Quantitative Biology 2020, 8(4): 295–311
https://doi.org/10.1007/s40484-020-0221-6

RESEARCH ARTICLE

Toward an understanding of the relation between gene regulation and 3D genome organization

Hao Tian[1], Ying Yang[1], Sirui Liu[1], Hui Quan[1], Yi Qin Gao[1,2,3,*]

[1] Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China
[2] Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing 100871, China
[3] Beijing Advanced Innovation Center for Genomics (ICG), Peking University, Beijing 100871, China
* Correspondence: gaoyq@pku.edu.cn

Received April 24, 2020; Revised June 9, 2020; Accepted July 12, 2020

***Background*: High-order chromatin structure has been shown to play a vital role in gene regulation. Previously we identified two types of sequence domains, CGI (CpG island) forest and CGI prairie, which tend to spatially segregate, but to different extent in different tissues. Here we aim to further quantify the association of domain segregation with gene regulation and therefore differentiation.**

***Methods*: By means of the published RNA-seq and Hi-C data, we identified tissue-specific genes and quantitatively investigated how their regulation is relevant to chromatin structure. Besides, two types of gene networks were constructed and the association between gene pair co-regulation and genome organization is discussed.**

***Results*: We show that compared to forests, tissue-specific genes tend to be enriched in prairies. Highly specific genes also tend to cluster according to their functions in a relatively small number of prairies. Furthermore, tissue-specific forest-prairie contact formation was associated with the regulation of tissue-specific genes, in particular those in the prairie domains, pointing to the important role of gene positioning, in the linear DNA sequence as well as in 3D chromatin structure, in gene regulatory network formation.**

***Conclusion*: We investigated how gene regulation is related to genome organization from the perspective of forest-prairie spatial interactions. Since unlike compartments A and B, forest and prairie are identified solely based on sequence properties. Therefore, the simple and uniform framework (forest-prairie domain segregation) provided here can be utilized to further understand the chromatin structure changes as well as the underlying biological significances in different stages, such as tumorgenesis.**

Keywords: CGI forest; CGI prairie; domain segregation; chromatin structure; gene regulation

Author summary: Growing evidence has revealed the vital role of genome architecture in gene regulation. In this paper, we systematically investigated how genome organization is associated with gene regulation from the perspective of spatial interactions between CGI forest and CGI prairie, two distinct sequence domains. Such structure-function relation appears to be a common phenomenon among all tissues we analyzed, indicating the important roles of forest-prairie domain segregation in gene regulatory network formation and differentiation.

INTRODUCTION

Over the last decade, with the development of chromosome conformation capture technologies, such as Hi-C [1] and ChIA-PET [2], people have gained much insight into the high order chromatin structure and a series of intriguing discoveries (for example, compartment [1], TAD [3] and chromatin loop [4]) have been made. In terms of the relation between structural organization and gene regulation, chromatin loops, frequently anchored by

CCCTC-binding factor (CTCF) and cohesin complex, can link promoter regions and distal regulatory elements such as enhancers to ensure normal transcription [4]. Such spatial interactions are often restricted into individual specific topological domains, named insulated neighbors that can be seen as a mechanistic basis of TADs [5,6]. TAD boundaries are often thought to be robust among different cell types [3] and the relationship between TAD formation and gene regulation has caused much research interest. For example, the disruption of specific TAD boundaries or insulated structures and consequently unexpected enhancer-promoter interactions can induce aberrant gene expression, especially that of disease-related genes [7–9]. Surprisingly, the global loss of TAD structure via depletion of cohesin leads to significant changes of expression level of only about 1000 genes [10], indicating a limited role of TAD structure in regulating gene transcription [11]. Moreover, multiplex promoter-centered chromatin interactions such as promoter-promoter interactions of different genes have been widely investigated using ChIA-PET targeting on RNA polymerase II (RNAPII) and are thought to play a vital role in transcription regulation [12]. A-B compartmentalization has been referred to efficiently participate in gene regulation possibly owing to the local enrichment of transcription-associated factors [13] and compartment shifts have been found to coincide with the activation or repression of specific genes [14,15]. However, a systematic understanding of regulation of tissue-specific genes (TSGs) from the perspective of chromatin organization is still in need.

