. These synthesized fragments (Sangon Biotech) were ligated into the PGL4.53 vector using the homologous recombination method. Their activity was detected by a luciferase reporter system.

2.7 Endogenous activation of enhancers and gene expression assay

The U6-BpiI-GFP vector was digested with BbSI, and the target fragment was recovered using a gel recovery kit (D4008, ZYMO). The first base of the gRNA sequence design starts with G, and the top three are selected for use. The rules for adding cohesive end bases in gRNA synthesis are as follows. F: CACC-gRNA, R: AAAC-gRNA, where gRNA represents the sequence given by the software during design. Dissolve gRNA primer with ddH2O at a concentration of 100uM, each take 1 μl, 6.5 μl ddH_2O, 1 μl 10 × 4 ligation buffer (NEB), 0.5 μl T4 PNK (NEB M0201S) to configure a 10 μl annealing reaction system and perform annealing

and phosphoric acid in a PCR machine change. Procedure: 37°C, 30 min, 95°C 5 min, gradient cooling to 25°C, 5°C/min. 1:200 dilution of the annealing product. Take 6 μl annealing product, add 1 μl linearized plasmid, 1 μl 10×4 ligation buffer, 1 μl T4 DNA ligase, 1μl ddH_2O, configure it into a 10 μl ligation reaction system, and ligate at room temperature for 4–6 h. We then proceeded to transformation, AMP plate coating, and clone sequencing with sequencing primer hU6-F: GAGGGCCTATTTCCCATGATT. The dSpCas9-VP64, U6-BpiI-gRNA-GFP were transfected into the cells at a ratio of 1:1, and RNA was collected at 24, 48 and 72 h to quantitatively detect the expression of target genes. If the transfection efficiency was low, the flow was enriched. The control group was transfected with dSpCas9-VP64 and U6-BpiI-GFP empty vectors. The mRNA was reverse-transcribed into cDNA using the reverse transcription kit (HiScript III RT SuperMix for qPCR, R323-01, Vazyme). The qPCR kit (AQ101, TransGen) was used for real-time quantitative PCR to detect the changes of target genes.

2.8 Establishment of an enhancer knockout cell

The ssDNA designing followed the following principles: (i) the cleavage site of Cas9 protein in the genome is taken as the center; (ii) upstream and downstream sequences of 90 bp length are intercepted as the left and right homologous arms of ssDNA respectively; (iii) enhancer is deleted in downstream sequence; (iv) the length of designed ssDNA is 180 bp; (v) the NGG of PAM sequence is mutated into NCG to prevent the repeated cleavage of Cas9; (vi) Hind III endonuclease site is designed to facilitate genotyping of knockout cells. Sequences of ssDNA related to Ce1 and Ce3 are as follows:

>ssDNA_Ce1

TTTAATGGCTGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
CTCTCTCTCTTTCAAAGCCACTTGGCCTCTTTATTAAGCTTCACGTGTCT
TGAGACTTGATTCCTGTCTAGATCCCTCTTATTCCAGAAATGACAGGAT
TTCCATGTACCTGCATCAATGCATCTTAGC

>ssDNA_Ce2

CCCACCAGTCCTGGGGCTCTGGTATTTATACTTTCTCCAGAATCCCCAG
AACTAAACCATCTGCAGCTGACAAAGATCATGCCCCGTCCTAAAGCTTT
GAGACAGTCACAATTAACGGCTGTAGACAAACTGAAGCAGTCCCATAT
ATATATTCATATAACATAACCGAAATCAAAATTC

>ssDNA_Ce3

GGAGTGTGGGTAATGAGTGACTGGATCAGTGTCGAAGGGAACAGTTTC
CGCATCTTTACTTATGGAAGGGTCTTAAACAAGCTTACTGTGTAGCCGT
TTGCACAGCCAGATTTAATTACTACTGCCTGTCTCCTCACGGGTGTTTA
AAATAAACTTGCATATTCTGTTACAGTTTAAGAA

Then the enhancer knockout cells were genotyped. The genomic DNA of enhancer knockout cells was used as a PCR template. The primer sequences are as follows:

Ce1_F: CTCAGACTCAGAGATCCCTC, Ce1_R: CTCTGCCTCTTGTAGGGAAG; Ce2_F: TCCATCAGATTGGCCTGTAG, Ce2_R: AACTGCTCTGAGCACCAGAC; Ce3_F: GTCAGTGCTCTTATCTGCTG, Ce3_R: GCAGAGCTGTCTACATAGTG.

The lengths of PCR products related to Ce1 to Ce3 are 888, 1029 and 1022 bp, respectively. They were digested by Hind III for genotyping detection. If the product is wild type, there is no Hind III restriction site in these sequences, which could not be cleaved by the enzyme. This was used as the criterion to determine the genotype of knockout cells.

2.9 Cell counting kit-8 assay

Cell proliferation was determined using the cell counting Kit-8 (C0037, Beyotime). Briefly, 3×10^4 cells were seeded in a well of a 24-well flat-bottomed plate and incubated for 12 h. We then replaced 1 ml of fresh medium, added 100μl CCK8 and incubated for 2 h in the incubator. The absorbance at 450 nm was measured using a microplate reader (Bio-Rad, Hercules, CA, USA). The experiments were repeated at least three times.

2.10 5-Ethynyl-2-deoxyuridine (EdU) assay

About 1.6×10^5 cells per well were inoculated in a 6-well plate. EDU (ST067, Beyotime) was added after 12 h of culture to make the final concentration 10 uM and the cells were incubated for 2 h in an incubator. After removing the medium, 1ml fixative (4% Paraformaldehyde) was added, and cells were fixed at room temperature for 15 min. After removal of fixative, 1ml permeabilization solution was added and cells were incubated at room temperature for 15 min. Later 0.5ml Click reaction solutions were added, and cells were incubated for 30 min in the dark. Finally, cells were stained with Hoechst. Results of the EdU assay were observed and analyzed under an EVOS M5000 microscope (Thermo Fisher Scientific).

2.11 Cas13d-mediated RNA interference

The Cas13d system was harnessed as a versatile tool for targeted gene interference. In a concise overview, the Cas13d plasmid was subjected to enzymatic digestion employing restriction endonuclease BbsI. Subsequently, primers designed to target the specific gene of interest were meticulously ligated onto the plasmid vector using the high-fidelity T4 DNA ligase. Following plasmid construction, cellular transfection was executed, and the ensuing impact on the expression of the target protein was assessed following a 48-hour interval.

2.12 RNA-seq library construction

RNA was isolated using Trizol (Invitrogen) and Kit (Mei5Bio,Cat No.MF167) from mESCs. Three libraries (biological duplicates) were constructed by the VAHTSTM Total RNA-seq (H/M/R) library prep kit (Vazyme Biotech, China). The library construction method was referred to in the instructions for the library prep kit and 0.5 μg total RNA was diluted with nuclease-free water in a nuclease-free PCR tube. Then the rRNA probe (H/M/R) was hybridized with a temperature gradient descending method (95–22°C, 0.1°C/sec). rRNA and genome DNA was digested with RNase H and DNase I successively (37°C for 30 min). The ribosomal-depleted RNA was purified with VAHTS RNA Clean Beads (in 2.2×) in a magnetic stand. Then the RNA sample was eluted with Frag/Primer buffer for fragmentation (For 150–200 bp insert: incubate at 94°C for 8 min). Double Strand cDNA was synthesized with 1st and 2nd strand enzymes and purified with VAHTS DNA Clean Beads (in 1.8x). After dA-tailing and RNA adapter ligation, the sample was purified, and size was selected with two steps of DNA Clean Beads (in 1x) incubation. PCR primer mix and amplification mix were used for library amplification. The quality of purified PCR product was determined by using an Agilent Technologies 2100 Bioanalyzer. Multiplexed massively parallel sequencing (paired ends of 150 bp) was conducted using the Illumina platform.

2.13 Mapping simulation

To figure out the optimal length of candidate DNA fragments to increase the library saturation and alignment accuracy, we used BEDTools (27) to randomly extract sequences with different lengths (20–60 bp, 1bp interval) from the human reference genome and align them to hg19 by bowtie (version 1.2.2) (28), bowtie2 (version 2.1.0) and bwa (version 0.7.5a-r405) (29), respectively. The unique mapping rate was counted for further analysis. In this study, each group had three repetitions, 1 million sequences with a specific length in each repetition.

2.14 Alignment of input and output libraries

MAE-seq data have been uploaded to the GEO database (GSE149028). The input library and output libraries raw data (raw reads) of fastq format were firstly trimmed by cutadapt (v1.16) (30) to remove adapter. Then, the remaining reads were aligned to the mm9/hg19 reference genomes using bwa (version 0.7.5a-r405). Uniquely mapped reads with 25 bp inserts and no more than two mismatches from either bwa (29) were kept. The genomic start and end coordinates of these reads were then used for downstream processing.

2.15 Normalization

The enrichments of input and output datasets are not uniform. The observation of fragments will be affected by library size and PCR cycles. Theoretically, the larger the library is, the greater the probability of observed frequency holds. Meanwhile, the 25 bp-fragments mapped to the genome are faultlessly aligned without transplacements. Due to this characteristic, the sequencing data cannot be removed from PCR duplicates, and the number of cycles will also influence the quantity. Therefore, we designed a quantitative function to normalize the fragment counts and compute the normalization coefficient:

$$f(x) = \frac{x}{N \times \log(1+p)^c}\omega$$

The observation is at a given location where the x is reads observation at a given location. N is the total number of reads. p is the PCR amplification efficiency (default = 1). c is the number of PCR cycles. ω is the zoom factor.

It is known that the smaller value of parameter in Poisson test will bring a lower significance, so we set the zoom factor as the compensation to fit the test.

The zoom factor ω is calculated by the following formula:

$$\omega = 10\ ([\log10\ (MIN\ (x, y))] + 1)$$

Where the MIN function is the minor value of the total input and output read counts, and x,y are the total read counts of input and output libraries, respectively.

2.16 Significant difference analysis and positive signal (MAE-seq enhancer) selection

To achieve significant difference analysis between input and output data, we first determined the data distribution of MAE-seq data. Obviously, the probability of a random 25 bp fragment that can be matched to a location in the reference genome is infinitesimal. Meanwhile, the number of repetitions is very large, and events are independent. Therefore, the MAE-seq data follows a Poisson distribution, and the probability function is:

$$P(x = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\lambda = n \times p$$

where k is represents the number of observations, n is the total number of reads and P is the probability of the observation.

Then, we applied the Poisson test to calculate the significant difference and used the Benjamini and Hochberg test to select the loci with Q values that are lower than 0.05 as effective signals. To avoid incorrect filtering, we set a compensation value for the loci appearing in the output dataset but not in the input dataset. This part was processed by R(4.0.2).

The compensation value is also playing a key role in the Poisson test. The higher the value, the more rigorous the result, and vice versa. There are two factors that determine the validity of this compensation value: (i) the non-overlap rate between MAE-seq signals and annotated regions; (ii) the remaining valid signals. To determine the value, we set 0.1–3 times the input data normalization coefficient as the compensation value. When the compensation value is equal to 1 normalization coefficient, the result takes into account both these factors (Supplementary Figure S1G).

2.17 Annotating enhancers in the genome

The coordinates of enhancers in the genome were annotated by HOMER (31). To annotate the enhancer location in terms of important genomic features, the genome was annotated in the following categories: TTS (Transcription Termination Site by default defined from −100 bp to +1 kb), intergenic, exon, intron, 5′UTR and 3′UTR. Each enhancer was sorted into a different category if its locus fell within the sequence of a specific category.

2.18 Endogenous chromatin environment of enhancer candidates

To study the endogenous chromatin environment of enhancer candidates, data from DNase-seq (ESC), ATAC-seq (293T), ChIP-seq for histones (H3K4me1 and H3K27ac of ESC and 293T) and transcription factors (Sox2, Nanog and Oct4 of ESC) were downloaded from the ENCODE or GEO database (Supplementary Table S2). Since most of the features, such as histone modifications, were found at nucleosomes near enhancer elements but not the exact location of the enhancer elements, we extended the region of peak by ±2500 bp. The proportion of overlap between the genome position of enhancers and the extended peaks was calculated by BEDTools (27). In addition, the enrichment of H3K4me1 around the MAE-seq enhancers was analyzed and shown by deepTools (32). The overlap between MAE-seq enhancers and H3K27ac, Nanog, Oct4, Sox2, DNase-seq were calculated by BEDTools and visualized by R.

2.19 RNA-seq data analysis

First, the quality control was conducted on the RNA-seq data by using FatQC (v0.11.5) (33). Then, to ensure the quality of RNA-seq data used in further analysis, rRNA sequences were first filtered out by SortMeRNA v2.0 (default parameter). RNA-seq data associated with this study are derived from the GEO database (GSE115750). Clean reads were obtained from the raw data by removing adapter-containing reads, poly-N-containing reads, and low-quality reads using Trimmomatic (version 0.36) (32). All downstream analyses were based on high quality, clean data. Clean data was mapped to the mm9 reference genome using TopHat v2.0.12 (34). The parameters –read-mismatches and –library-type were set to 5 and fr-firststrand, respectively. After that, the expression level FPKM (fragments per kilobase per million mapped fragments) of each gene was measured by Cufflinks (2.2.1) (http://cole-trapnell-lab.github.io/cufflinks/). In this study, the average FPKM was used to represent the expression levels of enhancer target genes.

2.20 Chromatin accessibility of enhancers

The chromatin accessibility of MAE-seq enhancers was determined by DNase-seq and ATAC-seq data. The MAE-seq enhancers that overlapped with peaks from DNase-seq and ATAC-seq were defined as in open chromatin regions, otherwise, they were considered as being in closed chromatin regions. The MAE-seq enhancers in open or closed chromatin regions were classified by BEDTools. Each enhancer was assigned to the one target gene by using the Activity-by-Contact (ABC) model (35). The expression levels (FPKM) of target genes, which were obtained from RNA-seq data, in open and closed chromatin regions were assigned by HOMER and visualized by R (version 4.0.2).

2.21 Escape enhancers detection

To verify if MAE-seq can detect the 'escape enhancers', the Hi-C data (GSE125656) and ChIP-seq data of histone marker H3K9me3 were used. The MAE-seq enhancers overlapping with H3K9me3 and that intersect with house-keeping genes were defined as 'escape enhancer' according to the report before. The mouse house-keeping genes were downloaded from https://housekeeping.unicamp.br/ and liftOver was used to transform mm10 coordinates to mm9. The overlap between MAE-seq enhancers, genes and loops were detected using BEDTools.

2.22 Defining the category of enhancers

Enhancer candidates were divided into two categories, according to the relationship between MAE-seq loci and known consensus enhancer regions (which included the EnhancerAtlas 2.0 dataset, DNase-seq, ATAC-seq and ChIP-seq of H3K4me1 and H3K27ac). Specifically, the enhancers whose coordinates overlap with EnhancerAtlas 2.0 enhancer regions, DNase-seq, ATAC-seq or ChIP-seq of H3K4me1 and H3K27ac peaks (±2500 bp) were represented as the known enhancers. Otherwise, if the genome positions of enhancers have never been reported in the EnhancerAtlas 2.0 dataset, DNase-seq, ATAC-seq or by ChIP-seq of H3K4me1 and H3K27ac, these loci were assigned to novel enhancers. Additionally, the distance to TSS for these two kinds of enhancers was computed and visualized.

2.23 Functional enrichment analysis of predicted target genes regulated by enhancers

Gene Ontology (GO) analyses of known and novel MAE-seq enhancers were performed by the clusterprofiler package (36) in R with the threshold of FDR < 0.05.

2.24 Analysis of the chromatin interaction related to novel enhancers

The significant chromatin interactions were identified from mESCs Hi-C (data associated with this study are derived from GEO database (GSE125656)) and by HOMER (31) and HiCCUP (32) respectively. The overlap between MAE-seq loci and chromatin interaction anchors was analyzed by BEDtools (27).

2.25 Statistical analysis

We used R (version 4.0.2) (37) and SPSS 19.0 (IBM, Chicago, IL, USA) for all statistical analysis. All results were expressed as the mean ± standard deviation (SD) of the number of experiments.

2.26 Overlapping enhancer and transcription factor motif regions

ChIP peak data for H3K27ac and H3K4me1 were sourced from ENCODE (accessions: ENCFF194TQD and ENCFF158GBZ), with peak coordinates mapped to mm9 using LiftOver. Employing Bedtools intersect, enhancer-ChIP peak overlap was determined, requiring at least 30% fraction overlap. Enhancers overlapping H3K27ac or H3K4me1, alongside randomly selected enhancers, were chosen for downstream analysis. FIMO, utilizing the JASPAR2020_CORE_vertebrates_non-redundant_pfms_meme database, scanned the mm9 genome for potential transcription factor binding motifs. Enhancers were subsequently overlapped with these motif regions, demanding at least 90% fraction overlap. Enumerated and sorted transcription factors associated with enhancer overlap were analyzed. Open chromosomal regions (ATAC peaks, GEO: GSE120376) were integrated using the same overlap methodology.

3 Results

3.1 MAE-Seq development

We developed MAE-seq based on random DNA libraries and fluorescence-assisted cell sorting (Figure 1A). To determine the optimal length of the DNA fragments and improve alignment accuracy, we conducted a bioinformatics simulation experiment. Random sequences of different lengths (20–60 bp, 1 bp interval) were extracted from the human genome (hg19). The result showed that, sequences longer than 25 bp had a unique mapping rate of approximately 80%, reaching a plateau stage, except for bowtie2 (Supplementary Figure S1B). Meanwhile, to verify whether the 25 bp DNA length has enhancer function, we intercepted 25 bp from the previously reported enhancer sequence for dual luciferase reporter assay. The results showed that the 25 bp sequence had enhancer activity. This result proves that the 25 bp DNA sequence can play an enhancer role as an independent functional unit. (Figure 1B). Thus, our results demonstrate that 25 bp fragments have advantages in both alignment accuracy and sensitivity, making them suitable for MAE-seq system development. 25 bp fragments are used to construct random DNA libraries and subsequently screen active enhancers.

In brief, 25 bp DNA random fragments are constructed into a 'classic' MPRA plasmid vector (38) with a fluorescent reporter (Supplementary Figure S1A) and then transfected into cultured cells (70%-80% confluent). If the insert fragment has enhancer activity, then it will interact with the mini promoter to make the downstream mCherry express red fluorescence. In the MAE-seq, the plasmid transfected cells are divided into two groups. One is used for flow fluorescence sorting. The enriched fluorescence cells are used for DNA extraction and later specific amplification. These fragments are applied for high-throughput sequencing (termed 'output library', Supplementary Figure S1C). These sequenced 25 bp fragments are considered to contain a high-ratio positive signal. To mitigate potential background noise and control for factors like multiple plasmids entering a single cell, the input library is constructed directly from the other group of transfected cells, without flow fluorescence sorting and sequenced. This is considered a control library. Similar to the ChIP-seq analysis (12), we design a model (see methods) based on Poisson distribution with input and output datasets to capture the true enhancer signal. For example, in mESCs, 5.5 million positive candidates were obtained from the output library. After model filtering, 626,879 valid enhancer candidates are identified (Supplementary Table S5). To evaluate the model's efficiency, we use enhancer histone markers (H3K4me1) to determine the result. This shows that although the number of positive candidates drops sharply (from 5,521,717 to 626,879, down ~89%) after filtering, the overlap with H3K4me1 only decreased by 13% (from 123,376 to 107,271). This indicates the filtering model works efficiently, and the identified enhancer candidates will be used for the further analyses.

