Computational and Structural Biotechnology Journal 17 (2019) 195–205

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Revealing Gene Function and Transcription Relationship by Reconstructing Gene-Level Chromatin Interaction

Li Liu [a,*], Qian-Zhong Li [a,*], Wen Jin [a], Hao Lv [b], Hao Lin [b,*]

[a] Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
[b] Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE INFO

Article history:
Received 7 December 2018
Received in revised form 24 January 2019
Accepted 28 January 2019
Available online 31 January 2019

Keywords:
Transcription factories
Co-transcription
Dynamic chromatin interaction
Gene function

ABSTRACT

Chromatin is hierarchically organized in human interphase nuclei. Dynamic chromatin interactions are thought to influence gene transcription and cell fate determination. A consensus concept is that genes may form transcription factories within nucleus by spatially interaction. However, it is still not well known whether the function-related genes co-locate in three-dimensional (3D) space for co-transcription. Especially, there is a lack of visualization method that directly reflect the relationship between gene spatial interaction, gene function and co-transcription. In this study, we constructed three kinds of matrices based on gene ontology annotations, high-through chromosome conformation capture (Hi-C) data and RNA-seq data from twenty human tissues and cell lines. The comparative analysis for gene pairs revealed that 3D genome organization influences gene transcription predominantly at local scale. We found that the local genes within family clusters have similar transcription patterns. We also found that spatial reorganization of a histone gene cluster could control gene transcription. These observations suggest that function-related genes are close in space and activated or repressed together. Our work provided a framework for genome-wide studying the relationship between gene function, co-transcription and spatial interaction.

1. Introduction

Why one genome sequence can give rise to so many different cell types in organism development, cell differentiation and disease occurrence? We now know that the driver is the temporal and spatial difference expression of genes, in which the gene transcriptional regulation plays a key role. Transcriptional regulation is a complex biological process involving a large number of interactions among DNA regulatory elements (promoters, enhancers, silencers, terminators etc), RNAs (mRNAs, tRNAs, non-coding RNAs etc) and proteins (transcription factors, histones, enzymes etc) [1–9]. The basic process of transcriptional control has been well described that transcription factors occupy specific DNA elements such as promoters or enhancers and then recruit RNA polymerase and cofactors to target genes [10–13]. Recent discoveries show that histone modifications, so-called "histone code", control the docking of regulatory factors with DNA elements by altering chromatin accessibility [14]. Furthermore, the discovery of dynamic and hierarchical chromatin structure added a new dimension to gene transcriptional regulation.

More and more evidences have proved that chromosomes are dynamically reorganized to suit the cell's needs [15–17]. They are compressed into compact bodies in mitosis, and decompressed in interphase to allow gene expression. Recently, some observations based on Hi-C method [1] coupled with super-resolution imaging technique [18] have revealed the hierarchical structure of chromatin organization for interphase chromosomes. The Hi-C-based methods were used to assess contacts for millions of loci simultaneously by averaging chromosome conformations from millions of nuclei. The microscopy-based methods directly image the spatial organization of single chromosome in single cell. These two methods have unraveled that chromatins are partitioned into active (A) and inactive (B) compartments in a polarized manner. The A/B compartments can change dynamically across cell types, and these changes are associated with gene transcription. Another kind of substructure within chromatin compartments called topologically associating domains (TADs) has been identified in mammalian genomes [19]. Unlike A/B compartments, the TADs are stable across different cell types and highly conserved across species, suggesting that TADs are the units of dynamic alterations in chromosome compartments. By analyzing the spatial organization of a 4.5-megabases region on the mouse X-chromosome, Nora et al. [20] found that the expression profiles of genes with promoters located within the same TAD were correlated. And their more detailed analysis of each domain suggested

* Corresponding authors.
E-mail addresses: liliu2010imu@163.com (L. Liu), qzli@imu.edu.cn (Q.-Z. Li), hlin@uestc.edu.cn (H. Lin).

https://doi.org/10.1016/j.csbj.2019.01.011

these genes are integrated into a similar cis-regulatory network, potentially sharing common cis-regulatory elements. Active transcription units can be clustered in the nucleus, in discrete sites called “transcription factories”. More studies have shown that transcription factories could help organize chromatin and nuclear structure, which contributes to both the formation of chromatin loops and the cluster of active and co-regulated genes [21,22]. For example, the RNA polymerase II (Pol II) concentrates in discrete sites to coordinate regulating gene transcription [23,24].

In fact, many works have tried to reveal the relationship between chromatin interactions and gene co-transcription in the genome of human and mouse [25–29]. Most of them statistically analyzed the correlation of gene expression level across a large collection of expression datasets in a broad range of conditions. However, the 4C or Hi-C datasets they used were generated only in one or two cell lines. For example, Dong et al. [25] tested the correlation between Hi-C interaction (observed from human GM06990 and K562 cells) and the mutual ranks of gene co-expression rates. Their results illustrated that co-expression is strongly associated with chromatin interaction. The inconsistency of the Hi-C data and RNA-seq data would prevent from achieving convincing conclusion because of the dynamics of chromatin organization. Indeed, relatively few Hi-C data in a small collection of primary cell lines restricts our knowledge of chromosome architecture.

Recently, Ren's lab [30] provided a rich resource of chromatin contact maps and gene transcription profiles across 21 well-annotated human tissues and cell types. Based on this dataset, we reconstructed gene-level chromatin interaction and co-transcription in twenty tissues and cell lines. We applied matrix visualization method designed to reveal the relationship between chromatin spatial organization, genes transcription and function cluster on genomics level. Total of 11,588 genes in 23 chromosomes were included and 23 function-similarity matrices, 23 co-transcription matrices and 460 contact matrices were generated for comparative analysis in this study. Finally, we focused on the local domains to study how the spatial reorganization of chromatin regulated gene transcription.

2. Materials and Methods

2.1. Gene annotations

The GENCODE release 19 for human genome was used as the initial gene annotations which is provided by the public research consortium named ENCODE. We collected 20,344 “protein coding” genes from the database and sorted them by chromosomal coordinates.

2.2. Gene function-similarity quantifying

The function “mgeneSim” of GOSemSim R package was used to calculate the pairwise GO semantic similarities between any two genes in each chromosome. This process provided a large-scale quantitative way to investigate functional similarities between genes. At first, those genes without GO annotations were filtered out automatically by the program. As a result, total of 11,588 genes were remained for 23 chromosomes (include chromosome 1–22 and X; Table S1). Subsequently, the graph-based [31] measure algorithm was utilized to measure the semantic similarity of two GO terms defined as follows.

Given two genes X and Y annotated with GO terms as $GO_X = \{go_{X1}, go_{X2}, \cdots, go_{Xm}\}$ and $GO_Y = \{go_{Y1}, go_{Y2}, \cdots, go_{Yn}\}$, respectively, the functional similarity between them is defined as,

$$S(X,Y) = \frac{\sum\limits_{1 \le i \le m} S(go_{Xi}, GO_Y) + \sum\limits_{1 \le j \le n} S\left(go_{Yj}, GO_X\right)}{m+n} \tag{1}$$

where

$$S(go, GO) = \max_{1 \le i \le k}(S_{GO}(go, go_i)) \tag{2}$$

The semantic similarity $S(go, GO)$ between one term go and a GO term set $GO = \{go_1, go_2, \cdots, go_k\}$ is defined as the maximum semantic similarity between term go and any of the terms in set GO. Accordingly, we can achieve a matrix with the element of functional similarity $S(X, Y)$ between two genes shown as,

$$\begin{bmatrix} S(gene_1, gene_1) & S(gene_1, gene_2) & \cdots & S(gene_1, gene_r) \\ S(gene_2, gene_1) & S(gene_2, gene_2) & \cdots & S(gene_2, gene_r) \\ \vdots & \vdots & \vdots & \vdots \\ S(gene_r, gene_1) & S(gene_r, gene_2) & \cdots & S(gene_r, gene_r) \end{bmatrix} \tag{3}$$

In fact, the GO term used in measurement can be restricted by assigning the corresponding parameter to “BP” (biological process), “MF” (molecular function) and “CC” (cellular component). Thus, we can obtain three matrices (Supplementary Fig. S4–S6). Because the three measurements can produce similar patterns, we only employed “MF” measurement for further study in this work.

