Computational and Structural Biotechnology Journal 27 (2025) 705–716

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Identifying associations between short tandem repeat sequences and gene expression in yeast reveals specific repeated motifs encoding transcriptional regulatory proteins

Zongyuan Yu [a,b,1], Yating Liang [b,1], Meida Xiang [b], Kang Xu [c], Xiang Xu [b], Dongyang Ran [d], Yawen Luo [b], Bijia Chen [b], Xiaochen Bo [b,*], Hebing Chen [a,b,**]

[a] School of Basic Medical Sciences, Anhui Medical University, Hefei 230032, China
[b] Academy of Military Medical Science, Beijing 100850, China
[c] School of Software, Shandong University, China
[d] School of Basic Medical Sciences, Henan University, Kaifeng, China

ARTICLE INFO

Keywords:
Short tandem repeats (STRs)
Saccharomyces cerevisiae
Coding sequences
Transcriptional regulation
Epigenetic characteristic

ABSTRACT

Tandem repeat sequences (TRs), a class of repetitive genomic elements, are broadly distributed in both coding and non-coding regions. Investigating the relationship between sequences and function is essential for understanding the genome. *Saccharomyces cerevisiae* serves as a vital model organism and is widely used as an engineered strain. Although the transcriptional regulatory functions of TRs in the promoters of *S.cerevisiae* have been elucidated, our understanding of their roles within coding sequences (CDS) remains limited. In this study, we integrate RNA-seq, ChIP-seq, ATAC-seq, Hi-C, and Micro-C data from *S.cerevisiae* to analyze the types and distribution of TRs, and their impact on gene expression. Our results indicate that genes containing short tandem repeats (STRs) in their CDS exhibit lower expression levels. Epigenetic analysis reveals that these regions are characterized by high levels of repressive histone modifications and low levels of activating marks, with reduced chromatin accessibility and fewer chromatin interactions. Furthermore, trinucleotide and hexanucleotide repeated motifs of STR are found primarily enriched in genes encoding transcriptional regulatory proteins. This study provides new insights into the functions and characteristics of STRs in the CDS of *S.cerevisiae*. The identification of key STR motifs offers potential targets for the design of transcriptional regulatory elements.

1. Introduction

Eukaryotic genomes contain a large number of transposable elements (TEs) and tandem repeats sequences (TRs) [1]. TRs exist in both noncoding and coding regions. It has been a common misconception that repetitive sequences were functionless [2]. Recent studies have demonstrated that TRs sequences play crucial roles in evolution, phenotypic diversity, gene expression regulation, gene silencing, and chromatin organization in human genome [3–7]. STRs can regulate gene expression by affecting the binding affinity of transcription factors [8]. Meanwhile, their mutational susceptibility significantly correlates with diseases [9–12]. *Saccharomyces cerevisiae* as a model eukaryotic organism and a commonly used engineered strain [13,14], investigating the relationship between TRs and transcriptional regulation, and their roles, will promote the understanding of DNA sequence function and facilitate the development and design of new regulatory functional elements.

The effects of TRs in promoter regions on transcriptional regulation have been well studied [15,16]. TRs in promoter have a high AT content. Variations in TR length and AT repeat frequency can affect transcription. These repeat sequences contribute to the formation of variable nucleosome-free DNA structures, directly impacting local chromatin structure. Furthermore, this can alter the accessibility of regulatory elements and the recruitment of transcription machinery, ultimately affecting gene expression [17–19]. Variable repeat sequences in promoter regions facilitate rapid evolution of gene expression and exhibit higher responsiveness to changing environmental conditions [20].

* Corresponding author.
** Corresponding author at: School of Basic Medical Sciences, Anhui Medical University, Hefei 230032, China.
E-mail addresses: boxc@bmi.ac.cn (X. Bo), chb-1012@163.com (H. Chen).
[1] Zongyuan Yu and Yating Liang contributed equally.

https://doi.org/10.1016/j.csbj.2025.02.003
Received 27 September 2024; Received in revised form 5 February 2025; Accepted 7 February 2025
Available online 14 February 2025

Although TRs in promoter regions are important for gene regulation, TRs in coding sequence (CDS) can also enhance the evolvability of proteins [21–23]. Variations within coding regions that alter protein function support many biological adaptations. One study reported a typical functional gene, FLO1, containing TRs adjacent to 3' end, which is associated with flocculation function in *S.cerevisiae* [24,25]. As for the characterization of TRs, previous studies primarily focused on physical features such as repeated motif type, motif length, and repetition frequency [15,26]. However, our understanding of TRs' function in coding regions and characteristics are still limited. Studying the specific protein function of TRs drives the understanding of the association between DNA sequences and function.

To accurately identify and characterize the functions of TRs in *S. cerevisiae*, we integrated RNA-seq, ATAC-seq, ChIP-seq, Hi-C and Micro-C data from *S.cerevisiae*. Through investigating the association of TRs with gene expression, we found that STRs with different motifs and locations have distinct functions, with trinucleotide and hexanucleotide repeats are enriched within genes encoding proteins that participate in transcriptional regulation. Next, given their suppressive effect on gene expression, we investigated the epigenetic characterization of STRs, including histone modifications, chromatin accessibility, nucleosome occupancy, and chromatin interactions. The study of STRs in CDS will help to understand the association of DNA repeat sequences with transcriptional regulation function, providing guidance for designing regulatory functional elements.

2. Materials and methods

2.1. Data source

High-throughput sequencing data of *S.cerevisiae* strain BY4741, including Hi-C, ATAC-seq and RNA-seq, are from our laboratory and have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number PRJNA1073072. Hi-C, ATAC-seq and RNA-seq for yZSJ025 (GSE219048, GSE168182), Micro-C for BY4741 (GSE85220), histone ChIP-Seq for BY4741 (GSE61888), and RNA-seq for human NHEK (GSE78594) are accessible from GEO datasets.

2.2. Extraction of TR

We employed the reference genome S288C for BY4741 strain. For yZSJ025 strain, the reference genome was sourced from the GEO DataSet with the identifier GSE168182. The human NHEK cell line utilized the reference genome version hg19. We utilized TRF (v4.09) to identify tandem repeats within the reference genomes [27]. By inputting a reference genome, a combination of sequence alignment and statistical methods is used to detect successive repetitive motifs. The parameters were set as follows: Match was set to 2, Mismatch to 7, Delta to 7, the PM (probability match) threshold for alignment was set at 80, PI was set to 10, the Minscore was set at 50, and the maximum repeat unit length in base pairs (Maxperiod) was set to 500.

2.3. Definition of entropy values for TRs

The entropy values for TRs were defined using TRF (v4.09) through an alignment process aimed at aligning for detection candidates. The consistency of sequences can be determined based on the magnitude of these entropy values. Each repeat is assigned an entropy value ranging from 0 to 2. Higher entropy values indicate greater sequence complexity, whereas lower entropy values signify greater sequence uniformity.

2.4. Hi-C data processing

Raw sequencing data were first pre-processed using TrimGalore (https://github.com/FelixKrueger/TrimGalore) and quality filtered using the following parameters: -q 20, –length 35 and -e 0.1. Filtered FASTQ files were then processed using HiC-Pro (v3.1.0) to align reads and correct for biases [28]. The configuration file settings included the enzyme recognition site GATCGATC, MAX_ITER = 100, FILTER_LOW_COUNT_PERC = 0.02, FILTER_HIGH_COUNT_PERC = 0, and EPS = 0.1.

2.5. Micro-C data processing

Quality control for the raw sequencing data followed the same procedures as for Hi-C data. The data processing for Micro-C utilized the pipeline provided by Micro-C (https://micro-c.readthedocs.io/en/latest/). This involved several steps, including: aligning reads to the reference genome using BWA [29], using Pairtool (parse) to identify proximity ligation events in the Micro-C library [30], recording the outermost (5') paired alignments and writing them to a pairsam file when a ligation event was identified in the alignment files, sorting the parsed fragments using Pairtools (sort), marking PCR duplicates using Pairtools (dedup), splitting the final pairsam file into bam and pairs files using Pairtools (split), sorting the bam file using Samtools (sort) [31], and finally, pairs files were converted to hic file format using juicertools for further analysis.

2.6. ATAC-seq data processing

Raw data quality control follows the same procedure as Hi-C. Reads were aligned to the reference genome using Bowtie2 [32], with the parameters –very-sensitive and -X 2000. The Sambamba markdup module was used for marking and de-duplication of BAM files [33]. Use samtools for further processing of bam files, including indexing and sorting. Finally, the bam file was transformed into a bw format file using bamCoverage, available at (https://github.com/gartician/deepTools-bamCoverage) with standardized use of CPM. The bw files were then used by the computeMatrix and plotHeatmap modules of deeptools to generate accessibility signal maps and heat maps of TR sites [34]. The callpeak module of MACS2 software was used to identify the peak sites of chromatin openness [35].