Using MAE-seq, we identified a total of 626,879 and 554,826 (Supplementary Table S6) 25 bp enhancers in mESCs and HEK 293T. To validate their enhancer activity, 40 MAE-seq sites from mESCs and HEK 293T were selected for dual luciferase assay (Supplementary Table S1). Remarkably, at least 80% of the MAE-seq sites showed enhancer activity. (Figure 1D, Supplementary Figure S1E). By constructing these loci into reporter vectors containing mCherry fluorescent genes and transfecting cells, fluorescence expression could be observed under the microscope. (Figure 1C, Supplementary Figure S1D,G). Further, we selected 20 sequences in mESCs that were from non-fluorescent cell populations, and only 1 sequence was shown to have enhancer activity by the dual-luciferase reporter system (Supplementary Figure S1F), demonstrating the effectiveness and reliability of the MAE-seq data.

3.2 Epigenetic characteristics and chromatin accessibility

To comprehensively annotate and analyze our MAE-seq datasets, we realized there were more diverse and experimental datasets generated in mESCs compared to

HEK 293T cells, making them more suitable for annotation and interpretation purposes (Epigenetic characteristics analysis of HEK 293T cells is shown in Supplementary Figure S6). Therefore, our following analysis focused on interpreting mESC datasets. In mESCs, we identified 626,879 enhancers through MAE-seq. We analyzed the genomic distribution of MAE-seq enhancers. The results show that over 90% of enhancers are located in intergenic or intron regions, which is consistent with predictions based on chromatin features (39, 40) (Figure 2A).

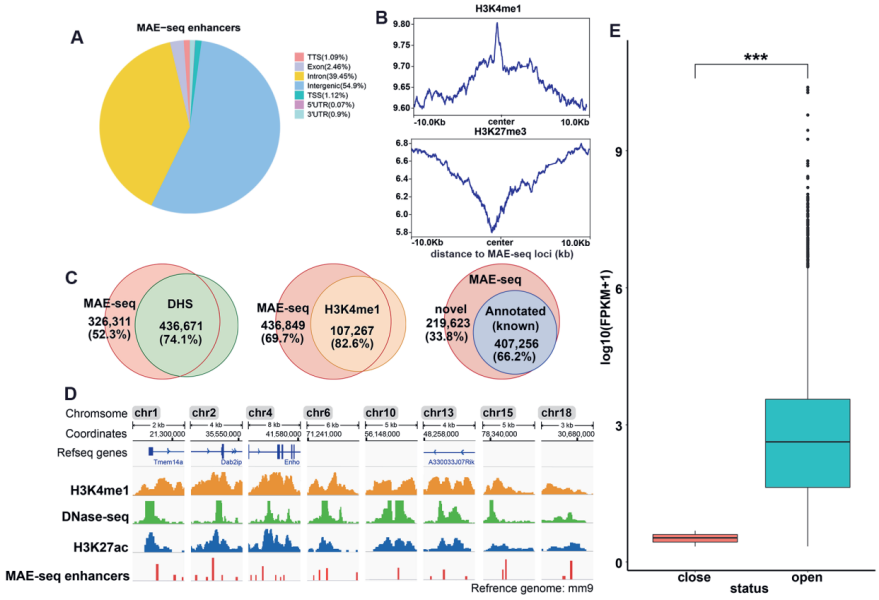


*Figure 2. Genomic annotation of MAE-seq enhancers. (A) Genomic distribution of MAE-seq enhancers. In mESCs, 626879 million loci have been identified by MAE-seq. Most of them are in non-coding regions such as intergenic and intron. (B) The enrichment of H3K4me1 and H3K27me3 marks around the MAE-seq loci. (C) The percentage of enhancer epigenetic markers(H3K4me1), DNase-seq and EnhancerAtlas 2.0 data base covered by MAE-seq sites. (D) Distribution of 25bp MAE-seq enhancers in histone modification regions(H3K27ac, H3K4me1, DNase-seq) of 8 chromosomes. (E) Expression of nearest genes related to MAE-seq enhancers which are in open and closed chromatin regions. Open, represent DHS; closed, represent none of DHS. *** P< 0.001.*

To further validate their genomic features, we performed a joint analysis with both known enhancer and repressed epigenetic data. The result showed that H3K4me1 (41) was significantly enriched at these loci, while the signals of the repressed chromatin mark, H3K27me3, vanished at these loci in mESCs (Figure 2B). To annotate these loci, we performed an overlap analysis with different active

genomic datasets. First, with the H3K4me1 enhancer marks, we found that 82.6% of them had MAE-seq signals (Figure 2C). It is consistent with the enhancer feature of the MAE-seq loci. Next, we overlapped them with the active chromatin datasets, the DNase I hypersensitive sites (DHSs), and we found that 74.1% of the DNase-seq sites can be annotated by MAE-seq loci (Figure 2C). This indicated MAE-seq loci mainly located at the open chromatin region (42). Lastly, we compared them with the Mouse Enhancer Atlas 2.0 database (43), which overlaps with MAE-seq loci. The result indicated that about 66.2% of the MAE-seq loci could be annotated (Figure 2C). Together, these confirm that the MAE-seq loci are enriched with enhancer and active chromatin signals and demonstrate the informativeness and reliability of our data. In addition, the ChIP-seq binding sites of various ESC-specific transcription factors (TFs), such as Sox2, Nanog and Oct4, had also been used to do the annotation, and the results are consistent with active genomic datasets (Supplementary Figure S2A). Again, these genomic datasets suggest our MAE-seq sites are informative and enriched with enhancer or active chromatin signals (Figure 2D). These further confirm the precision of the strategy.

Previous research on transcription networks and REs primarily focused on open chromatin regions of the genome. To address this, we analyzed the distribution of MAE-seq enhancers in both open and closed chromatin regions in mESCs. As expected, we found 74.27% of known MAE-seq enhancers presenting in open chromatin regions, which is 2.9-fold the percentage of enhancers in the closed chromatin regions (Supplementary Figure S2B, Supplementary Table S2). Next, following Arnold et al.'s study (44), we divided MAE-seq enhancers into open enhancers and closed enhancers according to whether they appear in open chromatin regions or not and assess the expression levels of their nearest genes. The results revealed that the target gene expression levels of closed enhancers are significantly lower (5.23-fold) than those of open ones (Figure 2E). These findings imply that these closed enhancers have enhancer sequence potential but may be silenced by the local endogenous chromatin environment.

In conjunction with the Hi-C data, we found some enhancers in heterochromatin (H3K9me2 and H3K9me3 markers) that interact with mouse housekeeping genes, including *Rrn3*, *Atf7ip*, *Mta3*. At the sites where we interact with these genes, luciferase experiments have shown that they have high enhancer activity (the average activity intensity is more than 400% higher in the control group) (Supplementary Figure S2C, Supplementary Table S2). The properties of these enhancers are similar to those of the 'escape enhancers' previously reported in the literature (45). This suggests that not all the enhancers would be silenced in their endogenous contexts. Still, quite a few enhancers retained their activity in closed chromatin regions. Their activities may be independent of the environment of endogenous chromatin. Additionally, we explored the relationship between

enhancer activity intensity and histone markers (DNase, H3K27ac and H3K4me1). Through luciferase experiments, we found that the luciferase intensity with enhancer histone mark sites was not significantly higher than that without histone mark sites, indicating that these chromatin signatures are not a good predictor of enhancer activity strength (Supplementary Figure S2D, Supplementary Table S2).

3.3 Determining the regulatory function of 25 bp fragments

The span of REs identified by MAE-seq is much shorter compared to those defined by ChIP-seq (H3K4me1, H3K27ac, Oct4 and Nanog) and DHS-seq (25 bp versus ~600 bp) (Figure 3A). To determine whether such a short sequence can function efficiently as the previously defined REs, we randomly selected five ChIP-seq or DNase-seq binding loci that overlap with the MAE-seq sequences (Supplementary Table S3). Firstly, we divided each binding locus into three fragments, referred to as 'L', 'E' and 'R'. The fragment, which overlapped with a fragment from MAE-seq data is referred to as 'E', the fragment to its left is referred to as 'L', and the fragment to its right is 'R'. We then synthesized these sequences using different combinations. As a result, for locus 1, when 'L' is missing, the remaining fragments (E&R) have no significant change in activity change. When the 'R' is missing, the remaining fragments (L&E) have no significant change in activity. As expected, when the 'E' is missing, the activity of all remaining combinations (L, R and L + R) shows a sharp decrease or disappears in the experiments (Figure 3B, C). We have performed 3 biological replicates for each of these experiments. Out of those 5 loci, 4 have shown similar results. This implies that these regulatory activities are heavily dependent on these 25 bp MAE-seq defined fragments. Similar results were also found in HEK293T (Supplementary Figure S3A,B). Together, these results suggest that ChIP-seq/DHS-seq defined loci have veiled sequence regions and need further annotation.

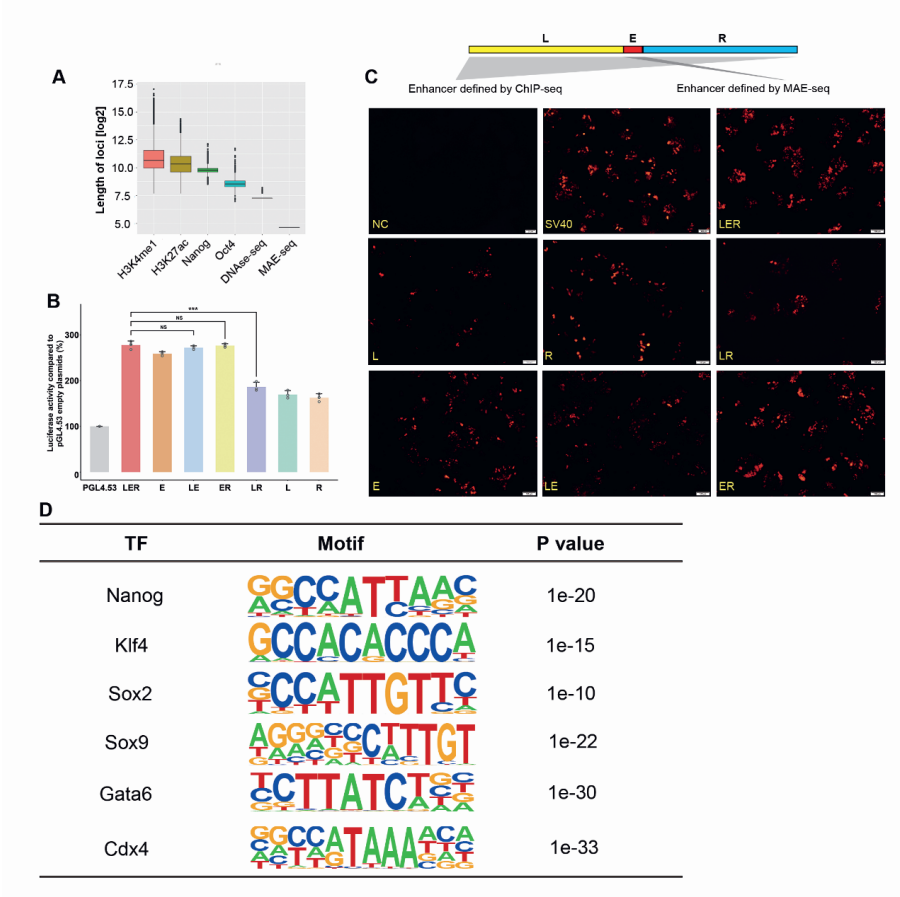


*Figure 3. Functional validation of MAE-seq enhancers. (A) The average length of enhancers annotated by epigenetic modification or transcription factor binding compared with the length of the MAE-seq enhancer. Data of sequence length is converted to log2. (B) Differential analysis of enhancer activity of different fragments in 'LER' experiment by luciferase. L, left flanking sequence; R, right flanking sequence; E, 25bp enhancer. empty pGL4.53 plasmid was used as the control for baseline luciferase activity. The y axis represents the percentage of luciferase activity compared to pGL4.53 empty plasmids in the respective cells (n = 3 biological independent samples; bars show mean value ± s.e.m.; ns: non-significant, *** P< 0.001, calculated using t-test). (C) mCherry fluorescence expression profile of different fragments in 'LER' experiments after transfection of mESCs. Different fragments were constructed into the reporter vector with mCher gene and transfected into cells, and the fluorescence expression was observed after 48 h. NC, empty reporter vector; SV40, SV40 enhancer; L, left flanking sequence; R, right flanking sequence; E, 25 bp enhancer. (D) The mESCs specific transcription factor motifs identified from mESCs MAE-seq data.*

Enhancers typically require binding TFs to perform their function (46). To further verify whether the MAE-seq data can enrich transcription factors with these 25 bp fragments, we used the findMotifsGenome.pl program in the software HOMER (31) to perform motif analysis with these 25-bp sequences, and the set parameter is -size (the default parameter is not set). As expected, these fragments could significantly enrich embryonic stem cell specific transcription factor binding sites, such as Sox2, Klf4 and Nanog (P value < 0.001) (Figure 3D). In summary, these imply that these 25 bp fragments could be used for motif analysis. Together, the defined 25-bp fragments could perform the similar regulatory functions as the ~270 bp chromatin-based loci and could also be used for bioinformatics analysis.

3.4 Identification and verification of novel candidates

EnhancerAtlas 2.0 is a comprehensive database that contains an atlas dataset from 241 cells, and tissue types in mice (47). Notably, 35.03% of MAE-seq enhancer loci have never been detected by DNase-seq or ChIP-seq (H3K27ac, H3K4me1, ESC-specific TFs) in any mouse cell type or tissue. These 35.03% MAE-seq sites as novel enhancers. To verify whether these novel sequences have enhancer activity, we randomly selected 10 novel enhancers and constructed them in a Luciferase reporter vector (Supplementary Table S1). The result showed that 80% of sites have enhancer activity (Figure 4B). They were constructed into a vector containing the mini promoter of the mCherry gene, and we could observe the expression of the fluorescent gene 48 h after transfection cells (Figure 4A). To further verify the effect of these elements on gene expression, we randomly selected six novel enhancers, using the dcas9-SunTag-VP64 system (48) to precisely activate these sites in the endogenous environment, and then the changes of distal regulatory genes expression were detected by QPCR combined with Hi-C data (Supplementary Table S4). The expression levels of the five distal regulatory genes were significantly upregulated (Figure 4C). We also examined potential off-target sites at each location, and the sequencing result showed that there are no mutations at these sites. These results are consistent with the previous finding that genes can be regulated by distant enhancers (49). Also, it implies that these novel elements do play critical roles in gene transcription regulation in living cells via distal associations. Therefore, this implies that MAE-seq is able to capture the omitted regulatory genomic information that is not elucidated by previous methods.

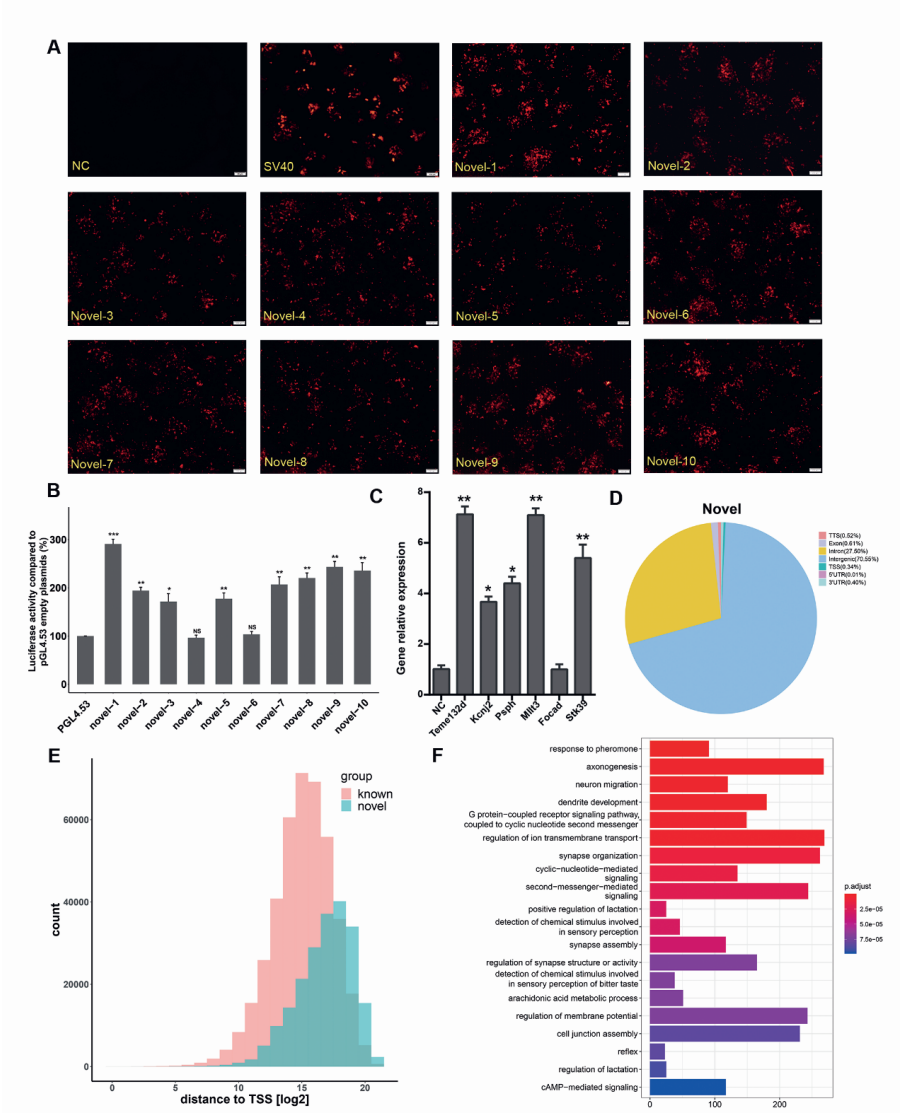


Figure 4. Genomic annotation and Functional validation of novel enhancers discovered by MAE-seq. (A) mCherry fluorescence expression of novel enhancers. Mouse enhancers annotated in EnhancerAtlas 2.0 database are used for classification of MAE-seq enhancers. 35.3% of MAE-seq enhancers that are not covered by the database are classified as novel enhancers. Ten of them were randomly collected and constructed into a reporter vector with mChery gene, and the fluorescent expression was observed 48 h after transfection of cells. NC, empty reporter vector; SV40, SV40 enhancer; Novel-1 to Novel-10, ten novel enhancers. (B) The enhancer activity of the 10 novel enhancers in (A) was quantified using a dual luciferase reporter system. empty pGL4.53 plasmid was used as the control for baseline luciferase activity. The y axis represents the percentage of luciferase activity compared to pGL4.53 empty plasmids in the

*respective cells (n = 3 biological independent samples; bars show mean value ± s.e.m.; ns: non-significant,**P< 0.01, ***P< 0.001, calculated using t-test). (C) The regulation effect of novel enhancers on the transcription level of the target genes. Six novel sites are selected for endogenous activation and the changes in the transcription level of their target genes are detected by QPCR (* P< 0.05; ** P< 0.01, calculated using t-test). (D) Genomic distribution of novel enhancers. (E) Statistics of distance to TSS of novel enhancers. Red column represents known mouse enhancers; the blue column represents novel enhancers. (F) GO enrichment analysis of nearest genes related to novel enhancers.*

To further characterize these novel elements, we analyzed their genomic distribution pattern. Similar to the known markers, the majority (70.55%) of these novel candidates are distributed to distal intergenic regions, followed by introns (Figure 4D). Next, we analyzed the distance to the nearest gene TSS of novel enhancers. Different from known ones, more novel enhancers prefer to be found in the intergenic region, far away from the TSS (Figure 4E). This suggests that these novel enhancers enact remote regulation of the target gene's expression. We tested this speculation by analyzing novel enhancers with Hi-C and Micro-C datasets (50). The results show that 122,199 novel enhancers may be involved in the formation of the chromatin loop, accounting for 55.40% of the total number of novel enhancers (Supplementary Figure S4A). Additionally, the average span of these interactions reaches 126 kb, implying that the distal interaction with promoters mediates the function of novel enhancers. To explore the roles of these types of enhancers in the biological process, GO analysis was conducted. The result shows that most of the nearest genes related to known enhancers play roles in the transport of ions and proteins across membranes, such as calcium ion homeostasis, localization, and transport of proteins (Supplementary Figure S4B). In contrast, genes linked to novel enhancers were significantly enriched in biological processes related to the neural correlates, such as axonogenesis and synapse organization (Figure 4F).

