Nucleic Acids Research, 2025, **53**, gkae1178
https://doi.org/10.1093/nar/gkae1178
Advance access publication date: 9 December 2024
Computational Biology

Motif distribution in genomes gives insights into gene clustering and co-regulation

Atreyi Chakraborty [1], Sumant Chopde[2] and Mallur Srivatsan Madhusudhan [1,2,*]

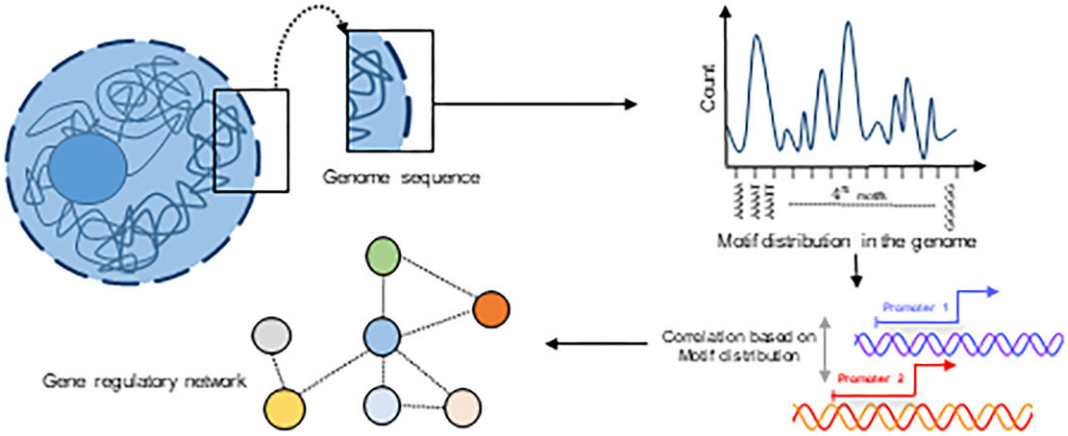
[1]Department of Biology, Indian Institute of Science Education and Research, Dr Homi Bhabha Rd, Pashan, Pune, Maharashtra 411008, India
[2]Department of Data Science, Indian Institute of Science Education and Research, Dr Homi Bhabha Rd, Pashan, Pune, Maharashtra 411008, India

[*]To whom correspondence should be addressed. Tel: +91 20 25908133; Fax: +91 20 2590 8001; Email: madhusudhan@iiserpune.ac.in

Abstract

We read the genome as proteins in the cell would – by studying the distributions of 5–6 base motifs of DNA in the whole genome or smaller stretches such as parts of, or whole chromosomes. This led us to some interesting findings about motif clustering and chromosome organization. It is quite clear that the motif distribution in genomes is not random at the length scales we examined: 1 kb to entire chromosomes. The observed-to-expected (OE) ratios of motif distributions show strong correlations in pairs of chromosomes that are susceptible to translocations. With the aid of examples, we suggest that similarity in motif distributions in promoter regions of genes could imply co-regulation. A simple extension of this idea empowers us with the ability to construct gene regulatory networks. Further, we could make inferences about the spatial proximity of genomic fragments using these motif distributions. Spatially proximal regions, as deduced by Hi-C or pcHi-C, were ∼3.5 times more likely to have their motif distributions correlated than non-proximal regions. These correlations had strong contributions from the CTCF protein recognizing motifs which are known markers of topologically associated domains. In general, correlating genomic regions by motif distribution comparisons alone is rife with functional information.

Graphical abstract



Introduction

The human nuclear genome comprises a vast sequence of 3.2 billion base pairs of DNA, organized into 23 or 24 distinct chromosomes. This sequence serves as the fundamental code that underpins the operation of the intricate biological machinery within the cell. Based on the T2T-CHM13 reference (1), a comprehensive annotation effort has identified a total of 63 494 genes within the human genome. This count includes both protein-coding genes, pseudogenes, hypothetical genes and functional RNA (long non-coding RNA, microRNA etc.) genes (2). Genic regions are however only a small fraction of the genome, which predominantly comprises non-coding regions (3–5). These non-coding parts of the genome harbour an abundance of repetitive DNA, such as transposable elements (6) and contain valuable information on disease-causing mutations (7), genetic variations (8,9) and evolutionary conservation (10,2). Analysing genomes for all of the characteristics mentioned above usually involves aligning sequence stretches to one another. In this study, we offer an alternative yet simple method to 'read' the genome.

Received: March 4, 2024. **Revised:** September 17, 2024. **Editorial Decision:** November 5, 2024. **Accepted:** November 15, 2024

The regulation of the transcriptome inside the nucleus is complex (11). For instance, genes and their corresponding regulators may be sequentially distant but spatially proximal (12–14) as the regulation is coordinated by proteins. Different techniques are used to investigate this spatial organization, protein-genome interaction and chromatin accessibility such as 3C techniques (15), Hi-C (9,16), Dam ID (17,18), Chip-Seq (19,20), ATAC Seq (21,22), super-resolution imaging (23) and integrative genome modelling (24,25).

The resolution of the information from the techniques mentioned above is usually in the range of 10–100 kb. Investigating genome stretches of size 10–100 kb, especially with the intent of gaining insights into genome organization/function poses several difficulties. Firstly, the 10–100 kb size range is very low resolution, given that DNA binding proteins engage directly with just a few base pairs. Secondly, even if one were to look for sequence conservation and/or patterns, 10–100 kb is a significantly large stretch compounded by the fact that the DNA alphabet consists of only four letters. Sequence alignments of such large stretches are often poor indicators of functional conservation. Even two randomly chosen sequence stretches of ∼10–100 kb could be ∼25% identical. Therefore, identifying and comparing sequence patterns necessitates a focus on shorter sequence sizes. Alignment-free methods, such as k-mer or word frequency estimation, offer alternatives where sequences are processed in moving windows of a specified word length (26,27). Notably, this approach does not hinge on positional information of bases in sequence. However, it proves valuable for comparing lengthy sequences, allowing the calculation of word density to determine the similarity or dissimilarity between two sequences (28,29). This method provides a practical means for assessing sequence patterns without relying on extensive alignments, particularly suited for analysing large genomic regions. In this study, we used a modification of the k-mer method to compare genome segments.

Binding patches on proteins typically recognize 5–6 bases of DNA (S. Nair and M.S. Madhusudhan, unpublished results: doi:10.1101/2022.05.19.492702). Our observation from all known DNA bound protein complexes in the Protein Data Bank (PDB) is that at the points where proteins contact DNA specifically, the interactions only span 5–6 nucleotides. However, there are several reports of larger DNA sequence motifs recognized by proteins (30,31). These happen in cases where the protein could oligomerize and hence recognize several, seemingly contiguous patches of DNA that are larger than 6 nucleotides. It could also happen when DNA wraps around a protein making contacts at different faces (Supplementary Figure S1). Our observation is that for every contact only 5–6 nucleotides are involved in base specific recognition such as hydrogen bonding. Hence, we focussed on sampling motifs of 5–6 bases of DNA and read the genome at that resolution. We used the term motif to represent an n-mer assuming that all n-mers are potential motifs for some cognate protein(s). In this study, we first analysed motif distributions of different sizes (from 2- to 6-mers). We then compared this to a randomized genome to show that patterns in real genomes are distinct. We examined patterns of motif distributions in whole genomes/chromosomes and in smaller regions such as centromeres and gene promoters. We correlated pattern distributions in gene promoter sites and established relationships between genes that are likely to be co-regulated. Sometimes when protein interact with DNA, the DNA could bend or alter its conformation from the regular B-form (32,33). That notwithstanding, in this study we are assuming that a particular DNA motif when binding to its cognate protein would always undergo the same recognition shape change.

We corroborated our findings, using three case studies. In the first case study, we selected three genes, LPHN1 (ADGRL1), CDK9 and TRIM8, that exhibited high correlation based on our motif abundance analysis and compared them to a network map using NetworkAnalyst (34,35). We identified two common transcription factors to all three genes, thus validating not just our method of constructing gene networks but also discovering the functional importance of such connections. For the second case study, we focused on a larger list of 19 genes coregulated by the transcription factor(s) Jun/Fos. The analysis here showed that we could predict differential gene regulation. In the final case study, we investigated whether gene co-regulation implied co-localization by comparing our results to Hi-C (36) and promoter capture Hi-C (37) data. We found a strong correlation between our scoring and the spatial positioning of genome segments, reinforcing the significance of our methods.

Materials and methods

Data of genomic sequence(s) and the gene annotation, for all calculations was obtained from NCBI RefSeq Human Genome assembly T2T-CHM13v2.0. (38,39)

Motif generation and distribution metric scoring

To assess the enrichment of specific motifs within certain regions, we employed the observed-to-expected ratio (OE ratio). The OE ratio is a measure of whether a feature is over or underrepresented in a given dataset. It is calculated by normalizing the observed frequency of the feature by the expected frequency based on the probability of occurrence. For a sequence stretch of base pairs, we can 'read' it using a window (motif size) of n base pairs. The number of times a motif occurs in the sequence stretch is recorded as the observed count (O). The expected count (E) of any motif is $(N-k+1)\prod_x f_x^{nx}$, where nx is the number of times a particular base occurs in a motif and whose probability is f_x and N is the size of the chromosome. The OE score is thus a normalized metric taking into account the AT/GC richness of a chromosome. Given that there are four bases (A/T/G/C), the number of motifs of size k, which we refer to as k-mer count, would be 4^k. Here we average the OE values for motifs and their reverse complements. These averaged OE ratios taken together are called the motif vector. We recorded the observed count of motifs and then computed the OE ratio for the following sequence stretches:

- $\mathbf{OE_{whole}}$: Here the sequence stretch is the whole chromosome.
- $\mathbf{OE_{nkb}}$: Chromosomes are divided into bins of n kb. The last bin of any chromosome may contain less than n kb, while all other bins are exactly n kb in length.
- $\mathbf{OE_{centromere}}$: Only the centromeric regions were considered. The boundaries of centromeric regions in the different chromosomes were taken as defined in the NCBI genome data viewer and UCSC genome browser (39,40) (Supplementary Table S1).