2.7. RNA-seq data processing

Firstly, genomeGenerate module of STAR was used to index the reference genome [36]. Subsequently, STAR software was used to compare the quality-controlled reads to the reference genome. Following this, the rsem-prepare-reference module of RSEM [37], a gene expression quantification software, was used to index the bam file. Finally, the rsem-calculation-expression module was used to quantify the gene expression.

2.8. Histone signal map

To generate the histone signal map, we used the computeMatrix reference-point module from the deeptools software package with the following parameters: –referencePoint center –missingDataAsZero -b 4000 -a 4000 –binSize 50, taking the histone data in bw file format as input to calculate the histone signal matrix for TR regions and their 4 kb upstream and downstream regions [34]. Subsequently, we utilized the plotProfile module to visualize the histone signal map centered on the TR regions.

2.9. Calculation of significant interactions

Using HiCPro2FitHiC.py script (https://github.com/nservant/HiC-Pro/releases), the transformation of HiC-Pro format to Fi-Hi-C recognizable format. These three converted Fit-Hi-C files are then used as input for Fit-Hi-C [38], with the parameter x set to All. The significant

interaction sites are visualized using the Circos software package [39]. For intrachromosome interactions, those with a count number greater than 500 and a significance P-value less than 0.001 are selected for visualization. Similarly, for interchromosome interactions, those with a count number greater than 50 and a significance P-value less than 0.001 are chosen for visualization.

2.10. Chromosome interaction circos plot

To illustrate the chromosome interactions using a circos plot, we employed R (v4.3.2). Initially, the RCircos package was loaded. Following this, we established the plotting area with RCircos.Set.Plot. Area, depicted the chromosome ideograms using RCircos.Chromosome. Ideogram.Plot, drew the FIRE signal curves with RCircos.Line.Plot, marked the positions of TRs via RCircos.Scatter.Plot, and finally represented both intra- and inter-chromosomal interactions through RCircos. Link.Plot.

2.11. Calculating FIRE score

The FIRE score was calculated based on the method described by Schmitt, AD [40]. First, the F_GC_MAP.file.sh (https://github.com/aylab/HiCnv) script was used to calculate the effective fragment length, GC content and mappability of the genome. Second, a sparseToDense.py (https://github.com/nservant/HiC-Pro/blob/master/bin/utils/sparseToDense.py) script was used for matrix transformation, and contact sums were calculated at a resolution of 2 kb for each bin within a distance of 200 kb. Finally, using the local interaction data and the F_GC_MAP files as inputs, the FIRE score is calculated using the HiCNormCis algorithm, which is based on Poisson regression normalization [41].

2.12. GO enrichment analysis

GO enrichment analysis was performed using the clusterProfiler software package based on R language [42], and the annotated data were from org.Sc.sgd.db (https://bioconductor.org/packages/release/data/annotation/html/org.Sc.sgd.db.html) Package, which provides the necessary background information. The significance threshold was set to 0.05.

2.13. Visualization of TR region aggregation matrices

TR Regions Aggregation Matrix Using FANC Kit (https://fan-c.readthedocs.io/en/latest/getting_started.html) [43]. The matrix uses Micro-C data, first pre-processing the TR regions, locating the TRs into a 2 kb bin as the TR-related regions, and using the processed.hic file and the TR-related regions file as the input of the fanc. The parameter settings include: –expected-norm to normalize the matrix to the expected value; –log: log2 transforms the normalized matrix; The –relative parameter is set to 4, which takes each region as a relative extension of the fraction of the region length. The final area in the image will be: <start- 4 * 2000 > to <end+ 4 * 2000 > .

2.14. Statistical analysis

All statistical analyses and graphs were generated using R (v4.3.2) and Linux. The statistical test method used the Wilcoxon rank-sum test and the Fisher's precision probability test, with a p-value of less than 0.05 considered statistically significant.

3. Results

3.1. Genes containing STRs in CDS exhibit lower expression compared to other TR types

To study the relationship between TRs and gene expression in *S. cerevisiae*, we selected both a wild-type strain BY4741 and a synthetic strain yZSJ025, which consists of six synthetic chromosomes with some deletions of transposable elements (TEs) [44–47]. Statistical analyses revealed a significant reduction in the count of TEs and a slight decrease in TR counts in synthetic chromosomes, leading to differences in chromosome size. Conversely, no differences were observed in non-synthetic chromosomes (Figure S1A-C). Both strains exhibited similar chromatin interaction patterns but with differing interaction frequencies (Figure S1D and E). Comparing the local chromatin structures of wild chromosome (chr4) and synthetic chromosome (chr5), chr4 showed similar local structures in both strains, while TE-depleted chr5 exhibited deviations in local interaction heatmaps (Figure S1F). Additionally, RNA-seq and ATAC-seq data were highly correlated between strains (Figure S1G and H). Therefore, despite the altered sequence of yZSJ025, TRs remained essentially unchanged, enhancing the robustness of our conclusions through comparative analysis.

TRs were categorized into four groups based on the length of their repeat motifs: Single base repeats (Single, 1 bp), Short Tandem Repeats (STR, 2–6 bp), Long Tandem Repeats (LongTR, 7–64 bp), and Super Long Tandem Repeats (sLongTR, >64 bp). The corresponding minimum repeat counts for these four categories (Single, STR, LongTR, sLongTR) were 25, 5, 2, and 2, respectively, while the average repeat counts were 33, 17, 4, and 4 (Figure S2A). One characteristic of the TRF algorithm we used lies in its capability to determine the consensus pattern of the smallest repeating unit within tandem repeats [27]. This implies that the TRF software is able to recognizing the fundamental, shortest repetitive units that constitute larger repetitive regions. For instance, the sequence GATGATGATGAT is described as four repeats of the trinucleotide GAT, rather than two repeats of the hexanucleotide GATGAT. To gain insights into the multi-omics landscape of these TR regions, we conducted an integrative analysis using Hi-C, ATAC-seq, RNA-seq, and histone modification datasets (Fig. 1A). Our quantitative analysis revealed LongTRs to be the most abundant, followed by STRs (Fig. 1B). The minor variations in TRs numbers between the two strains indicated the robustness of our TRs screening method. In addition, considering the enrichment of repetitive sequences in telomeres and centromeres [48–50], we specifically analyzed TRs in these regions and found LongTRs to be particularly abundant (Figure S2B). Since gene expressions in telomeres and centromeres were generally lower than in other genome regions [51,52], we excluded genes in these regions to avoid potential interference.

Previous studies have reported that different types of TRs regulate gene expression at both transcriptional and post-transcriptional levels through cis- and trans-regulatory mechanisms [53–55]. We first compared the expression of genes containing various TRs types and discovered that gene expression in TR regions was significantly reduced, particularly for STRs (Fig. 1C and S2C). To verify the conservation of these findings, we scanned TRs in NHEK cell lines and compared gene expression, finding TRs significantly enriched in coding regions with lower gene expression (Figure S2D and E). As TRs occur across a wide spectrum of genic and intergenic regions [56], we further investigated their impact on gene expression by comparing their genomic distributions and observed significant differences among various TR types. Specifically, Single-base repeats were relatively scarce in CDS and primarily enriched in promoter regions, whereas sLongTRs were almost exclusively enriched in CDS (Figure S3A). We classified TRs based on their location relative to CDS into Up-CDS (upstream of CDS), Down-CDS (downstream of CDS), and In-CDS (within CDS), and compared their gene expression. STRs within CDS regions showed significantly lower expression compared to non-TR genes (Fig. 1D and S2F). Due to the small number and uneven distribution of Single base