In a recent study, it was demonstrated that the majority of noncoding regions in the human genome lack annotated elements and lack epigenomic or protein binding signals. Interestingly, they selectively knocked out a region called hub_22_7, which significantly affected the cell phenotype (7). Among the sites they verified, 444 (46.25% of 960) coincided with the MAE-seq sites of human cells. In order to verify the function of these sites, we selected nine sites for luciferase verification in HEK 293T cells. The results showed that these sites showed enhancer activity (Supplementary Figure S4C). Further experiments showed that activation of these sites could significantly increase the expression level of their target genes (Supplementary Figure S4D). This discovery further emphasizes the ability of MAE-seq to define novel enhancers from a different perspective. It suggests that the new tools, such as our MAE-seq, are increasingly required for precise and functional annotation of regulatory elements throughout the genome.

3.5 Multiple refined elements hierarchically regulate *Cdh1* expression

To evaluate the potential impact of MAE-Seq enhancers on cell phenotype, we focused on the gene of interest, *Cdh1*, for further validation. Direct interactions between three MAE-seq enhancers (*Cdh1* enhancer1(Ce1), *Cdh1* enhancer2(Ce2) and *Cdh1* enhancer3(Ce3)) (Supplementary Table S4) and the promoter of the *Cdh1* gene were found in Hi-C datasets (Figure 5A). Among them, Ce1 and Ce3 are known enhancers, and Ce2 is a novel enhancer. All these interaction loci have histone mark activities (Figure 5A). To explore the biological function, CRISPR/Cas9 system (Supplementary Figure S5A) was employed to assess their requirement for gene expression in genomic loss-of-function models. We successfully generated homozygous (-/-) knockout mESC lines for single knockout, double knockout, and triple knockout. To investigate whether these three enhancers can affect cell phenotype individually, they are deleted individually in mESCs (Supplementary Figure S5B). We found that when knocking out any single one of them, the transcription level of the *Cdh*1 gene did not change significantly (Figure 5B). To amplify the effect, double knockouts of different enhancer combinations (Ce1&Ce2, Ce2&Ce3 and Ce1&Ce3) were conducted. Our result shows that only Ce2&Ce3 double knockout cells slightly downregulate the expression of the *Cdh1* gene (Figure 5B). Notably, when we knocked out all three enhancers, the transcription level of Cdh1 decreased significantly (Figure 5B). These results imply that proper expression needs coordination, and it may be hierarchically regulated.

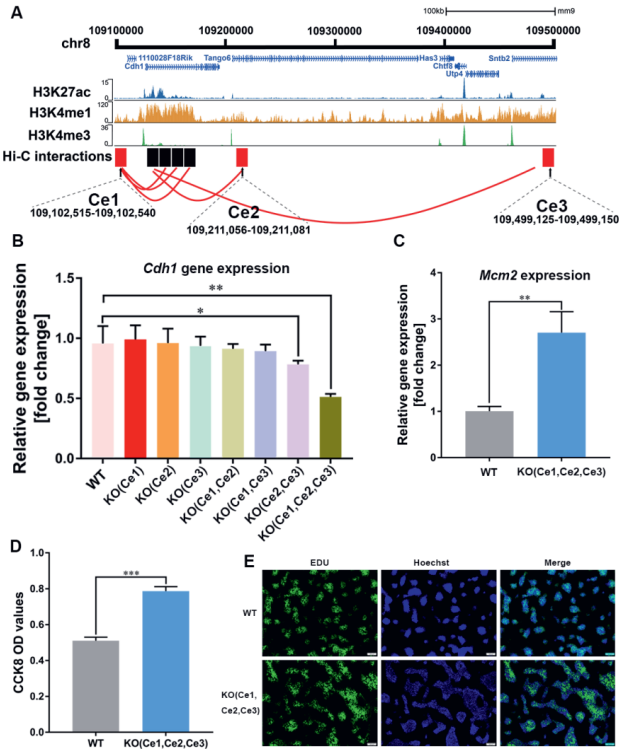


*Figure 5. Endogenous functional validation of MAE-seq enhancers targeted to Cdh1 gene. (**A**) Chromatin interaction between the Cdh1 gene and the MAE-seq enhancers (Ce1, Ce2 and Ce3). Red represents the location of the enhancers, and black represents the region of the Cdh1 gene that interacts with the enhancer. (**B**) Expression of Cdh1 gene in enhancer knockout cells. Expression level of Cdh1 gene is significantly downregulated in Ce2 & Ce3 double knockout and Ce1& Ce2 & Ce3 triple knockout cell compared with wild type cell. Each group has three technical replications. * P< 0.05; ** P< 0.01. (**C**) Mcm2 expression of triple knockout cell. ns: not significant; ** P< 0.01,WT, wildtype cell; KO, triple knockout cell. (**D**) CCK8 detection of triple knockout cells. *** P< 0.001, WT, wildtype cell; KO, triple knockout cell. (**E**) EDU proliferation experiment of triple knockout cell. WT, wildtype cell; KO, triple knockout cell.*

In a previous study, downregulation of the *Cdh1* gene affected the proliferation of cancer cells (51). We explored this phenomenon in mESCs with the Ce1, Ce2 and Ce3 triple knockouts. When we inoculated the same density of wild-type and knockout cells in a 6-well plate for 48 h, we observed a significantly higher number of cells in the knockout cell line compared to the wild-type (Supplementary Figure S5C). To detect the cell cycle difference more directly between the knockout cell line and the wild-type cell line, we tested the expression levels of cell cycle-related genes *Mcm2* and *Pcna*. The expression levels of the

knockout cell lines' *Mcm2* and *Pcna* were significantly upregulated (Figure 5C, Supplementary Figure S5D). Next, we used the CCK8 method to detect the proliferation ability of the knockout cell line more directly, and the result showed that the number of viable cells in the knockout cell line was significantly higher than that of the wild type (Figure 5D). To observe the difference in the number of DNA replication cells more intuitively, we used EDU to stain the two cell lines. It shows that the number of cells stained by EDU in the knockout cell line was significantly higher than that of the wild type (Figure 5E). These results reveal that the novel enhancers identified by MAE-seq play an important role in regulating gene expression and cell phenotype.
Overall, with the high-resolution interactome data and MAE-seq loci, we were able to identify the *Cdh1'*s multiple enhancers-one promoter regulation pattern. This knowledge may represent a critical step towards understanding its precision biology.

3.6 Identification of cell-specific enhancers and accurate annotation of super enhancers

To assess the capability of MAE-seq in recognizing cell-specific enhancers, we applied this technique to identify 554,826 enhancers within C2C12 cells (Supplementary Table S7). Comparative analysis involving these enhancers and those found in mESCs revealed that merely 12.69% of C2C12 enhancers overlapped with mESCs enhancers (Figure 6A). Impressively, the remaining non-overlapping enhancers encompassed >70% of cell-specific ATAC-seq, H3K4me1, and H3K27ac ChIP-seq data (Figure 6B). The heightened resolution of 25 bp enhancers facilitated precise one-vs-one analysis of enhancers and TFs, which contrasted with longer enhancer sequences. Notably, a substantial portion (61.1%) of MAE-seq enhancers in mESCs bound solely to a single TF on their sequence (Supplementary Figure S7A), in contrast to longer sequences like super enhancers (Figure S7B). Strikingly, analysis of TF types bound by these enhancers unveiled a prevalence of cell-specific TFs, including Klf4, Esrrb, Sox2 and Oct4 (Figure 6C). Parallel observations were evident in C2C12 cells, featuring specific TFs such as MyoD1 and Myog (Figure 6C). Expression levels of these enhancers target gene notably exceeded the average level (Figure S7C), accentuating their functional relevance. To corroborate the impact of cell-specific TFs on enhancer activity, we employed Cas13d (52) to target *Oct4* and *Cdh1*. Notably, fluorescence quantification demonstrated their substantial downregulation (Figure S7D), an observation confirmed by diminished protein expression (Figure 6D). Dual luciferase assays corroborated the attenuation or disappearance of enhancer activity upon interference (Figure 6E). Intriguingly, the most enriched TFs within mESCs-specific enhancers were *Esrrb* and *Klf4*, aligning with superenhancer

findings (21). Notably, the integration of superenhancer sites and MAE-seq sites affirmed the precise annotation of TF and histone binding sites within superenhancers (Figure 6F). Furthermore, removal of these MAE-seq sites from superenhancers led to a rapid decline in their activity (Figure 6G), highlighting MAE-seq's accuracy in annotating superenhancers, while underscoring that superenhancers comprise a compilation of multiple cell-specific MAE-seq enhancers.

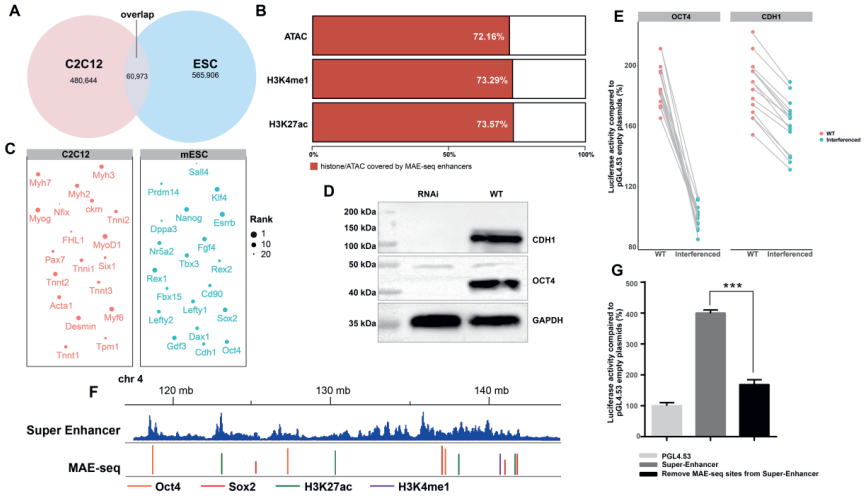


*Figure 6. MAE-seq can identify cell-specific enhancers and accurately annotate super enhancers. (A) The overlap between C2C12 and MAE-seq enhancers identified by mESCs. (B) The coverage percentage of MAE-seq enhancers identified in C2C12 to ATAC-seq, ChIP-seq (H3K3me1, H3K27ac) data in C2C12. (C) The types of single specific TFs bound to enhancer sequences in C2C12 and mESCs. The size of the point represents the ranking of the degree of enrichment. (D) Cas13d interfered with the expression levels of corresponding proteins in cells after 48 h of OCT4 and CDH1 genes. WT, wildtype cell; RNAi, Cas13d RNAi cell. (E) The changes of enhancer activity in corresponding cells after OCT4 and CDH1 interference. WT, wildtype cell; interferenced, Cas13d RNAi cell. (F) Comparison of peaks of Super-Enhancer and MAE-seq sites. A Super-Enhancer of chr 4 is shown in the figure. (G) The activity comparison between super enhancer and super enhancer knocking out MAE-seq site; empty pGL4.53 plasmid was used as the control for baseline luciferase activity. The y axis represents the percentage of luciferase activity compared to pGL4.53 empty plasmids in the respective cells (n = 3 biological independent samples; bars show mean value ± s.e.m.; ns: non-significant,***P< 0.001,calculated using t-test).*

3.7 Differential cofactor dependencies linked to enhancer chromatin environment

The orchestration of enhancer-mediated transcription necessitates cofactors. To discern differences in novel and known enhancer reliance on cofactors, we employed small molecule inhibitors to target *CDK9* and *BRD4* genes. Quantitative PCR results evidenced substantial downregulation 48 h post-interference (Figure 7A). To probe the impact on enhancer activity, dual luciferase assays demonstrated significant reductions in 90% of known enhancers (Figure 7B), with diverse *CDK9* and *BRD4* perturbation effects. In contrast, only the sixth novel enhancer experienced *CDK9*-mediated activity reduction, suggesting context-dependent responses (Figure 7B). Given the enclosed chromatin state of novel enhancers, we probed the influence of chromatin environment on cofactor dependence. The results of interference with SMARCC1 and SMARCB1 showed that the expression levels were down-regulated (Figure 7C). In terms of the effect on enhancer activity, the activity of novel enhancers decreased significantly or disappeared, while the activity of known enhancers was not affected (Figure 7D). These findings suggest that novel enhancers depend on the CRCs family, while known enhancers rely on the COFs family, contingent on chromatin state. Additionally, novel enhancer sequences exhibited significantly higher TF binding compared to known enhancers (Figure 7E). These observations culminate in a proposed model whereby known enhancers bind a TF and subsequently recruit COFs for target gene interaction. Notably, within closed chromatin regions, collaboration between TFs drives the CRCs family to compete with nucleosomes for DNA binding, ultimately binding to novel enhancer sequences (Figure 7F), akin to the 'billboard' model of enhancer collaborative binding (53).

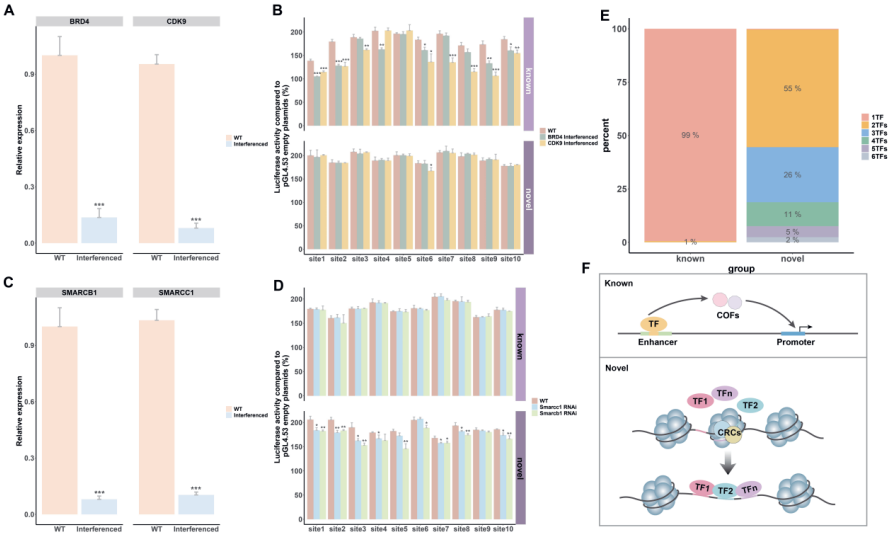


*Figure 7. Differential cofactors are related to the chromatin environment of the enhancer. (A) After 48 h of small molecule inhibitors interfering with BRD4 and CDK9 genes, QPCR was used to detect the changes of corresponding gene expression. ***P< 0.001, calculated using t-test. (B) The changes of known and novel enhancer activity after BRD4 and CDK9 interference. The activity of 10 novel and known enhancers was detected in the cells after BRD4 and CDK9 interference. WT, wildtype cell. (C) After 48 h of small molecule inhibitors interfering with SMARCC1 and SMARCB1 genes, QPCR was used to detect the changes of corresponding gene expression. ***P< 0.001, calculated using t-test. (D) The changes of known and novel enhancer activity after SMARCC1 and SMARCB1 interference. The activity of 10 novel and known enhancers was detected in the cells after BRD4 and CDK9 interference. WT, wildtype cell. (E) The number of TFs combined on the known and novel enhancer sequences is shown in the graph as a result of ≤6 TFs. (F) Cofactors and TFs regulate known and novel enhancer regulatory mechanism models.*

4 Discussion

In the past few decades, scientists have made significant efforts in defining REs across the genome, employing experimental and bioinformatics analyses. However, most experimental methods are based on chromatin, and struggle to precisely pinpoint the loci of these elements. The length of defined REs is typically several hundred base pairs (~270 bp in average) (54). Recently, the length of STARR-seq defined REs has been analyzed and is similar to the results of the above methods. In contrast, the motif analysis of these binding sites has revealed that the core sequences of these REs are typically 6–14 bp long. However, studies have also indicated that these motif sequences might not function independently, as their activities depend heavily on the flanking sequences at both ends (55,56). These flanking sequences can facilitate efficient transcription factor binding or modulate binding affinity to influence gene transcription regulation.

A recent study investigated the activity of REs with different fragment lengths and found that long fragments, besides being less repetitive, displayed advantages in enhancer activity. The researchers speculated that long fragments may provide more position for transcription factor binding (19). In contrast, our data present a different perspective. We propose that the flanking sequences of some long fragments may also bind some transcriptional repressors and influence enhancer activity. We believe that within long fragments, there exists a core fragment that binds the most crucial transcription factors. This core fragment acts as a functionally independent element and this core fragment is a functionally independent element, and its deletion directly affects the activity of the entire sequence. A study by Jason *et al.* also mentioned that deletion of the middle 354bp from a long fragment (678 bp) resulted in the loss or reduction of activity for the entire fragment. In our study, we observed a similar effect with a 25 bp deletion through the 'LER' experiment and demonstrated by knockout experiment that this

25-bp deletion can affect cell phenotype. These experimental results support our hypothesis to some extent, bringing us closer to precise genome annotation.

Over the past two decades, ENCODE Project and Epigenome Roadmap Consortiums have invested tremendous efforts in generating comprehensive predictive annotations of REs located in non-coding regions of the genome (57,58). According to the latest results, 2.1 million REs (enhancers) have been proposed or predicted across the human genome, ~567 million bp (18% of the whole genome, 40% of non-repetitive genome sequences) in total (59). However, with our MAE-seq data, while these annotations cover most of the known REs, they suggest that only approximately 52.5 million base pairs are functional sequences. The remaining ~514.5 million base pairs may have other potential functions that need further investigation.