- $OE_{promoter}$: We sampled three different sequence stretches in the 5' untranslated region (5′-UTR) that we define as promoter proximal control regions. For simplicity we refer to this as promoter regions in the rest of the text. These were 1, 2 and 6 kb upstream of the gene transcription start site and denoted as $OE_{1\ kb\_kmer}$, OE_{2kb_kmer} and OE_{6kb_kmer}, where motif size k = 5 or 6. We obtained the gene start and stop coordinates from the NCBI RefSeq gene annotation (39,40).
- OE_{random} : For different computations (as expounded in the results), it was necessary to scramble a sequence stretch and rerecord the observed count. To do this, the sequences of entire chromosomes were shuffled (randomized). The procedure for randomization is mentioned in the 'Materials and methods' section.

Note that in all these computations, the observed count is taken from the different sequence stretches while the expected count is computed based on chromosome size (N) and is the same across all computations (for a particular chromosome).

Correlating motif vectors

The correlation between two motif vectors, X and Y, is computed as:

$$\rho X,\ Y = \frac{cov\,(X,\ Y)}{\sigma X \times \sigma Y} \text{ where } cov\,(X,\ Y) = \sum\ (X - \mu X) \times\ (Y - \mu Y) \quad (1)$$

Where $cov(X,\ Y)$ is the covariance, (μX and μY) and (σX and σY) are the means and standard deviations of OEs of motif vectors X and Y, respectively.

Generating random genome

To randomly generate DNA sequences of different chromosomes, we used the random module of python3 and ensured weighted random nucleotide selections. This procedure generates random sequences of chromosomes of the same size and base probability as the wild type. The random genome constructed thus has the same nucleotide composition as the original, just scrambled in sequence.

Finding Hi-C contacts with genes

Hi-C data were obtained from GEO accession number GSE18215 (in .mcool format), where chromosomes were divided into uniform bins of different sizes. We chose to use data from the bins of the smallest size (highest resolution), 10 kb. We extracted pairs of 10 kb bins from the data where cross links were recorded between chromosomes 18 and 19.

For the promoter capture Hi-C (pcHi-C), we extracted the chromosome contact start-stop coordinates from the .bedpe files from GEO accession number: GSM1704495. From the NCBI gene annotation for the corresponding reference assembly used in mapping the contacts, we obtained the gene coordinates that overlap with the contact regions between the two chromosomes. For the gene pairs that fall in the region of contacts, we obtained the promoter correlations according to our motif vector calculations at the different aforementioned promoter sizes.

Results

Motif preferences as assessed by the OE ratio

We had previously established that protein binding stretches in genomic DNA span 5–6 nucleotides (41) (Supplementary Figure S1). So, in this study, we have read the genome at this length scale, i.e. at biologically relevant DNA motif sizes. To begin with, we read the genome with motif sizes of 2–6 nucleotides. The aim here was to check the effective size of motifs that would confer specificity to DNA–protein binding. We noticed that as the motif size increased from 2 to 6, the number of motifs with high OE_{whole} ratio values increased (Figure 1A and Supplementary Figure S2). While the trend is monotonic, we did not explore beyond 6-mers as this is the optimal protein recognition size and we believe that all specificity discriminations are likely to occur at this length. The inference here is that the larger the variation in the OE_{whole} ratio values, the more pronounced the patterns of occurrences of these motifs. This in turn has implications on how different proteins would engage with different parts of the genome.

For the rest of this study, we concentrate on 5- and 6-mer motifs. We next looked at the OE_{whole} ratio of 5- and 6-mers for all individual chromosomes (Figure 1B and C). While considering OE ratios we averaged the values of a motif and that of its reverse complement as we are looking at double stranded DNA. Hence, we got 512 different motifs of 5-mers and 2080 motifs of 6-mers. [The number of 6-mer motifs and their reverse complement pairs is not exactly half of all possible motifs (4096) because palindromic motifs are identical to their reverse complements.] There are two appreciable patterns to discern here: (i) Certain motifs are less abundant in the whole genome while others are present in large abundance, as inferred from their OE_{whole} ratio values. For many motifs, these patterns of abundance hold across the different chromosomes. For instance, the 6-mer motif CCCAGC has high abundance in all chromosomes. The **CCCAGC**AG is a binding site for the zinc finger, ZNF143 (42), a protein involved in 3D genome construction. (ii) There are variations in the OE_{whole} ratio values for the same motif across chromosomes. We clustered the chromosomes based on their motif distribution similarity (represented as dendrograms in Figure 1B and C). The gene density (number of genes per Mbps of a chromosome) of individual chromosomes is independent of the chromosome clustering based on the motif OE ratios (Figure 1B and C and Supplementary Figure S3). We find that the clustering of chromosomes when considering 5-mers or 6-mers is similar (with minor rearrangements). Chromosomes 2–8, 10–12, 14, 18, 20 and X are all more closely related to one another than to the others. Among the others, (17, 19, 1, 15, 16), and (13, 21, 22) form smaller sub-clusters of similar chromosomes. Chromosomes 9 and Y are the two most distinct ones and bear the least resemblance to the others (see Supplementary Figure S4A for pairwise correlations).

On closer inspection, it is clear that each chromosome has its unique pattern of motif abundance. For instance, it is plain to see that the 5-mer and 6-mer motif abundances of chromosome 9 and Y are starkly different from the other chromosomes (Figure 1B and C and Supplementary Figure S4A). Chromosome 9 is known for its significant structural variability, housing the largest autosomal block of heterochromatin (43). Within this region, gene locations exhibit a distinctive pattern, with individual genes scattered among stretches of repeated sequences (44,45). The case is similar to chromo-

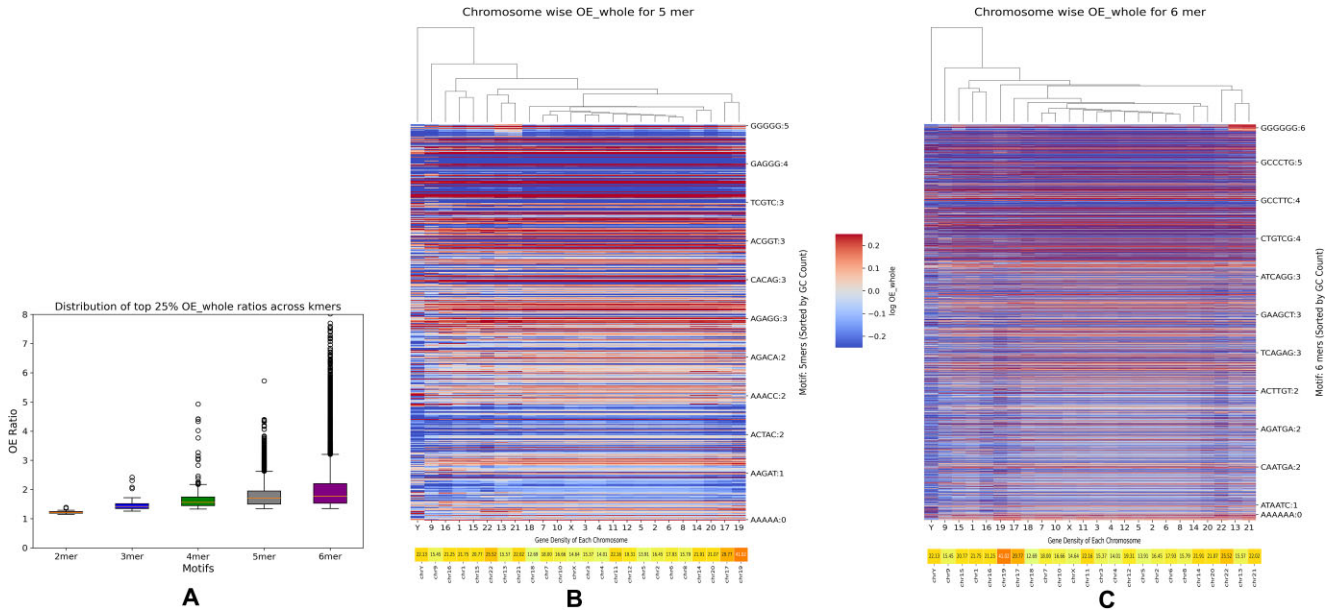


Figure 1. The distribution of OE_{whole} ratios for motif size 2–6 (**A**) The OE ratios are limited to the range from 0 to 8 and motifs of sizes 2, 3, 4, 5 and 6 are represented in different colours. Heat maps of OE_{whole} ratios for motif size 5–6 [panels (**B**) and (**C**), respectively] across the different chromosomes shown in log scale. The log (OE) ratios are coloured red through to blue (colour legend) where red and blue indicate values of >1 and <1, respectively. The darker the shade of the red (or blue) the higher (or lower) the ratio. The 512 and 2080 motifs for 5-mers and 6 mers respectively are arranged according to GC content, which increases going from bottom to top. In panels (**B**) and (**C**), only a few representative motifs (and their reverse complements) are labelled. Shown alongside the motif labels are their GC contents. The dendrogram represents the clustering of chromosomes based on the similarity in motif distribution. A colour bar of gene density is shown at the bottom of panels (**B**) and (**C**). The denser is the gene content the redder is the colour bar.

some Y, which is known to have the maximum occurrences or repeat regions (46). These results suggest the possibility that different parts of the genome are recognized differentially, presumably, by different proteins (47). These recognition events in turn could differentially regulate gene expression and other molecular functions (48).

The distribution of motifs in chromosomes is non-random

From the results above it is clear that motifs are present with different abundances in the different chromosomes. What is also clear is that while some broad conservation patterns may exist, the motif distribution in individual chromosomes is distinct. Within each chromosome, sub-regions have different distributions of the motifs, as seen in the 6-mer motif distribution along 100 kb stretches of chromosome 18 (Figure 2A). Details of the distribution of 5- and 6-mer motifs for all chromosomes are presented in supplementary data. Chromosome 18 has ∼80.3 Mb, which were divided into regions of 100 kb. We retained the same composition of bases and randomly scrambled the sequence of the whole chromosome (Figure 2B). The distinct patterns visible in the real distribution ($OE_{100\ kb}$) are no longer apparent in the scrambled sequence (OE_{random}). We did the randomization multiple times) and in each such attempt, the distinctness of the pattern of the real chromosome (in 100 kb bins) was absent (Supplementary Figures S5 and S6). The only discernible pattern in the randomized chromosome is the similarities of motif distributions of distinct GC content. It is interesting to note that the bin wise motif distribution is distinctly different in regions that correspond to the centromere (Supplementary Figure S5). The centromeres can be easily distinguished in each of the chromosomes. This can be attributed to the satellite repeat regions in the centromeres (49).