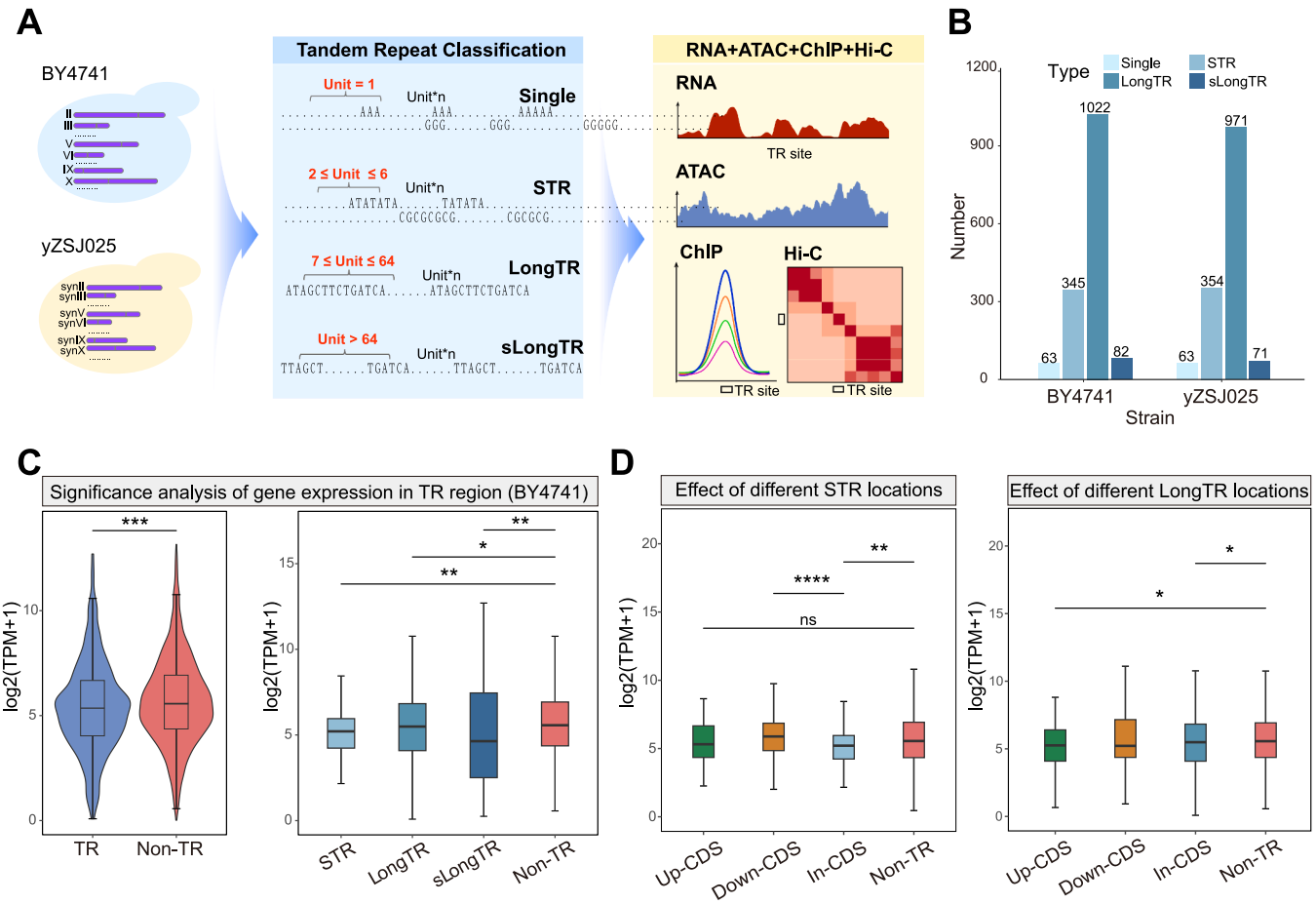


Fig. 1. Low expression of STR regions. (A) Investigation framework of this study. (B) Number of various TRs types in two strains. (C) Significance analysis of gene expression levels between genes containing various types of TRs and genes without any TRs in BY4741 strain. (D) Comparing expression levels of genes with STRs and LongTR located in Up-CDS, Down-CDS, In-CDS, and Non-TR regions. STR (left); LongTR (right). *P < 0.05, * *P < 0.01, * **P < 0.001, Wilcoxon test.

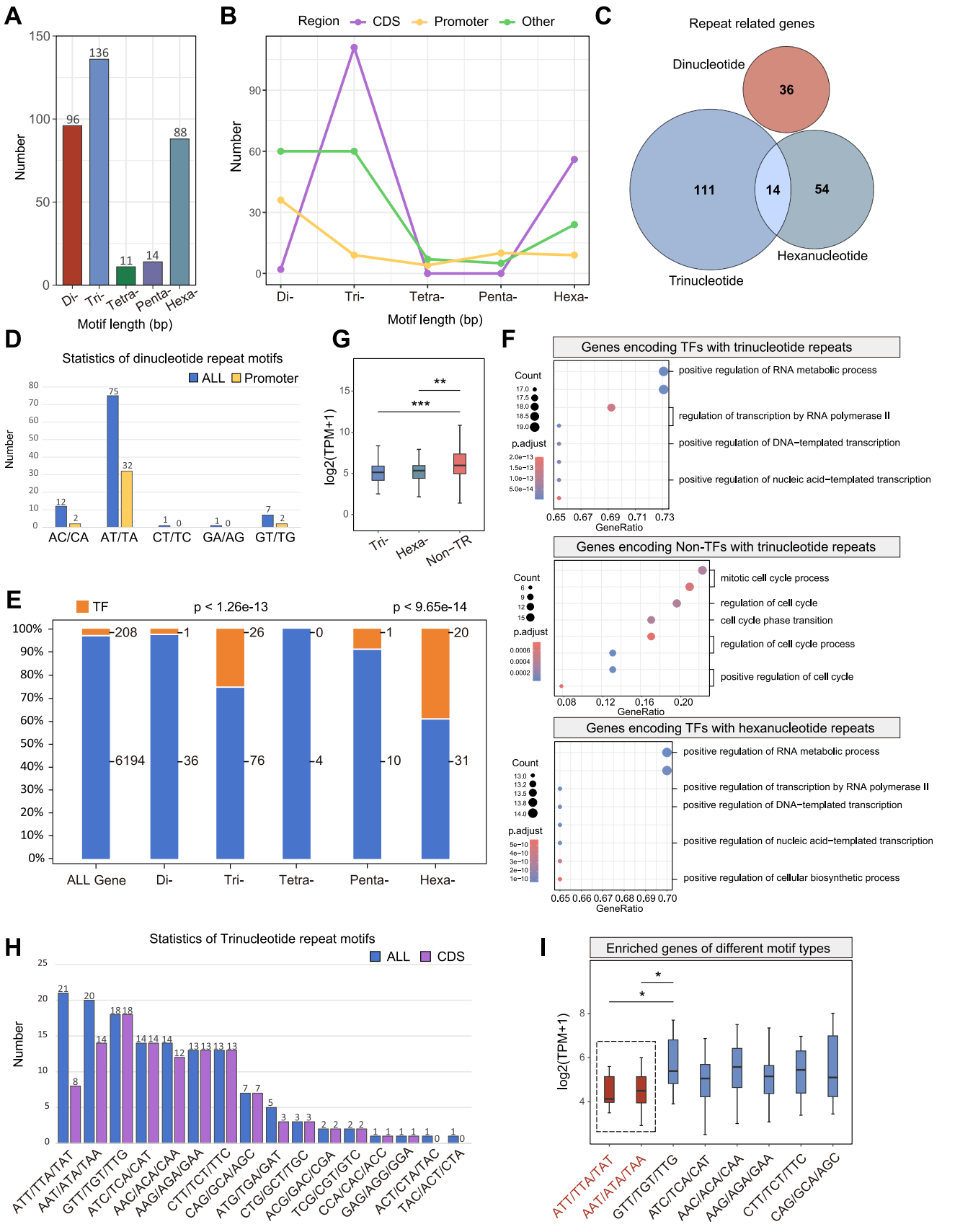
repeats and sLongTRs, we only compared STRs and LongTRs here. The varying gene expression among different types of TRs located in distinct positions indicated that TRs' influence on gene expression is position dependent. When STRs were located within the CDS, they exhibited the most significant suppressive effect on gene expression.

Based on the above findings, we explored the biological functions of these low-expressing genes containing various types of TRs. Genes with STRs in their CDS were primarily involved in transcriptional regulation mediated by RNA polymerase II, regulating RNA synthesis, and transcriptional activity. LongTRs exhibited similar functions to STRs. However, genes with sLongTRs in the CDS displayed certain differences, being closely associated with the structural dynamics and stability of DNA (Figure S3B-E). The above results indicated that genes containing STRs in their CDS had lower expression levels and played unique roles in transcriptional regulation. Next, we further investigated the functions and characteristics of STRs specifically within CDS regions.