At present, most putative REs are identified via chromatin-based assays, followed by high-throughput sequencing such as ChIP-seq, DNase-seq and FAIRE-seq (12–14). Usually, these candidates need validation experiments to determine their specific function through various analyses. Recent studies have indicated that such RE annotation might be incomplete. For example, sequences labeled with P300 and H3K27ac show a high false positive rate on enhancer validation, and fragments without binding capacity can also show regulatory activities (60). This has also been observed that histone marks previously associated with active genes do not directly cause transcriptional activation (61). In our study, we surprisingly find that a considerable number of our identified enhancers (219,623, ~35.03%) have no signal on either histone or annotation, based on current knowledge. This finding aligns with a recent study (7) and our validations demonstrated that most of these novel loci exhibited regulatory activity. This suggests that epigenetic modifications alone are insufficient for genome-wide RE annotation. Instead, it likely depends heavily on the local chromatin environment, three-dimensional (3D) genome folding, and the relative concentrations of various downstream effector proteins (62–64). Our results indicate that numerous sequences might function in this manner.

The biological function of non-coding sequences has always been an intriguing subject. Unfortunately, there are few reports in this area. For example, the deletion of the enhancer sequence of *Sox9* can lead to gender reversal in mice (65). A recent study reported that non-coding sequences without any epigenetic characteristics can affect cell apoptosis (7). In our study, we performed knockout experiments on three short enhancers interacting with *Cdh1*, two of which were known and one was *de novo* (lacking epigenetic characteristics) identified in our study. CRISPR/Cas9-mediated knockout experiments confirmed that these sequences can regulate cell phenotype. This implies an elegant ‘multiple-enhancer-one promoter’

regulation model. Moreover, we notice that there are more than 2.1 million known enhancers and only 22,000 genes annotated in the human genome (59,66). Based on our findings, the number of REs could be expected to increase significantly. Therefore, our data support the prevalent of 'multiple elements-one target gene' pattern (67,68) that surpasses initial expectations. In this study, we used *Cdh1* as a model to further explore this hypothesis. We found that among the 3 tested enhancers, none of them was redundant in single enhancer knock-out experiment (69). This study verifies the involvement of the novel *Cdh1* enhancer in hierarchical transcriptional regulation and confirms the biological function of short non-coding sequences. It might represent a prevalent model for precise transcription regulation.

Cell-specific transcription factors (TFs) are hallmarks of cellular identity, and prior enhancer studies often faced challenges in achieving precise 'one-vs-one' enhancer-TF analysis and identifying enhancers exclusively binding to cell-specific TFs. Remarkably, most MAE-seq enhancer sequences exclusively bind a single TF, enabling identification of numerous cell-specific TF-binding enhancers. Astonishingly, these enhancers effectively annotate functional regions within super enhancer sites. Although our MAE-seq approach doesn't encompass the entire genome and thus did not annotate all enhancers (231) reported in the literature (21), this direction might remain a potential avenue for the future work. The role of enhancers necessitates the recruitment of cofactors by TFs. While previous studies highlighted the dependence of enhancers on COFs family cofactors, our study indicates that epigenetic variations underlie differential cofactor (70) requirements for novel and known enhancers. This implies intricate regulatory mechanisms governing enhancers within the genome.

MAE-seq stands out as an innovative and robust method for the precise annotation of functional enhancers within the genome. Its versatile nature enables a wide spectrum of applications in various biological domains, encompassing fields like comparative genomics, disease biology, and synthetic biology. Nonetheless, it is imperative to recognize the limitations inherent in MAE-seq. This method operates within a plasmid system, preventing the study of enhancers in their native chromatin context. Researchers should be mindful of this constraint when planning experiments and interpreting the results. Furthermore, it's worth noting that MAE-seq can be relatively resource-intensive in terms of both cost and time. Nevertheless, the advantages of MAE-seq, such as its high resolution and accuracy, often outweigh these limitations.

While ChIP-seq and other tools can identify genome-wide REs, their spans typically range from several hundred base pairs or longer. Therefore, in the capture of most known REs, MAE-seq is a useful method for precisely identifying active

REs, including those without epigenetic marker. It enables re-evaluating the genome annotation and refines our understanding of REs, great facilitating the experimental design for gene function analyses. Hence, MAE-seq provides a valuable tool for unraveling or precisely annotating the 'dark matter' across the genome.

Data availability

All the sequencing data and processed data of this MAE-seq research are available on Gene Expression Omnibus (GEO) with GEO accession: GSE193494; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE193494.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Qinglei Xu for assisting cell screening experiments and Dr. Jun Yin & Jacob Short for text editing.

Funding

This work was supported by the National Key Research and Development Program of China [SQ2023YFF1000085, 2018YFA0903201]; National Natural Science Foundation of China [31970592]; The Agricultural Science and Technology Innovation Program; The Elite Young Scientists Program of Chinese Academy of Agricultural Sciences [CAASQNYC-KYYJ-41]; National Natural Science Foundation of China (No. 32002173); Natural Science foundation of Guangdong Province, China [2018A0303130009]; the Projects Funded by China Postdoctoral Science Foundation (No. BX2021367 and 2021M703543); the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515010766); the Shenzhen Science and Technology Program (Grant No. KCXFZ20201221173205015 and RCBS20210609104512021).

Conflict of interest statement. None declared.

5 References

1. Umarov R., Li Y., Arakawa T., Takizawa S., Gao X., Arner E. ReFeaFi: genome-wide prediction of regulatory elements driving transcription initiation. PLoS Comput. Biol. 2021; 17:e1009376.
2. Graur D., Zheng Y., Price N., Azevedo R.B., Zufall R.A., Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol. 2013; 5:578–590.
3. Parenteau J., Durand M., Véronneau S., Lacombe A.A., Morin G., Guérin V., Cecez B., Gervais-Bird J., Koh C.S., Brunelle D.et al. . Deletion of many yeast introns reveals a minority of genes that require splicing for function. Mol. Biol. Cell. 2008; 19:1932 – 1941.
4. Vallania F., Schiavone D., Dewilde S., Pupo E., Garbay S., Calogero R., Pontoglio M., Provero P., Poli V. Genome-wide discovery of functional transcription factor binding sites by comparative genomics: the case of Stat3. Proc. Natl. Acad. Sci. U.S.A. 2009; 106:5117–5122.
5. Karlić R., Chung H.R., Lasserre J., Vlahovicek K., Vingron M. Histone modification levels are predictive for gene expression. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:2926–2931.
6. Whitfield T.W., Wang J., Collins P.J., Partridge E.C., Aldred S.F., Trinklein N.D., Myers R.M., Weng Z. Functional analysis of transcription factor binding sites in human promoters. Genome Biol. 2012; 13:R50.
7. Ding B., Liu Y., Liu Z., Zheng L., Xu P., Chen Z., Wu P., Zhao Y., Pan Q., Guo Y.et al. . Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability. Sci. Adv. 2021; 7:eabi6020.
8. Salzberg S.L. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019; 20:92.
9. Buecker C., Wysocka J. Enhancers as information integration hubs in development: lessons from genomics. Trends Genet. 2012; 28:276–284.
10. Doane A.S., Elemento O. Regulatory elements in molecular networks. Wiley Interdiscip. Rev. Syst. Biol. Med. 2017; 9:https://doi.org/10.1002/wsbm.1374.
11. Bentley G.A., Lewit-Bentley A., Finch J.T., Podjarny A.D., Roth M. Crystal structure of the nucleosome core particle at 16 A resolution. J. Mol. Biol. 1984; 176:55–75.
12. Johnson D.S., Mortazavi A., Myers R.M., Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316:1497–1502.
13. Boyle A.P., Davis S., Shulha H.P., Meltzer P., Margulies E.H., Weng Z., Furey T.S., Crawford G.E. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132:311–322.
14. Gaulton K.J., Nammo T., Pasquali L., Simon J.M., Giresi P.G., Fogarty M.P., Panhuis T.M., Mieczkowski P., Secchi A., Bosco D.et al. . A map of open chromatin in human pancreatic islets. Nat. Genet. 2010; 42:255–259.
15. Muerdter F., Boryń Ł.M., Arnold C.D STARR-seq - principles and applications. Genomics. 2015; 106:145–150.
16. Peng T., Zhai Y., Atlasi Y., Ter Huurne M., Marks H., Stunnenberg H.G., Megchelenbrink W. STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. Genome Biol. 2020; 21:243.
17. Bardet A.F., Steinmann J., Bafna S., Knoblich J.A., Zeitlinger J., Stark A. Identification of transcription factor binding sites from ChIP-seq data at high resolution. Bioinformatics. 2013; 29:2705–2713.
18. Lee B.H., Rhie S.K. Molecular and computational approaches to map regulatory elements in 3D chromatin structure. Epigenetics Chromatin. 2021; 14:14.

19. Klein J.C., Agarwal V., Inoue F., Keith A., Martin B., Kircher M., Ahituv N., Shendure J. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods. 2020; 17:1083–1091.
20. Hu H., Li X.Feng J., Fu W., Sun F. Transcription factor binding site identification by phylogenetic footprinting. Frontiers in Computational and Systems Biology. 2010; LondonSpringer London113–131.
21. Whyte W.A., Orlando D.A., Hnisz D., Abraham B.J., Lin C.Y., Kagey M.H., Rahl P.B., Lee T.I., Young R.A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013; 153:307–319.
22. van Groningen T., Koster J., Valentijn L.J., Zwijnenburg D.A., Akogul N., Hasselt N.E., Broekmans M., Haneveld F., Nowakowska N.E., Bras J.et al. . Neuroblastoma is composed of two super-enhancer-associated differentiation states. Nat. Genet. 2017; 49:1261–1266.
23. Barakat T.S., Halbritter F., Zhang M., Rendeiro A.F., Perenthaler E., Bock C., Chambers I. Functional dissection of the enhancer repertoire in Human embryonic stem cells. Cell Stem Cell. 2018; 23:276–288.
24. Allen B.L., Taatjes D.J. The Mediator complex: a central integrator of transcription. Nat. Rev. Mol. Cell Biol. 2015; 16:155–166.
25. Reiter F., Wienerroither S., Stark A. Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. 2017; 43:73–81.
26. Nakagawa T., Yoneda M., Higashi M., Ohkuma Y., Ito T. Enhancer function regulated by combinations of transcription factors and cofactors. Genes Cells. 2018; 23:808–821.
27. Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842.
28. Langmead B., Trapnell C., Pop M., Salzberg S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25.
29. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760.
30. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760.
31. Heinz S., Benner C., Spann N., Bertolino E., Lin Y.C., Laslo P., Cheng J.X., Murre C., Singh H., Glass C.K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell. 2010; 38:576–589.
32. Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30:2114–2120.
33. Andrews S. FastQC A quality control tool for high throughput sequence data. 2014;
34. Trapnell C., Pachter L., Salzberg S.L. TopHat: discovering splice junctions with RNA-seq. Bioinformatics. 2009; 25:1105–1111.
35. Fulco C.P., Nasser J., Jones T.R., Munson G., Bergman D.T., Subramanian V., Grossman S.R., Anyoha R., Doughty B.R., Patwardhan T.A.et al. . Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat. Genet. 2019; 51:1664–1669.
36. Yu G., Wang L.G., Han Y., He Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16:284–287.
37. RC T. R: a language and environment for statistical computing. MSOR Connections. 2014; 25:1105–1111.

38. Patwardhan R.P., Hiatt J.B., Witten D.M., Kim M.J., Smith R.P., May D., Lee C., Andrie J.M., Lee S.I., Cooper G.M.et al. . Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. 2012; 30:265–270.
39. Kharchenko P.V., Alekseyenko A.A., Schwartz Y.B., Minoda A., Riddle N.C., Ernst J., Sabo P.J., Larschan E., Gorchakov A.A., Gu T.et al. . Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature. 2011; 471:480–485.
40. Roy S., Ernst J., Kharchenko P.V., Kheradpour P., Negre N., Eaton M.L., Landolin J.M., Bristow C.A., Ma L., Lin M.F.et al. . Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330:1787–1797.
41. Creyghton M.P., Cheng A.W., Welstead G.G., Kooistra T., Carey B.W., Steine E.J., Hanna J., Lodato M.A., Frampton G.M., Sharp P.A.et al. . Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U.S.A. 2010; 107:21931–21936.
42. Bonn S., Zinzen R.P., Girardot C., Gustafson E.H., Perez-Gonzalez A., Delhomme N., Ghavi-Helm Y., Wilczyński B., Riddell A., Furlong E.E. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat. Genet. 2012; 44:148–156.
43. Liu Y., Yu S., Dhiman V.K., Brunetti T., Eckart H., White K.P. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. Genome Biol. 2017; 18:219.
44. Arnold C.D., Gerlach D., Stelzer C., Boryń Ł.M., Rath M., Stark A Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013; 339:1074–1077.
45. Leemans C., van der Zwalm M.C.H., Brueckner L., Comoglio F., van Schaik T., Pagie L., van Arensbergen J., van Steensel B. Promoter-intrinsic and local chromatin features determine gene repression in LADs. Cell. 2019; 177:852–864.
46. Yáñez-Cuna J.O., Kvon E.Z., Stark A. Deciphering the transcriptional cis-regulatory code. Trends Genet. 2013; 29:11–22.
47. Gao T., Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic. Acids. Res. 2020; 48:D58–d64.
48. Liu P., Chen M., Liu Y., Qi L.S., Ding S. CRISPR-based chromatin remodeling of the endogenous Oct4 or Sox2 locus enables reprogramming to pluripotency. Cell Stem Cell. 2018; 22:252–261.
49. Zhang Y., Wong C.H., Birnbaum R.Y., Li G., Favaro R., Ngan C.Y., Lim J., Tai E., Poh H.M., Wong E.et al. . Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013; 504:306–310.
50. Hsieh T.S., Cattoglio C., Slobodyanyuk E., Hansen A.S., Rando O.J., Tjian R., Darzacq X. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. Mol. Cell. 2020; 78:539–553.
51. Dong L.L., Liu L., Ma C.H., Li J.S., Du C., Xu S., Han L.H., Li L., Wang X.W. E-cadherin promotes proliferation of human ovarian cancer cells in vitro via activating MEK/ERK pathway. Acta Pharmacol. Sin. 2012; 33:817–822.
52. Konermann S., Lotfy P., Brideau N.J., Oki J., Shokhirev M.N., Hsu P.D. Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. Cell. 2018; 173:665–676.
53. Arnosti D.N., Kulkarni M.M. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards?. J. Cell. Biochem. 2005; 94:890–898.
54. Moore J.E., Purcaro M.J., Pratt H.E., Epstein C.B., Shoresh N., Adrian J., Kawli T., Davis C.A., Dobin A., Kaul R.et al. . Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020; 583:699–710.

55. Cohen D.M., Lim H.W., Won K.J., Steger D.J. Shared nucleotide flanks confer transcriptional competency to bZip core motifs. Nucleic Acids Res. 2018; 46:8371–8384.
56. Yella V.R., Bhimsaria D., Ghoshdastidar D., Rodríguez-Martínez J.A., Ansari A.Z., Bansal M. Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. Nucleic. Acids. Res. 2018; 46:11883–11897.
57. Leads D.P.A.L., Investigators P, Production NG.D An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74.
58. Kundaje A., Meuleman W., Ernst J., Bilenky M., Yen A., Heravi-Moussavi A., Kheradpour P., Zhang Z., Wang J., Ziller M.J.et al. . Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330.
59. Boix C.A., James B.T., Park Y.P., Meuleman W., Kellis M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature. 2021; 590:300–307.
60. Dickel D.E., Zhu Y., Nord A.S., Wylie J.N., Akiyama J.A., Afzal V., Plajzer-Frick I., Kirkpatrick A., Göttgens B., Bruneau B.G.et al. . Function-based identification of mammalian enhancers using site-specific integration. Nat. Methods. 2014; 11:566–571.
61. Morgan M.A.J., Shilatifard A. Reevaluating the roles of histone-modifying enzymes and their associated chromatin modifications in transcriptional regulation. Nat. Genet. 2020; 52:1271–1281.
62. Henikoff S., Shilatifard A. Histone modification: cause or cog. Trends Genet. 2011; 27:389–396.
63. Shilatifard A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annu. Rev. Biochem. 2012; 81:65–95.
64. Murray S., Lorenz P., Howe F., Wouters M., Brown T., Xi S., Fischl H., Khushaim W., Rayappu J.R., Angel A.et al. . H3K4me3 is neither instructive for, nor informed by, transcription. 2019; bioRxiv doi:19 July 2019, preprint: not peer reviewedhttps://doi.org/10.1101/709014.
65. Gonen N., Futtner C.R., Wood S., Garcia-Moreno S.A., Salamone I.M., Samson S.C., Sekido R., Poulat F., Maatouk D.M., Lovell-Badge R. Sex reversal following deletion of a single distal enhancer of Sox9. Science. 2018; 360:1469–1473.
66. Levine M., Cattoglio C., Tjian R. Looping back to leap forward: transcription enters a new era. Cell. 2014; 157:13–25.
67. Krijger P.H., de Laat W. Regulation of disease-associated gene expression in the 3D genome. Nat. Rev. Mol. Cell Biol. 2016; 17:771–782.
68. Long H.K., Prescott S.L., Wysocka J. Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell. 2016; 167:1170–1187.
69. Osterwalder M., Barozzi I., Tissières V., Fukuda-Yuzawa Y., Mannion B.J., Afzal S.Y., Lee E.A., Zhu Y., Plajzer-Frick I., Pickle C.S.et al. . Enhancer redundancy provides phenotypic robustness in mammalian development. Nature. 2018; 554:239–243.
70. Neumayr C., Haberle V., Serebreni L., Karner K., Hendy O., Boija A., Henninger J.E., Li C.H., Stejskal K., Lin G.et al. . Differential cofactor dependencies define distinct types of human enhancers. Nature. 2022; 606:406–413.

Chapter 5

Identification of enhancer core sequence from MAE-seq data by multiple kernel learning classifier

Unpublished

Qitong Huang†, Xiusheng Zhu†, Jing Luo1, Martien A.M. Groenen, Ole Madsen and Yubo Zhang
†These authors contributed equally to this work

Abstract

Background: The resolution of current enhancer identification methods is often too indistinct to capture specific fine-grained functional elements, particularly in case of the 25bp core enhancer sequences obtained with MAE-seq (Massive Active Enhancers by Sequencing) to identify mammalian enhancers. Integrating machine learning with MAE-seq data offers a promising solution for a better identification of core enhancer elements.

Methods: To address the dual challenge of interpreting conserved motifs and capturing complex sequence interactions, we developed a hybrid-kernel Support Vector Machine (SVM) model. This model strategically combines:

1) A linear kernel (30% weight) to maintain interpretability for conserved regulatory motifs.

2) A radial basis function (rbf) kernel (70% weight) to model the nonlinear interactions within gapped k-mer (Gkm)-encoded 25 bp sequences.