As introduced above, one of the important factors affecting gene expression is the compartmentalization [13–15], which was found to be strongly affected by the DNA sequence [16]. Based on the uneven distribution of CpG islands (CGIs), the whole genome was divided into two types of domains, named CGI forest (F) and CGI prairie (P), respectively. The former is enriched in CGI and possesses high gene (especially housekeeping gene) density and active histone marks, whereas the latter is characterized by low CGI and gene densities, and strong signals of repressive histone marks. More importantly, different extent of domain segregation between forests and prairies was observed in different cell types and forests and prairies are mainly composed of compartment A and B, respectively. Intriguingly, the functions of prairie genes in compartment A (pA) are found to be cell-type specific and prairie genes residing in compartment B (pB) harbors cell-type specificity in a complementary way to pA, indicating that prairie genes may be more cell-specifically regulated.

Besides regulation of individual genes, the correlation between the expression levels of different genes in connection with their positions in 3D space provides additional information on biological function regulation. The likelihood of transcriptional co-regulation during cell differentiation was found to be maximal at the TAD level [17] and paralog genes tend to be co-regulated and co-localized within TADs [18]. However, it was also reported that although gene pairs within the same TADs possess higher co-expression level than those positioned in different TADs, the co-expression domains have poor correlations with TADs [19]. Besides, highly co-regulated gene pairs are found to share similar contact profiles [19] and tend to physically contact with each other [20]. Accordingly, in recent years, phase separation models, describing a phenomenon in which specific proteins, such as transcription factors (TFs), co-activators and RNAPII, can accumulate and self-organize into condensates, have been proposed to contribute to the formation of 3D chromatin structure as well as gene regulation [21–23]. Such a model is accordant with observations on the co-location of TF-binding sites and corresponding genes [24,25]. Given that gene pair expression coordination is closely associated with 3D chromatin structure, which is strongly affected by the linear DNA sequence, it is thus interesting to further investigate how gene co-regulation is related to forest-prairie domain segregation, especially in tissues showing distinct forest-prairie spatial interaction patterns.

Based on these previous findings, here we further evaluate the difference of tissue-specific gene enrichment between forest and prairie domains and quantitatively investigate gene regulation in terms of the forest-prairie spatial interaction. We first explore the distribution of genes of different tissue specificities in the linear genome, in particular their different distribution patterns in forest and prairie domains. Accordant with our previous study [16], prairies are more likely than forests to contain genes of high tissue specificity. In addition, we found here that a small number of individual prairie domains significantly encompass large numbers of genes that are specifically highly expressed in a tissue-specific way, which are functionally related. Namely, the tissue-specific genes can significantly cluster in the linear genome and, in particular, in prairie domains. Next, we found that genes of higher tissue specificity in prairies but not forests are prone to be characterized by lower gene body CpG density. Based on earlier findings on the negative correlation between GC content (or gene body CpG density) and repressive histone mark H3K9me3 [26,27] and our discovery on the positive association between gene body CpG density and compartment index (a parameter used to evaluate the compartmentalization degree of genes, high (positive) value indicates genes reside in the interior of compartment A), we validated that prairie genes of higher tissue specificities tend to reside in repressed (heterochromatin-like) environment unless they

are highly expressed (activated) in a tissue-specific manner [16]. Furthermore, how TSGs are activated in corresponding tissues from the perspective of spatial interactions between forest and prairie domains is discussed. We examined the idea that the achievement of cell identity is realized under the help of spatial contacts formed between prairie and forest, through the assistant of (tissue-specific) TFs. To further explore the association between gene pair expression coordination (co-regulation) and genome organization, we constructed gene co-regulation networks based on expression correlation calculation, and found that in these networks genes tend to correlate with other genes in the same compartment, consistent with earlier studies [18,28]. We then examined the spatial interaction patterns of co-regulated genes through constructing gene networks based on a combined usage of Hi-C [29] and RNA-seq [30] data. From these two kinds of networks, we observed that the F-P gene expression correlation patterns are indeed consistent with the overall chromatin structure features. In particular, the connections between genes in these networks do reflect the varied overall domain segregation among different tissues.