3.2. Trinucleotide and hexanucleotide repeated motifs enriched in CDS of genes encoding transcriptional regulation correlated proteins

Differences in the repeat motifs and frequencies of TRs influenced their functional diversity. Based on their motif lengths, STRs were classified into dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide repeat sequences, with counts of 96, 136, 11, 14, and 88 respectively. This indicated that STRs in *S.cerevisiae* were mainly composed of dinucleotide, trinucleotide, and hexanucleotide motif lengths (Fig. 2A). Different motif lengths of STRs exhibited a non-random distribution. Trinucleotide repeats and hexanucleotide repeats were primarily localized in CDS regions, whereas dinucleotide repeat motifs were enriched in promoter regions (Fig. 2B). We examined whether genes enriched with different motifs overlap. Only 14 overlapping genes were found between those containing trinucleotide and hexanucleotide motifs. We selected genes containing a dinucleotide repeat motif in the promoter as a dinucleotide repeat-related gene and found no overlap with genes containing trinucleotide and hexanucleotide motifs (Fig. 2C). The genes containing different length motifs differed, making it suitable for studying their individual functions.

In addition, as the same length motif has different base types, we next examined the functions of genes corresponding to different STR motif types. AT/TA constituted the major of dinucleotide repeats (Fig. 2D), this aligned with previous studies that AT enriched in promoter regions [15,57]. According to the Saccharomyces Genome Database description, genes containing trinucleotide and hexanucleotide repeats were found primarily associated with transcriptional regulation, encoding large proportion of transcription factors (TFs) and transcriptional regulation correlated proteins (Supplementary Table 2). For instance, YKL028W containing trinucleotide repeats encoded the large subunit of TFIIE, recruiting RNA polymerase II to promoter [58]. YNL027W with hexanucleotide repeats encoded a TF activating stress-responsive genes [59]. YBR289W encoded a SWI/SNF chromatin remodeling complex subunit [60]. YOL051W encoded an RNA polymerase II mediator complex subunit [61]. Therefore, we compared TFs counts in genes with varying motifs lengths. Therefore, we compared the counts of TFs in genes with different motif lengths and found almost no genes encoding TFs associated with dinucleotide, tetranucleotide, and pentanucleotide repeats. In contrast, 26 out of 102 genes with trinucleotide repeats



(caption on next page)

Fig. 2. Different STR motifs serves distinct functions. (A) STRs are categorized into five types based on motif length (2–6 bp). The bar graph illustrates the counts for each STR motif type. Distinct colors represent different motif lengths. (B) Distribution of different STRs motif lengths in CDS, promoters, and other genomic regions. Colors indicate different regions. (C) Overlap of genes associated with dinucleotide repeats, trinucleotide repeats, and hexanucleotide repeats. (D) Statistical count of dinucleotide repeat motif types. The counts are located at the top of the bar graph. The blue color represents the number of all motifs, and the yellow color represents the number of motifs located in promoters. (E) Proportion of genes encoding TFs in different STRs motif types. Bar graphs are labeled with the number of genes on the right side. (F) Enrichment analysis of genes encoding TFs with trinucleotide repeats (top), enrichment analysis of genes encoding other proteins with trinucleotide repeats (middle), and enrichment analysis of genes encoding TFs with hexanucleotide repeats (bottom). Bubble size indicates the number of enriched genes, with color changes reflecting P.adjust significance levels. Functional descriptions marked in red are discussed in the article. (G) Comparative analysis of gene expression of trinucleotide repeats, and hexanucleotide repeats, and Non-TR. * *P < 0.01, * **P < 0.001, Wilcoxon test. (H) Statistics of trinucleotide repeat types. Counts are located at the top of the bar graph. Blue color represents all counts for that motif, and purple color represents the counts for that motif located in the CDS region. (I) Gene expression analysis of different trinucleotide repeat types. *P < 0.05, Wilcoxon test.

(p < 1.26e-13, Fisher's precision probability test) and 20 out of 51 genes with hexanucleotide repeats (p < 9.65e-14, Fisher's precision probability test) encoded TFs (Fig. 2E). Statistical analyses of amino acid sequences encoded by TRs in genes related to transcription regulation revealed a significant enrichment of Q (glutamine) and N (asparagine) (Figure S4A and Supplementary Table 3). After querying the SGD database, we found that 42 out of 46 transcription regulatory proteins had amino acids encoded by TRs that belonged to intrinsically disordered regions (IDRs) (Supplementary Table 3). The enrichment of glutamine and asparagine has been reported to correlate with IDRs of proteins. These regions lack a fixed three-dimensional structure but perform crucial cellular functions [62]. For instance, IDRs often serve as hubs in protein interaction networks, interacting with other proteins and playing a key role in regulating signaling pathways and essential cellular processes like transcription, translation, and the cell cycle [63–67]. Recent study reported that IDRs mediate specific interactions between transcription factors [68]. We hypothesized that IDRs within transcription factors encoded by trinucleotide and hexanucleotide repeat sequences may facilitate interactions among transcription factors, which require further validation.

Comparing the functions of TF-encoding and non-TF-encoding genes containing trinucleotide and hexanucleotide repeats (Fig. 2F), genes with hexanucleotide repeats that did not encode TFs showed no enriched functions, indicating that genes with trinucleotide and hexanucleotide repeats are mostly associated with transcriptional regulation. Additionally, genes containing trinucleotide and hexanucleotide repeats exhibited significantly lower expression levels compared to other genes without TRs (Fig. 2G). We then quantified the types of trinucleotide and hexanucleotide repeats motifs (Fig. 2H and S4B). Eight prominent types were identified among trinucleotide repeats, while hexanucleotide repeats types were diverse. We then compared the expression of these eight prominent types and found that genes containing ATT/TTA/TAT and AAT/ATA/TAA motifs exhibited the lowest expression levels (Fig. 2I). As previously noted, genes with trinucleotide and hexanucleotide repeats are primarily involved in transcriptional regulation. These genes may potentially serve as transcriptional regulatory elements, modulating transcriptional regulation by altering the corresponding motifs.

Additionally, we compared the distribution and function of TRs in the human genome. STRs were relatively enriched within genes (Figure S4C). Compared to yeast, dinucleotide and tetranucleotide repeats constituted the highest proportion, while dinucleotide and trinucleotide repeats were significantly enriched in exon regions (Figure S4D and E). GO enrichment analyses revealed that genes enriched with dinucleotide sequences play crucial roles in neural signal transmission and regulation, while genes enriched with trinucleotide sequences were closely related to transcriptional regulation, consistent with our findings in yeast (Figure S4F and G). This suggested that genes containing trinucleotide repeats exhibit a certain degree of functional conservation.

3.3. STRs exhibit high-repression and low-activation histone mark patterns

Previous study had reported that methylation of TRs can lead to the formation of repressive chromatin and transcriptional silencing, indicating that epigenetic modifications affect gene expression patterns [69,70]. Given the low expression levels of STR-associated genes, we investigated the epigenetic modification features of these motifs. Histone modification data were obtained from Nir Friedman et al. [71]. Initially, we mapped the signals of the repressive and activating histone modifications, H3K36me3 and H4K5ac, respectively, across different types of TRs regions (Fig. 3A). We found that H4K5ac signals were consistently lower across various types of TRs, whereas H3K36me3 signals were enriched in TRs regions except for single-base repeat regions. Interestingly, H3K36me3 was notably enriched in trinucleotide and hexanucleotide repeats. This suggested that the differential enrichment of histone modifications may be associated with gene expression levels in these regions. We further selected repressive histones including H3K79me3, H4R3me2s, H3K36me3, H4K20me, and H3K4me, along with activating histones such as H4K8ac, H4K12ac, H4K5ac, H2AK5ac, and Htz1. The signals of these ten histone modifications were mapped across TR and STR regions. Our findings revealed that repressive histone marks were enriched in TR regions, particularly in trinucleotide and hexanucleotide repeat regions (Figure S5A-C), consistent with our previous observation of lower gene expression levels associated with trinucleotide and hexanucleotide repeats. The length and type of TRs motifs might affect the extent of epigenetic modification. Furthermore, when comparing repeat sequence complexity using entropy (see method), we found that lower entropy (I to V) correlated with lower repressive and activating signals, which are more likely to consist of Single and dinucleotide repeats, supporting our earlier hypothesis (Figure S5D).

Comparing the genomic histone signal tracks at STR locations clearly demonstrated the enrichment of repressive histone marks and the depletion of activating histone marks at STR positions (Fig. 3B and S6A). Consistent enrichment patterns were observed in randomly genomic regions, further confirming the robustness of our results (Figure S6B). To illustrate the robustness of our analysis, we performed a significant analysis of histone signals in regions containing trinucleotide and hexanucleotide repeats. The results showed higher repressive histone signals and lower activating histone signals in trinucleotide and hexanucleotide repeat regions compared to non-TR regions (Fig. 3C). This result was also apparent across all TRs regions (Figure S7A).