2.3. Assessing gene co-transcription

The FPKM values of all genes in 20 tissues and cell types were obtained from Ren's re-analyzing RNA-seq dataset [30]. The accession number for the processed sequencing data is GEO:GSE87112. In order to facilitate comparative analysis, only 11,588 corresponding genes were retained. As Eisen has done [32], the Pearson correlation coefficient (PCC) analysis was performed to characterize the co-transcription.

Let G_t equal the FPKM value (log-transformed) for gene G in tissue t. For any two genes X and Y observed over a series of N tissues, a similarity score can be calculated by:

$$P(X,Y) = \frac{1}{N}\sum\nolimits_{t=1}^{N}\left(\frac{X_t - X_{\text{offset}}}{P_X}\right)\left(\frac{Y_t - Y_{\text{offset}}}{P_Y}\right) \tag{4}$$

where

$$P_G = \sqrt{\sum\nolimits_{t=1}^{N}\frac{\left(G_t - G_{\text{offset}}\right)^2}{N}} \tag{5}$$

When G_{offset} is set to the mean of observations on G, then P_G becomes the standard deviation of G, and $P(X, Y)$ is exactly equal to the PCC of the observations of X and Y. When the observations unchange across all tissues, P_G equals to 0, then $P(X, Y)$ is set to “NA” with the white region in heatmap. The PCC matrix is constructed using “cor” function in R.

2.4. Generating gene-level contact maps from Hi-C library

Previous Hi-C maps were created by dividing a genome into fixed resolution loci (e.g., 40 kb, 1 Mb) and counting the number of cross-linked DNA fragments between two loci. Thus, one locus can include many genes and one gene can reside in several loci. To get the contacts count between a pair of genes, we converted the loci-based Hi-C matrix to a gene-based Hi-C matrix by using Babaei's method [26].

In detail, we represented the interactions of gene pairs by the contacts between two loci where the transcription start sites (TSS) of genes locate on. Let h_{XY} represent the interactions between gene pairs (X, Y) and H_{ab} represent the interactions between two loci (a, b). Then, we can get the following equation:

$$h_{XY} = H_{ab}, \text{when } \; TSS(X) \in a \text{ and } \; TSS(Y) \in b \tag{6}$$

It is reasonable due to the transcriptional regulating is always taken place at the promoter region. Then, the 40 kb resolution Hi-C maps were

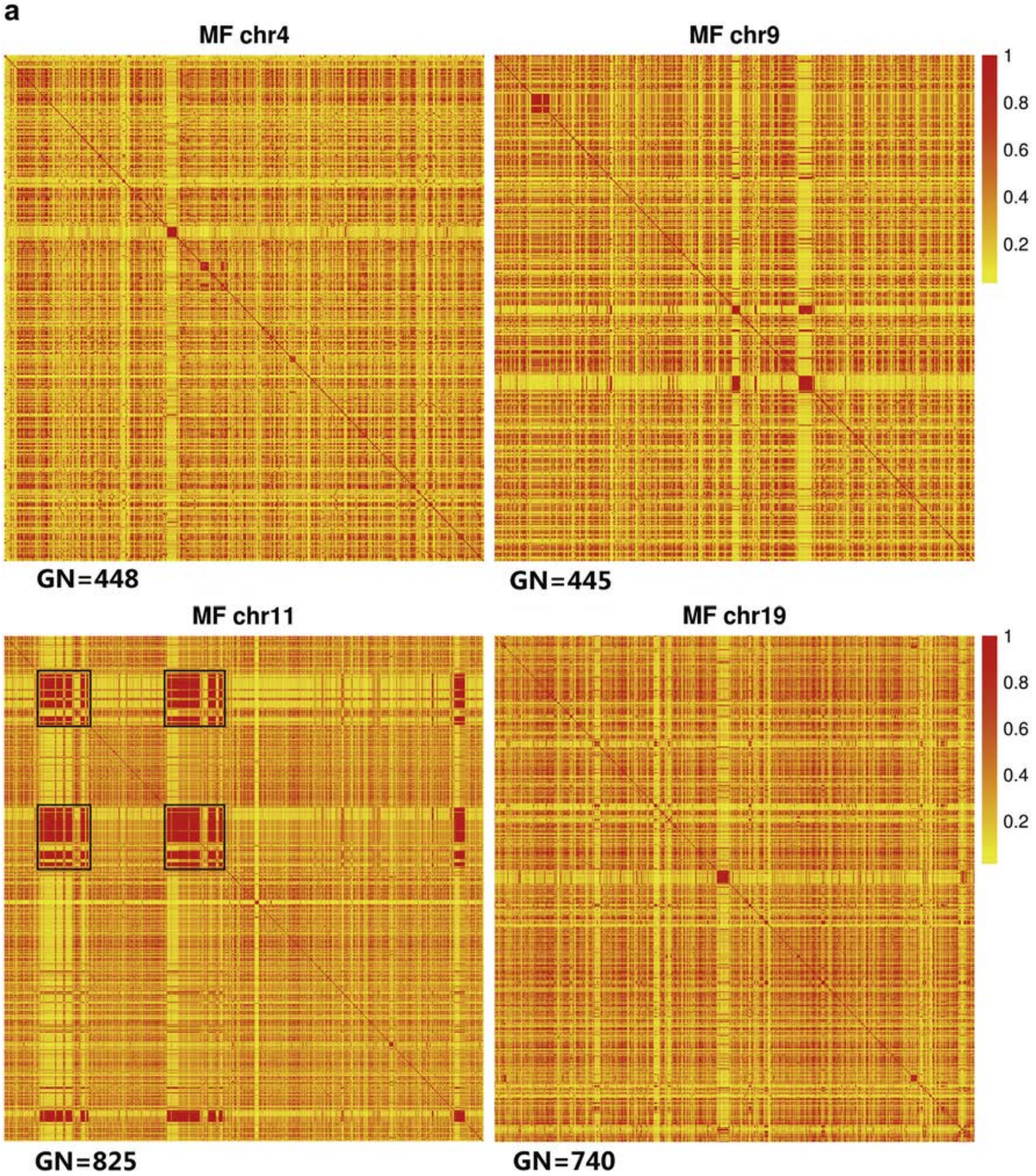

Fig. 1. Intra-chromatin gene function similarity. **a** Heat maps of molecular function similarity for chromosome 4, 9, 11 and 19. The similarity is quantified by "GOSemSim" in each individual chromosome with the parameter "MF". The normalized output values are between 0 (yellow) and 1 (red). "GN" in the lower left of each panel indicate the number of genes in that chromosome. **b** Heat maps for chromosome 6 with the parameter "BP", "CC" and "MF", respectively. The heat map at the lower right indicate the focus region of histone gene cluster in chromosome 6.

converted to gene-level interaction maps. We only focused on intra-chromatin contacts in this study. Therefore, an interaction matrix was obtained for each chromosome in each tissue or cell line. Like Hi-C maps, they can also be visually represented by a heatmap, with intensity indicating contact frequency. The contact frequency was performed log transformation to make patterns more visible here.

2.5. Histone modification

The histone modifications signal used in this study were derived from the NIH Roadmap Epigenomics Project [33]. We selected histone H3 lysine 4 trimethylation (H3K4me3) and histone H3 lysine 27 acetylation (H3K27ac) which have been well known as the markers of active promoters. The average signal intensity of the [−500 bp, +500 bp] region flanking the TSS was calculated for each gene in each cell line.