As mentioned above, the pattern of motif abundances varied not just across chromosomes but also within chromosomes. While this is evident for 100 kb bins across the length of chromosome 18 ($OE_{100\ kb}$ distribution in Figure 2A), we next investigated centromere regions on all chromosomes. The boundaries of centromeric regions of the different chromosomes were taken as defined in the NCBI genome data viewer and UCSC genome browser (39,40) (Supplementary Table S1). Here again, it is clear that the centromeric regions display distinct motif distributions. The $OE_{centromere}$ distributions across the different chromosomes of the genome are different (Figure 2C). For instance, the acrocentric chromosomes 13, 14, 15, 21 and 22 have a distinctly different $OE_{centromere}$ pattern (40) (Figure 2C and Supplementary Figure S4B). It is known that the short arms of these acrocentric chromosomes are abundant with segmental duplications (50), which could be the explanation for our observation.

We next compared the centromere regions to the same regions in the randomly scrambled chromosome sequences (Figures 2D–F). Here too, the difference between the motif distribution patterns in the real centromere ($OE_{centromere}$) and the centromere from the scrambled chromosome is clear – the centromere from the scrambled sequences are not distinct across chromosomes. We triplicated this result for confidence (Figures 2D–F).

We have established that motif distributions in chromosomes are not random and that such distributions are also different in different sequence segments within each chromosome. This gives further credence to the point made earlier that the distinct motif distributions ensure differential binding by the different protein partners.

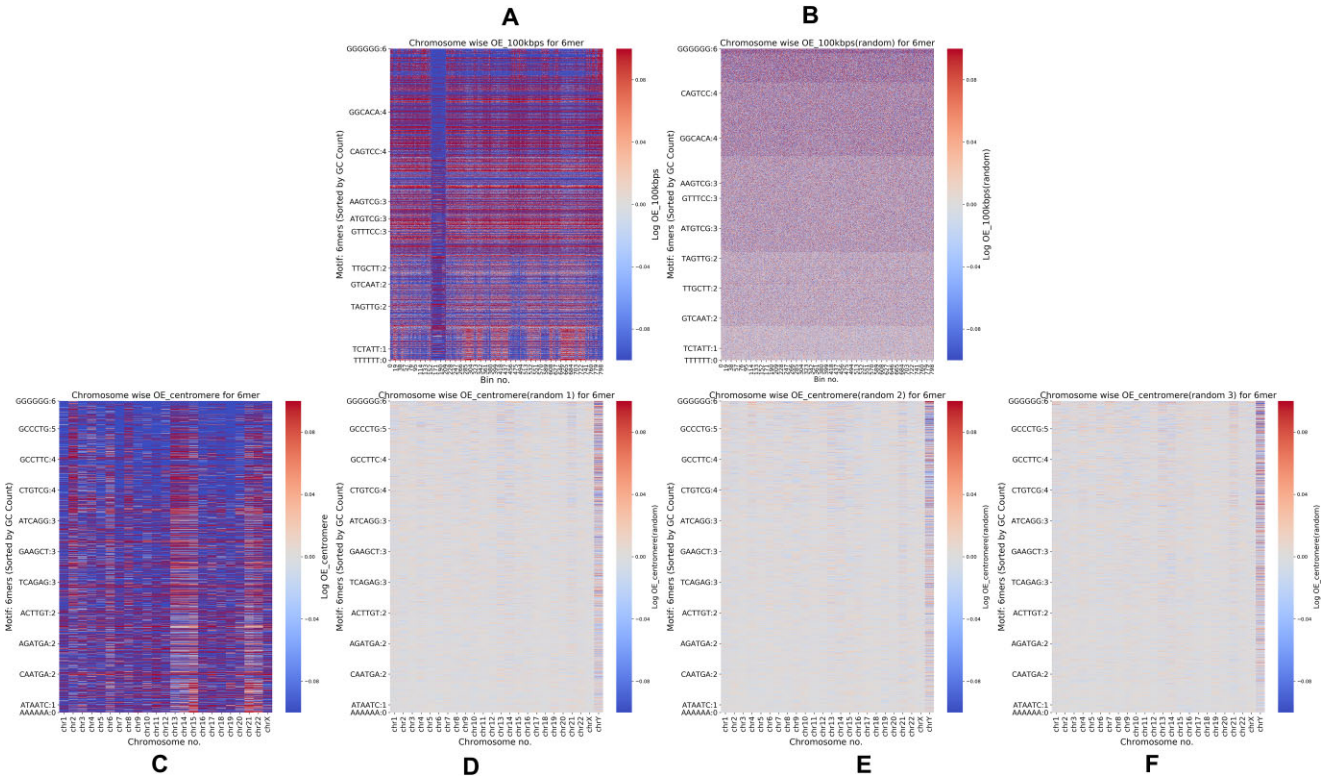


Figure 2. Panels (**A**) and (**B**) show the $OE_{100\ kb}$ and OE_{random} heat maps for chromosome 18 in log scale respectively. Heat maps of the log values of motif $OE_{centomere}$ ratios of 6-mers for centromeric regions in all chromosomes [panel (**C**)]. The randomized sequences of the same centromeres are shown in panels (**D**)–(**F**). As in Figure 1, only representative motifs and their reverse complements are labelled and their GC content is shown alongside The colour coding and the arrangement of motifs is the same as in Figure 1.

Gene correlations deduced from promoter motif distributions

In the next few sections, we concentrate on the 63 494 genic regions, as identified by the T2T CHM13 reference assembly. We variously defined promoter proximal control regions (promoters for short) associated with a gene as the regions 1, 2 or 6 kb upstream of the gene transcription start site. These promoter sizes were chosen considering a conservative estimate (1 kb), a maximum known effective size (6 kb) (51) and a value in between (2 kb).

We considered the relative abundances of all motifs (separately for 5-mers and 6-mers) as motif vectors. The coefficients of the vectors are the 512 and 2080 $OE_{Xkb\_k\text{-}mer}$ values for k-mers (k = 5 and 6) respectively, where X is the size of the upstream regions considered as promoters. All versus all gene promoter motif vectors were correlated (equation 1) and the results were collated chromosome-wise (Figure 3). Of the ∼4 billion possible gene–gene motif vector correlations, about ∼1 million have correlation coefficients ≥0.9. (The numbers are of a similar order for 6-mers and for different promoter sizes – 2 and 6 kb; see supplementary data of promoter correlations above 0.9.) We decided to impose this stringent cut-off to ensure that the genes with correlated promoters are chosen stringently and practically this would mean that we only deal with a small fraction (∼0.25%) of a very large dataset. As the promoter size increases from 1 to 6 kb the baseline of correlations changes while the trends remain the same (Supplementary Figure S7A–E and supplementary data of promoter correlations above 0.9). The baseline notwithstanding, some features of chromosome-chromosome correlations stand out. Most of the high correlations between gene promoter motif vectors are from within the same chromosome. This is understandable as genes that are functionally related are often found in the same chromosome and in close proximity to one another. On average the number of intra-chromosome gene promoter correlations above 0.9 is ∼920 per chromosome pair (when considering 5-mers with 1 kb promoters). This signifies that on average 920 gene pairs from two different chromosomes are correlated with coefficients ≥0.9. The maximum number for inter-chromosomal gene promoter correlations is 36, 408 between chromosomes 13 and 21, while the minimum is 0 between chromosomes Y with 1 and 19 (supplementary data of promoter correlations above 0.9).

These correlation data are vast and we are only going to present a figment of it in this study. These data can be interrogated to answer various questions about gene associations. What is of interest to us here are the inter chromosome interactions. There are genes from different chromosomes whose promoters have similar motif vectors. This implies that these genes could be co-regulated. The high correlations, especially ones with coefficients ≥0.9, could be symptomatic of interactions.

Once again the acrocentric chromosomes (13–15, 21 and 22) show strong correlations to one another. We found several hundreds of genes in each of the acrocentric chromosomes whose promoter regions were correlated to one another (see supplementary data of promoter correlations above 0.9). Of these, there are 1768, 2216, 2072, 992 and 1310 genes in chromosomes 13, 14, 15, 21 and 22, respectively, that are corre-