3.4. Low chromatin accessibility and nucleosome density in TR regions

The enrichment of repressive histone signals in STR regions was closely associated with low gene expression levels, which may relate to nucleosome density and chromatin accessibility to TFs. Employing MNase-seq data to assess nucleosome density, we observed high density of nucleosomes in STR regions compared to non-TR regions (Fig. 3D). This was consistent in all TR regions (Figure S7B). Previous studies had demonstrated a strong correlation between open chromatin and gene promoter activity [72]. Our analysis revealed a notable reduction in the number of open chromatin peaks within promoter regions as gene expression levels decreased, suggesting a correlation between gene expression and chromatin accessibility (Figure S7C). We further profiled the chromatin accessibility signals across different types of TRs regions.

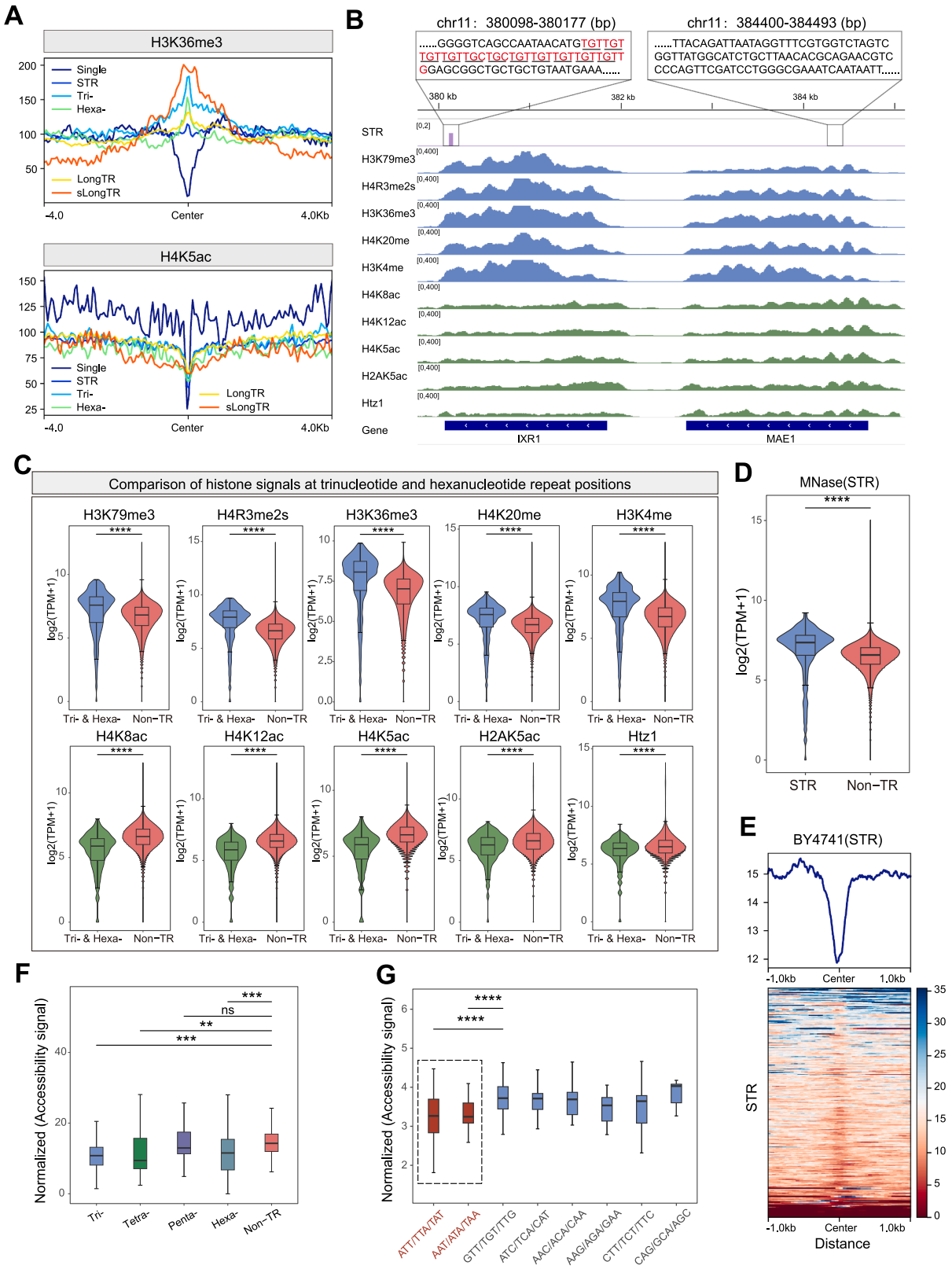


Fig. 3. Analysis of histone modifications and chromatin accessibility in TR regions. (A) Histone signal profiles of Single, STR, trinucleotide repeats, hexanucleotide repeats, LongTR, and sLongTR regions. Histones include H3K36me3 (top) and H4K5ac (bottom). (B) Histone signals distribution tracks. Height of STR signals indicates entropy values between 0 and 2. Blue tracks indicate inhibitory histone marks, green tracks indicate activating histone marks. Genes are located at the bottom. The sequence of the chr11:380098–380177 bp region is labeled on the upper left. The sequence in the region chr11:384400–384493 bp is labeled on the upper right. (C) Significance analysis of 10 histone mark signals in trinucleotide and hexanucleotide repeat regions versus Non-TR regions. (D) Comparison of MNase signals in STR region and Non-TR region. (E) Profiles and heatmaps of chromatin accessibility signals at STR sites in BY4741 strain. (F) Comparison of chromatin accessibility in STR (with different motifs) and Non-TR regions. * *P < 0.01, * **P < 0.001, Wilcoxon test. (G) Comparison of chromatin accessibility in different 3 bp motif types. * *P < 0.01, * **P < 0.001, * ** *P < 0.0001, Wilcoxon test.

The results showed low chromatin accessibility signals for all TRs, whether Single, STR, LongTR, or sLongTR in BY4741 and yZSJ025 (Fig. 3E, S7D and E).

These findings highlighted the low chromatin accessibility of TR regions, consistent with reduced gene expression in these areas. Based on our previous classification of STRs, we examined chromatin accessibility levels in STR regions with different motif lengths and types. The results showed that trinucleotide and hexanucleotide repeat regions had lower chromatin accessibility. Additionally, the previously discussed trinucleotide repeats (ATT/TTA/TAT and AAT/ATA/TAA), associated with reduced gene expression, also displayed lower chromatin accessibility (Fig. 3F and G). Our study emphasized the low chromatin accessibility feature of STRs.

3.5. Low chromatin interaction pattern within a 10 kb range in STR regions

TRs have been reported to alter local chromatin structure [19]. Given the strong relationship between chromatin interactions and gene expression [73,74], we further investigated chromatin interactions within TR regions. Using Hi-C data, we first identified significant intra- and inter-chromosomal interaction sites in BY4741 and evaluated the strength of these chromatin interactions using FIRE (frequent interaction region) scores. Circos plots revealed that significant interaction sites were often distant from regions with high entropy TRs [39], and most of these regions exhibited low FIRE signals (Fig. 4A and B). Frequent interactions between different telomeres are associated with the Rabl configuration in *S.cerevisiae* [44,75]. We computed interactions around trinucleotide and hexanucleotide repeat regions. The results showed that trinucleotide and hexanucleotide repeat regions lack both intra- and inter-chromosomal interactions, with the number of interactions increasing with distance from these regions (Fig. 4C). Similar results were observed for all TR regions except Single and sLongTR (Figure S8A). These findings suggest that TR regions exhibit distinct patterns of chromatin interactions in three-dimensional genome architecture compared to other genomic regions. By comparing interactions within TR and non-TR regions, we found that the number of interactions within TR regions was significantly lower than those in non-TR regions (Figure S8B).

Since Micro-C reflects interactions at the nucleosomal level, to study chromatin interactions in STR regions, we utilized Micro-C data provided by Rando, O. J., et al., for BY4741 strain [76], obtaining a high-resolution chromatin interaction matrix. We aggregate the interaction matrix for the upstream and downstream 10 kb region at trinucleotide and hexanucleotide repeat regions. The result shows that chromatin interactions in the trinucleotide and hexanucleotide repeat regions are significantly lower than in the surrounding regions, and this effect extends to a spatial extent of 8–10 kb (Fig. 4D). Interactions were also visualized for Single, STR, and LongTR regions (Figure S8C), sLongTR was excluded due to its limited number. These findings underscored the chromatin interaction characteristics of TR regions.