3. Results

3.1. Gene function-similar analysis

We downloaded the human gene annotations from ENCODE/GENCODE [34] and obtained 20,344 protein coding genes. After filtering out the genes which had no gene ontology (GO) annotations (see Methods for details), total of 11,588 genes were retained. The "GOSemSim" [35] R package was used to compute semantic similarities between genes in the same chromosome. For each chromosome, a matrix with linear arrangement of protein coding genes was constructed

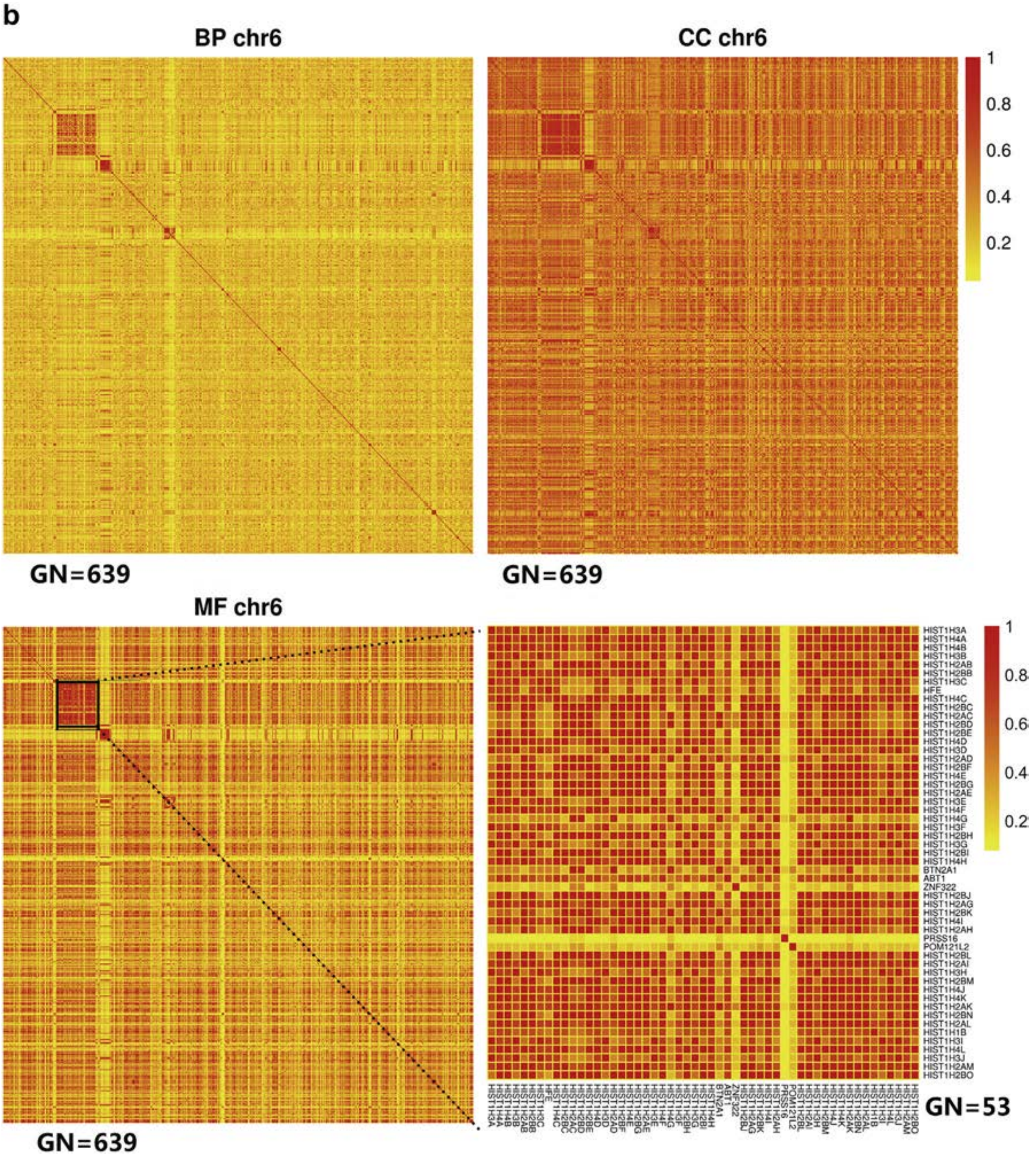

Fig. 1 (*continued*).

and shown in Fig. 1. Each entry in the matrix represents the semantic similarity value of a gene pair. From this figure, we found that function-similar genes are frequently separated by unrelated one. And many gene family clusters with special function have very low correlation with other genes. Here, we showed an example of the olfactory receptors (OR) which are members of the class A rhodopsin-like family of G protein-coupled receptors (GPCRs) [36]. The OR gene family is the largest one consisting of around 800 genes in human genome. In the function-similar map, it is easy to find two large OR gene clusters occur in chromosome 11(Fig. 1a).

3.2. Gene co-transcription analysis

The aim of assessing gene co-transcription is to find out the genes that have similar transcription profiles. One accepted viewpoint is that co-transcription genes are controlled by a same transcriptional regulatory program [37]. These genes usually display correlation on function and are members of the same pathway or protein complex. Here, the transcription correlations between gene pairs across all 20 tissues and cell types were measured by the PCC. Transcription levels of the 11,588 genes were obtained from seven cultured cell types and thirteen human adult tissues. Supplementary Fig. S1 shows the transcription correlation of these genes among 23 human chromosomes (except for chromosome Y and M). Similarly, co-transcription maps were used to display how similarly the changes of genes' expression on the same chromosome. These maps could reveal chromosomal domains of gene expression like Cohen's discovery in the *Saccharomyces cerevisiae* genome [38].

Three representative maps for chromosomes 4, 6, 11 have been shown in Fig. 2. Adjacent groups of correlated genes were depicted as