RESULTS

Forest and prairie gene features

As mentioned earlier, we divided the whole genome into two types of domains, forests and prairies, solely based on DNA sequence properties. Forests and prairies tend to segregate in space, leading to a strong sequence dependence in chromatin compartmentalization. In order to further clarify the underlying biological significances of domain segregation (*e.g.*, the activation of specific genes), we first examined the distribution of genes of different tissue specificities (see "Materials and methods") as well as the enrichment of TSGs between the two types of linear domains. High tissue specificity for gene *i* in tissue *t* indicates high expression level of gene *i* in tissue *t* relative to the baseline and we consider gene *i* to be a TSG if its tissue specificity is bigger than 2 in a given tissue. Our previous results [16] revealed that the variances of prairie gene expression are higher than those in forest, similar to this, it can be seen from Table 1 that TSGs belonging to one tissue tend to be enriched in prairies rather than forests in most tissues we considered (notably, ten tissues listed in Table 1 were mainly used in this work because of the availability of both Hi-C (GSE87112) and RNA-seq experimental data (https://zenodo.org/record/838734)), except for those possessing very small numbers of TSGs, such as aorta, left ventricle and ovary. The heterogeneity of the number of TSGs belonging to different tissues results in the big statistical difference (for instance, the number of liver and aorta TSGs is 312 and 14, respectively, a robustness test can be found in Supplementary Table S1.) The distribution of gene tissue specificity in these two domain types is also distinctly different (Fig. 1A and Supplementary Fig. S1). The proportion of genes of high (positive) and extremely negative tissue specificity (a gene possessing a negative tissue specificity is highly expressed in some other tissue (s) rather than the reference one) is larger in prairies than in forests, corresponding to the enrichment of TSGs in the relevant and other tissues, respectively. In contrast, forests are abundant in genes, the tissue specificity distribution of which is centered at zero (not tissue-specific).

It is known that genes of related functions can cluster along the linear genome [31,32]. To quantify the clustering of function-related genes within individual forest or prairie domains, we calculated the Pearson correlation coefficient (PCC) of the tissue specificity files of two forest/prairie genes within the same domain (each gene possesses 38 tissue specificities corresponding to 38 tissues, see Supplementary File 1). As a control, we also calculated the PCC between forest and prairie genes as well as all gene pairs regardless of the distinction of forest and prairie. It was found that forest genes within the same domain, forest-prairie genes and all gene pairs regardless of their attribution are all significantly less correlated in tissue specificity than prairies (Fig. 1B), indicating that when tissue type changes, the expression variation of nearby forest genes tend to be independent to each other, whereas prairie genes display a much stronger tendency of co-activation/repression with its nearby prairie genes. Therefore, prairie genes of similar expression profiles (and thus probably highly related biological functions) prefer to cluster in adjacent genome regions of low CGI density. They are likely positioned for synergistic regulation, as shown below. Along this line, a number of prairie domains are identified to significantly enrich TSGs of similar functions and we term these domains as functional modules. A functional module is defined based on the condition that the number of relevant TSGs (TSGs belonging to one certain tissue, *e.g.*, liver TSGs) in this prairie domain is equal to or greater than the maximum number of relevant TSGs in individual forest domains (Supplementary Fig. S2A), noting that forests possess the great majority (78.5%) and thus a much higher density of genes. For instance, spleen functional module, the 31st prairie domain (PD31) of chr19 (genomic location: chr19 54711241–55450974, chr19 has 32 forest domains and 34 prairie domains), contains 27 genes and 12 out of 27 genes (p-value $< 10^{-18}$ by Fisher's exact test) are specific to spleen; 24 genes in this domain are related to immune function. As such, this particular prairie domain is heavily involved in immune response and maintenance of the normal physiology of immune system. Corresponding