The kernel weighting was directly informed by MAE-seq-derived enhancer activity profiles, supporting that the model is guided by empirical biological data.

Results:

Our hybrid-kernel approach directly translated into significant increased performance. The model achieved an 18% higher precision in identifying functional 25-bp core sequences compared to conventional single-kernel methods. Crucially, the adaptive nature of the RBF kernel (γ-adaptivity) effectively resolved MAE-seq's inherent short-read mapping biases, particularly at core sequence boundaries, leading to more accurate and stable predictions.

Conclusion: By successfully integrating a biologically guided, dual-kernel SVM with MAE-Seq data, our method bridges the technique's high-throughput advantage with the need for precise core-element recognition. This provides a powerful computational toolkit for dissecting enhancer architecture, moving from broad identification to functional fine-mapping.

Keywords

Multiple kernel learning, MAE-seq, transcriptional regulatory element, enhancer

1 Introduction

Enhancers are crucial transcriptional regulatory elements that are involved in increased transcription of genes in eukaryotes [1]. The first identification of an enhancer came from an SV40 tumor virus genome region, which was shown to significantly enhance the transcription of a heterologous human gene [2]. In recent years, the ascending level of genomic sequencing has enabled a better understanding of the relationship between mutations in enhancers and specific diseases [3]. It has been reported that blood disorders are related to a mutation in the *GATA2* enhancer region [4]. Moreover, in human prostatic carcinoma cells, roughly 88% of the SNPs within the known prostate cancer loci are located within putative enhancer regions [5]. However, identification of enhancers and

determination of their function is still a challenge.

Currently, a wide range of methods are used for the genome-wide identification of enhancers, resulting in a large variety of data types. These include evolutionary conservation data [6], histone marks [7]，open chromatin [8], transcription factor-binding sites (TFBSs) [9], eRNA expression from CAGE data [10] and enhancer-screening data by STARR-seq [11]. Advantages of these data are their ease to compute (evolutionary conservation data), high accuracy (eRNA data and Enhancer-screening data) and provision of cell-line-/tissue-specific information that characterize enhancers (Histone marks and TFBSs) [12]. Bioinformatics approaches for the identification of enhancers include Hidden Markov Models (HMMs) [13], the combination of DNase I hypersensitive site (DHS) with TFBSs [8], Dynamic Bayesian Networks (DBNs) [14], Support Vector Machine (SVMs) [15], and eRNA expression analysis [10]. A drawback of these existing methods is the limited resolution, with sizes of active enhancers ranging from around 200 to 1000 bp [16, 17]. This size of experimentally defined regulatory enhancer elements contradicts the bioinformatics computation of protein binding motifs which only span ~6-14bp [18].

To facilitate a better resolution of enhancer elements, MAE-seq [18] was developed to explore the core region of enhancers. This technology revealed that 25 bp sequences can act as a functional enhancer unit. This significantly narrows down the range of enhancer sequences that need to be explored, while simultaneously substantially reducing both the computational cost and processing time required for enhancer identification using machine learning algorithms.

Here, we provide a multiple kernel learning (MKL) classifier to identify the enhancer core region based on MAE-seq data. First, positive training sequences derived from known enhancer-associated MAE-seq data were converted into a numerical feature space via a simple vectorization scheme. Subsequently, an MKL classifier was trained on this positive set to discern the discriminative patterns of enhancer sequences. To construct a balanced training dataset, a set of negative control sequences was generated in a randomized manner, ensuring no overlap with the positive sequences in both composition and genomic origin. We applied this classifier to the whole genome sequence of mouse (http://hgdownload.soe.ucsc.edu/goldenPath/mm9/chromosomes/) to identify potential enhancer core regions. We find that around 82.3% of the identified enhancers intersect with annotated enhancer loci demonstrating the validity of our method.

2 Results

2.1 Multiple Kernel Learning Algorithm

Support vector machine (SVM) is a supervised learning algorithm whose core idea is to find an optimal hyperplane that separates data of different classes while

maximizing the distance between the hyperplane and the closest data points (i.e., support vectors). Traditional SVM typically uses a single kernel function, while MKL extends this concept by allowing the combination of multiple kernel functions. Compared to single-kernel SVM, MKL offers the following advantages: greater flexibility, improved performance and interpretability, and the ability to handle heterogeneous data [19]. Additionally, compared to other machine learning methods such as logistic regression and random forests, SVM is a reliable and efficient approach for cross-species enhancer prediction [20]. Existing studies [21] have shown that SVM exhibits more robust performance when predicting enhancers in species not part of the training set. Therefore, we adopted MKL for the classification and identification of MAE-seq data. Based on single-kernel computational performance (Figure 1), we selected linear and radial basis function (rbf) as the computational kernels for MKL. We found the optimal weighting by comparing the AUC and AP values of the weights for different kernels. This resulted in a hybrid model combining linear (30% weight) and Gaussian (70% weight) kernels which demonstrated significantly better performance compared to using individual kernels (Figure 2). Based on the comparison of accuracy results, the MKL approach achieved higher accuracy than SVM.

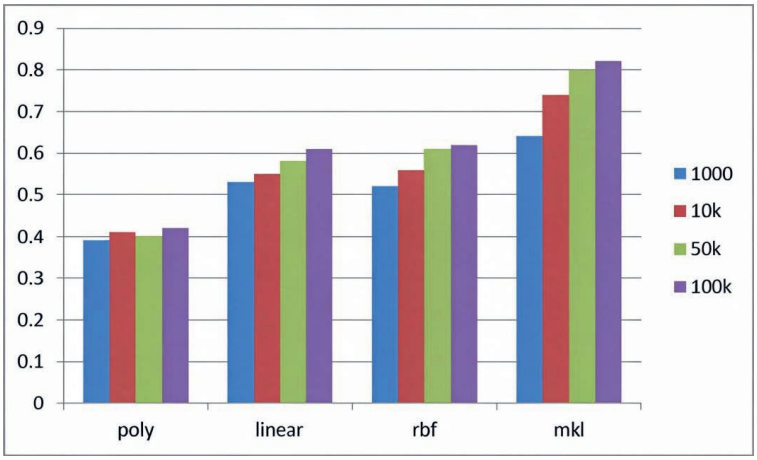


Figure 1. Comparison of different kernel functions. The numbers represent the sizes of the datasets. Poly, linear, and rbf are all single-kernel SVMs, where "poly" denotes the polynomial kernel, "linear" the linear kernel, and "rbf" the radial basis function kernel, i.e., the Gaussian kernel. MKL is multiple kernel learning.

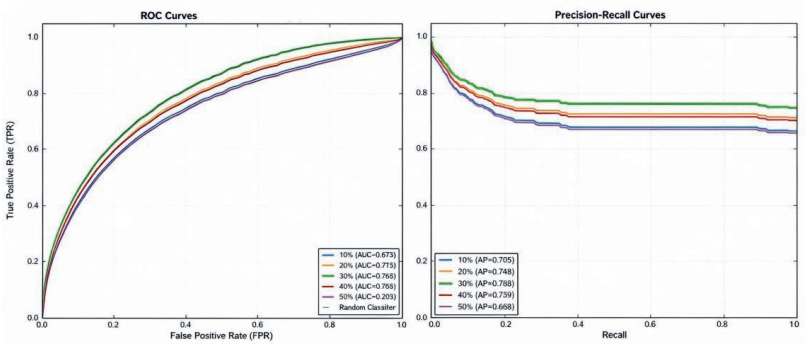


Figure 2. Optimal weight value comparison. The green line represents the curve when the linear kernel accounts for 30% and Gaussian kernel for 70%, achieving the best AUC (0.755) and AP value (0.78).

2.2 Vectorization methods for DNA sequences

The vectorization methods for DNA sequences must balance biological significance and computational efficiency, while enabling flexible integration of multimodal features through different kernel functions. For various encoding approaches, we compared their characteristics (Table 1, see supplementary 1).

Table 1. Comparison: One-Hot Encoding vs. gkm + PWM

Issue	**One-Hot Encoding**	**gkm + position weight matrix (PWM)**
Dimensional Efficiency	High-dimensional, sparse (100 dimensions for 25bp)	Low-dimensional, dense (e.g., 20 – 50 dimensions for gkm)
Feature Informativeness	Only encodes base presence (A/T/C/G)	Captures higher-order patterns (motifs, TF binding, conservation)
Compatibility with MKL	Poor (hard to combine with other kernels like epigenetic or structural data)	Natural fit (e.g., PWM kernels for TF binding + gkm for sequence motifs)
Performance on Small Datasets	Prone to overfitting (noisy, sparse features)	More robust (biologically meaningful features)
Interpretability	Low (no direct biological relevance)	High (features map to known biology, e.g., TF motifs)
Computational Cost	Moderate (requires large matrices)	gkm-SVM uses efficient hashing

PWM represents the Position Weight Matrix. D is the abbreviation for Dimension.

The results indicate that one-hot encoding, commonly used in artificial neural networks, is not well-suited for shallow learning methods. As demonstrated by the comparative experiments across varying dataset sizes, the gapped k-mer model +

position weight matrix (gkm + PWM) combination achieves higher accuracy than one-hot encoding when the sample size exceeds 10k (Figure 3).

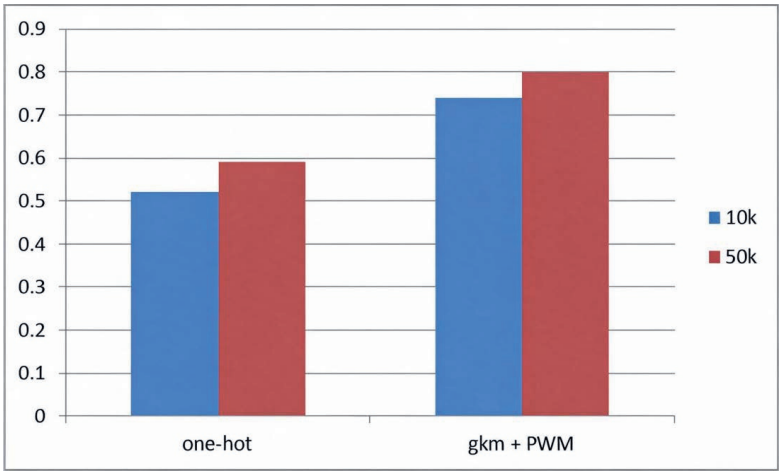


Figure 3. Comparison of different vectorization methods

2.3 Novel enhancer sequence generation and validation

Lastly, we randomly selected 1 million 25 bp sequences from the mouse genome (excluding sex chromosomes) and screened them using our trained MKL classifier. Among the sequences predicted as "true" positives by the classifier (23,711 sequences), 82.3% showed overlapping signals with ChIP-seq enhancer peaks demonstrating high accuracy in enhancer prediction by our MKL classifier.

3 Discussion

Our study employed a multiple kernel learning framework to dissect sequences that are determinants of enhancer activity. The model performance underscores that nucleotide order and occurrence frequency are crucial elements for accurate enhancer prediction, with sequence vectorization and kernel function selection being critical methodological factors. This finding aligns with the established view that enhancer grammar relies on both the presence of specific motifs and their precise spatial arrangement [22].

The effective integration of gkm and PWM features, which atomize short sequences to quantify sequential dependencies and frequency distributions, supports a conceptual analogy between enhancer sequences and linguistic structures. The 25 bp enhancer cores appear to operate under a "syntax" where the arrangement of sequence elements defines their regulatory "semantics." This perspective is consistent with emerging frameworks that treat cis-regulatory elements as a combinatorial code [23] [24]. To fully decipher this grammar, future work must construct functional nucleotide combinations that define precise

fundamental regulatory "words". However, substantial differences distinguish this nucleotide "language" from natural language. Enhancer grammar is notably ambiguous and context-dependent, as the same sequence module can confer divergent regulatory outputs in different cellular or chromatin contexts [25].
Regarding the architecture of our models, the optimal weighting (70% Gaussian kernel, 30% linear kernel) is explainable considering the sequence features of enhancers. The dominance of the Gaussian (RBF) kernel reflects the prevalence of local, non-linear interaction patterns captured after gkm-based vectorization, which effectively "textualizes" the sequences into complex, locally defined features. On the other hand, the linear kernel provides computational efficiency and preserves global linear separability but has limited capacity to model complex patterns [26]. The Gaussian kernel compensates for this limitation by mapping data into a high-dimensional space, capturing intricate local structures via the γ parameter. Thus, this hybrid approach balances the need to model complex feature interactions while maintaining parsimony, a strategy supported by prior work in genomic sequence classification [27].

Our findings directly address a persistent challenge in the field: reliably predicting the functional impact of subtle sequence variations. As noted by Stormo and Zhao [22], the effects of single-nucleotide variants in enhancers remain difficult to forecast. In a focused analysis [18], we found that introducing as few as two substitutions (edit distance=2) within a 25 bp core sequence led to a dramatic loss of enhancer classification for the majority of tested variants (94 out of 100). This steep functional sensitivity suggests that enhancer activity within these short cores can exquisitely depend on precise nucleotide identity and context. It corroborates studies showing that minor sequence alterations can disrupt transcription factor binding and cofactor recruitment, leading to binary on/off switches in enhancer function [28]. Therefore, our model not only predicts enhancer locations but also implies high sequence constraint within the identified cores, offering a computational tool to assess the potential disruptiveness of sequence variations.

4 Conclusion

In this study, we designed and evaluated the MKL algorithm for classifying and predicting core enhancer sequences, leveraging the advantages of MAE-seq short reads to significantly reduce computational costs and processing time. Compared to the 200-1000 bp length range of ChIP-seq signals, the 25 bp reads demonstrate remarkable advantages, enabling faster genome-wide identification of core enhancer sequences.
Furthermore, our MKL-based system facilitates non-experimental approaches to enhancer identification prediction. By randomly capturing genomic sequences for validation, we can generate extensive novel core sequence datasets, which can then

be verified through the MAE-seq system, substantially reducing experimental workload.

5 Materials and Methods

5.1 Processed data sets used

We obtained core regulatory elements of genome wide enhancer data of MAE-seq from mESCs [18]. In short, DNA fragments were randomly synthesized and integrated into a fluorescent reporter vector with a mini promoter. Following transfection of reporter vector library, flow cytosorting was applied to collect positive cells. Then the plasmid DNA was extracted for high-throughput sequencing [29]. The library was sequenced on a Illumina platform using 50 bp single end reads strategy. After MAE-seq data analysis, we set the sequences with significant enhancer activities as the positive dataset.

The negative dataset consisted of randomly generated 25-50 bp DNA fragments which follow the length distribution of next generation sequencing result. Meanwhile, we filtered for duplicate sequences as in the positive dataset.

We obtained epigenome data of ChIP-seq and DNase-seq from the mouse ENCODE consortium for mouse embryonic stem cell (E14) [18]. We used narrow peaks for DNase-seq (GSM1014154) and broad peaks for H3K4me1 (GSE136454) and H3K27ac (GSM1000126) histone modifications. We directly obtained the peaks files from mouse ENCODE project (accessions: ENCFF194TQD and ENCFF158GBZ). We obtained mouse enhancer regions from EnhancerAtlas 2.0 dataset (http://www.enhanceratlas.org/) from 241 cells and tissue types in mice.

5.2 MAE-seq data analysis

Sequence raw reads were trimmed using cutadapt (1.8.dev0) [31]. Fastp [32] was used to filter out low quality reads with default arguments. For short DNA fragments mapping, BWA was applied to mm9 (https://hgdownload.soe.ucsc.edu/goldenPath/mm9/liftOver/) with samse algorithm [33]. Then we kept the uniquely mapped reads and removed PCR duplicates using samtools [34]. We selected the reads which were from 25 to 50 bp as the positive dataset.

5.3 Sequence vectorization

For sequence vectorization, we employed an optimized k-mer Hash coding method base on B-tree algorithm [35]. First, we transformed the base into binary code:

$$\mathrm{Code(x)} = \begin{cases} 00 & x = C \\ 01 & x = T \\ 10 & x = A \\ 11 & x = G \end{cases}$$

Then the Hash function f:$\sum k \to N$ was defined as:

$$f(s[1..k]) = \sum_{j=1}^{k} 4^{k-j} code(S_j)$$

where $s[1..k]$ is the length of DNA fragment(k-mer).
Further details on the principles of vectorization can be found in supplementary file 1.

5.4 MKL model development

MKL model was applied to classify the sequences. We randomly chose 100,000 reads from the positive and from the negative dataset, respectively, for MKL model training. We used another 20,000 reads from the positive dataset and 10,000 reads from the negative dataset for model testing. We compared different kernel functions for this model and selected the linear kernel as the function. We evaluated the performance of the model on test data by generating the ROC curve by plotting true positive rate versus false positive rate and the precision recall curves (PRC). We ultimately established a hybrid model combining a linear kernel (weight: 30%) and a Gaussian kernel (weight: 70%).

Linear Kernel：

$K\text{linear}(\mathbf{x}i,\mathbf{x}j)=\mathbf{x}i\top\mathbf{x}j+c$

Where xi,xj：Eigenvector，c：Optional constant term (usually set to 0)

RBF Kernel：

$K\text{Gaussian}(\mathbf{x}i,\mathbf{x}j)=\exp(-\gamma||\mathbf{x}i-\mathbf{x}j||2)$

Where γ=2σ21：Control the width of the nucleus（σ is the standard deviation）

Combination formula:

$2K\text{linear}(\mathbf{x}1,\mathbf{x}2)=\mathbf{x}1\top\mathbf{x}2$

$K\text{Gaussian}(\mathbf{x}1,\mathbf{x}2)=\exp(-\gamma||\mathbf{x}1-\mathbf{x}2||2)$

$K\text{hybrid}=0.3\times K\text{linear}+0.7\times K\text{Gaussian}$

5.5 ChIP-seq data analysis

Given the limited coverage of the narrow-peak ChIP-seq dataset, the peaks were extended upstream and downstream（±500bp）to encompass the full enhancer regions.

Data availability

The MAE-seq datasets have been uploaded into the GEO database (GSE149028).
The code is in https://gitee.com/agiswurhqt/enhMKL.

Competing interests
The authors declare that they have no competing interests

Funding
This work was supported by the National Natural Science Foundation of China [31970592]; National Key Research and Development Program of China [2018YFA0903201]; The Agricultural Science and Technology Innovation Program; The Elite Young Scientists Program of Chinese Academy of Agricultural Sciences [CAASQNYC-KYYJ-41]; National Natural Science Foundation of China (No. 32002173); Natural Science foundation of Guangdong Province, China [2018A0303130009].

Authors' contributions

Qitong Huang conceived the study, designed the model, and performed all bioinformatics analyses. Xiusheng Zhu conducted all wet-laboratory experiments. Jing Luo assisted with figure preparation. Martien A.M. Groenen provided critical discussions and contributed to manuscript revision. Yubo Zhang and Ole Madsen jointly supervised the project, reviewed, and edited the manuscript. The manuscript was written by Qitong Huang and Ole Madsen.