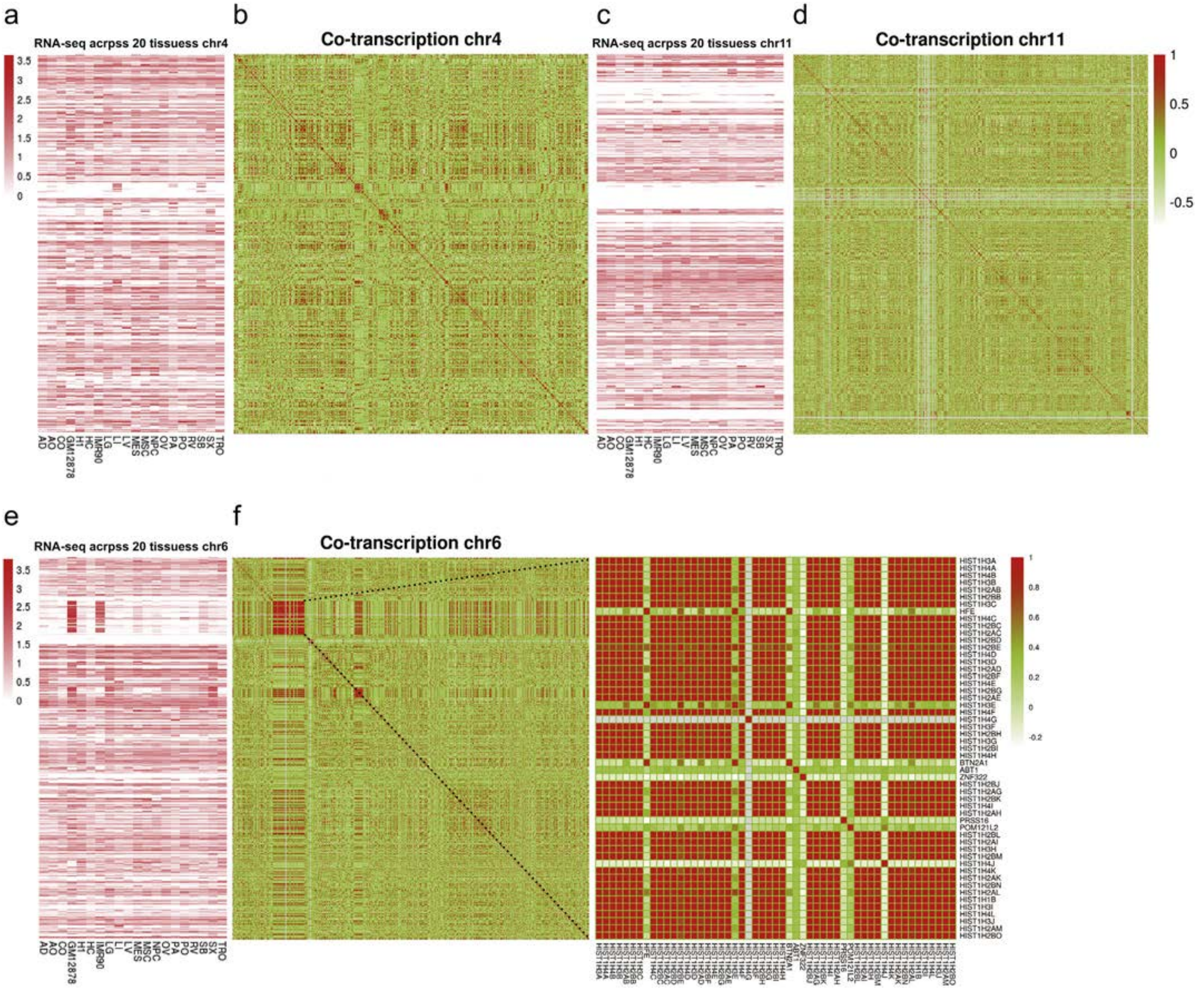

Fig. 2. Gene co-transcription. **a**, **c** and **e** are the asymptotic red heat maps for the FPKM values across 20 tissues for genes of chromosome 4, 6 and 11 respectively. **b**, **d** and **f** are the green-red heat maps for pairwise co-transcription quantified by Pearson correlation coefficient (white line: standard deviations equal to zero).

blocks of red squares centered on the diagonal. Co-transcription maps can only determine whether the transcript level of gene pairs rise and fall together across samples, but cannot tell us when they rise or fall. Therefore, we also compared gene transcription profiles across different tissues and cell types and found 47 clustered histone genes in chromosome 6 do not transcribe or transcribe a little across most tissues and cell types, but upregulate in GM12878 and IMR90 cell types (Fig. 2).

3.3. Gene spatial interaction analysis

Ren's data provided systematic characterization of chromosome architecture across 14 primary human tissues and 7 cell types [30]. The statistical significances of contacts in Hi-C data were identified at 40-kb resolution in their research. Based on Ren's data, we assessed the relationships between genome organization and gene co-transcription. The gene-resolution contact matrices were constructed to investigate the interactions between genes by the contacts number between the chromatin bins (here, 40 kb) where genes locate on. We focused only on intra-chromatin gene pairs. Then, 460 contact matrices and corresponding maps were generated (the contact maps for each chromosome in GM12878 and IMR90 cell lines were listed in Additional file 2: Supplementary Fig. S2 and Additional file 3 Supplementary Fig. S3). From the contact maps, we observed the plaid patterns of spatial compartment and the TADs along the diagonal. TADs are highly conserved across different tissues and cell types, which are agree with previous results [15,19,30]. In contrast, spatial compartmentalization is not always the same in different tissues and cell types. For example, in IMR90 cell type, the genes spatial organization compartmentalized highly. However, it was not obviously found in others (Fig. 3a). The gene expressions in IMR90 are usually upregulated, suggesting that spatial compartmentalization is conducive to gene transcription.

Earlier research using fluorescence in situ hybridization (FISH) in several tissues of mouse has revealed that spatial genome organization is tissue-specific [39]. Recently, results from Hi-C experiments in multiple cell lines during stem cell differentiation [15,30] and primary fibroblasts over 56-h time course [40] have exhibited genome dynamic and tissue-specific organization in human. Here, we counted the numbers of gene interaction in different tissues or cell lines and used the diagram (Fig. 3b) to display the dynamic of gene interplay among different tissues. We found that total of 274,755 interactions only occurred in one tissue or cell type. The retained 10,912 interactions appeared in all 20 tissues and cell types in chromosome 1 (Fig. 3c). These

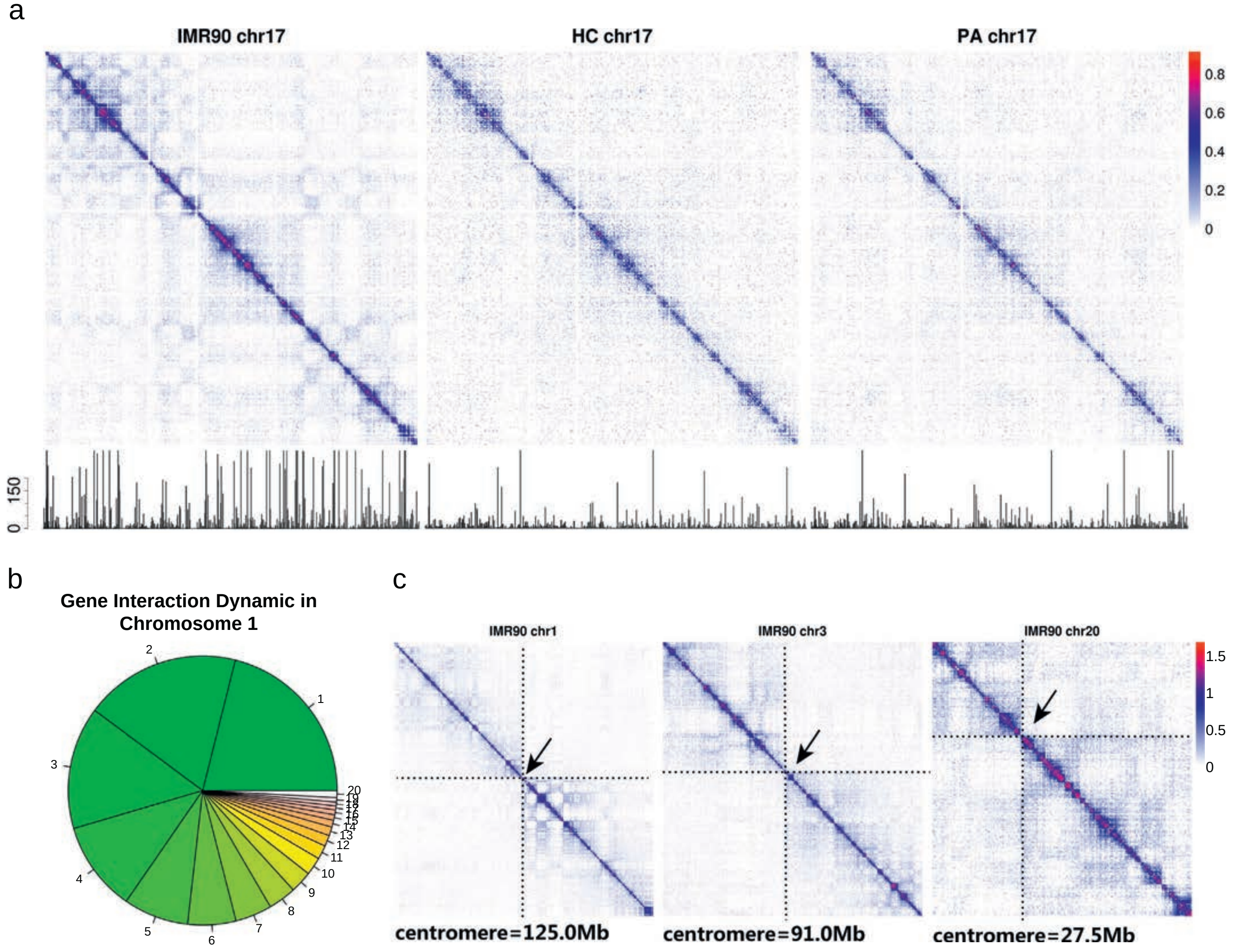

Fig. 3. Chromatin spatial organization. **a** The degree of chromatin spatial compartmentalization influences gene expression. The heat maps are gene-level contact maps of chromosome 17 in IMR90, Hippopotamus (HC) and Pancreas (PA) respectively. The bar plots directly below show the FPKM values of each gene in corresponding cell line or tissue. **b** Gene interaction dynamic in chromosome 1. Pie plot shows the proportion of gene interaction which occur in N tissues or cell lines. **c** Human genome displays polarized organization. The polarized organization is apparent on gene-level contact maps of chromosome 1, 3 and 20 in IMR90 cell line. The arrows above maps are pointing to the position of centromere.

observations show that the interactions between genes are dynamic and tissue-specific, but the basic structure units, TADs, are conserved.