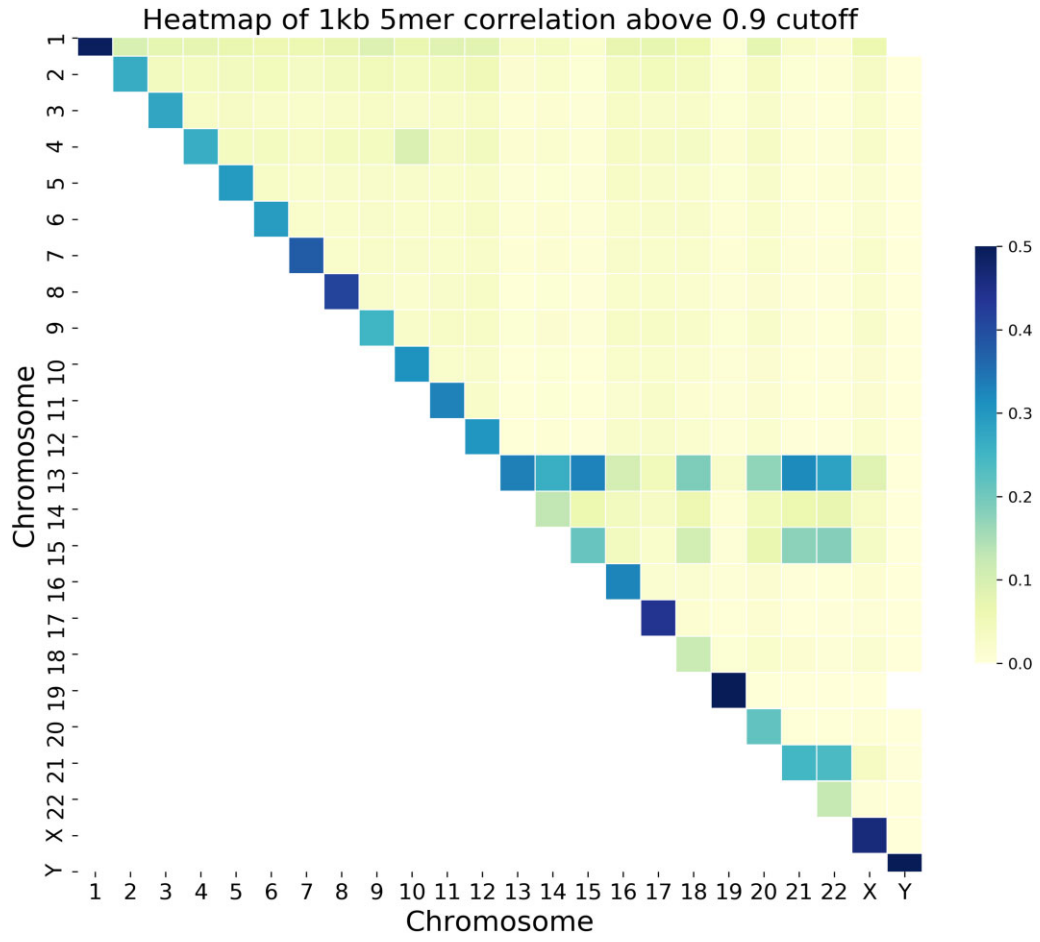


Figure 3. An inter-chromosome motif vector correlation map of frequency of gene pairs with correlation coefficients >0.9 for 5 mers with promoter sizes of 1 kb. The proportion of correlated genes with a correlation coefficient ≥0.9 are colour coded on a white–yellow–green–blue gradient. The white end represents no correlation while the darker the shade of blue the larger the number of correlated gene pairs.

lated to at least two promoters from genes of other acrocentric chromosomes. This is interesting, as not only do the centromeric regions of these genes show high correlations (see the 'The distribution of motifs in chromosomes is non-random' section above), but the promoters of genes in these chromosomes show high correlations too. This we believe may be the basis of Robertsonian translocations (52,53), where the long arms of these chromosomes are joined to one another (52,53). If the motif abundance is similar in two regions, it is conceivable that recombination could sometimes result in interchanges, resulting in translocations. Ribosomal RNA (rRNA) genes constitute ∼4–13% (Supplementary Table S2A) of the total genes in chromosomes 13, 14, 15, 21 and 22. Additionally, we also checked what fraction of the total high correlations (Pearson's correlation coefficient ≥0.9) were contributed by the rRNA gene pairs from both chromosomes. We observed ∼81–92% of the high correlations are contributed by the rRNA genes from both the chromosomes for chromosomes 13, 14, 15, 21 and 22 (Supplementary Table S2B and supplementary data of rRNA gene correlations). Thus, such tandem repeats of rRNA genes constitute the conserved patterns across these chromosomes. Additionally, we compared the similarity in the promoter sequence of the highly correlated genes (Pearson's correlation coefficient ≥0.9) in chromosomes 13 and 14 (chromosomes reported to be involved in translocations) versus chromosomes 18 and 19. A total of 20% of the highly correlated gene pairs between chromosomes 13 and 14 have a sequence identity of at least 80%. However, there are no gene promoter pairs between chromosomes 18–19 that have sequence identity higher than 37% (Supplementary Figure S8 and supplementary data of identity versus correlations). This suggests that translocations occur in regions where there is strong motif vector correlation as well as high sequence identity.

Another notable gene promoter correlation was between chromosomes 4 and 10 where 1017 (supplementary data of rRNA gene correlations) gene promoter pairs had a correlation coefficient above 0.9 (Figure 3). A translocation between these two chromosomes is also known in certain types of leukaemia (54,55) and Wolf–Hirschhorn syndrome (56). Unlike acrocentric chromosomes, the gene promoter pairs does not include rRNA genes. However the reason for the translocation could be the local similarity in the motif distribution.

As a control, we repeated the analysis in three pairs of chromosomes for regions 1 kb downstream of the gene and also for 1 kb regions chosen randomly from the chromosome. In each of these controls, the number of selected regions was the same as the number of gene promoter regions. We selected the following chromosome pairs: (6, 7, 12, 13) and (11, 18). The pair selections were done to include chromosomes of different sizes and gene densities (Supplementary Figure S3). We obtained the correlations of 5- and 6-mer motif vectors in

Table 1A. Comparison of gene-pair correlation above 0.5 in promoter and downstream versus random bins

1 kb genomic regions	Chr6–chr7	Chr12–chr13	Chr11–chr18
Promoter	0.98%	1.97%	0.99%
Downstream	0.59%	0.40%	0.41%
Random	0.23%	0.18%	0.15%

Table 1B. Distribution of genes with overlapping coordinates in different chromosomes

Chromosome	Total genes	% overlap
Chr 6	3086	22.65%
Chr 7	2891	21.90%
Chr 11	2995	20.50%
Chr 12	2575	24.82%
Chr 13	1768	18.27%
Chr 18	1022	22.20%

gene promoters, downstream regions and randomly selected regions.

For the three pairs of chromosomes chosen, there are between 892 167 and 4 552 601 possible gene–gene correlations. Of these, between 1–2% of the promoter correlations are ≥0.5 for these three pairs. Interestingly, the correlation between downstream regions is ∼0.5% in all three pairs (Table 1A and supplementary data of promoter versus downstream versus random correlations). This can be partially explained by the fact that in each of the selected chromosomes, ∼20% of the genes overlap with one another (Table 1B). It is likely therefore that promoters of one gene could be the downstream region of another and *vice versa*. Also many closely located genes may be controlled by a single promoter region. All this notwithstanding, there are however a significant number of high correlations between downstream regions that warrant further investigations. The random regions of 1 kb show the poorest correlations among all three pairs, accounting for less than ∼0.2% of all correlations. In fact, for the chromosome pair (12,13) none of the randomly chosen 1 kb regions correlate with a coefficient of ≥0.5. It is clear from these data that correlations between promoter regions, and to a smaller extent regions downstream of the gene, are distinctly different from correlating randomly selected 1 kb regions. We also compared the motif vectors correlations (Pearson's correlation) in the promoter regions of all genes between chr12–chr17 and chr18–chr19 with Spearman's rank correlation. We observed that the trend was similar for both chromosome pair 12–17 and 18–19 with R^2 values 0.83 and 0.79, respectively, for 1 kb 5-mer motif vector comparisons (Supplementary Figure S9 and supplementary data of Spearman's versus Pearson' s correlations).

We also identified that the high abundance (OE ≥ 5) of motifs of size 5 and 6 in the promoter regions holds true for a majority (∼50%) of the genes (Supplementary Table S3). The large number of genes whose promoters have a high abundance of motifs is suggestive that these motifs contribute to the recognition by some cognate protein regulators. On inspecting the motif distribution in promoter regions of 1 kb, we observed that on average ∼332 motifs out of 512 5-mer motif pairs and ∼1962 out of 2080 6-mer motif pairs have $OE_{1\ kb\_5/6\text{-}mer} \geq 5$ in at least one promoter region. These motifs having $OE_{1\ kb\_5/6\text{-}mer} \geq 5$ across different promoter-proximal control regions contribute significantly to the correlations. We have also identified seven 5-mers and sixteen 6-mers that have $OE_{1\ kb\_5/6\text{-}mer} \geq 10$ in at least 1% of the total genes in each chromosome (Supplementary Figure S10). These motifs show $OE_{1\ kb\_5/6\text{-}mer} \geq 10$ consistently across all the chromosomes. The seven 5-mer conserved motifs are a subset of the sixteen 6-mer motifs thus suggesting the conservation of the base specific recognition by the regulatory partners. The seven 5-meric motifs (CGGGG, GCGCG, GCGGG, GGCGG, GGGGC, GGGGG, GGAGG) are all GC motifs (with the exception of GGAGG). Even though ∼165 and ∼1178 different combinations of 5-mer and 6-mer motifs show a high abundance ($OE_{1\ kb\_5/6\text{-}mer} \geq 10$), only seven and sixteen of them are conserved across promoter proximal control regions. This implies that for differential regulation, different combinations of these high abundance motifs contribute to the high correlations.

Probing the significance of high motif vector correlations

The relevance of high correlations

We looked at three genes whose promoter motif vectors had high correlations (>0.9) to one another. These genes – LPHN1 (ADGRL1) (adhesion G protein-coupled receptor L1), CDK9 (cyclin dependant kinase 9) and TRIM8 (tripartite motif containing 8) – were from chromosomes 19, 9 and 10, respectively. The correlations between the motif vectors of (LPHN1 and CDK9), (LPHN1 and TRIM8) and (CDK9 and TRIM8) were 0.905, 0.903 and 0.910, respectively (Supplementary Table S4). The reason for the high correlation between these motif vectors is because of a few motifs that are common to all three with high OE ratio values (Supplementary Table S5). We next employed NetworkAnalyst (35,57) to determine if there were common transcription factors to these three genes (Figure 4A). Among the many transcription factors associated with these genes, two – SP1 and TFAP2A – were common to all three genes. We use JASPER (58,59) to get the consensus sequence recognized by these transcription factors (Supplementary Table S5).

The TF consensus for SP1 and TFAP2A from JASPAR can be interpreted as NNGGNNN[G/T][G/C/A][T/A] and GCCNNN[G/A][G/A/T][G/C] respectively. For SP1, a subset of three 5-mer motifs that can be derived from the consensus (GGCGG, CGGGG, GCCGG) have OE values above 10. Similarly, for TFAP2A, seven 5-mer motifs (GCGGC, GGGCG, GCCCG, GGGCC, GGCCG, GCCGG, GGGGC) derived from the consensus have OE values above 10 in all three gene promoters (Supplementary Table S5). It is important to note here that there were other 5-mers apart from those matching with the consensus that also had high OE values (Supplementary Table S5). From these data, it is clear that both the common transcription factors have an abundant number of binding sites on the promoters of all three genes.