To further validate the chromatin interaction characteristics of TRs, we compared FIRE scores between TR and non-TR regions. The results indicated that FIRE scores in TR regions of both BY4741 and yZSJ025 strains were significantly lower than in non-TR regions (Figure S8D). This trend was consistent across different types of TRs (Figure S8E). Checking FIRE scores at 2 kb intervals around TR regions revealed consistently lower scores extending 10 kb from the center of TRs (Figure S8F). Analysis of the correlation between FIRE scores and gene expression also demonstrated a strong positive correlation between frequent chromatin interactions and transcription levels (Fig. 4E). Thus, we concluded that low gene expression levels in TR regions are associated with reduced chromatin interaction levels. Additionally, we analyzed FIRE scores for different motif lengths and types of STRs. The previously discussed trinucleotide repeats (ATT/TTA/TAT, AAT/ATA/TAT) displayed lower FIRE scores (Fig. 4F and G).

In summary, this study characterized STRs at the epigenetic and chromatin structural level, demonstrating that STRs within CDS are characterized by increased nucleosome density, enrichment of repressive histone modifications, and a lack of chromatin interactions within these regions. We illustrated the epigenetic features of genes containing STRs in a schematic diagram (Fig. 4H).

4. Discussion

The relationship between sequence and function has always been a topic of interest. Our research focused on the relationship between TRs and gene expression. Since STRs have a more pronounced impact on gene expression, we investigated the functions and characteristics of STRs in CDS. Genes containing trinucleotide and hexanucleotide repeat primarily encode proteins involved in transcriptional regulation and exhibit lower expression levels. This provide new ideas on the function of repetitive sequences in transcriptional regulatory mechanisms, and also provides new guidance for the design of transcriptional regulatory elements for synthetic biology and precision medicine. Additionally, these regions were marked by high repressive and low activating histone modifications. The characteristics of Single, STRs, LongTRs, and sLongTRs were summarized in a schematic diagram (Fig. 5). Upon their epigenetic characteristics, researchers can utilize epigenome editing technologies to modify and design these functional elements in the future [77], thereby achieving the goal of regulating transcription processes.

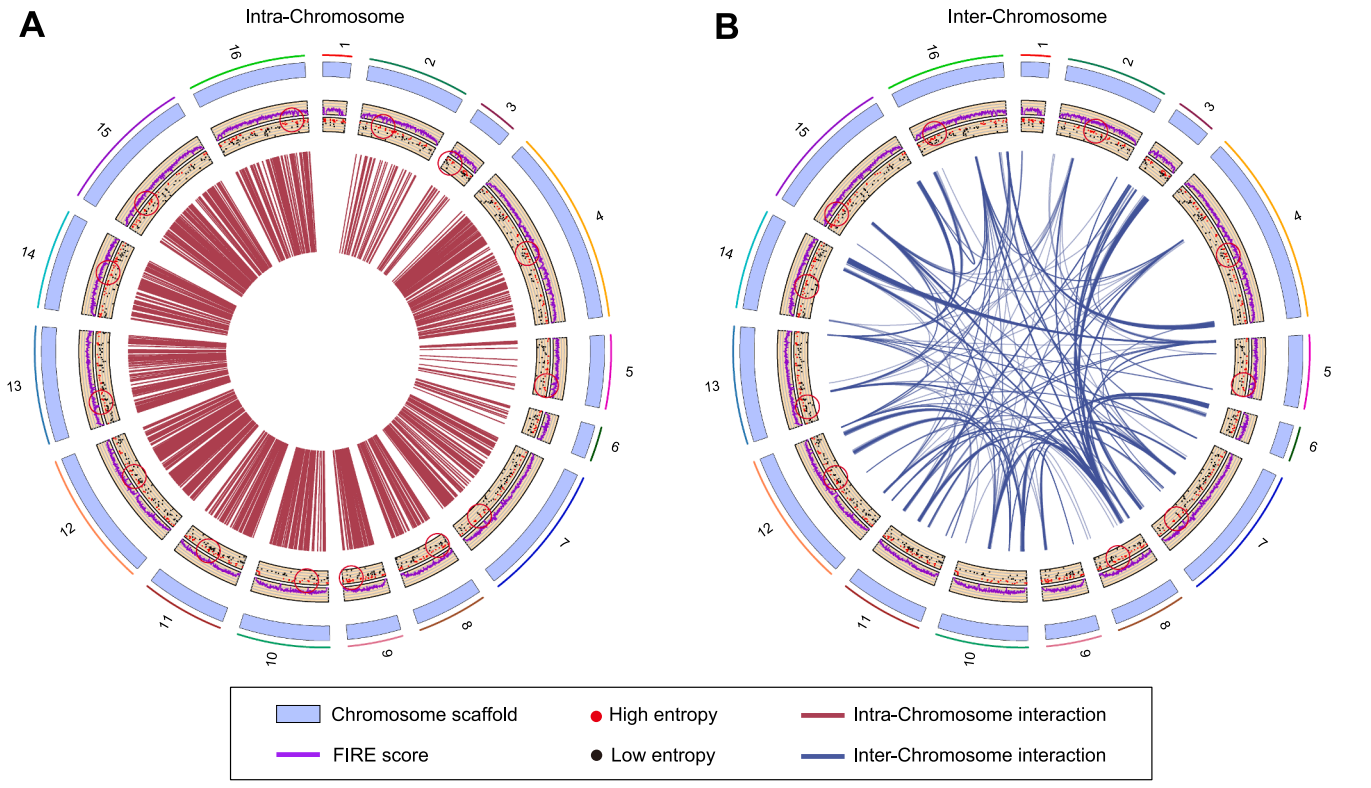
Despite the varied distribution patterns of different TRs, TR-enriched regions are generally characterized by low chromatin accessibility, few chromatin interactions, and reduced gene expression. We speculate that low chromatin accessibility impedes the binding of TFs and transcriptional regulatory proteins to promoter regions, potentially inhibiting transcription initiation. The low levels of chromatin interactions observed within a 10 kb range of TRs are probably due to the repressive histone modifications. In human genome, TRs are closely linked with the formation of heterochromatin, often featuring numerous repressive histone modifications and a high density of nucleosomes, which block the activation of genes near these TR regions. Altogether, these factors exert complex effects on gene expression, and mutations can disrupt the transcriptional regulatory network, affecting gene expression.

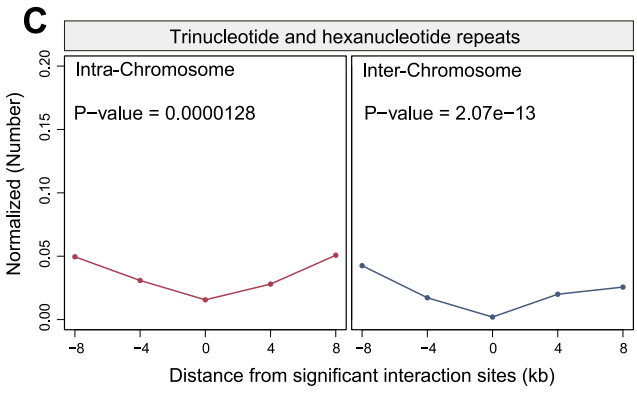
In particular, genes with trinucleotide and hexanucleotide repeats encode a large number of transcription factors and transcriptional regulation correlated proteins. Perhaps there is a “code” between TR-rich transcription factors and transcriptional regulatory proteins, as well as their binding to promoter regions, perhaps the repeat sequences are the key for decoding. Studies have reported the impact of motifs on transcription factor binding affinity [8]. This finding gives new inspiration to study the mechanisms of transcriptional regulation and also provides guidance for the design of regulatory functional elements, which has important implications for synthetic biology. Next, we will continue to investigate the role of TRs in transcriptional regulatory proteins.

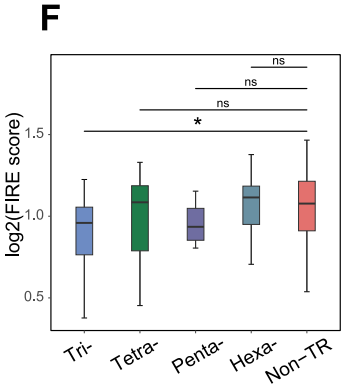
Comprehensive analysis of TRs in *S.cerevisiae* drives the understanding of the relationship between sequence and function, particularly in transcriptional regulation. Future research will explore the functional role of STRs in the structure of transcriptional regulatory proteins and how these functional elements can be applied to the artificial manipulation of complex regulatory networks.