Supplementary 1

The gkm + PWM features for shallow classifiers
The superior performance of the gkm + PWM representation over one-hot encoding in shallow learning models can be attributed to the nature of the extracted sequence features. One-hot encoding provides a raw, position-specific representation of nucleotides without capturing any contextual or higher-order dependencies between bases. Such sparse and discrete features are poorly suited for shallow classifiers, which as they lack the capacity to infer complex sequence structures from raw input. In contrast, the gkm + PWM approach incorporates biologically meaningful information by quantifying both local compositional patterns (gapped k-mers) and regulatory motif preferences (position weight matrices). This representation effectively embeds "syntactic" and "semantic" features of enhancer sequences into a continuous feature space, allowing the classifier to better distinguish functional sequences. The advantage becomes more pronounced with larger datasets (>10k samples), where gkm + PWM features provide richer and more informative patterns for model generalization.

Mathematical Basis of gkm + PWM Features

To represent enhancer sequences for classification, we employed a combination of gapped k-mer (gkm) and position weight matrix (PWM) features, capturing both sequence structural patterns and regulatory motifs.

Gapped k-mer (gkm) features quantify the occurrence of nucleotide subsequences of length k while allowing a limited number of mismatches or gaps. Formally, for sequences Si and Sj, the gkm kernel is defined as the inner product of their feature vectors:

$$K_{\text{gkm}}(S_i, S_j) = \langle \phi_{\text{gkm}}(S_i), \phi_{\text{gkm}}(S_j) \rangle,$$

Where Φgkm(S) represents the vector of gapped k-mer frequencies in sequence S. This approach captures local compositional patterns and partial sequence dependencies, analogous to n-gram features in natural language.

Position weight matrix (PWM) features represent the nucleotide preference at each position of a motif. For a sequence S of length L and a PWM M, the sequence is scored using a log-likelihood function relative to background probabilities bx:

$$\text{Score}(S \mid M) = \sum_{i=1}^{L} \log \frac{P_{sl,l}}{b_{sl}},$$

where s_l is the nucleotide at position l and p_{sl},l is the corresponding PWM probability. High scores indicate a strong match to biologically relevant motifs.

Finally, in a multiple kernel learning framework, the gkm and PWM kernels are combined as a weighted sum:

$$K_{\text{combined}}(S_i, S_j) = \alpha K_{\text{gkm}}(S_i, S_j) + (1 - \alpha) K_{\text{PWM}}(S_i, S_j),$$

where α is optimized via cross-validation. This combination integrates local sequence patterns ("syntax") with motif-based regulatory signals ("semantics"), providing a rich feature representation for enhancer classification.

6 References

1. Pennacchio, L.A., et al., *Enhancers: five essential questions.* Nat Rev Genet, 2013. **14**(4): p. 288-95.
2. Atchison, M.L., *Enhancers: mechanisms of action and cell specificity.* Annu Rev Cell Biol, 1988. **4**: p. 127-53.
3. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
4. Soukup, A.A., et al., *Single-nucleotide human disease mutation inactivates a blood-regenerative GATA2 enhancer.* J Clin Invest, 2019. **129**(3): p. 1180-1192.
5. Wang, Z., et al., *HEDD: Human Enhancer Disease Database.* Nucleic Acids Res, 2018. **46**(D1): p. D113-D120.
6. Visel, A., J. Bristow, and L.A. Pennacchio, *Enhancer identification through comparative genomics.* Semin Cell

Dev Biol, 2007. **18**(1): p. 140-52.
7. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.
8. Boyle, A.P., et al., *High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells.* Genome Res, 2011. **21**(3): p. 456-64.
9. Rye, M., et al., *Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements.* BMC Biol, 2011. **9**: p. 80.
10. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.
11. Kwasnieski, J.C., et al., *High-throughput functional testing of ENCODE segmentation predictions.* Genome Res, 2014. **24**(10): p. 1595-602.
12. Kleftogiannis, D., P. Kalnis, and V.B. Bajic, *Progress and challenges in bioinformatics approaches for enhancer identification.* Brief Bioinform, 2016. **17**(6): p. 967-979.
13. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization.* Nat Methods, 2012. **9**(3): p. 215-6.
14. Hoffman, M.M., et al., *Unsupervised pattern discovery in human chromatin structure through genomic segmentation.* Nat Methods, 2012. **9**(5): p. 473-6.
15. Fletez-Brant, C., et al., *kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W544-56.
16. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
17. Gaulton, K.J., et al., *A map of open chromatin in human pancreatic islets.* Nat Genet, 2010. **42**(3): p. 255-9.
18. Zhu, X., et al., *MAE-seq refines regulatory elements across the genome.* Nucleic Acids Res, 2024. **52**(2): p. e9.
19. Wilson, C.M., et al., *Multiple-kernel learning for genomic data mining and prediction.* BMC Bioinformatics, 2019. **20**(1): p. 426.
20. MacPhillamy, C., et al., *Cross-species enhancer prediction using machine learning.* Genomics, 2022. **114**(5): p. 110454.
21. Fernandez, M. and D. Miranda-Saavedra, *Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines.* Nucleic Acids Res, 2012. **40**(10): p. e77.
22. Stormo, G.D. and Y. Zhao, *Determining the specificity of protein-DNA interactions.* Nat Rev Genet, 2010. **11**(11): p. 751-60.
23. Zabidi, M.A. and A. Stark, *Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors.* Trends Genet, 2016. **32**(12): p. 801-814.
24. Haberle, V. and A. Stark, *Eukaryotic core promoters and the functional basis of transcription initiation.* Nat Rev Mol Cell Biol, 2018. **19**(10): p. 621-637.
25. Ong, C.T. and V.G. Corces, *Enhancer function: new insights into the regulation of tissue-specific gene expression.* Nat Rev Genet, 2011. **12**(4): p. 283-93.
26. Ghandi, M., et al., *Enhanced regulatory sequence prediction using gapped k-mer features.* PLoS Comput Biol, 2014. **10**(7): p. e1003711.
27. Ben-Hur, A., et al., *Support vector machines and kernels for computational biology.* PLoS Comput Biol, 2008. **4**(10): p. e1000173.
28. Kircher, M., et al., *Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution.* Nat Commun, 2019. **10**(1): p. 3583.
29. Zhu, X., et al., *Start from Scratch: Precisely Identify Massive Active Enhancers by Sequencing.* bioRxiv, 2019.
30. Kong, S., et al., *Exonuclease combinations reduce noises in 3D genomics technologies.* Nucleic Acids Res, 2020. **48**(8): p. e44.
31. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* Embnet Journal, 2011.
32. Chen, S., et al., *fastp: an ultra-fast all-in-one FASTQ preprocessor.* Bioinformatics, 2018. **34**(17): p. i884-i890.
33. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
34. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
35. WANG Shulin, W.J., CHEN Huowang, ZHANG Dingxing, *Research on Counting Algorithm of k-mer Occurrence in DNA Sequence.* Computer Engineering, 2007.

Chapter 6

General Discussion

General Discussion

Accurately identifying transcriptional regulatory elements contributes to unveiling the molecular mechanisms of gene expression regulation. This thesis conducts innovative practices in identifying transcriptional regulatory elements by utilizing various analytical techniques and omics data, along with designing algorithmic models. This includes the identification of chromatin loops based on Hi-C data (Chapter 2), data mining of promoters using single-cell TSS-seq data (Chapter 3), and the discovery and validation of core enhancer sequences using MAE-seq data (Chapters 4 and 5). In the discussion section, I will explore the existing challenges in current Hi-C analysis and the prospects of applying large language models to the identification of sequence features of transcriptional regulatory elements.

6.1 Overcoming Limitations in Identifying Chromosomal Loops using Hi-C Technologies

With the continuous advancement of Hi-C technology, numerous algorithms and computational tools have been developed for the analysis of Hi-C data. These encompass hierarchical structural analysis (such as A/B compartments, TADs, chromatin loops) [1], prediction of 3D features, analysis and detection of structural variations (CNV and structural variant detection) [2], de novo genome scaffolding, and 3D modelling [3]. Among these, one of the most challenging tasks enabled by Hi-C technology is the identification of chromatin loops. Spatial interaction information of chromatin can be extracted by identifying chromatin loops, thereby investigating the interactions between promoters and enhancers. Presently, in light of the characteristics of Hi-C data, methods for identifying chromatin loops primarily rely on utilizing local signal distributions to estimate backgrounds or employ global (chromosome-scale or genome-wide) approaches for modelling [1, 4-6]. However, numerous challenges remain to be addressed, such as stringent resolution requirements and issues concerning the uniformity of algorithm design.

6.1.1 Hi-C Data Resolution Constraints and Challenges with Processing Large Datasets

The identification of chromatin loops typically aims at a resolution of 5 kb or less, which however requires large amounts of data for achieving sufficient high resolution [1]. Software tools like HiCCUPS [5] and FastHiC [7] are designed for analysis of high-resolution datasets. The currently highest achievable resolution in mammalian genomes is approximately 1 kb, but at the cost of sequencing billions of reads [8]. At the moment, there are three main strategies for improving resolution: 1) increase sequencing depth, 2) optimize experimental techniques to

reduce biases, and 3) utilize deep learning to predict high-depth information from low-depth datasets. Among these, higher sequencing depth translates to higher sequencing costs and greater computational resource requirements. In terms of technical optimization, techniques such as eHi-C [9] and DLO Hi-C [10] have been developed, which reduce data noise and experimental costs. Additionally, there are techniques like Capture-HiC (cHi-C) [11], which can identify interactions limited to promoter and regulatory element pairs or specific loci (e.g., SNPs). However, these techniques differ from the traditional Hi-C "all versus all" pattern as they only involve an asymmetric detection capturing "many versus all" interactions and are unable in detecting novel interactions. Furthermore, due to varying capture efficiencies, they are also susceptible to experimental biases. These experimental biases includes conventional Hi-C (e.g., in situ Hi-C) where loop detection is biased by restriction site density and by contact decay with distance, Micro-C, using MNase fragmentation, which provides uniform genome-wide coverage but requires high sequencing depth and HiChIP/PLAC-seq that captures protein-associated interactions with high sensitivity but exhibits strong protein-centric bias and may miss protein-independent structural contexts while being susceptible to antibody-related noise. To address such biases, CHiCAGO mitigates experimental bias effects by employing a composite background distribution (negative binomial and Poisson), specific implicit normalization, and a multiple-testing correction method based on weighted p-values [12]. Deep learning is another approach to enhance resolution. The HiCPlus tool employs convolutional neural networks to enhance the resolution of Hi-C data, using mean squared error as a loss function [13]. It achieves this by merging portions of smoothed Hi-C maps into smaller segments. HiCNN is another computational method for resolution enhancement. A modified version of the HiCPlus method utilizes an extremely deep convolutional neural network [14]. It transforms Hi-C matrices of regular resolution into highly detailed yet sparse matrices at higher resolution. DeepHiC, based on a network of generative adversarial networks (GANs), enhances Hi-C data by preserving the content while maintaining the same bin size. DeepHiC reconstructs content from approximately 1% of the original data [15]. HIFI addresses sparsity by utilizing dependencies between adjacent restricted fragments, with the best performance demonstrated by a Markov random field [16]. The evolution of these tools has progressed from simple regression-based imputation (HiCPlus) to image super-resolution (HiCNN) followed by distribution-level generative adversarial networks (DeepHiC). Deep learning has shifted Hi-C enhancement from "numerical imputation" to "distribution learning," with the core objective transitioning from "seeing clearly" to "seeing truthfully." These tools can now generate high-quality matrices with realistic statistical properties from very low amounts of data (Table 1), but they remain highly dependent on training data quality and cannot predict unseen structures. Future

trends point toward multi-modal integration and single-cell Hi-C enhancement. The expansion in depth and breadth of multimodal integration involves jointly modeling Hi-C data with microscopic imaging (e.g., FISH) and intranuclear spatial coordinate information to construct a unified view spanning from one-dimensional sequences to three-dimensional spatial localization. This also includes integrating time-resolved Hi-C, transcriptomic, and epigenomic data to develop models capable of inferring how dynamic changes in the three-dimensional genome drive temporal gene expression programs. This will directly address the core question of whether structural changes are the cause or the consequence. Furthermore, leveraging large foundational models to learn a unified low-dimensional representation of different modal data - such as chromatin interactions, epigenetic markers, and gene expression - enables cross-modal retrieval, comparison, and prediction within a shared "semantic space."

As mentioned, increasing sequencing depth and reducing bias fundamentally rely on data quantity to achieve higher resolution. On the other hand, deep learning simulates high-resolution outcomes with lower experimental costs, but its accuracy cannot be ensured across all datasets. Currently, the optimal objective involves selecting the best algorithm to improve the identification rate and accuracy of chromatin loops, while ensuring an adequate amount of data. Loop calling tools should further explore the characteristics of Hi-C interaction signals to optimize the analysis model. In that respect we developed Tloop (Chapter 2) to identify and analyse the local signal attenuation caused by differences in signal density in different regions and the unknotted effect. Furthermore, the significance signal is based on local features, by calibrating the expected abundance of interacting sites, enabling rapid assessment of numerous expressions characterized by subtle distal interactions. This approach holds the potential to mitigate the decline in recall resulting from insufficient resolution.

Table 1. Comparison of four deep learning methods for Hi-C data enhancement, each with distinct approaches and trade-offs.

Method	**Enhancement Approach**	**Key Strength**	**limitation**
HiCPlus	Infills sparse high-resolution matrices into dense ones	Directly addresses data sparsity; conceptually intuitive	Predictions may be oversmoothed, lacking sharp details
HiCNN	Predicts high-resolution matrices from low-resolution inputs	Genuinely increases resolution numerically; adds high-frequency details	Input loses sparse signals from original high-resolution data

DeepHiC	Generates realistic matrices from extremely low-depth data using a GAN framework	High-quality generation; extremely efficient data usage (only ~1% data needed)	GAN training is complex and unstable; does not create new bins
HIFI	Imputes data by modelling dependencies between adjacent fragments	Incorporates interpretable biophysical prior knowledge	Relatively traditional model; limited ability to capture complex patterns

6.1.2 Lack of Unified Standards in Hi-C analysis

Recognition of chromosomal loops based on Hi-C data still lacks unified standards. I have previously discussed the principles and characteristics of different loop-calling algorithms in chapter 2. In loop-calling we still need to strike a balance between recall and precision rates. The lack of unified standards presents a challenge: when multiple callers generate computational results, do we consider their intersection or union? In many biological studies, the primary concern often lies in the identification of interaction sites of interest. As a result, an algorithm is typically evaluated based on performance metrics such as noise resistance, sequencing depth robustness, speed, consistency, and detection rate [16]. This has led to an evaluative approach driven by inferred results, rather than formulating principles to determine computational standards. Consequently, situations arise where an algorithm performs 'exceptionally well' due to the characteristics of a specific dataset, even in the absence of sufficient datasets for testing. For example, the preference dependence of restriction enzyme sites in the analysis will affect the expected interaction in that region. The expected interaction signal for interactions not in that region will be reduced, leading to the generation of false positive loops [17].

6.1.3 Further Exploration Needed for Local Features of Hi-C Signals

Current identification of chromosomal loops generally follows a globally uniform principle, hereby lacking classification of local features. FastHiC [5] applies a Hidden Markov Random Field for loop calling, simulating spatial dependencies between neighbouring loci. PSYCHIC [16] considers TAD structures to identify significant interactions with respect to TAD-specific background models. These two local categories belong to external features. The external features of Hi-C signals refer to the relationship between interaction signals and neighboring signals within the structural characteristics. Thus, these features are not determined by enrichment. In chapter 2, I focused on features such as local signal density and data affinity for loop identification, reducing the computational load of global signals. This approach shares similarities with DBSCAN clustering-based methods like LASCA [17] and cLoops2 [18], which also emphasize the local distribution

characteristics of Hi-C signals when calling loops. It can be anticipated that, when computational resources permit, direct clustering of large matrices will swiftly yield significant interaction domains. However, the current strategy still centers on segmented processing of matrices. Machine vision also serves as a means of capturing signal features; where tools like Chromosight [19] and SIP [20] learn and classify patterns in Hi-C signals from matrices. How to synthesize and generalize these local features to assist loop calling is one of the exciting directions in the future evolution of loop callers. Learning the characteristics of loop signal distribution in a heatmap directly through machine vision is similar to face recognition. These features are directly summarized and used to scan and match the input heatmap.

6.2 Accurate Identification of Transcriptional Regulatory Elements:

6.2.1 New Definition for the Core Region of Transcriptional Regulatory Elements

With the application of MAE-seq data (chapter 4), we have greatly refined the identification of the core regions of transcriptional regulatory elements. This has provided a fresh perspective on the mechanisms of transcriptional regulation. Differing from techniques like ChIP-seq, MAE-seq offers high-throughput capabilities. Through the combination of flow cytometry and random synthesis techniques, we have accurately identified enhancer of 25 bp. We categorized these enhancers based on the number of transcription factors binding to them. For the first time, we discovered that a single transcription factor is sufficient to activate enhancers, and we described the characteristics of these diverse enhancers. Finally, based on these characteristics, we proposed a mechanistic model for enhancer action (A single transcription factor is capable of activating the promoter, cell-specific enhancer). The prevailing model of enhancer activation emphasizes the collaborative effort of multiple transcription factors to displace nucleosomes and recruit the transcriptional machinery [18]. While certain 'pioneer factors' are acknowledged for their ability to bind silent chromatin, their action is still framed as initiating a cascade that requires subsequent partners for full activation [19]. Our discovery that single specific TFs are sufficient to activate diverse enhancers challenges this paradigm of obligatory cooperativity. This capability aligns with, and significantly extends, rare prior suggestions that master regulators like OCT4 may exert dominant control in specific contexts [20]. The mechanistic model we propose - wherein a single master TF establishes a permissive chromatin state and then orchestrates further specialization - integrates our novel finding with established principles of combinatorial regulation. It explains how a single factor can dictate cell-specific outcomes while accounting for the observed diversity in enhancer characteristics.

6.2.2 Twenty-five bp is not the limit

The choice of 25 bp for enhancer detection was based on the optimal alignment length obtained from the current alignment software programs. However, this does not imply that the core enhancer sequence is strictly limited to 25 bp, shorter core sequences remain plausible. MAE-seq technology can further validate the selected positive 25 bp sequences. Using the “LER” approach described in chapter 4, the length of each 25 bp fragment is progressively shortened until the fluorescent signal is no longer detectable. Yet, due to the limitations of current alignment algorithms, verifying shorter sequences incurs significant computational costs. Currently, whether it is BWA [24], Bowtie2 [25], Minimap2 [26], or Segemehl [27], the shortest alignment reaches 20 bp. However, once it goes below 25 bp, a substantial amount of multiple mapping occurs. The good news is that with MAE-seq technology, we can validate specific core sequences of shorter effective lengths, albeit limited to a very small number of sites of interest. Simultaneously, we are also highly interested in understanding the relationship between core sequences, protein binding sites, and the sequence features of regulatory element motifs. Whether the base sequence and frequency of occurrence are correlated or even causally related warrants further investigation.