An inspection of the tissues in which these three genes and their common two transcription factors are expressed reveals an interesting fact. The genes are almost ubiquitously expressed and the same is true for the transcription factor SP1 (https://www.ncbi.nlm.nih.gov/gene/6667). TFAP2A is however expressed in fewer tissue types (https://www.ncbi.nlm.nih.gov/gene/7020) with a high expression in skin, placenta and salivary gland. For instance, LPHN1, CDK9 and TRIM8 are all implicated in functions related to fat (60–64). TFAP2A

is expressed at low levels in adipose tissue. We conjecture that in fat, SP1 is the common regulator of all three genes. In other tissues, both SP1 and TFSP2A could assist in transcription, perhaps redundantly.

We also examined the relative abundances of motifs in the promoters for TFs that were common to two of the three proteins. We looked for motifs that were abundant in the promoters of two of the genes while being poorly represented in the other (Supplementary Table S5). The TFs E2F1, HINFP and RELA are the TFs that are common between (LPNH1 and TRIM8), (CDK9 and LPNH1) and (CDK9 and TRIM8) respectively. The JASPAR consensus motifs for these TFs can be interpreted as TTT[G/C][G/C]CG[C/G], [A/G]C/G]GTCCGC and [G/C/T][G/T]G[A/G]NTTTCC, respectively. Though there are TF consensus derived motifs that have higher OE ratio values in the promoters of the two genes sharing the same TF than the one not in common, the absolute values of the abundances are somewhat low. We noticed this for all three TFs E2F1, HINFP and RELA where 3, 5 and 1 motifs showed such relative abundances respectively. This may be indicative of the fact that differential regulation may not always involve TFs alone. As pointed out in the case of SP1 and TFAP2A, there are other motifs that are common to the promoters and present in high abundance but not linked to the TF. These are likely to be motifs that are recognized by other regulatory elements.

The illustration (Figure 4) with the three genes and two common transcription factors exemplifies the mechanism of differential control of gene regulation. To ensure that regulators, such as transcription factors, bind to the promoter region there is an abundance of such binding sites on the promoter. In the 1 kb promoter regions of LPHN1, CDK9 and TRIM8 there are (22, 22, 31, 16) and (16, 20) binding sites for (SP1, TFAP2A) respectively. In general, the larger the number of such binding sites in the promoter region, the greater the probability of binding of the regulatory element. The redundancy in binding sites also ensures robustness in the case of mutations. The composite set of motifs in promoter regions is such that their extent of correlation to other promoter regions varies. On one end of the spectrum, there are high correlations, where the same set of motifs are repeated several times in a pair (Supplementary Table S3 and Supplementary Figure S10). Then there are correlations that are just high enough to be above a threshold. In such cases, it is likely that a pair shares a few common motifs of high OE ratios while there are others with high OE ratios in one promoter and not the other (Supplementary Table S5). A single promoter could hence have motifs in common with several other promoters which in turn may not be correlated to one another. This would form the basis of differential regulation (65).

Constructing a gene regulatory network

Following up on the theme of how differential regulation can be affected in cells, we looked at the promoter regions of all genes that were known to be transcribed by the same transcription factor(s) – Jun/Fos (66,67). We selected a set of 19 genes whose transcription is controlled by the transcription factors Jun/Fos (68). The genes were identified in mouse while our analysis considered the human homologous of the mouse genes.

An all-versus-all correlation was done using the promoter motif vectors over this gene set (Figure 4B). Some of the genes are more closely correlated with a few of the other genes and the pattern of clusters that emerge has three major groups of 10, 4 and 6 genes. The genes were clustered into three groups based on the genes which have higher correlations within themselves compared to others in the list of genes. Within each cluster, the genes are better correlated to one another than they are to the genes in other clusters. There is a nuanced patterning of the high-scoring common motifs among these gene clusters that lead to higher intra-cluster correlation values. An interesting observation here is that the Fos/Jun binding consensus motif TGACTCA is not among the motifs with high OE ratio values. The motifs with high OE here that are in common to several genes (such as motif CGCGG present in genes BCL2, BCL2L11, HBEGF and EGFR) could be binding sites of some gene regulators (identity unknown) (supplementary data of AP1 network). This example shows how one could construct a gene regulatory network by clustering together genes whose promoters share the same motifs.

Spatially proximal genes have strong motif vector correlations

We rationalized that genes whose promoters are correlated are likely to be coregulated. Our next investigation was to check how often correlation/coregulation implied colocalization and whether we could predict it. Experimentally, colocalization is inferred using chemical cross linking such as in Hi-C (9,16). Data from Hi-C experiments are at the resolution of a few kb or even tens of kb. Our computations however are at a higher resolution as our inferences are drawn from 5/6 base motifs. More precisely, we are interested in the abundances of all 5/6-mer motifs in a given sequence stretch, variously considered as 1, 2 and 6 kb immediately upstream of genes. We also looked at promoter capture Hi-C data (pcHi-C), which captures distal promoter interacting regions for all promoters (69). We looked at the correlation of genes that overlap with these regions detected by Hi-C experiments and how they are correlated at different resolutions. In this study, we used two datasets: (i) Hi-C data of long-range chromatin contacts in HCT116 colon cancer cells (36) (accessible under GEO accession number GSE18215) and (ii) pcHi-C data, which looks at novel gene contacts associated with autoimmune risk loci (37) (accessible under GEO accession number: GSM1704495).

Three important caveats apply to our analysis: (i). Hi-C records cross links across the genome without any prejudice to any region(s). Our data, in this study, is restricted to correlations in gene promoter regions only. This implies that we would only be able to compare our correlation data to a subset of Hi-C data that cross link promoter regions of genes. We make a higher number of predictions as we consider all possible gene pair correlations between the two chromosomes. (ii). The resolution of data for the both the Hi-C and the pcHi-C experiments are in the order of ∼10 kb, whereas our predictions, that examine 5/6 bp, are of higher resolution. (iii). Different Hi-C experiments use different genome reference assemblies to establish the genomic coordinates between two chromosomes that are in proximity. The gene annotations vary with each reference assembly (Supplementary Table S6). We have used all the gene annotations according to the latest T2T reference assembly. Thus, it is likely that we are missing out on some gene correlations, because of discrepancies in the annotations across different reference assemblies.

Here we have compared our motif vector correlations to Hi-C data for a couple of chromosomes – 18 and 19 – that have 1022 (spread over ∼80.5 Mbps) and 2531 genes (spread over ∼61.7 Mbps), respectively. These two chromosomes

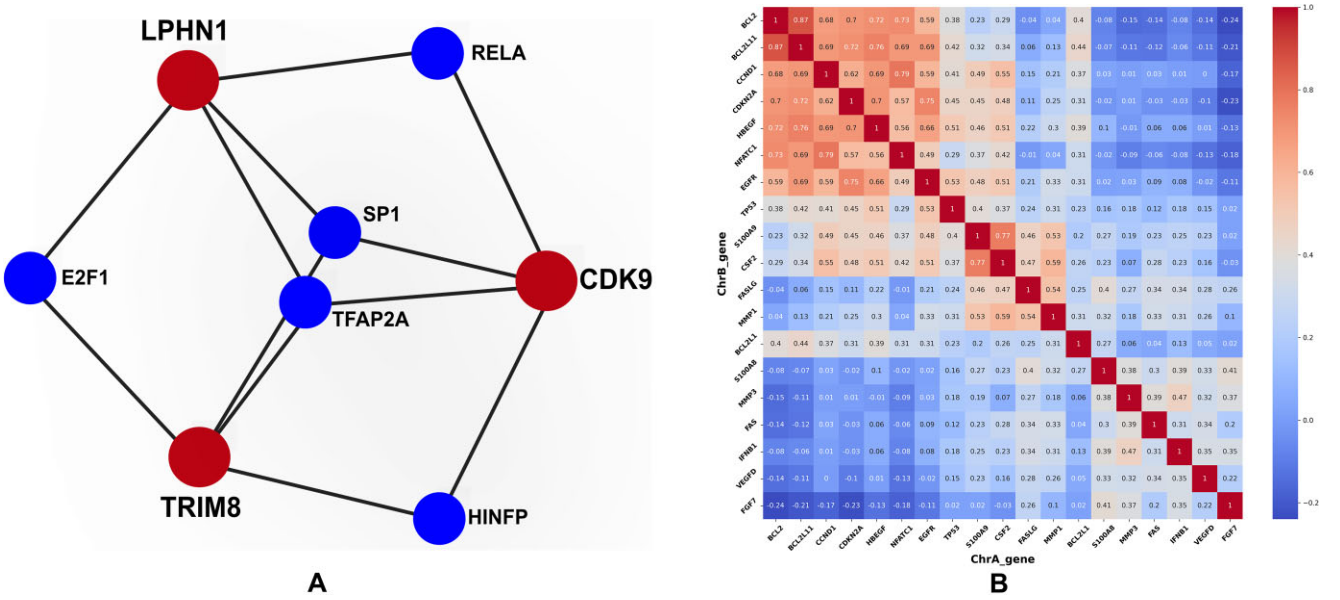


Figure 4. Panel (**A**) is a graph representation of a gene regulatory network involving three genes LPHN1, CDK9 and TRIM8 (shown as red nodes). Each such red node is connected via edges to transcription factors controlling their expression (cyan nodes). Panel (**B**) shows a gene–gene interaction heat map of $OE_{1\ kb\_5\text{-}mer}$ for 18 genes that are all regulated by the transcription factor Jun-Fos. The higher (or lower) the correlation the darker the shade of red (or bluer) is the grid.

were chosen because they localize differently in the nucleus – 18 is peripheral while 19 is mostly internalized (70,71) and have contrasting gene densities (72) (Supplementary Figure S3). The Hi-C data provides the contact links for different parts of the two chromosomes. We obtained the coordinates of the contacts between the two chromosomes and obtained the genes whose coordinates overlap with the Hi-C contacts. The extraction of the gene information was done using the respective reference assembly that was used to generate the Hi-C data.