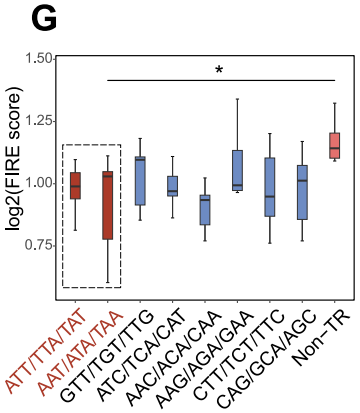
Funding

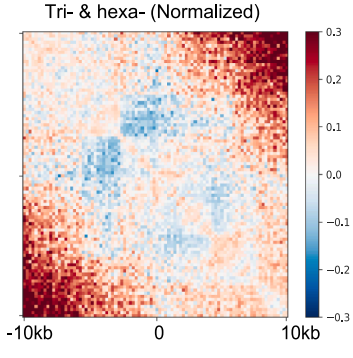
This work was supported by the National Natural Science Foundation of China (No. 62422318 and No. 62173338 to H.C.), and National Key Research and Development Program of China (No. 2023YFF0725500 to H.C.).

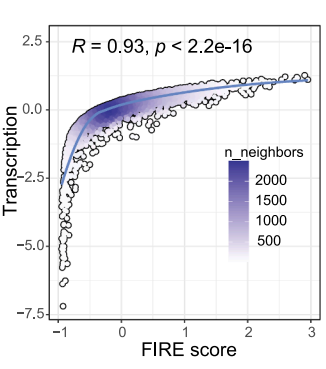


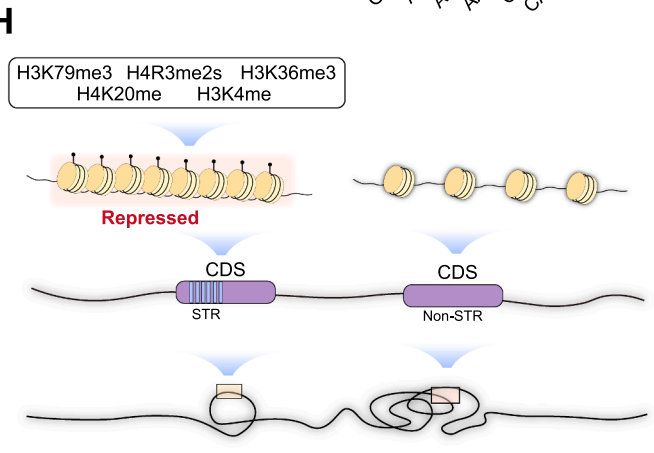












(caption on next page)

Fig. 4. Chromatin interactions within TR regions. (A-B) Circos plots illustrating significant intra- and inter-chromosomal interactions. The blue bar outlines represent chromosomes; the purple tracks denote FIRE scores; red dots signify the positions of high-entropy TRs; black dots indicate low-entropy TRs; the innermost red lines represent significant intra-chromosomal interactions; blue lines signify inter-chromosomal interactions. Data consistency areas are marked with red circles. (C) Distribution of the number of significant interactions within the trinucleotide and hexanucleotide repeat regions. The x-axis displays the regions 8 kb upstream and downstream of the trinucleotide and hexanucleotide repeats; the y-axis indicates the normalized counts of significant interactions. Red indicates intra-chromosomal interactions (left); blue indicates inter-chromosomal interactions (right). $P < 0.05$, Fisher's exact test. (D) An aggregation interaction matrix of 10 kb upstream and downstream regions of trinucleotide and hexanucleotide repeat regions, is visualized after normalization to expected values and logarithmic transformation. Blue color indicates low chromatin interactions in the region. (E) Correlation analysis between FIRE signals and transcription levels. Pearson correlation coefficient 0.93, p-value < 2.2e-16. (F) Comparison of FIRE scores for different motif length regions and Non-TR regions. (G) Comparison of FIRE scores for different motif types. *$P < 0.05$, Wilcoxon test. (H) Model of genes containing STRs.

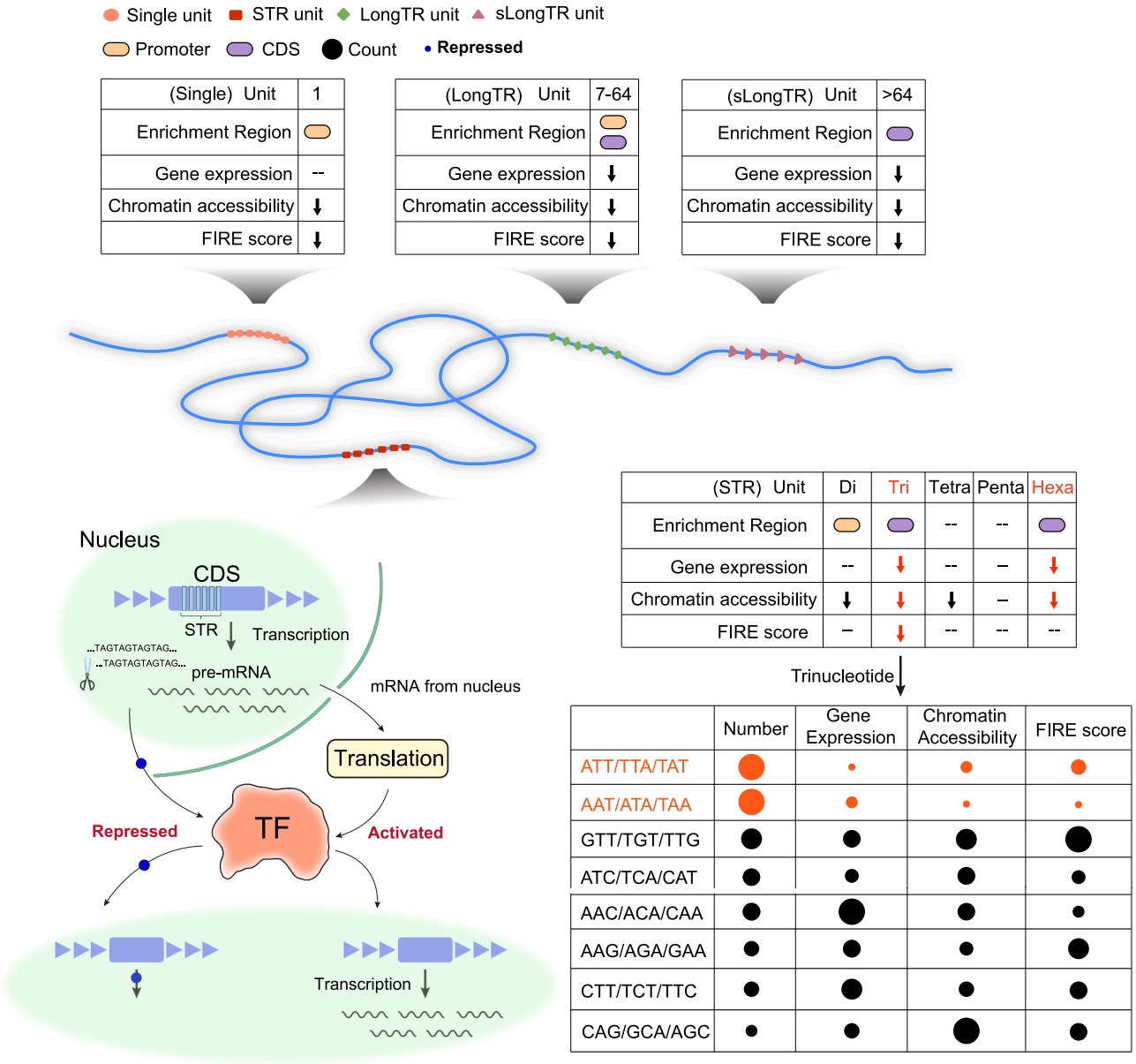


Fig. 5. Comprehensive descriptions of Single, STR, LongTR, and sLongTR. A schematic diagram illustrating the influence of STRs within CDS on transcriptional regulation is depicted left.

CRediT authorship contribution statement

Yu Zongyuan: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Liang Yating:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Data curation, Conceptualization. **Chen Hebing:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Xu Xiang:** Validation, Formal analysis. **Ran Dongyang:** Validation, Formal analysis. **Xiang Meida:** Validation, Formal analysis. **Xu Kang:** Validation, Formal analysis. **Luo Yawen:** Validation, Formal analysis. **Chen Bijia:** Validation, Formal analysis. **Bo Xiaochen:** Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

There is no conflict of interest in submitting this manuscript. All authors have read and approved this manuscript. Neither the entire manuscript nor any part of its content has been published or accepted elsewhere. All the listed authors have approved the enclosed manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.02.003.