6.2.3 Improvement Directions for Quantitative Methods

For quantitative analysis of MAE-seq, it is important to realize that the sequences are highly stochastic. Consequently, we adopted specific normalization methods for quantification and Poisson testing for significance analysis. Due to limited datasets, many interfering factors in quantification were not considered, such as GC content of short sequences and sequence similarity evaluation. Reproducibility has consistently been a drawback of MAE-seq data [21]. Therefore, we continue to expand the MAE-seq data to enhance coverage across the target genome and hereby increasing the likelihood of detecting the target signal. By employing a data accumulation method were data identified as "true" after each analysis are stored into a target database, we can avoid redundant computation and experimentation. Additionally, using confirmed sequences to identify and evaluate new data enhances the accuracy of the model. The construction of such a database relies on the introduction of large-scale data, and our current data volume still needs to be expanded to achieve thus. Our approach is aimed at achieving a more accurate assessment of the quantitative characteristics of peak counts in MAE-seq data. Simultaneously, in the analysis of MAE-seq data, we did not perform error-tolerant analyses or similarity assessment for short sequences. This, because even slight variations in bases within 25 bp sequences can lead to changes in alignment results. Additionally, among highly similar sequence groups, sequences with similar expression levels cannot be directly merged. Due to these two reasons, we opted for the most stringent sequence filtering strategy where we prioritized precision

over recall.
The major limitation of MAE-seq is its poor inter-batch reproducibility, which demands substantial downstream experimental validation to confirm the reliability of the findings. In addition, the high randomness of the signals leads to variable coverage uniformity across sequencing libraries [21]. While existing normalization methods perform adequately when library size differences are minimal, they lack robustness in scenarios involving substantial disparities in library sizes. In addition to increasing the overall data volume, randomly sampling sequences from the target species' genome to boost alignment rates and produce more effective data is a promising strategy. However, the quantitative relationship between enhanced alignment efficiency and effective signal enrichment remains to be validated.

6.3 Machine Learning in Transcriptional Regulatory Element Recognition – Advances and Limitations

6.3.1 Selection of DNA Sequence Vectorization

DNA sequences need to be encoded to be recognized by machine learning programs. Generally, three types of methods are commonly used for sequence encoding: ordinal encoding, one-hot encoding, and k-mer encoding. As I described in Chapter 5, ordinal encoding demonstrates an advantage in training time, while both methods perform comparably in encoding performance [28]. The performance of one-hot encoding remains consistent across different datasets, but it requires suitable convolutional neural network (CNN) configurations to perform well. Ordinal encoding using matrix representation performs best in certain evaluated algorithms, like Decision Tree and Random Forest [22]. Currently, one-hot encoding is the most widely used approach. DeepBarcoding employs DNA barcodes for species classification using deep learning [29]. It inputs raw sequences and represents them using one-hot encoding in the form of a one-dimensional image. Majdandzic et al. [30] effectively reduced noise in deep neural networks through statistical correction in DNA methylation analysis. They utilized a convolutional-recurrent neural network architecture with attention mechanisms for systematic analysis and accurate identification of DNA N4-methylcytosine sites [31], all of which are achieved using one-hot encoding. K-mer encoding is commonly used for encoding long sequences. DNAshapeR utilizes k-mer encoding to predict DNA shape features, like minor groove width, helix twist, propeller twist and roll, from genomic sequencing data in an ultra-fast and high-throughput manner [32]. Compared to Ordinal Encoding, while it retains the sequential order of bases and conserves storage space, it also disregards the similarity and distinctions between bases, thus failing to capture intricate sequence patterns. On the other hand, k-mer encoding with shorter k values may lead to the curse of dimensionality, requiring substantial storage space and being unsuitable

for MAE-seq data. Building on k-mer features, gkm combined with PWM (gapped k-mer + position weight matrix) is a DNA sequence representation method that integrates the pattern-capturing power of gapped k-mers with the position-specific preferences provided by PWMs. Gapped k-mers allow the extraction of sub-sequences with possible gaps, making the approach robust to local mutations and motif variations, while PWMs quantify base preferences at each position, offering biological interpretability. Together, gkm + PWM can effectively capture complex sequence motifs and local dependencies, making it particularly useful for applications such as transcription factor binding site prediction and regulatory element identification. Compared with one-hot, ordinal, and standard k-mer encoding, gkm combined with PWM can more effectively capture complex sequence patterns and local dependencies. While one-hot preserves base positions but requires convolutional kernels to detect motifs, and ordinal encoding reduces storage but ignores base relationships, and k-mer captures fixed-length local motifs but is insensitive to gapped or variant motifs, gkm + PWM allows for gapped k-mers, making it robust to motif variations and local mutations, while PWM quantifies base preferences at each position, providing a powerful and biologically interpretable motif representation [23, 24]. During our research, we also attempted a quaternary encoding method, where sequences were mapped to quaternary digits. However, this approach could not address the evaluation of sequence similarity and was subsequently discarded.

6.3.2 Characteristics and Improvement Directions of Multi-Kernel Learning for Transcriptional Regulatory Element Recognition

Currently, many machine learning methods as well as deep learning methods have been applied to the extraction and classification of DNA sequence features. Multiple Kernel Learning (MKL) can enhance the interpretability of models, as it integrates multiple kernel functions or similarity measures to capture various aspects of data across different feature spaces [25]. This diversity can aid in extracting crucial features from the data, thereby augmenting the understanding of model decisions. By judiciously selecting and weighting kernel functions, it is possible to enable the model's interpretation across different dimensions and features, facilitating the revelation of relationships within the data. However, the interpretability of MKL is also influenced by the complexity of the kernel functions. Certain kernel functions might introduce intricate mappings in high-dimensional spaces, challenging the interpretations in the original feature space [26]. Additionally, with MKL more parameters and complexity could be introduced, thereby elevating the model's intricacy and subsequently diminishing its interpretability. To integrate the results from the MAE-seq data using MKL while preserving the interpretability of conserved motifs and capturing the nonlinear interactions in the 25 bp core sequences encoded by gkm, we employed a

hybrid approach combining linear and Gaussian kernels. The following considerations was the basis for the selected kernels:

1). Kernel Selection for Interpretability and Complexity:

The linear kernel was chosen explicitly to maintain direct interpretability of feature weights, allowing for the attribution of importance to specific k-mers or gapped k-mers corresponding to known or novel transcription factor binding motifs. This aligns with the established practice of using linear kernels for motif discovery and importance analysis in sequence models [24]. Conversely, the Gaussian (RBF) kernel was incorporated to model the complex, non-linear dependencies between nucleotide positions within the short 25 bp core, which are often missed by linear models but are crucial for accurate enhancer prediction [27].

2). MKL Framework and Weight Optimization:

We adopted the SimpleMKL framework [28] for its efficacy in learning an optimal convex combination of kernels. This method automates the critical step of balancing the contributions of the interpretable linear kernel and the powerful Gaussian kernel during training, ensuring neither dominates unless justified by the data. The optimization objective inherently performs a form of data-driven feature and kernel selection, a strategy recommended for genomic data fusion where the relevance of different information sources (e.g., motif conservation vs. spatial interaction) is not known a priori [29].

6.4 Artificial Intelligence in the Research and Applications of Three-Dimensional Genomes and Transcriptional Regulatory Elements

In recent years, artificial intelligence (AI) has been extensively applied in the field of three-dimensional genomics, encompassing hierarchical structure recognition, resolution enhancement, and corresponding biological mechanisms. Particularly, in the study of chromosomal structural changes in cancer, AI has contributed to refining the regulatory maps of three-dimensional genomes [30] [2]. Additionally, as various models continue to advance, it is to be expected that a more significant role for AI in biological information analysis will emerge. However, it is essential to acknowledge the current limitations of AI in the field of bioinformatics and consider future directions.

6.4.1 Current Artificial Intelligence Approaches

Broadly speaking, AI encompasses machine learning and deep learning. Deep learning has achieved remarkable innovation in Hi-C data analysis. As described in section 6.1.1, AI holds a significant advantage in Hi-C data enhancement by Deep learning tools. Utilizing image-based processing methods along with generative adversarial networks, higher-resolution structures can be predicted with relatively limited data, such as with hicGAN [34] and DeepHiC [13]. Simultaneously,

machine learning plays a pivotal role in the identification of transcriptional regulatory elements. Current methods integrate various features of different markers of transcriptional regulatory elements, including chromatin accessibility (ATAC-seq), histone marks [34], transcription factor binding site information [35], and sequence features [36], to make judgments and predictions regarding experimental outcomes. Through the integration of multi-omics features in machine learning analysis, the ultimate "experimental outcome" transforms raw, massive, and fragmented high-throughput sequencing data into biologically meaningful, interpretable, and verifiable "functional hypotheses" and "regulatory rules". It not only answers the question of "where the regulatory elements are" but also delves deeper into core biological questions such as "under what conditions they are activated", "by what factors they are regulated", and "what consequences arise from variations". This systematically reveals the causal logic of gene expression regulation, marking a profound shift in the field from descriptive data mining to mechanism-driven discovery. At the same time, this approach not only improves the accuracy of distinguishing active versus inactive elements but also enables the inference of potential transcription factors regulating these regions. By prioritizing high-confidence candidates, it effectively guides downstream experimental validation, reducing time and resource consumption. Moreover, the model provides insights into the underlying regulatory mechanisms and is readily scalable to different cell types, conditions, or species, thereby facilitating comprehensive and high-throughput analyses of transcriptional regulation [23] [31].

6.4.2 Feasibility of large language models in DNA sequence analysis

Treating DNA sequences as a form of natural language, with k-mers as tokens, allows large language models (LLMs) based on BERT/Transformer architectures to capture rich contextual features of nucleotide sequences [32]. This approach offers several notable advantages, including end-to-end learning, modelling of complex long-range dependencies, and applicability across diverse tasks such as sequence classification, functional annotation, and generation. Models such as megaDNA [33], a multiscale transformer, demonstrate the feasibility of applying LLMs to genome-wide analyses, enabling prediction and annotation of proteins and regulatory elements at the whole-genome scale. Nonetheless, several challenges remain. Processing extremely long sequences presents significant computational and memory demands, motivating the development of specialized architectures such as HybriDNA [34] and JanusDNA [35]. Furthermore, interpretability and biological relevance are still limited, as the patterns learned by LLMs (motifs, sequence context, dependencies) are not easily mapped to underlying regulatory mechanisms. Finally, biases in training data and tokenization strategies (k-mer, single nucleotide, or hybrid) may affect generalization across

species, atypical sequences, or genomic variants. Addressing these limitations in the future will be critical for fully realizing the potential of LLMs in genomics.

Whether MAE-seq data are suitable for LLMs is also a topic worth discussing. MAE-seq sequences are only 25 bp long, a feature that reduces the computational demands associated with long-range dependencies, while MKL classification results indicate that these sequences possess certain sequential and semantic characteristics. I adopted the principles of LLMs and simulated modelling using an n-gram language model on a subset of MAE-seq data. The training set consisted of 1,000 positive sequences from the MAE-seq data, and 10,000 new sequences were generated for MKL prediction. In this simulation, I used k=3, context_size = 4, and temperature sampling between 0.8 and 1.2. The newly generated sequences were aligned to the mouse genome, resulting in an alignment rate of 21.13%. Among these 2,113 aligned sequences, 1,226 (58.02%) were classified as true by MKL. The simulation results suggest that, although MAE-seq sequences are relatively short (25 bp) and thus limit the capture of long-range dependencies, they still contain sufficient sequential and semantic features for partial learning by language models. The moderate alignment rate to the mouse genome (21.13%) and the MKL-based validation (58.02% of aligned sequences classified as true) indicate that LLM principles can capture some biologically relevant patterns in enhancer sequences. These findings imply that short DNA sequences, such as those from MAE-seq, can serve as meaningful inputs for language models, particularly when supplemented by appropriate sequence encoding and modelling strategies. However, achieving robust predictive performance and fully leveraging LLM capabilities will likely require larger datasets, more diverse sequence contexts, and optimized model parameters. Overall, this analysis demonstrates the potential of applying LLMs to DNA sequence analysis, highlighting both the opportunities for learning sequence patterns and the challenges posed by data sparsity, sequence length, and model complexity.

6.4.3 Limitations of LLMs in DNA sequence analysis

Despite the promising potential of LLMs for DNA sequence analysis, several limitations remain. First, traditional transformer-based architectures face substantial computational and memory demands when processing long sequences, such as full genomic regions, which restricts their scalability [36]. Second, the performance of LLMs heavily depends on the quantity and diversity of training data; datasets with limited sequence length or coverage, such as MAE-seq, may result in underfitting or limited generalization across species and sequence contexts [37]. Third, the interpretability of learned patterns remains challenging, as motifs, context dependencies, or sequence features captured by the model are not easily mapped to biological mechanisms such as transcription factor binding or enhancer activity. Additionally, tokenization strategies, whether k-mer, single-nucleotide, or

hybrid, can influence the model's ability to capture sequence context and long-range dependencies, potentially affecting consistency across tasks. Addressing these challenges - through larger and more diverse training datasets, optimized sequence representations, and interpretable model designs - will be essential to fully realize the benefits of LLMs in genomics.

7 References

1. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
2. Dixon, J.R., et al., *Integrative detection and analysis of structural variation in cancer genomes.* Nat Genet, 2018. **50**(10): p. 1388-1398.
3. Burton, J.N., et al., *Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions.* Nat Biotechnol, 2013. **31**(12): p. 1119-25.
4. Ay, F., T.L. Bailey, and W.S. Noble, *Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.* Genome Res, 2014. **24**(6): p. 999-1011.
5. Durand, N.C., et al., *Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.* Cell Syst, 2016. **3**(1): p. 95-8.
6. Ron, G., et al., *Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains.* Nat Commun, 2017. **8**(1): p. 2237.
7. Xu, Z., et al., *FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data.* Bioinformatics, 2016. **32**(17): p. 2692-5.
8. Pal, K., M. Forcato, and F. Ferrari, *Hi-C analysis: from data generation to integration.* Biophys Rev, 2019. **11**(1): p. 67-78.
9. Kong, S., et al., *Exonuclease combinations reduce noises in 3D genomics technologies.* Nucleic Acids Res, 2020. **48**(8): p. e44.
10. Lin, D., et al., *Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture.* Nat Genet, 2018. **50**(5): p. 754-763.
11. Mifsud, B., et al., *Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C.* Nat Genet, 2015. **47**(6): p. 598-606.
12. Cairns, J., et al., *CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data.* Genome Biol, 2016. **17**(1): p. 127.
13. Zhang, Y., et al., *Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus.* Nat Commun, 2018. **9**(1): p. 750.
14. Liu, T. and Z. Wang, *HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data.* Bioinformatics, 2019. **35**(21): p. 4222-4228.
15. Hong, H., et al., *DeepHiC: A generative adversarial network for enhancing Hi-C data resolution.* PLoS Comput Biol, 2020. **16**(2): p. e1007287.
16. Cameron, C.J., J. Dostie, and M. Blanchette, *HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution.* Genome Biol, 2020. **21**(1): p. 11.
17. Yaffe, E. and A. Tanay, *Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.* Nat Genet, 2011. **43**(11): p. 1059-65.
18. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nat Rev Genet, 2012. **13**(9): p. 613-26.
19. Zaret, K.S. and J.S. Carroll, *Pioneer transcription factors: establishing competence for gene expression.* Genes Dev, 2011. **25**(21): p. 2227-41.
20. Whyte, W.A., et al., *Master transcription factors and mediator establish super-enhancers at key cell identity genes.* Cell, 2013. **153**(2): p. 307-19.

21. Zhu, X., et al., *MAE-seq refines regulatory elements across the genome.* Nucleic Acids Res, 2024. **52**(2): p. e9.
22. Tyagi, N., N. Vahab, and S. Tyagi, *Genome language modeling (GLM): a beginner's cheat sheet.* Biol Methods Protoc, 2025. **10**(1): p. bpaf022.
23. Alipanahi, B., et al., *Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.* Nat Biotechnol, 2015. **33**(8): p. 831-8.
24. Ghandi, M., et al., *Enhanced regulatory sequence prediction using gapped k-mer features.* PLoS Comput Biol, 2014. **10**(7): p. e1003711.
25. Briscik, M., et al., *Supervised multiple kernel learning approaches for multi-omics data integration.* BioData Min, 2024. **17**(1): p. 53.
26. Briscik, M., M.A. Dillies, and S. Dejean, *Improvement of variables interpretability in kernel PCA.* BMC Bioinformatics, 2023. **24**(1): p. 282.
27. Liu, Q., M. Gan, and R. Jiang, *A sequence-based method to predict the impact of regulatory variants using random forest.* BMC Syst Biol, 2017. **11**(Suppl 2): p. 7.
28. Liu, X., et al., *Absent Multiple Kernel Learning Algorithms.* IEEE Trans Pattern Anal Mach Intell, 2020. **42**(6): p. 1303-1316.
29. Niu, W., et al., *Efficient Multiple Kernel Learning Algorithms Using Low-Rank Representation.* Comput Intell Neurosci, 2017. **2017**: p. 3678487.
30. Akdemir, K.C., et al., *Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer.* Nat Genet, 2020. **52**(3): p. 294-305.
31. Zhou, J. and O.G. Troyanskaya, *Predicting effects of noncoding variants with deep learning-based sequence model.* Nat Methods, 2015. **12**(10): p. 931-4.
32. Sarumi, O.A. and D. Heider, *Large language models and their applications in bioinformatics.* Comput Struct Biotechnol J, 2024. **23**: p. 3498-3505.
33. Shao, B. and J. Yan, *A long-context language model for deciphering and generating bacteriophage genomes.* Nat Commun, 2024. **15**(1): p. 9392.
34. Ma, M., et al., *HybriDNA: A Hybrid Transformer-Mamba2 Long-Range DNA Language Model.* ArXiv, 2025. **abs/2502.10807**.
35. Qihao Duan, B.H., Zhenqiao Song,Irina Lehmann, Lei Gu, Roland Eils, Benjamin Wild, *JanusDNA: A Powerful Bi-directional Hybrid DNAFoundation Model.* arXiv, 2025.
36. Dai, Z., et al., *Transformer-XL: Attentive Language Models beyond a Fixed-Length Context.* ArXiv, 2019. **abs/1901.02860**.
37. Zhang, D., et al., *DNAGPT: A Generalized Pre-trained Tool for Multiple DNA Sequence Analysis Tasks.* bioRxiv, 2024.

Summary

Over the past two decades, advances in chromosome conformation capture technologies have revealed that the genome is organized into a complex three-dimensional hierarchy, including chromosome territories, A/B compartments, TADs, and chromatin loops, which play important roles in gene regulation by bringing distal regulatory elements into proximity with their target promoters. Meanwhile, transcriptional regulatory elements have been extensively studied using chromatin-based assays and functional reporter systems, and various machine learning and deep learning approaches have been applied to predict these elements from diverse genomic features. Despite these advances, several important challenges remain unresolved, including the accurate identification of chromatin loops from Hi-C data, the precise localization of transcriptional regulatory elements at single-base resolution, and the effective computational characterization of short functional sequences. To address these challenges, this thesis presents four major contributions including improved loop calling algorithms, single-cell TSS detection, high-resolution enhancer refinement, and machine learning-based sequence classification.