Previous studies in Drosophila have shown the limited nature of interactions between genes or chromatin regions on different chromosome arms [41–44]. In plants, after mitosis, a polarized genome organization called "Rabl" has been observed [45]. In the polarized genome, each chromosome occupies a territory with centromere at one nuclear pole and telomeres on the opposite side of the nucleus. Such organization is also apparent in our study, behaving as many maps display two big blocks breaking at the centromere (Fig. 3c). It suggests that the contact frequencies between two arms of chromosome are drastically reduced.

3.4. The relationship between co-transcription, interaction and function

We further comparatively analyzed the relationship between co-transcription, interaction and function based on the three matrices. As there are a large number of tissue-specific interactions, we firstly calculated the average Hi-C scores of chromosome 6 across 20 tissues and cell lines. The average Hi-C maps show the conserved TADs along the diagonal more clearly (Fig. 4a). Then, the transcription correlation distribution of the gene-pairs was calculated and shown in Fig. 4b according to different average Hi-C scores. This treatment differs from previous study which calculated correlation between Hi-C interaction and co-transcription in only one cell line [25]. Next, we calculated the functional similarity distribution of the gene-pairs in various thresholds of the average Hi-C scores (Fig. 4c). As the interactions are tissue-specific, the Hi-C scores >0.4 were considered as the stable interactions by filtering out a majority of tissue-specific interactions. The results indicate that stably interacting gene pairs are more likely function-related. The intersection of the gene co-transcription, functional similarity and interaction was shown in Fig. 4d. In chromosome 6, there are 41,977 gene pairs that are co-transcription (PCC ≥ 0.5), and 109,973 gene pairs are molecular function similar (GOSemSim score ≥ 0.5) and 7265 gene pairs have stably interaction. About 41.0% (2980/7265) of interacting gene pairs are co-transcription and 7.1% (29,80/41,977) of co-transcription gene pairs are also interacting gene pairs. In addition, 36.3% (15,252/41,977) of co-transcription gene pairs and 53.9% (3919/7265) of interaction gene pairs display functional similarity. These results indicate that the interaction and function similarity can contribute to but not determine gene co-transcription, which is in accordance with previous conclusion [4,6,46].

Although the gene correlation is weak at whole genome wide, we still observed that some function-related gene clusters were physical

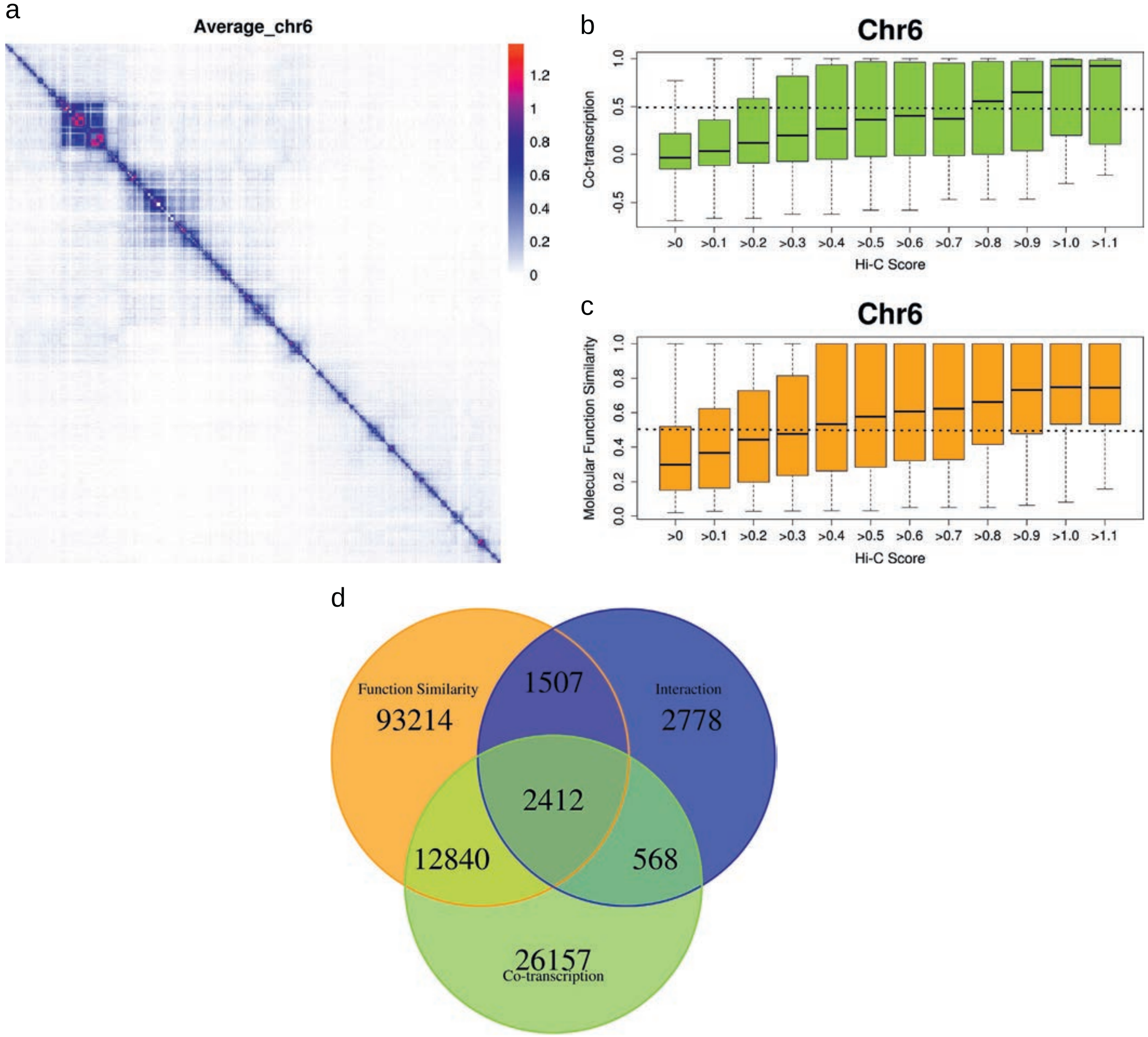

Fig. 4. The global relevance between co-transcription, interaction and function similarity. **a** Average contact map of chromosome 6. **b** The green boxes plot shows the distributions for co-transcription gene pairs which interact at least in N tissues or cell lines, whereas the yellow boxes are distributions for function-similar gene pairs which interact at least in N tissues or cell lines. **c** The Venn plot showing the intersection of the three. The interaction is included when the entry is greater than zero at least in five Hi-C contact matrices. The number of gene pairs is calculated when Pearson correlation coefficient of gene expression is equal to or >0.5. Similarly, two genes are regarded as function-similar when GOSemSim score is equal to or >0.5.

proximity and displayed similar transcription patterns. The typical representative is histone gene family in chromosome 6 which include 47 histone genes (from HIST1H3A Chr6:26,020,717–26,021,186 to HIST1H2BO chr6:27,861,202–27,861,669 along the line genome). These histone genes interact with each other to form a local structure domain in a chromatin region spanning 1.8 Mb. In this region, the correlation of the three (gene co-transcription, functional similarity and interaction) is strong. For example, there are 73.7% (1693/2298) of interaction gene pairs that are co-transcription and 87.5% (1693/1935) of co-transcription gene pairs interact with each other. 93.2% (1803/1935) of co-transcription gene pairs and 81.9% (1881/2298) of interacting gene pairs display functional similarity (Fig. 5a). The interaction frequency in the adjacent region described above displays a hot spot on the heatmap (Fig. 5b). The frequent contacts imply that megabase chromatin compress tightly, resulting in a substructure in the three-dimensional space of the nucleus. The transcription correlation patterns are well characterized through the PCC heatmap (Fig. 5b). As this histone gene cluster is repressed across 18 tissues and activated in GM12878 and IMR90 cell lines (particularly, the FPKM (fragments per kilobase of exon model per million fragments mapped) value run up to 1000). Histone is the key parts of nucleosome which was a basic unit of DNA packaging in eukaryote, consisting of a ~147 bp of DNA wound in sequence around eight histone protein cores and widely distributes along genomics. Thus, these histone genes must express together to satisfy the requirement of nucleosome. However, the biology importance of extreme tissue-specific expression for these histone genes still need to be revealed through dynamic chromatin organization.