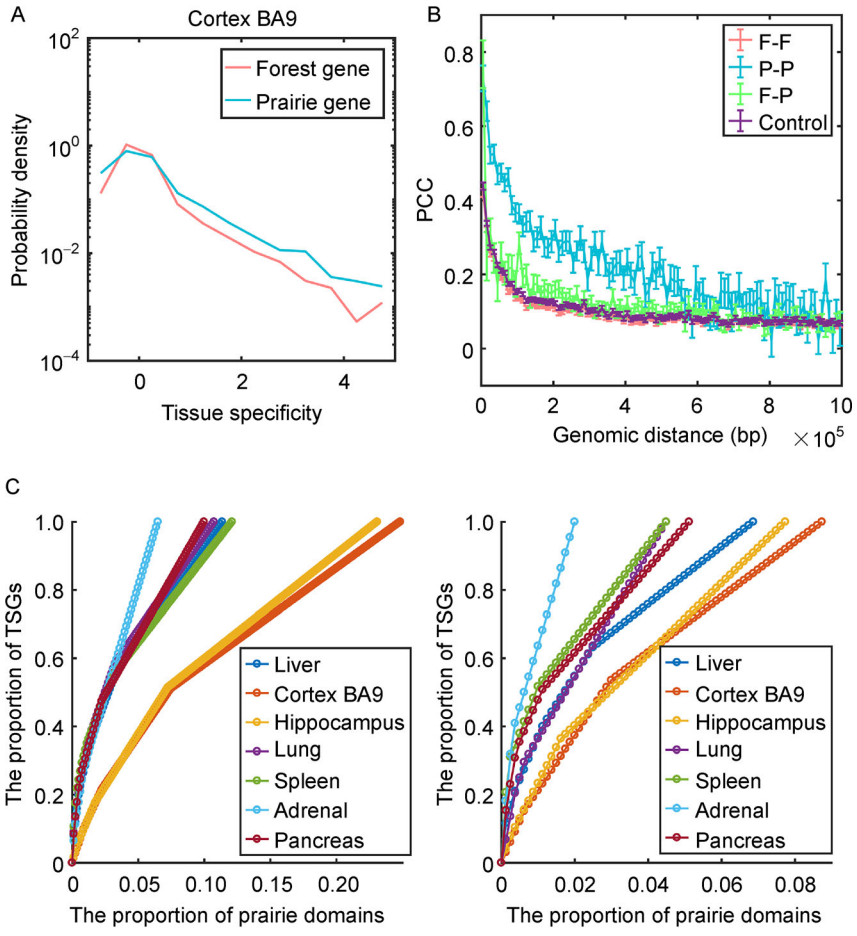

Figure 1. Forest and prairie gene features. (A) The probability density distribution profile of gene tissue specificities for forests and prairies in cortex BA9. The number of forest and prairie genes is 15070 and 3343, respectively. Genes located in autosomes were considered. (B) Pearson correlation coefficient (PCC) calculated between the tissue specificity files (among different tissues) of two forest or prairie genes within the same domain. As a comparison, PCC for all gene pairs regardless of their attribution (labeled as "control") and forest-prairie genes was also calculated. Each data point represents the mean value of PCC derived from a group of gene pairs, the genomic distance of which lies within one 10-kb-long range, *e.g.*, [0, 10,000 bp), [10,000 bp, 20,000 bp). *p*-value = 1.9e–23 (prairie gene-prairie gene vs forest gene-forest gene), 2.2e–19 (prairie gene-prairie gene vs forest gene-prairie gene) and 3.1e–23 (prairie gene-prairie gene vs control) by t-test. Error bar: mean±SE (standard error). (C) Distribution of prairie TSGs at individual domain level (see methods), left, $s_i^t > 1$; right, $s_i^t > 2$. Briefly, the vector composed of the number (non-zero) of relevant prairie TSGs (TSGs belonging to one certain tissue) in each prairie domains was sorted in a descending order and the *i*-th element was converted to the ratio between the summation of the first *i* elements and the number of all relevant prairie TSGs.

functional modules were also identified for liver, adrenal and pancreas (Supplementary Fig. S2A and Table S2). If a less strict criterion (the tissue specificity of gene *i* in tissue *t*, s_i^t, is larger than 1 instead of 2) in identifying TSGs is adopted, one such functional module (Supplementary Fig. S2B) is also identified for lung, which is again PD31, indicating the importance of immune related gene expressions in lung. In contrast, no functional module can be identified in cortex BA9 or hippocampus under either conditions, although they encompass considerable numbers of relevant prairie TSGs (Supplementary Fig. S2A and S2B). Our previous analysis revealed that unlike liver and spleen, which display a strong F-P domain segregation and long-range P-P aggregation, cortex BA9 chromatin exhibits a significant mixing between the forest and prairie domains [16]. The lack of brain-related functional module and thus the largely uniform distribution of brain-related genes in the genome

Table 1 TSG distribution in forest and prairie domains

	Liver	Cortex	Hippo	Lung	LV	Spleen	Ovary	Adrenal	Aorta	Panc
F_t	217	193	169	67	40	106	43	55	10	128
F_{nt}	14853	14877	14901	15003	15030	14964	15027	15015	15060	14942
P_t	95	99	77	44	8	58	11	22	4	65
P_{nt}	3248	3244	3266	3299	3335	3285	3332	3321	3339	3278
p-value	1.1e–7	9.7e–11	6.4e–7	9.5e–8	–	1.3e–7	–	0.025	–	2.4e–7

F_t, F_{nt}, P_t and P_{nt} represent the number of forest TSG, forest non-TSG, prairie TSG and prairie non-TSG in corresponding tissue, respectively. Fisher's exact test was performed and "–" indicates the corresponding *p*-value is larger than 0.05. Hippo = hippocampus, LV = left ventricle, Panc = pancreas.

sequence (Fig. 1C and Supplementary Figs. S2 and S3) are consistent with such a chromatin 3D structural feature, which will be further discussed later.