Chapter 2 presents TLoops, a novel computational tool for accurate and rapid chromatin loop identification from Hi-C data. Most existing loop callers are based on the fractal globule theory, which assumes chromatin forms a knot-free continuous fractal path. However, recent studies have demonstrated that approximately 50% of chromosomes contain topological knots, challenging this assumption and suggesting that loop extrusion facilitates both dynamic loop formation and knot resolution. TLoops implements a three-step strategy to address these challenges. First, it employs a dynamic sliding window that replaces fixed binning, consisting of an inner window that slides horizontally and an outer concentric window that expands outward, using the Poisson test to detect significant density differences. Second, it applies affinity propagation clustering (AP-Cluster), a graph-based algorithm that automatically determines the optimal number of clusters without pre-specifying parameters. Third, it performs statistical testing with a modified mathematical expectation that accounts for unknotted TAD effects, adjusting the power-law decay exponent for loops in regions where the original TAD structure has been weakened. Benchmarking against HiCCUPS, Fit-Hi-C, and HOMER on mouse Hi-C datasets showed that TLoops achieved comparable loop counts while demonstrating superior performance in aggregated peak analysis, indicating higher signal-to-noise ratios. The tool exhibited linear sensitivity across varying sequencing depths ($R^2 > 0.961$) and successfully identified approximately 83% TAD-associated loops and 17% independent untangling-associated loops. This work provides a valuable tool that explicitly models the relationship between topological constraints and chromatin loop

formation, offering improved detection of loops that other methods may overlook.

Chapter 3 describes scTSS-seq, a single-cell genomic technology developed for genome-wide identification of alternative transcription start sites. Alternative TSS usage affects over 50% of genes in mammals and plays critical roles in gene regulation, yet bulk-cell methods like CAGE only reveal average signals across cell populations and cannot detect cellular heterogeneity. The scTSS-seq method integrates the Fluidigm C1 system for single-cell isolation, SMARTer chemistry for full-length cDNA synthesis, and dual priming oligonucleotides for specific 5'-end enrichment. Applied to naïve and IFN-β-pretreated mouse embryonic fibroblasts, the method demonstrated robust performance, detecting a median of 5,002 TSSs per cell. Quality filtering removed low-quality cells, and the remaining high-quality cells showed strong correlation with bulk RNA-seq data (Spearman rho = 0.65–0.71). t-SNE analysis clearly separated the two cell populations, confirming the method's ability to detect cellular heterogeneity. Remarkably, 81% of expressed genes contained multiple TSSs, substantially exceeding the 58% predicted by CAGE and 54% by empirical determination. Importantly, memory interferon-stimulated genes exhibited significantly larger numbers of alternative TSSs in pretreated cells compared to naïve cells ($p < 0.001$), suggesting that alternative TSS usage contributes to transcriptional memory by providing a platform for rapid RNA Polymerase II recruitment. This study provides a validated experimental protocol and a comprehensive resource of alternative TSS annotations for future investigations.

Chapter 4 introduces MAE-seq (Massive Active Enhancers by Sequencing), an experimental method that identifies functional regulatory elements at 25-bp resolution, dramatically refining genome annotation compared to conventional techniques. The lengths of regulatory elements defined by ChIP-seq, DNase-seq, and STARR-seq typically range from 200–1000 bp, yet transcription factor binding motifs span only 6–14 bp, indicating substantial sequence redundancy. MAE-seq addresses this by synthesizing random 25-bp DNA fragments (selected as the optimal length based on alignment simulation showing ~80% unique mapping rate) and cloning them into a fluorescent reporter vector with a minimal promoter. Following transfection, fluorescence-activated cell sorting enriches mCherry-positive cells, and a Poisson-based statistical model identifies significantly enriched fragments as active enhancers. Applied to mouse embryonic stem cells, MAE-seq identified 626,879 25-bp enhancers, with at least 80% validated by dual-luciferase assays. The "L-E-R" dissection experiment demonstrated that the 25-bp core was both necessary and sufficient for enhancer activity. Strikingly, 35% of MAE-seq enhancers were classified as "novel" because they lacked any overlap with known epigenetic marks or annotations, yet 80% showed enhancer activity in validation assays. Integration with Hi-C data revealed that 55.4% of novel enhancers participate in chromatin loops.

Functional analysis of three enhancers targeting the Cdh1 gene revealed a hierarchical regulatory model where only the triple knockout significantly reduced Cdh1 expression and altered cell proliferation. The method also enabled precise one-to-one enhancer-TF analysis, revealing that most enhancers bind a single cell-specific transcription factor, and demonstrated differential cofactor dependencies: known enhancers depend on COFs family members while novel enhancers rely on chromatin remodeling complexes.

Chapter 5 presents a multiple kernel learning (MKL) classifier developed to identify enhancer core sequences from MAE-seq data. Support vector machines were selected as the base classifier due to their robust cross-species prediction performance. Comparisons of DNA sequence vectorization methods revealed that gapped k-mer combined with position weight matrix (gkm+PWM) features substantially outperformed one-hot encoding for shallow classifiers, capturing both local compositional patterns and regulatory motif preferences. The MKL model combined a linear kernel (30% weight) to preserve interpretability for conserved motifs with an RBF kernel (70% weight) to model complex non-linear interactions, achieving superior performance (AUC = 0.755, AP = 0.78) compared to single-kernel SVMs. When applied to 1 million randomly sampled 25-bp sequences from the mouse genome, the classifier predicted 23,711 candidate enhancers, of which 82.3% overlapped with known ChIP-seq enhancer peaks, demonstrating high predictive accuracy. Strikingly, introducing as few as two substitutions within a 25-bp core led to loss of enhancer classification for 94 out of 100 tested variants, revealing strong sequence constraint. The MKL-based system enables rapid genome-wide identification of core enhancer sequences, substantially reducing computational costs compared to methods analyzing longer regions, and provides a non-experimental approach for generating novel candidate sequences for subsequent experimental validation.

Chapter 6 synthesizes the key findings and discusses broader implications, limitations, and future directions. In Hi-C data analysis, chromatin loop identification requires high resolution but demands massive sequencing depth, with deep learning methods (HiCPlus, HiCNN, DeepHiC) offering promising alternatives to increase resolution from low-depth data. The lack of unified standards for loop calling remains a challenge, as different algorithms produce different results and are often evaluated by inferred outcomes rather than principled standards. TLoops contributes to addressing these issues through its local feature-based approach that explicitly models topological untangling effects. MAE-seq's major advance is refining regulatory elements to 25-bp resolution, but limitations include poor inter-batch reproducibility, high signal randomness, and extensive downstream validation requirements. Strategies to address these limitations include accumulating confirmed sequences into

a target database and expanding data volume. The chapter explores the potential of large language models for DNA sequence analysis, noting that treating DNA as a natural language offers advantages including end-to-end learning and modeling of complex dependencies, with a preliminary n-gram simulation showing promising results on MAE-seq sequences. Current limitations of large language models in genomics include computational demands, dependence on large training datasets, poor interpretability, and sensitivity to tokenization. Promising future directions include multi-modal integration of Hi-C with microscopic imaging and spatial coordinates, time-resolved integration with transcriptomics and epigenomics to determine whether structural changes are causes or consequences of gene expression, and leveraging large foundation models to learn unified representations across different data modalities for cross-modal prediction.

Appendices

ABOUT THE AUTHOR

Qitong Huang was born in Hanzhong, Shaanxi Province, China. From the moment he entered the computer science program at university, he developed a passion for applying computational and mathematical knowledge to solve problems across various computational methods. His undergraduate thesis focused on embedding textual information within the header files of image files, involving the design and implementation of related encoding and decoding schemes. This experience later proved highly inspiring when he began working with DNA sequences. He earned his Bachelor of Engineering degree from Shenzhen University in 2013, followed by a Master of Science degree from Hong Kong Baptist University. During his master's studies, he gained extensive exposure to machine learning algorithms and applied them in project work to optimize and deploy models.

After completing his master's degree, he joined the Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, as a bioinformatics engineer, where he was primarily responsible for analyzing various types of omics data, with a particular focus on three-dimensional genomics. Throughout his research and analysis, he continuously sought to leverage diverse algorithms to address challenges in omics analysis—an endeavor he found akin to an exciting adventure.

In 2019, he enrolled in a joint doctoral program between Wageningen University and the Chinese Academy of Agricultural Sciences, joining the Animal Breeding Group. His doctoral research centered on interpreting three-dimensional genomics data and designing algorithms for identifying transcriptional regulatory elements, using multiple types of omics data as a foundation to decipher the remarkable codes embedded within DNA sequences. This thesis presents the findings of his doctoral research.

During the period spanning his professional career and doctoral studies, his two sons were born, and he remains deeply grateful for the arrival of both children. To balance work and family, he shuttled between China and the Netherlands. Meanwhile, he also weathered changes within his team in China, and health issues among family members caused him considerable concern. Fortunately, both his promoter and supervisor offered unwavering support for his academic pursuits and personal life. His doctoral journey deepened his passion for his field and reinforced his commitment to a lifelong career in bioinformatics.

LIST OF PUBLICATIONS

Zhu X†, **Huang Q†**, Huang L†, Luo J, Li Q, Kong D, Deng B, Gu Y, Wang X, Li C, Kong S, Zhang Y. MAE-seq refines regulatory elements across the genome. Nucleic Acids Res. 2024 Jan 25;52(2):e9. doi: 10.1093/nar/gkad1129. PMID: 38038259; PMCID: PMC10810209.

Peng Y†, **Huang Q†**, Liu D, Kong S, Kamada R, Ozato K, Zhang Y, Zhu J. A single-cell genomic strategy for alternative transcript start sites identification. Biotechnol J. 2024 Mar;19(3):e2300516. doi: 10.1002/biot.202300516. PMID: 38472100.

Huang Q, Li Q, Zhang Y. Linking chromatin conformation to gene function. Yi Chuan. 2020 Jan 20;42(1):1-17. doi: 10.16288/j.yczz.19-257. PMID: 31956093.

Zhu X†, **Huang Q†**, Luo J, Kong D, Zhang Y. Mini-review: Gene regulatory network benefits from three-dimensional chromatin conformation and structural biology. Comput Struct Biotechnol J. 2023 Feb 16;21:1728-1737. doi: 10.1016/j.csbj.2023.02.028. PMID: 36890880; PMCID: PMC9986247.

Li D†, **Huang Q†**, Huang L, Wen J, Luo J, Li Q, Peng Y, Zhang Y. Baiting out a full length sequence from unmapped RNA-seq data. BMC Genomics. 2021 Nov 27;22(1):857. doi: 10.1186/s12864-021-08146-4. PMID: 34837950; PMCID: PMC8626966.

Huang Q†, Siyuan Kong†, Jing Luo†, Martien A.M. Groenen, Ole Madsen and Yubo Zhang. Accurate and fast loop calling from Hi-C data by TLoops. Submitted.

Qitong Huang†, Xiusheng Zhu†, Jing Luo1, Martien A.M. Groenen, Ole Madsen and Yubo Zhang. Submitted.

TRAINING AND SUPERVISION PLAN

Training and Supervision Plan (TSP)

Graduate School WIAS

Section 1. GENERAL INFORMATION AND LEARNING GOALS

Name PhD candidate	Qitong Huang	
Project title	3-dimensional transcriptional regulation in mammalian cells and their application in animal breeding	
Group	Animal Breeding and Genetics	
Supervising team:		
	Name:	Daily supervisor?
Intended (co)promotor	Madsen Ole (Dr.)	☑
Intended promotor	Groenen Martien(Dr. Prof.) Yubo Zhang(Dr.	☐
PhD study advisor	Janneke van Seters	
Project term	from 12-06-2019 until 22-6-2023	
Submitted	Sep-19first plan	

To start	**Dowe?**
Introduction interview with WIAS secretery:	☑
Introduction interview with WIAS education coordinator and PhDconfidant:	☑
Competence assesement (MyTalent):	☑
Evaluation of learning goals with PhD study advisor at start of PhD training:	☑
Evaluation of learning goals with PhD study advisor at mid-term:	☑

Personal learning goals

What do you want to learn during your PhD study at WIAS

Formulate at least 3 detailed and specific learning goals in the TSP Appendix Personal Learning Goals.
Submit this separate document together with the TSP

Section 2. SUPERVISION

Agreements made between daily supervisor(s), supervisor(s) and PhD student
Discuss with your supervisor(s) and fill in at least the questions below; extend with potential more agreements
Meetings with daily supervisor will take place every week
Urgent questions will be answered within one day
Meetings with supervisor / supervisory committee will take place every 6 months
Feedback on manuscripts will be given within 2 weeks
Progress and (adjustment of) planning will be discussed at least 6 months

In month 14 of the study, progress must be evaluated, resulting in a go/no go decision before month 18.

EDUCATION AND TRAINING (minimum 30 credits)

A. The Basic Package	year	credits *
WIAS Introduction Day **(mandatory)**	2019	0.3
Course on philosophy of science and/or ethics **(mandatory)**	2019	1.5
Course on essential skills *(recommended)*	2019	1.2
Subtotal Basic Package		**3**

B. Disciplinary Competences *(minimum 2 courses)*	year	credits
Research proposal	2019	6.0
PhD and postdoc training school on CHIP-seq (wet-lab) and basic functional animal genome analysis	2019	1.5
WIAS/PE&RC advanced statistics course Design of Experiments	2019	0.8
R and Big Data	2019	0.6
Basic Statistics	2019	1.5
Introduction to R for Statistical Analysis	2019	0.6
Statistical Uncertainty Analysis of Dynamic Models	2019	1.5
Subtotal Disciplinary Competences		**13**

C. Professional Competences *(minimum 2 courses)*	year	credits
Scientific Publishing	2019	0.3
Personal Leadership The Basis of Effective and Efficient Academic Development	2018	3.0
Brain friendly working and with	2019	0.3
Project and Time Management	2019	1.5
Brain Training	2019	0.3
Research Methodology: From topic to proposal	2019	4.0
Competence Assessment	2019	0.3
Career assessment	2019	0.3
Subtotal Professional Competences		**9**

D. Presentation Skills *(maximum 4 credits)*		year	credits
MAE-seq refines regulatory elements across the genome, BioSB 2023 Conference , Egmond, the Netherlands	Poster	2021	1.0
Exploring the Three Dimensional Genome of the NLRP3 Gene in diabetic cardiomyopathy, New Adivances in Traditional Chinese and Western Medicine Diagnosis and Treatment of Diabetes and Its Complications conference, Shenzhen, China	Oral	2025	1.0
Subtotal presentations			**2**

E. Teaching competences *(max 6 credits)*	year	credits
Supervising 2 MSc major students	2021-2022	4.0
Subtotal Teaching competences		**4**

Education and Training Total (minimum 30 credits)*	**31**

**One ECTS credit equals a studyloed of approximately 28 hours*

ACKNOWLEDGEMENTS

As I begin to write this acknowledgment, my thoughts play back like a film, scene by scene, from 2019 to the present, and my PhD journey is now drawing to a close. When I first entered bioinformatics, I saw it as an independent programmer processing data. Experimentalists were the producers of data, and supervisors were the validators of analytical results. Fortunately, my PhD life corrected this misconception, teaching me that scientific research requires constant communication and mutual learning—we are not isolated individuals processing data, but parts of an organic whole. This thesis is no exception, and I would like to express my heartfelt gratitude to everyone who has encouraged and helped me throughout my doctoral journey.

First and foremost, I would like to express my deepest gratitude to my supervisors.

It is a great honor to be a member of Professor Dr. Martien Groenen's group. Before I set foot in the Netherlands, you took the time to share your research journey and insights with me during our meeting in Beijing. Your encouragement to think independently and your unwavering support have helped me stay on course throughout this process. Every word of guidance you've given me has shaped the way I approach my research. For all of it, I am deeply grateful.

Likewise, I sincerely appreciate my co-promotor, Dr. Ole Madsen. You cared for my academic progress and personal life in every possible way. You were patient with my shortcomings, pointed them out promptly, and guided me towards the right direction. Whenever I came cross any difficulties, you were like a strong hand supporting me from behind. You read every one of my papers carefully and helped me identify problems, guiding me toward improvement. Although I rarely expressed my emotions, I have always been grateful in my heart for your care and dedication.

At the same time, I would also like to thank my Chinese promoter, Prof. Dr. Yubo Zhang. It was you who brought me into the vast world of bioinformatics. Your full trust in me gave me the freedom to explore and inspired me to keep improving. I feel so lucky to have crossed paths with you.

I would also like to thank my colleagues in the Animal Breeding and Genomics (ABG) group—your enthusiasm and erudition have helped me enormously. Siyuan X, Langqing, Zhou, Ibrahim, Shuwen, and Lim, you were my close companions when I first arrived at ABG, helping me overcome all kinds of difficulties in a foreign land. Lucas, Helene, Yun, Xiaofei, You, Wenye, Aixin, Rui, Zhuoshi, Xue,

Liyan, Jifan, Junxin, Huifang—discussing problems with you was always thought-provoking, and thank you also for the parties. My sincere thanks also go to our chair group members, secretaries, and the graduate school staff, especially Fadma and Paddy. You gave me tremendous support in administrative and various other matters.

I would also like to thank my teachers and classmates at the Chinese Academy of Agricultural Sciences. Lijing was a colleague I worked with back when I was still a bioinformatics engineer—I will never forget the email with my first interview invitation that you sent me. I feel fortunate that we also studied and researched together in the Netherlands. Lei, Yanling, Siyuan K, Dongwei, Xiusheng, Qing, Jing, Qiujia, Yajuan, Xueyan, Shuai, Danli—thank you for your care and help. Shanshan, Jiaqi, Cong, Weijia, Mengshuai, Yi, Xiao, Jianing, Li, Yening, Chenglong, Haozhou, Zhenbiao, Xingjuan, Wenyuan—I am honored to have been your classmate and to have grown together with you.

Wageningen is a quiet place, and that quiet gave me more time to think and to settle myself. I feel fortunate that in such a quiet place, there is a group of warm-hearted mentors and friends who encouraged me and helped me grow. This is a period of time I will never forget.

最后，我想把最深挚的感谢送给我的家人。我的父母给了我最坚实的支持，是我起航时源源不断的动力，是我无力时永远能回望的港湾。有幸能成为你们的孩子。然后是我的妻子，也是我从高中就相识的同学。感谢你一路的理解与陪伴，陪我面对生活中的一个个挑战。我同样要感谢我的岳父岳母，你们的支持、付出与信任同样是我坚实的后盾。最后，十分感激我的两个儿子维岳和维烈，感谢你们的到来，你们是我不断向前的动力。

希望你们永远健康平安！

This work was supported by the National Key Research and Development Program of China [SQ2023YFF1000085, 2018YFA0903201]; National Natural Science Foundation of China [31970592]; The Agricultural Science and Technology Innovation Program; The Elite Young Scientists Program of Chinese Academy of Agricultural Sciences [CAASQNYC-KYYJ-41]; National Natural Science Foundation of China (No. 32002173); Natural Science foundation of Guangdong Province, China [2018A0303130009]; the Projects Funded by China Postdoctoral Science Foundation (No. BX2021367 and 2021M703543); the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515010766); the Shenzhen Science and Technology Program (Grant No. KCXFZ20201221173205015 and RCBS20210609104512021).

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Qitong Huang
Printed by ProefschriftMaken.nl