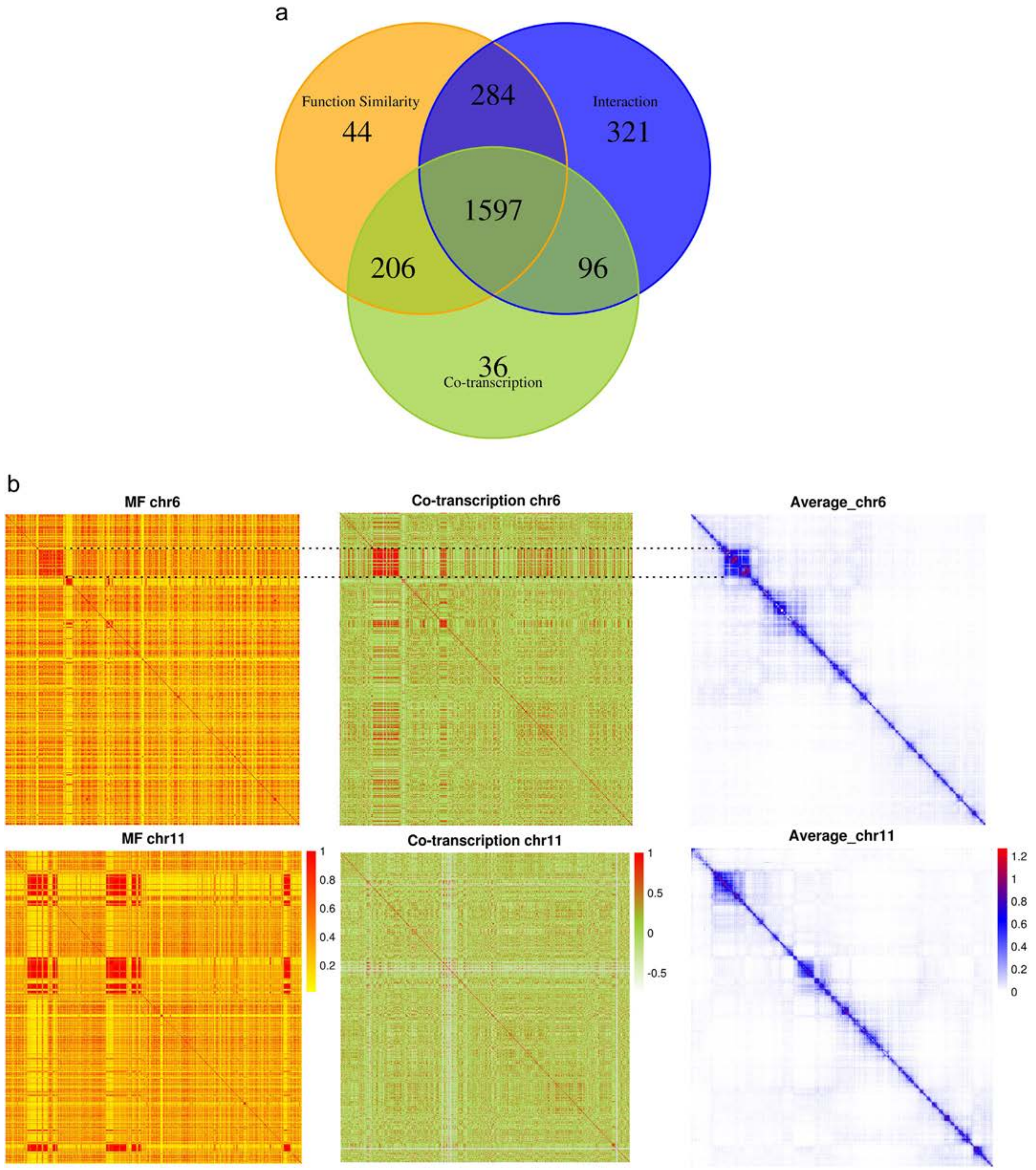

Fig. 5. The local relevance between co-transcription, interaction and function similarity. **a** The Venn plot showing the intersection of the three when focus on the local region of chromosome 6. **b** Three heat maps (yellow for function similarity, green for co-transcription, blue for interaction) comparison illustrating the strong correlation in local region of chromosome 6 and 11. Dotted line marked the strong correlation region on the maps.

We next analyzed chromatin dynamic in the hot spot and adjacent region (from HIST1H3A Chr6:26,020,717–26,021,186 to chr6:32,780,539–32,784,825). It is divided into four structure domains based on the contact map (Fig. 6a). Domain A (from HIST1H3A Chr6:26,020,717–26,021,186 to HIST1H2BO chr6:27,861,202–27,861,669) is composed of 48 histone genes and 5 other genes. Domain B (from OR2B2 chr6:27,878,962–27,880,174 to OR2H1 chr6:29,424,957–29,432,105) includes 14 olfactory receptor genes and 4 other genes. Domain C (from MAS1L chr6:29,454,473–29,455,738 to NOTCH4 chr6:32,162,619–32,191,844) contains various genes such as