According to the 10 kb bin long range chromatin contact Hi-C data, there are 1 327 617, 699 290 and 23 688 Hi-C contacts intra-chromosome-18, intra-chromosome-19 and inter-18–19, respectively, according to the GRCh38 reference assembly. Of these, only 12 790, 115 595 and 926 are contacts overlap with regions having genes in intra-18, intra-19 and inter-18–19, respectively. Overlap with gene pairs is established when both Hi-C contact bins have a gene (in whole or part). Expectedly, the number of inter-18–19 contacts is small in comparison to the intra-chromosome contacts given their different localizations. The aim here was to determine if the gene pairs whose promoter motif vector are correlated are also physically proximal. Of the overlaps determined, only 12 213, 111 730 and 890 gene-pairs find a match with the current T2T gene annotation list which is our reference for the motif vector correlations (supplementary data of Hi-C validation).

We identify the genes having motif vector correlation scores ≥0.5 from the inter and intra chr18, chr19 Hi-C contacts. We see enrichment in the motif vector correlations in the Hi-C contacts for all the inter and intra-Hi-C contacts at all resolutions of motif vector correlations at different sizes (Table 2). We also verified if the high correlations between the gene promoter are a result of sequential proximity between the genes/promoters. As a control, we plotted the distance between the genes against the correlation of their promoter regions. We observe that high correlations between gene promoter regions are not dependent on the sequential proximity of the genes (Supplementary Figure S11). It is to be noted that this comparison was done only across intra-gene correlations as we cannot determine the sequential proximity of inter-chromosome correlations.

Table 2. Comparison of gene-pair correlation inter- and intra-chromosome 18 and chromosome 19 in Hi-C contacts versus all

Chr-pair	Gene pairs with motif vector correlation ≥0.5		Gene pairs in Hi-C with motif vector correlation ≥0.5	
	1 kb 5-mer	1 kb 6-mer	1 kb 5-mer	1 kb 6-mer
Chr 18–chr 19	9.92%	0.47%	14.38%	1.01%
Chr 18–chr 18	9.96%	1.18%	18.84%	7.72%
Chr 19–chr 19	15.10%	0.90%	22.18%	3.74%

We repeated this analysis using pcHi-C data using chromosome pairs 18 and 19 and obtained 3655 contacts. Of these, 123 gene coordinates overlap with the pcHi-C contacts according to the GRCh37 reference assembly (Supplementary Table S7 and supplementary data of Hi-C validations). 118 of the 123 gene-pairs match with our correlations using current T2T reference assembly gene annotations (Table 3 and Supplementary Table S6).

Of all the 2 586 682 possible gene pairs of chromosomes 18 and 19 256 623 (9.92% of 1 kb_5-mer) have motif vector correlations of ≥0.5 (Supplementary Table S7). A total of 128 (14.32%) gene pairs with Hi-C contacts have a motif vector correlation ≥0.5, an enrichment of ∼5%. This number is 27 (∼23%%) for pcHi-C. Interestingly there is no overlap between the Hi-C and the pcHi-C data (Supplementary Figure S12). Cumulatively, the motif vector correlations of ≥0.5 identify 155 pairs of genes that are experimentally linked to one another, which is an enrichment of 37%. Given that the two different types of Hi-C have no overlaps, it is possible that

Table 3. Distribution of gene pairs contributing to high correlation in Hi-C and pcHi-C

Dataset	Total correlations			Correlations ≥0.5 (1 kb 5-mer)		
	18-19	18	19	18-19	18	19
All gene correlations	2 586 682	1022	2531	256 623 (9.92%)	885	2238
Hi-C 10 kb bins with genes	890[a]	412[a]	655[a]	128 (14.38%)	100	112
pcHi-C bins with genes	118[a]	20[a]	94[a]	27 (22.88%)	12	23

[a]the numbers are with reference to genes matching with the T2T reference assembly.

many more of the gene pairs identified by our correlations could be spatially proximal under certain conditions.

Topologically associated domains (TADs) are well defined chromosome segments that make more frequent genomic contacts to one another and are hence purportedly in spatial proximity (73). It has been established that binding sites for the protein CTCF, characterized by the motif CCCTC, mark the TAD boundaries (74,75). The CCCTC motif is among the top 50 most abundantly present motifs in the human genome (43rd on a list of 512) with an average $OE_{1\ kb\_5me}$ value of 1.85 (supplementary data of OE_{whole} rations). On closer inspection, we observed that 349 (39.2%) and 67 (56.8%) out of the 890 and 118 gene pairs in pcHi-C and Hi-C respectively show an $OE_{1\ kb\_5\text{-}mer}$ of the CTCF motif >3 in both of the genes in contact (supplementary data of Hi-C validation). This number is enriched when we look at gene pairs with correlation of ≥0.5. We observed that 93 (72.6%) and 23 (85.2%) out of the 128 and 27 gene pairs with correlation ≥0.5 in pcHi-C and Hi-C respectively show an $OE_{1\ kb\_5\text{-}mer}$ of the CTCF motif >3 in both of the genes. Interestingly, the $OE_{1\ kb\_5\text{-}mer}$ for CCCTC in the promoter regions of gene pairs that have a motif vector correlation of ≥0.5 and are proximal to Hi-C or the pcHi-C contacts in chromosomes 18 and 19 has a maximum value of ∼20, with a mean of ∼3.5–4.5. (Supplementary Figure S13 and supplementary data of Hi-C validation). In contrast, the $OE_{1\ kb\_5\text{-}mer}$ for the other motifs in the genes in proximity, do not exceed a maximum of 2.5, with their means ranging from ∼1 to 1.5. Thus, the CTCF motif is one of the major contributors to the high correlation (correlation coefficient ≥0.5) for gene pairs in proximity.

The motif vector correlations are high among gene pairs that also have Hi-C contacts. This opens up the possibility that Hi-C contacts could be predicted from gene correlations. There is no overlap between the 123 and 926 gene pairs identified by Hi-C and pcHi-C respectively. Though these data involve 15 and 33 genes from chromosomes 18 and 19, respectively, they are always matched with different partner genes in the two sets of data (Supplementary Figure S12).

Discussions

In this study, we have attempted to read parts of the genome just as they would be perceived by the proteins that interact with them, i.e. 4 to 6 base pairs at a time. These are data we have obtained by observing DNA–protein complexes. For our analysis, we simplistically and separately contend that DNA binding proteins recognize their cognate sites on genomes by making interaction with either 5 or 6 bases per domain. Some databases list consensus binding sites that are longer – but these are usually in cases where the DNA binding domains dimerise.

How are 5/6 mers of DNA distributed in the genome? There are a total of 512 and 2080 unique 5-mers and 6-mers respectively that were obtained by averaging the values of motifs and their corresponding reverse complements. We have shown that the distribution of these 5/6-mers is non-random. In real genomes, there are regions of the genome where certain motifs are over or under-represented. We have quantified this using the OE ratios. The non-randomness of the motif distributions happens at the scale of the whole chromosome or even at the level of smaller segments of the genome such as 100 kb stretches or even around centromeric regions across all chromosomes. Using the simple metric of $OE_{centromere}$ ratios, we establish a unique pattern of motif distributions conserved across the different acrocentric chromosomes (13–15,21) and (22). The centromeres of these chromosomes are known to have segmental duplications and our analysis using only the motif distributions in these regions also identifies similarities.

Having established that the motif distributions follow non-random patterns we next investigated its possible relevance. For this, we first correlated (Pearson correlation) the motif distributions in promoter regions of genes. To do this we constructed motif vectors of the 5/6-mer motifs where the coefficients of the vectors are the OE ratios of the motifs. After establishing that the motif distributions follow non-random patterns we next investigated its possible relevance. For this, we first correlated (Pearson correlation) the motif distributions in promoter regions of genes. To do this we constructed motif vectors of the 5/6-mer motifs where the coefficients of the vectors are the OE ratios of the motifs. We observed that certain 5- and 6-mer motifs have high abundance across the majority of the promoter regions of genes (∼50%). We identified seven 5-mers and sixteen 6-mers that are conserved across the promoter regions in different chromosomes. These motifs of high abundance suggest identification by some cognate protein regulators. When the correlation values of all genes are taken together chromosome-wise, we observed that there were some chromosome pairs with very strong correlations (≥0.9). We have highlighted two such clusters of high correlations – between chromosomes (13–15, 21 and 22) and (4,10). In both these cases there are documented instances of chromosome translocations amongst cluster members. The (13–15, 21, 22) clusters are the same acrocentric chromosomes discussed above. The mix and match of chromosomal segments between these genes is termed Robertsonian translocations and is implicated in genetic diseases such as Patau syndrome (76) and Down syndrome (77). Translocations between chromosomes 4 and 10 are associated with certain types of leukaemia (54,55) and Wolf–Hirschhorn syndrome (56). Our somewhat simplistic motif vector correlations also pick up on these phenomena. Translocations may only be possible when there is a high degree of sequence similarity/identity between the regions of translocation. However, not all regions sharing

high similarity/identity may be susceptible to translocations as it may require the presence of certain motif(s). What motif(s) and their distributions can only be gauged with a focused analysis of the translocating segments. The high abundances of promoter correlations are also an attribute of the tandem repeats in the promoter regions of the rRNA genes across these chromosomes. The patterns of high correlation across gene promoters are also distinct from that of downstream regions of genes or random regions in the chromosomes.

From the level of chromosomes, we went down to the level of individual genes, more precisely to the promoter region of genes. We showed that when two genes have high correlations between their promoter vector motifs there is usually a functional implication. For instance, the genes LPNH1 from chromosome 19, CDK9 from chromosome 9 and TRIM8 from chromosome 10 are all strongly correlated to one another (correlation coefficient >0.9). The reason for the high correlations there is a common set of motifs that are over/under-represented in all their promoter regions. Examining the over-abundant motifs, we found that the motifs GGCGG and CCGGG were among the ones with the highest OE ratios in all their promoters and these were the binding sites of transcription factors SP1 and TFAP2A. These transcription factors are differentially expressed in different tissue types, for instance, TFAP2A is not found in fat while SP2 is. This, we speculate is the basis of differential gene regulation in different cell/tissue types. Further, we could construct an entire gene regulatory network by associating gene promoter regions with significant correlations. While we are constructing this network, it is too vast and somewhat beyond the scope of the current investigation to be reported here.