References

[1] Biémont C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 2010;186:1085–93.
[2] Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nat Rev Genet 2010;11:559–71.
[3] Sun, J.H. et al. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. (2018).
[4] Liao XY, et al. Repetitive DNA sequence detection and its role in the human genome. Commun Biol 2023;6:21.
[5] Zhang XY, Meyerson M. Illuminating the noncoding genome in cancer. Nat Cancer 2020;1:864–72.
[6] Mehrotra S, Goyal V. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genom Proteom Bioinforma 2014;12:164–71.
[7] Plohl, M.J.P.B. Those mysterious sequences of satellite DNAs. **112**, 403-410 (2010).
[8] Horton CA, et al. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. 1304-+ Science 2023;381. 1304-+.
[9] La Spada AR, Taylor JP. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet 2010;11:247–58.
[10] McMurray CT. Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet 2010;11:1.
[11] Mirkin SM. Expandable DNA repeats and human disease. Nature 2007;447:932–40.
[12] Pearson CE, Edamura KN, Cleary JD. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 2005;6(10):729–42.
[13] Andrews LB, Nielsen AAK, Voigt CA. Cellular checkpoint control using programmable sequential logic. 1217-+ Science 2018;361. 1217-+.
[14] Nielsen J. Yeast cell factories on the horizon. Science 2015;349:1050–1.
[15] Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K.J.J.S. Unstable tandem repeats in promoters confer transcriptional evolvability. **324**, 1213-1216 (2009).
[16] Iyer, V. & Struhl, K.J.E.J. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. **14**, 2570-2579 (1995).
[17] Winter DJ, et al. Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloe festucae. PLoS Genet 2018;14:29.
[18] Soyer JL, et al. Epigenetic control of effector gene expression in the plant pathogenic fungus leptosphaeria maculans. PLoS Genet 2014;10:19.
[19] Haws SA, Simandi Z, Barnett RJ, Phillips-Cremins JE. 3D genome, on repeat: Higher-order folding principles of the heterochromatinized repetitive genome. Cell 2022;185:2690–707.
[20] Fotsing SF, et al. The impact of short tandem repeat variation on gene expression. 1652-+ Nat Genet 2019;51. 1652-+.
[21] Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 2008;18:1011–9.
[22] Yi H, et al. The tandem repeats enabling reversible switching between the two phases of β-lactamase substrate spectrum. PLoS Genet 2014;10:11.
[23] Bulik-Sullivan B, et al. An atlas of genetic correlations across human diseases and traits. 1236-+ Nat Genet 2015;47. 1236-+.
[24] Yue F, Du Z, Guo X, He X, Zhang B. Effect of tandem repeats adjacent to 3′-terminal of FLO1 on the flocculation function of Saccharomyces cerevisiae. Weishengwu Xuebao 2013;53:1276–84.
[25] Verstrepen KJ, Jansen A, Lewitter F, et al. Intragenic tandem repeats generate functional variability. Nat Genet 2005;37(9):986–90.
[26] Liang, Y. et al. Statistical Genomics Analysis of Simple Sequence Repeats from thePaphiopedilum MalipoenseTranscriptome Reveals Control Knob Motifs Modulating Gene Expression. **11**.
[27] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573–80.
[28] Servant N, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015;16:11.
[29] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
[30] Abdennur N, et al. Pairtools: from sequencing data to chromosome contacts. PLoS Comput Biol 2024;20:17.
[31] Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
[32] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–U354.
[33] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;31:2032–4.
[34] Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 2014;42:W187–91.
[35] Jeon H, Lee H, Jang BI, Roh T-Y, Jang I. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. Genom Inform 2020;18:e42.
[36] Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.
[37] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinforma 2011;12:16.
[38] Kaul A, Bhattacharyya S, Ay F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc 2020;15:991–1012.
[39] Zhang HE, Meltzer P, Davis S. RCircos: an R package for Circos 2D track plots. BMC Bioinforma 2013;14:5.
[40] Schmitt AD, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep 2016;17:2042–59.
[41] Deng L, et al. Dissection of 3D chromosome organization in Streptomyces coelicolor A3(2) leads to biosynthetic gene cluster overexpression. Proc Natl Acad Sci USA 2023;120:10.
[42] Yu GC, Wang LG, Han YY, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 2012;16:284–7.
[43] Kruse K, Hug CB, Vaquerizas JM. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. Genome Biol 2020;21:19.
[44] Mercy G, et al. 3D organization of synthetic and scrambled chromosomes. Science 2017;355:8.
[45] Richardson, et al. Design of a synthetic yeast genome. Science 2017;355:1040–4.
[46] Kang JP, et al. Enhancement and mapping of tolerance to salt stress and 5-fluorocytosine in synthetic yeast strains via SCRaMbLE. Synth Syst Biotechnol 2022;7:869–77.
[47] Zhou S, et al. Dynamics of synthetic yeast chromosome evolution shaped by hierarchical chromatin organization. Natl Sci Rev 2023;10.
[48] Lander ES, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[49] Cheung J, et al. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 2003;4:10.
[50] Thakur J, Packiaraj J, Henikoff S. Sequence, chromatin and evolution of satellite DNA. Int J Mol Sci 2021;22:28.
[51] Paldi F, et al. Convergent genes shape budding yeast pericentromeres. 119-+ Nature 2020;582. 119-+.
[52] Peng J, Zhou JQ. The tail-module of yeast Mediator complex is required for telomere heterochromatin maintenance. Nucleic Acids Res 2012;40:581–93.
[53] Liang YY, et al. Statistical genomics analysis of simple sequence repeats from the paphiopedilum malipoense transcriptome reveals control knob motifs modulating gene expression. Adv Sci 2024;11:15.
[54] Lu JY, et al. Genomic repeats categorize genes with distinct functions for orchestrated regulation. 3296-+ Cell Rep 2020;30. 3296-+.
[55] Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front Genet 2016;7:15.
[56] Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet 2018;19:286–98.
[57] Yuan et al. Genome-Scale Identification of Nucleosome Positions in S. cerevisiae. (2005).
[58] Lee TI, Young RA. Transcription of eukaryotic protein-coding genes. Annu Rev Genet 2000;34:77–137.
[59] Stathopoulos AM, Cyert MS. Calcineurin acts through the CRZ1/TCN1-encoded transcription factor to regulate gene expression in yeast. Genes Dev 1997;11:3432–44.
[60] Peterson CL, Zhao Y, Chait BT. Subunits of the yeast SWI/SNF complex are members of the actin-related protein (ARP) family. J Biol Chem 1998;273(37):23641–4.
[61] Guglielmi B, et al. A high resolution protein interaction map of the yeast Mediator complex. Nucleic Acids Res 2004;32:5379–91.
[62] Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol 2015;16:18–29.
[63] Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets - the roles of intrinsic disorder in protein interaction networks. Febs J 2005;272:5129–48.
[64] Kim PM, Sboner A, Xia Y, Gerstein M. The role of disorder in interaction networks: a structural analysis. Mol Syst Biol 2008;4:7.
[65] Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323:573–84.
[66] Liu JG, et al. Intrinsic disorder in transcription factors. Biochemistry 2006;45:6873–88.
[67] Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry 2008;47:7598–609.
[68] Fu X, et al. A foundation model of transcription across human cell types. Nature 2025;33.
[69] Cabrera A, et al. The sound of silence: transgene silencing in mammalian cell engineering. Cell Syst 2022;13:950–73.
[70] Alhaji SY, Ngai SC, Abdullah S. In: Mayes S, editor. Biotechnology and Genetic Engineering Reviews. London: Taylor & Francis Ltd; 2019. p. 1–25.
[71] Weiner A, et al. High-resolution chromatin dynamics during a yeast stress response. Mol Cell 2015;58:371–86.

[72] Kubik S, et al. Opposing chromatin remodelers control transcription initiation frequency and start site selection. Nat Struct Mol Biol 2019;26:744.
[73] Deng L, et al. Dissection of 3D chromosome organization in Streptomyces coelicolor A3(2) leads to biosynthetic gene cluster overexpression. Proc Natl Acad Sci USA 2023;120:10.
[74] Lioy VS, et al. Multiscale structuring of the E. coli chromosome by nucleoid-associated and condensin proteins. 771-+ Cell 2018;172. 771-+.
[75] Duan Z, et al. A three-dimensional model of the yeast genome. Nature 2010;465:363–7.
[76] Hsieh TH, et al. Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell 2015;162:108–19.
[77] McCutcheon SR, Rohm D, Iglesias N, Gersbach CA. Epigenome editing technologies for discovery and medicine. Nat Biotechnol 2024;42:1199–217.