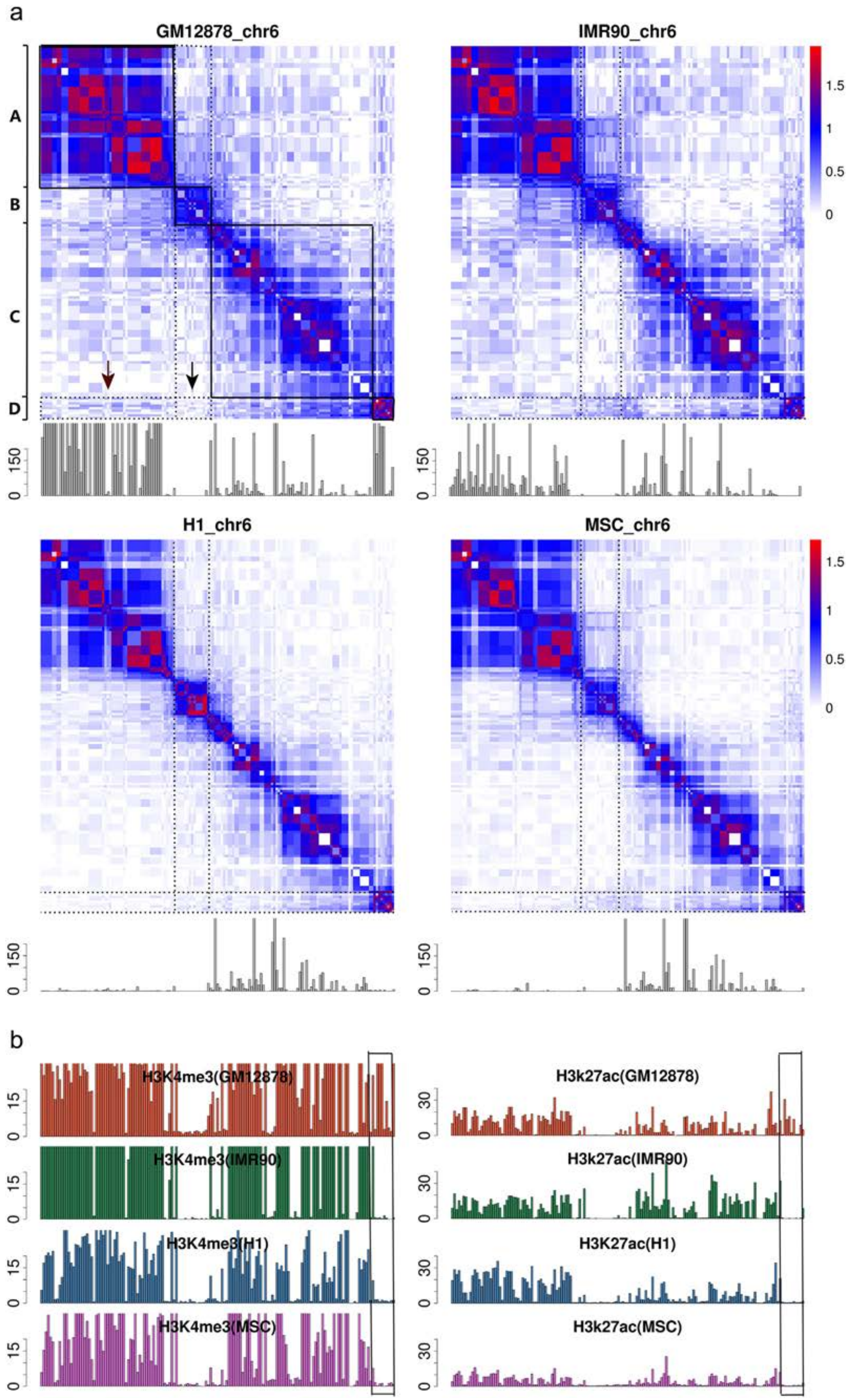

Fig. 6. Chromatin dynamical organization. **a** Tissue-specific interaction regulates gene expression across GM12878, IMR90, Embryonic Stem Cell (H1) and Mesenchymal Stem Cell (MSC). Boxes indicate the boundaries of Domain A, B, C and D. The arrows are pointing to the tissue-special interaction regions. The bar plots directly below show the FPKM values of each gene in corresponding cell lines.**b** The changes of spatial position are accompanied by the changes of histone modification. The left bar plots show the distributions for H3K4me3 in GM12878, IMR90, H1 and MSC cell lines, and the right bar plots for H3K27ac. The boxes highlight the significantly different region between four cell lines.

HLA gene family and TRIM gene family and so on. Domain D (from HLA-DRA chr6:32,407,618–32,412,823 to HLA-DOB chr6:32,780,539–32,784,825) is composed of 8 HLA genes. The genes in Domain C express stably across all the 20 tissues and cell lines. Most of the olfactory receptor genes in Domain B are silent in the 20 tissues and cell lines. The genes in Domain A and Domain D express only in few cell lines but are co-expression. In GM12878 and IMR90 cell lines, the contacts between Domain A and Domain C increase, resulting in histone genes being activated sharply. The FPKM value increases hundreds of times even ten thoughts of times. In IMR90, H1 and MSC cell lines, the Domain D interacts frequently with silent Domain B, resulting in the HLA genes being repressed. In GM12878 cell line, the Domain D is away from Domain B and closing to Domain A and Domain C which have been activated. Furthermore, we also compared the distributions of two histone modifications (H3K4me3 and H3K27ac) at the promoter regions (Fig. 6b). One may notice that the signal intensities of the two modifications in Domain D correspond to the domains it connected with. This observe extends Rao's observation in GM12878 and IMR90 cells for H3K36me3 and H3K27me3 [47]. We demonstrate that interaction dynamic is accompanied by changes of histone marks. This conclusion is consistent with the others' observation that 3D domains correlate strongly with the 1D epigenomic information along the genome [48]. In summary, these results indicate chromatin spatial organization indeed affects gene expression. The genes within stable domains are co-regulated by the spatialization of the domains.

Another example is the biggest olfactory receptor gene cluster which locate on chromosome 11. Approximately 200 OR genes are divided into three groups and scatter across chromosome 11 (Fig. 5b). These genes are always silent and form independent structure domains with self-organizing. They contact each other frequently but seldom interact with other genes. It is consistent with previous study that chromatins are partitioned into active and inactive compartments.

4. Discussion

In prokaryotes, a cluster of function-related structural genes are organized into an operon which contains only a single promoter, which make their expression easily being coregulated [49]. The genes organization in eukaryotic genomes is more complicated than prokaryotes. Besides histone modification, transcription factor and super enhancer, the three-dimensional genome organization has been increasingly considered as an important regulator of gene transcription.

Do the function-related genes close physically proximity to each other for being coregulated? To address this question, we characterized three types of relationships (co-transcription, spatial interaction and function similarity) of pairwise genes intra-chromatin wide and made a comparative analysis. Different from previous observations [25] in only two cell lines (human GM06990 and K562), we failed to find significant correlation on global level. Only 10.5% of interaction gene pairs (Hi-C score > 0) are co-transcription and 27.2% are function-related. A large number of tissue-specific gene interactions result in the phenomena. However, we could still observe the strong correlations in the conserved and stable interactions (within substructures). In particular, some hot spot regions where gene family clustered in are highly coincided on three relationships, implying the contribution of substructure for gene co-transcription. The most representative instance is histone gene family in chromosome 6 which displays the coregulation of frequently interacting genes. The expressions of these genes are highly tissue-specific. Based on the visualized maps, we observed the influence of genome dynamic organization on gene co-transcription. Further analysis on the hot spot region reveals that the chromatin could arrangement across various tissues and some special structure domains may be recruited to a transcription factory where the appropriate transcription and processing factors are highly concentrated, thereby facilitating the expression of those genes.

Our discovery is similar with previous studies on the mouse globin genes in erythroid tissues [50]. Using 4C and DNA FISH techniques, they showed that specialized transcription factories boost the expression of clustered and co-regulated genes. These researches [39,50] thought that preferential associations in transcription factories substantially affected higher order chromosomal conformations and were a major driving force in tissue-specific chromosome positioning. Whether transcription is causative or a consequence of higher order chromatin organization is still a matter of debate. But our discovery and previous examples in various model organisms [51–54] demonstrated that spatial association between co-regulated genes is a widespread principle of nuclear organization.

In fact, our discovery is taking advantage of the most comprehensive Hi-C maps in human tissues. Although Hi-C method is a powerful tool to offer a global view of chromatin interactions within a single experiment, it is costly to sequence to sufficient depth to provide enough resolution to capture gene-gene loops. Recently, a high-resolution capture Hi-C method that map long-range promoter contacts has been developed [55]. It can achieve fragment enrichment up to hundreds of fold, greatly improving the detection of local chromatin interaction of the genome regions of interest. In addition, chromatin interaction analysis by paired-end tag (ChIA-PET) [56] and protein-centric chromatin conformation assay (HiChIP) [57] are developed for capturing chromatin interactions mediated by specific proteins such as Pol Ⅱ. In the future, along with these interatomic data increasing in diverse tissues and cell types even in single cell [58], we can learn more knowledge of transcriptional co-regulation using this framework.

5. Conclusions

In this study, we provided a framework for calculating and analyzing the functional similarities, transcription correlation and intra-chromatin interaction between gene pairs in each chromosome and their relationship. We found that the correlation between chromatin spatial structure, gene transcription and function cluster is weak at global scale, but strong at local domain scale. Some super gene clusters, such as histone gene family in chromosome 6 and olfactory receptor gene family in chromosome 1, 6, 7, 9, 11 and 14, are always close to each other in space and more likely to co-transcription across these tissues and cell types. These observations coincide with transcription factories theory. It suggests that function-similar genes are close in space and have similar transcription mechanism. In addition, our framework allows the integration of various genome-wide datasets for transcriptional regulation analysis in gene-resolution and is easy to apply to other species. In the future, we hope the data mining techniques [59-63] could be applied in this fields.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2019.01.011.

Funding

This work was supported by a grant from the National Natural Science Foundation of China (Nos.31460234, 11747315, 61772119).

Conflict of interesting

None declared.

References

[1] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326:289–93.
[2] Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell 2009;137:1194–211.
[3] Rajapakse I, Groudine M. On emerging nuclear order. J Cell Biol 2011;192:711–21.
[4] Cavalli G, Misteli T. Functional implications of genome topology. Nat Struct Mol Biol 2013;20:290–9.