Having introduced the possibility of obtaining gene regulatory networks, we illustrate this with an example of promoter correlations between all co-regulated genes, i.e. genes transcribed by a common transcription factor, Jun/Fos. Here, we found some gene clusters to be strongly correlated amongst cluster members but not necessarily with members of other clusters. The reason for this is again the fact that the correlations are strong because of a common set of motifs having similar abundances (OE ratios) across promoter regions. These motifs are then of functional importance as established in the case of LPHN1, CDK9 and TRIM8. In this illustrative network, however, the common high abundance motifs are not transcription binding sites (even though some of the pairs within the network share common transcription factors other than Jun/Fos). The Jun/Fos binding motifs themselves are not among the motifs with the highest ratios. We speculate that the over-abundant common set of motifs could be binding sites for common regulatory elements, perhaps hitherto undiscovered. Since the high-abundance motifs in promotors within a single cluster are not identical between cluster members, it opens up the possibility of differential regulation.

Our final investigation was to examine whether high gene promoter correlation also implied physical proximity. Here the results are interesting and suggestive of the predictive power of these motif vector correlations. We compared our results to Hi-C data that show physicalproximity between genic regions using cross-linking. Promoter pairs that overlapped with Hi-C data had an enrichment for high motif vector correlations (coefficients ≥0.5). This enrichment was consistent when the analysis was repeated with pcHi-C. While processing this result, we should bear in mind that the number of gene–gene promoter correlations with coefficients >0.5 is at least an order of magnitude greater than the amount of Hi-C data available to us. While it is unlikely that all strongly correlated gene–gene promoter regions imply physical proximity, it is possible that many of them do, even though there is no Hi-C data to support the claim. The reason for this could be that Hi-C data is typically of a resolution of 10 kb (in many cases >10 kb), while we are examining the genome at a higher resolution (length scale of 5/6 base pairs). The promoter regions of the genes in Hi-C contacts also showed a high abundance of the CTCF binding motifs that regulates the TAD organization. One significant observation we made was the lack of overlap between the different Hi-C data. The gene pairs present in the bins with the Hi-C contacts do not overlap with the pcHi-C data. This can be an attribute of the difference in the reference assembly used to generate the Hi-C and also the possible different crosslinks in different cell types. Our observations on high correlations are made with the latest reference assembly, and are independent of cell type. One other reason why some of our high correlations are not validated by Hi-C data could be that even though the genes are in close proximity, they may not have been close enough to be detected by Hi-C, but could be close enough to be co-regulated. We are also aware that spatial proximity may not always follow from strong motif vector correlations even though the genes are co-regulated. For instance, transcription factors that have peripheral locations on the nucleus could co-regulate two (or more) genes that are spatially distant but are also somewhat close to the nuclear outer periphery. These results open up the possibility of predicting, at a high resolution, 3D proximities within genomes. Potentially, this could give us a more nuanced map of the genomic arrangement within the nucleus and how it changes from one cell type to another or even during different cell phases.

Typically, analysis of genes/promoters/genomes etc. involves sequence alignments. While the benefits of alignments are undeniable, this study shows that it is insightful and significant to read the genome 5/6 bases at a time, just as proteins that bind DNA do. From a rather simplistic consideration of the distribution of 5/6-mers in the chromosome or smaller stretches of the genomes (such as promoter regions), we can make fundamental connections between genes and chromosomes. For instance, given that we can detect patterns where translocations are likely to occur, predict the proximity of correlated (co-regulated) genes, and give ourselves the ability to construct regulatory networks, we see great benefit in examining motif abundances in the genome.

Data availability

Data has been deposited in Zenodo at https://doi.org/10.5281/zenodo.14055970.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We would like to acknowledge Prof. Chandra Verma, Prof. Kundan Sengupta, Prof. Richa Ricky, Prof. Leelavati Narlikar, S. Mukundan, Avadhoot Jadhav and other COSPI lab members for constructive criticisms. We also acknowledge the support and computational resources provided by the PARAM

Brahma facility under the National Supercomputing Mission, Government of India at the Indian Institute of Science Education and Research Pune.

Author contributions: Idea conception: M.S.M. with help from A.C. Data generation and analysis: A.C. with help from S.C. Manuscript writing: A.C. and M.S.M.

Funding

Indian Institute of Science Education and Research Pune; Department of Biotechnology, Government of India [BT/PR40262/BTIS/137/38/2022]. Funding for open access charge: DBT [BT/PR40262/BTIS/137/38/2022].

Conflict of interest statement

None declared.

References

1. Aganezov,S., Yan,S.M., Soto,D.C., Kirsche,M., Zarate,S., Avdeyev,P., Taylor,D.J., Shafin,K., Shumate,A., Xiao,C., *et al.* (2022) A complete reference genome improves analysis of human genetic variation. *Science*, **376**, eabl3533.
2. Nurk,S., Koren,S., Rhie,A., Rautiainen,M., Bzikadze,A.V., Mikheenko,A., Vollger,M.R., Altemose,N., Uralsky,L., Gershman,A., *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.
3. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
4. Bejerano,G., Haussler,D. and Blanchette,M. (2004) Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics*, **20**, i40–i48.
5. Guturu,H., Doxey,A.C., Wenger,A.M. and Bejerano,G. (2013) Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **368**, 20130029.
6. Fedoroff,N.V. (2012) Transposable elements, epigenetics, and genome evolution. *Science*, **338**, 758–767.
7. Vitsios,D., Dhindsa,R.S., Middleton,L., Gussow,A.B. and Petrovski,S. (2021) Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat. Commun.*, **12**, 1504.
8. Sémon,M. and Duret,L. (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.*, **23**, 1715–1723.
9. Schneider,V.A., Graves-Lindsay,T., Howe,K., Bouk,N., Chen,H.C., Kitts,P.A., Murphy,T.D., Pruitt,K.D., Thibaud-Nissen,F., Albracht,D., *et al.* (2017) Chow. *Genome Res.*, **27**, 849–864.
10. Church,D.M. (2022) A next-generation human genome sequence. *Science*, **376**, 34–35.
11. Cozzolino,F., Iacobucci,I., Monaco,V. and Monti,M. (2021) Protein-DNA/RNA interactions: an overview of investigation methods in the omics era. *J. Proteome Res.*, **20**, 3018–3030.
12. Dekker,J. and Mirny,L. (2016) The 3D genome as moderator of chromosomal communication. *Cell*, **164**, 1110–1121.
13. Comfort,N.C. (2001) From controlling elements to transposons: barbara McClintock and the Nobel Prize. *Trends Genet.*, **17**, 475–478.
14. Misteli,T. (2010) Higher-order genome organization in human disease. *Cold Spring Harb. Perspect. Biol.*, **2**, a000794.
15. de Wit,E. and de Laat,W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev.*, **26**, 11–24.
16. Tjong,H., Li,W., Kalhor,R., Dai,C., Hao,S., Gong,K., Zhou,Y., Li,H., Zhou,X.J., Le Gros,M.A., *et al.* (2016) Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl Acad. Sci. U.S.A.*, **113**, E1663–E1672.
17. Van Steensel,B. and Henikoff,S. (2000) Identification of *in vivo* DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat. Biotechnol.*, **18**, 424–428.
18. Guelen,L., Pagie,L., Brasset,E., Meuleman,W., Faza,M.B., Talhout,W., Eussen,B.H., De Klein,A., Wessels,L., De Laat,W., *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948–951.
19. Solomon,M.J., Larsen,P.L. and Varshavsky,A. (1988) Mapping protein-DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, **53**, 937–947.
20. Orlando,V. (2000) Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci.*, **25**, 99–104.
21. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
22. Henikoff,S., Henikoff,J.G., Kaya-Okur,H.S. and Ahmad,K. (2020) Efficient chromatin accessibility mapping *in situ* by nucleosome-tethered tagmentation. *eLife*, **9**, e63274.
23. Nguyen,H.Q., Chattoraj,S., Castillo,D., Nguyen,S.C., Nir,G., Lioutas,A., Hershberg,E.A., Martins,N.M.C., Reginato,P.L., Hannan,M., *et al.* (2020) 3D mapping and accelerated super-resolution imaging of the human genome using *in situ* sequencing. *Nat. Methods.*, **17**, 822–832.
24. Nir,G., Farabella,I., Pérez Estrada,C., Ebeling,C.G., Beliveau,B.J., Sasaki,H.M., Lee,S.H., Nguyen,S.C., McCole,R.B., Chattoraj,S., *et al.* (2018) Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.*, **14**, e1007872.
25. Boninsegna,L., Yildirim,A., Polles,G., Zhan,Y., Quinodoz,S.A., Finn,E.H., Guttman,M., Zhou,X.J. and Alber,F. (2022) Integrative genome modeling platform reveals essentiality of rare contact events in 3D genome organizations. *Nat. Methods*, **19**, 938–949.
26. Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison—a review. *Bioinformatics*, **19**, 513–523.
27. Reinert,G., Chew,D., Sun,F. and Waterman,M.S. (2009) Alignment-free sequence comparison (I): statistics and power. *J. Comput. Biol.*, **16**, 1615–1634.
28. Sarkar,B.K., Sharma,A.R., Bhattacharya,M., Sharma,G., Lee,S.S. and Chakraborty,C. (2021) Determination of k-mer density in a DNA sequence and subsequent cluster formation algorithm based on the application of electronic filter. *Sci. Reports*, **11**, 13701.
29. Saw,A.K., Raj,G., Das,M., Talukdar,N.C., Tripathy,B.C. and Nandi,S. (2019) Alignment-free method for DNA sequence clustering using Fuzzy integral similarity. *Sci. Rep.*, **9**, 3753.
30. Wingender,E. (1988) Compilation of transcription regulating proteins. *Nucleic Acids Res.*, **16**, 1879–1902.
31. Wingender,E., Schoeps,T., Haubrock,M., Krull,M. and Dönitz,J. (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.*, **46**, D343–D347.
32. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.
33. Chiu,T.P., Xin,B., Markarian,N., Wang,Y. and Rohs,R. (2020) TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **48**, D246–D255.
34. Zhou,G., Soufan,O., Ewald,J., Hancock,R.E.W., Basu,N. and Xia,J. (2019) NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.*, **47**, W234–W241.