[5] Smallwood A, Ren B. Genome organization and long-range regulation of gene expression by enhancers. Curr Opin Cell Biol 2013;25:387–94.
[6] Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell 2014;14:762–75.
[7] Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet 2016;17:772.
[8] Sexton T, Cavalli G. The role of chromosome domains in shaping the functional genome. Cell 2015;160:1049–59.
[9] Mercer TR, Mattick JS. Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Res 2013;23:1081–8.
[10] Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell 2013;152:1237–51.
[11] Lelli KM, Slattery M, Mann RS. Disentangling the many layers of eukaryotic transcriptional regulation. Annu Rev Genet 2012;46:43–68.
[12] Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 2011;12:283–93.
[13] Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 2012;13:613–26.
[14] Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009;459:108–12.
[15] Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature 2015;518:331–6.
[16] Cope NF, Fraser P, Eskiw CH. The yin and yang of chromatin spatial organization. Genome Biol 2010;11:204.
[17] Therizols P, Illingworth RS, Courilleau C, Boyle S, Wood AJ, Bickmore WA. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science 2014;346:1238–42.
[18] Wang S, Su JH, Beliveau BJ, Bintu B, Moffitt JR, Wu CT, et al. Spatial organization of chromatin domains and compartments in single chromosomes. Science 2016;353:598–602.
[19] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485:376–80.
[20] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation Centre. Nature 2012;485:381–5.
[21] Papantonis A, Cook PR. Transcription factories: genome organization and gene regulation. Chem Rev 2013;113:8683–705.
[22] Rieder D, Trajanoski Z, McNally JG. Transcription factories. Front Genet 2012;3:221.
[23] Chen X, Wei M, Zheng MM, Zhao J, Hao H, Chang L, et al. Correction to study of RNA polymerase II clustering inside live-cell nuclei using Bayesian nanoscopy. ACS Nano 2016;10:4882.
[24] Razin SV, Gavrilov AA, Pichugin A, Lipinski M, Iarovaia OV, Vassetzky YS. Transcription factories in the context of the nuclear and genome organization. Nucleic Acids Res 2011;39:9085–92.
[25] Dong X, Li C, Chen Y, Ding G, Li Y. Human transcriptional interactome of chromatin contribute to gene co-expression. BMC Genomics 2010;11:704.
[26] Babaei S, Mahfouz A, Hulsman M, Lelieveldt BP, de Ridder J, Reinders M. Hi-C chromatin interaction networks predict co-expression in the mouse cortex. PLoS Comput Biol 2015;11:e1004221.
[27] Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res 2012;22:490–503.
[28] Kaufmann S, Fuchs C, Gonik M, Khrameeva EE, Mironov AA, Frishman D. Inter-chromosomal contact networks provide insights into mammalian chromatin organization. PLoS One 2015;10:e0126125.
[29] Homouz D, Kudlicki AS. The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes. PLoS One 2013;8:e54699.
[30] Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep 2016;17:2042–59.
[31] Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007;23:1274–81.
[32] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.
[33] C. Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M.J. Ziller, V. Amin, J.W. Whitaker, M.D. Schultz, L.D. Ward, A. Sarkar, G. Quon, R.S. Sandstrom, M.L. Eaton, Y.C. Wu, A.R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R.A. Harris, N. Shoresh, C.B. Epstein, E. Gjoneska, D. Leung, W. Xie, R.D. Hawkins, R. Lister, C. Hong, P. Gascard, A.J. Mungall, R. Moore, E. Chuah, A. Tam, T.K. Canfield, R.S. Hansen, R. Kaul, P.J. Sabo, M.S. Bansal, A. Carles, J.R. Dixon, K.H. Farh, S. Feizi, R. Karlic, A.R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T.R. Mercer, S.J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R.C. Sallari, K.T. Siebenthall, N.A. Sinnott-Armstrong, M. Stevens, R.E. Thurman, J. Wu, B. Zhang, X. Zhou, A.E. Beaudet, L.A. Boyer, P.L. De Jager, P.J. Farnham, S.J. Fisher, D. Haussler, S.J. Jones, W. Li, M.A. Marra, M.T. McManus, S. Sunyaev, J.A. Thomson, T.D. Tlsty, L.H. Tsai, W. Wang, R.A. Waterland, M.Q. Zhang, L.H. Chadwick, B.E. Bernstein, J.F. Costello, J.R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J.A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes, Nature, 518 (2015) 317–330.
[34] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res 2012;22:1760–74.
[35] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 2010;26:976–8.
[36] Gaillard I, Rouquier S, Giorgi D. Olfactory receptors. Cell Mol Life Sci 2004;61:456–69.
[37] Weirauch MT. Gene Coexpression Networks for the Analysis of DNA Microarray Data. Wiley-VCH Verlag GmbH & Co. KGaA; 2011.
[38] Cohen BA, Mitra RD, Hughes JD, Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet 2000;26:183–6.
[39] Parada LA, McQueen PG, Misteli T. Tissue-specific spatial organization of genomes. Genome Biol 2004;5:R44.
[40] Chen H, Chen J, Muir LA, Ronquist S, Meixner W, Ljungman M, et al. Functional organization of the human 4D Nucleome. Proc Natl Acad Sci U S A 2015;112:8002–7.
[41] Marshall WF, Dernburg AF, Harmon B, Agard DA, Sedat JW. Specific interactions of chromatin with the nuclear envelope: positional determination within the nucleus in Drosophila melanogaster. Mol Biol Cell 1996;7:825–42.
[42] Lowenstein MG, Goddard TD, Sedat JW. Long-range interphase chromosome organization in Drosophila: a study using color barcoded fluorescence in situ hybridization and structural clustering analysis. Mol Biol Cell 2004;15:5678–92.
[43] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 2012;148:458–72.
[44] Li Q, Tjong H, Li X, Gong K, Zhou XJ, Chiolo I, et al. The three-dimensional genome organization of Drosophila melanogaster through data integration. Genome Biol 2017;18:145.
[45] Cowan CR, Carlton PM, Cande WZ. The polar arrangement of telomeres in interphase and meiosis. Rabl organization and the bouquet. Plant Physiol 2001;125:532–8.
[46] Misteli T. Self-organization in the genome. Proc Natl Acad Sci U S A 2009;106:6885–6.
[47] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159:1665–80.
[48] Chen Y, Wang Y, Xuan Z, Chen M, Zhang MQ. De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. Nucleic Acids Res 2016;44:e106.
[49] Jacob F, Perrin D, Sanchez C, Monod J. The Operon: A Group of Genes Whose Expression is Coordinated by an Operator. Compte Rendu De Lacademie Des Sciences; 1960; 1727–9.
[50] Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet 2010;42:53–61.
[51] Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, et al. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell 2011;144:214–26.
[52] Denholtz M, Bonora G, Chronis C, Splinter E, de Laat W, Ernst J, et al. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. Cell Stem Cell 2013;13:602–16.
[53] Vieux-Rochas M, Fabre PJ, Leleu M, Duboule D, Noordermeer D. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. Proc Natl Acad Sci U S A 2015;112:4672–7.
[54] Schoenfelder S, Sugar R, Dimond A, Javierre BM, Armstrong H, Mifsud B, et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat Genet 2015;47:1179–86.
[55] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 2015;47:598–606.
[56] Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 2009;462:58–64.
[57] Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods 2016;13:919–22.
[58] Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 2017;544:59–64.
[59] Song J, Li F, Leier A, Marquez-Lago TT, Akutsu T, Haffari G, et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics 2018;34:684–7.
[60] Manavalan B, Subramaniyam S, Shin TH, Kim MO, Lee G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. J Proteome Res 2018;17:2715–26.
[61] Feng CQ, Zhang ZY, Zhu XJ, Lin Y, Chen W, Tang H, et al. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics 2018. https://doi.org/10.1093/bioinformatics/bty827.
[62] Basith S, Manavalan B, Shin TH, Lee G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput Struct Biotechnol J 2018;16:412–20.
[63] Zhu XJ, Feng CQ, Lai HY, Chen W, Lin H. Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl Based Syst 2019;163:787–93.