35. Xia,J., Gill,E.E. and Hancock,R.E.W. (2015) NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.*, **10**, 823–844.
36. Spracklin,G., Abdennur,N., Imakaev,M., Chowdhury,N., Pradhan,S., Mirny,L.A. and Dekker,J. (2022) Diverse silent chromatin states modulate genome compartmentalization and loop extrusion barriers. *Nat. Struct. Mol. Biol.*, **30**, 38–51.
37. Martin,P., McGovern,A., Orozco,G., Duffus,K., Yarwood,A., Schoenfelder,S., Cooper,N.J., Barton,A., Wallace,C., Fraser,P., *et al.* (2015) Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.*, **6**, 10069.
38. O'Leary,N.A., Cox,E., Holmes,J.B., Anderson,W.R., Falk,R., Hem,V., Tsuchiya,M.T.N., Schuler,G.D., Zhang,X., Torcivia,J., *et al.* . (2024) Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data*, **11**, 732.
39. Rangwala,S.H., Kuznetsov,A., Ananiev,V., Asztalos,A., Borodin,E., Evgeniev,V., Joukov,V., Lotov,V., Pannu,R., Rudnev,D., *et al.* . (2021) Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). *Genome Res.*, **31**, 159–169.
40. Nassar,L.R., Barber,G.P., Benet-Pagès,A., Casper,J., Clawson,H., Diekhans,M., Fischer,C., Gonzalez,J.N., Hinrichs,A.S., Lee,B.T., *et al.* (2023) The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.*, **51**, D1188–D1195.
41. Nair,S. and Madhusudhan,M.S. (2022) JEDII: juxtaposition enabled DNA-binding Interface Identifier. bioRxiv doi: https://doi.org/10.1101/2022.05.19.492702, 20 May 2022, preprint: not peer reviewed..
42. Gonzalez,D., Luyten,A., Bartholdy,B., Zhou,Q., Kardosova,M., Ebralidze,A., Swanson,K.D., Radomska,H.S., Zhang,P., Kobayashi,S.S., *et al.* (2017) ZNF143 protein is an important regulator of the myeloid transcription factor C/EBPα. *J. Biol. Chem.*, **292**, 18924.
43. Humphray,S.J., Oliver,K., Hunt,A.R., Plumb,R.W., Loveland,J.E., Howe,K.L., Andrews,T.D., Searle,S., Hunt,S.E., Scott,C.E., *et al.* (2004) DNA sequence and analysis of human chromosome 9. *Nature*, **429**, 369–374.
44. Sinclair,D.A.R., Schulze,S., Silva,E., Fitzpatrick,K.A. and Honda,B.M. (2000) Essential genes in autosomal heterochromatin of *Drosophila melanogaster*. *Genetica*, **109**, 9–18.
45. Eberl,D.F., Duyf,B.J. and Hilliker,A.J. (1993) The role of heterochromatin in the expression of a heterochromatic gene, the rolled locus of *Drosophila melanogaster*. *Genetics*, **134**, 277–292.
46. Rhie,A., Nurk,S., Cechova,M., Hoyt,S.J., Taylor,D.J., Altemose,N., Hook,P.W., Koren,S., Rautiainen,M., Alexandrov,I.A., *et al.* (2023) The complete sequence of a human Y chromosome. *Nature*, **621**, 344.
47. Stadhouders,R., Vidal,E., Serra,F., Di Stefano,B., Le Dily,F., Quilez,J., Gomez,A., Collombet,S., Berenguer,C., Cuartero,Y., *et al.* (2018) Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat. Genet.*, **50**, 238.
48. Groves,A.K., George,K.M., Tissier-Seta,J.P., Engel,J.D., Brunet,J.F. and Anderson,D.J. (1995) Differential regulation of transcription factor gene expression and phenotypic markers in developing sympathetic neurons. *Development*, **121**, 887–901.
49. Hartley,G. and O'neill,R.J. (2019) Centromere repeats: hidden gems of the genome. *Genes (Basel)*, **10**, 223.
50. Vollger,M.R., Guitart,X., Dishuck,P.C., Mercuri,L., Harvey,W.T., Gershman,A., Diekhans,M., Sulovari,A., Munson,K.M., Lewis,A.P., *et al.* (2022) Segmental duplications and their variation in a complete human genome. *Science*, **376**, 6588.
51. Hurst,L.D., Sachenkova,O., Daub,C., Forrest,A.R.R. and Huminiecki,L. (2014) A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol.*, **15**, 413.
52. Saeed,S., Hassan,J., Javed,S.M., Shan,S. and Naz,M. (2022) A familial case of Robertsonian translocation 13;14: case report. *Cureus*, **14**, e29430.
53. Spinner,N.B., Conlin,L.K., Mulchandani,S. and Emanuel,B.S. (2013) Deletions and other structural abnormalities of the autosomes. *Emery Rimoin's Princ. Pract. Med. Genet.*, https://doi.org/10.1016/B978-0-12-383834-6.00051-3.
54. Harris,M., Shuster,J., Carroll,A., Look,A., Borowitz,M., Crist,W., Nitschke,R., Pullen,J., Steuber,C. and Land,V. (1992) Trisomy of leukemic cell chromosomes 4 and 10 identifies children with B-progenitor cell acute lymphoblastic leukemia with a very low risk of treatment failure: a pediatric oncology group study. *Blood*, **79**, 3316–3324.
55. Wong,K.F. and So,C.C. (2001) Acute myeloid leukemia with concomitant trisomies 4 and 10: a distinctive form of myeloid leukemia? *Cancer Genet. Cytogenet.*, **127**, 74–76.
56. Goodship,J., Curtis,A., Cross,I., Brown,J., Emslie,J., Wolstenholme,J., Bhattacharya,S. and Burn,J. (1992) A submicroscopic translocation, t (4;10), responsible for recurrent Wolf–Hirschhorn syndrome identified by allele loss and fluorescent *in situ* hybridisation. *J. Med. Genet.*, **29**, 451.
57. Xia,J., Benner,M.J. and Hancock,R.E.W. (2014) NetworkAnalyst - Integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic Acids Res.*, **42**, W167–W174.
58. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.Y., Chou,A., Ienasescu,H., *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
59. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A., Perez,N.M., *et al.* (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **50**, D165–D173.
60. Martinez,A.F., Muenke,M. and Arcos-Burgos,M. (2011) From the black widow spider to human behavior: latrophilins, a relatively unknown class of G protein-coupled receptors, are implicated in psychiatric disorders. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, **156B**, 1.
61. Itkonen,H.M., Poulose,N., Walker,S. and Mills,I.G. (2019) CDK9 inhibition induces a metabolic switch that renders prostate cancer cells dependent on fatty acid oxidation. *Neoplasia*, **21**, 713–720.
62. Yan,F.J., Zhang,X.J., Wang,W.X., Ji,Y.X., Wang,P.X., Yang,Y., Gong,J., Shen,L.J., Zhu,X.Y., Huang,Z., *et al.* (2017) The E3 ligase tripartite motif 8 targets TAK1 to promote insulin resistance and steatohepatitis. *Hepatology*, **65**, 1492–1511.
63. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S., *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **50**, D20–D26.
64. Fagerberg,L., Hallstrom,B.M., Oksvold,P., Kampf,C., Djureinovic,D., Odeberg,J., Habuka,M., Tahmasebpoor,S., Danielsson,A., Edlund,K., *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics*, **13**, 397–406.
65. Shabalina,S.A., Spiridonov,A.N., Spiridonov,N.A. and Koonin,E.V. (2010) Connections between alternative transcription and alternative splicing in mammals. *Genome Biol. Evol.*, **2**, 791.
66. Shaulian,E. and Karin,M. (2002) AP-1 as a regulator of cell life and death. *Nat. Cell Biol.*, **4**, E131–E136.
67. Zenz,R. and Wagner,E.F. (2006) Jun signalling in the epidermis: from developmental defects to psoriasis and skin tumors. *Int. J. Biochem. Cell Biol.*, **38**, 1043–1049.
68. Zenz,R., Eferl,R., Scheinecker,C., Redlich,K., Smolen,J., Schonthaler,H.B., Kenner,L., Tschachler,E. and Wagner,E.F. (2008) Activator protein 1 (Fos/Jun) functions in inflammatory bone and skin disease. *Arthritis Res. Ther.*, **10**, 201.

69. Schoenfelder,S., Javierre,B.M., Furlan-Magaril,M., Wingett,S.W. and Fraser,P. (2018) Promoter capture Hi-C: high-resolution, genome-wide profiling of promoter interactions. *J. Vis. Exp.*, **2018**, 57320.
70. Croft,J.A., Bridger,J.M., Boyle,S., Perry,P., Teague,P. and Bickmore,W.A. (1999) Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.*, **145**, 1119.
71. Boyle,S., Gilchrist,S., Bridger,J.M., Mahy,N.L., Ellis,J.A. and Bickmore,W.A. (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.*, **10**, 211–220.
72. Vitalini,M.W., Dialynas,G., Wallrath,L.L., Mackey,S.R. and Stainbrook,S.C. (2021) Nuclear organization, chromatin structure, and gene silencing. *Encycl. Biol. Chem. Third Ed.*, **5**, 393–397.
73. Sefer,E. (2022) A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinformatics*, **23**, 127.
74. Nanni,L., Ceri,S. and Logie,C. (2020) Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. *Genome Biol.*, **21**, 197.
75. Liu,Y. and Dekker,J. (2022) CTCF–CTCF loops and intra-TAD interactions show differential dependence on cohesin ring integrity. *Nat. Cell Biol.*, **24**, 1516–1527.
76. Kuznetsova,M.A., Zaytseva,G.V., Zryachkin,N.I., Makarova,O.A. and Khmilevskaya,S.A. (2023) Patau Syndrome. *Vopr. Prakt. Pediatr.*, **10**, 90–93.
77. Ganguly,B.B (2022) In: *Genetics and Neurobiology of Down Syndrome*. Academic Press.

Received: March 4, 2024. **Revised:** September 17, 2024. **Editorial Decision:** November 5, 2024. **Accepted:** November 15, 2024