Furthermore, the forest and prairie domains are shown to have a different dependence of gene tissue specificity on gene body CpG density (Fig. 2A). For prairie but not forest TSGs, as their tissue specificity increases, the gene body CpG density decreases (Spearman coefficient = –0.13, *p*-value = 2.9e–14). Related to this finding, several earlier studies have revealed that the repressive histone mark H3K9me3 is strongly enriched in genes characterized by low CpG density in gene body [26] and the H3K9me3 density negatively correlates with GC content [27]. Here, we also observed a positive correlation (forest: PCC = 0.21, *p*-value = 1.14e–152; prairie: PCC = 0.0526, *p*-value = 0.0023) between compartment index (see "Materials and methods") and gene body CpG density (Fig. 2B and Supplementary Figs. S4 and S5). These results indicate that prairie genes possessing higher tissue specificities tend to be CpG poor in gene body and reside in compartment B, unless they are actively transcribed in their corresponding tissue, presumably with the help of specific TFs. On the other hand, prairie HKGs are on average CpG rich (Fig. 2A), although their transcriptional level is still lower than their counterparts in forests (Supplementary Fig. S6).

3D chromatin structure in the regulation of prairie genes

It is thus interesting to understand how prairie TSGs, especially those characterized by very low CpG densities in gene body, become highly transcribed in their corresponding tissues. Previous studies have shown that compartment A-B switch was associated with regulation of specific genes [14,15] and prairie regions located in compartment A tend to be cell-specific based on gene function classification [16]. Notably, here we found that unlike forest TSGs, the majority of prairie TSGs reside in compartment B (the corresponding eigenvector is smaller than zero, see "Materials and methods") even in tissues where they are highly transcribed (Fig. 3A). To examine whether the activation of prairie TSGs can be related to chromatin structural organization, we used forest index (f_i, see "Materials and methods") to evaluate the local spatial contacting environment of prairie TSGs. A high value of f_i implies that a gene is embedded in a spatial environment rich in forest domains. Our analysis revealed that the forest indices of relevant prairie TSGs (*i.e.*, prairie TSGs belonging to one tissue, *e.g.*, liver prairie TSGs) in the relevant tissue (*e.g.*, liver prairie TSGs in liver) are generally more positive than those of the same genes when they are in other tissues (Fig. 3B and Supplementary Fig. S7A), in agreement with the changes of compartment index (Supplementary Figs. S7C and S8A), one parameter used to describe the compartmentalization degree of genes, high (positive) value indicates gene resides in the inner of compartment A (see "Materials and methods"). Furthermore, as one changes from one tissue type (*e.g.*, spleen) to another (*e.g.*, liver), the associated changes of forest and compartment indices of prairie TSGs specific to the latter (*e.g.*, liver) are generally higher and more positive (Fig. 3C and Supplementary Figs. S7B, S7D, S8B and S8C) than genes that are specific to other tissues (*e.g.*, non-liver prairie TSGs). These observations indicate that the prairie TSGs of a given tissue are in a more forest-surrounded and more active (compartment A) 3D structural environment in that particular tissue than in other tissues. Besides, as previously noted, the TSGs are enriched in prairie domains compared to forests and genes of similar expression behaviors (co-activation/repression) are more likely to cluster in a limited number of prairie regions. These particular prairie domains serve as functionally specified warehouses of TSGs. Collective activation or repression of these genes can then be achieved through concerted and tissue-specific 3D chromatin reorganization of the prairie domains they occupy, by specifically intermingling with or separating from forest domains, respectively.

Our previous results [16] showed that forest and prairie regions are largely overlapped with compartment A and B, respectively, therefore, from the relation between prairie gene regulation and genome organization, the behavior of these prairie genes from the perspective of conventional A/B resembles that of forest/prairie. What's the relation between forest/prairie, two distinctly different sequence domains, and compartment A/B, identified

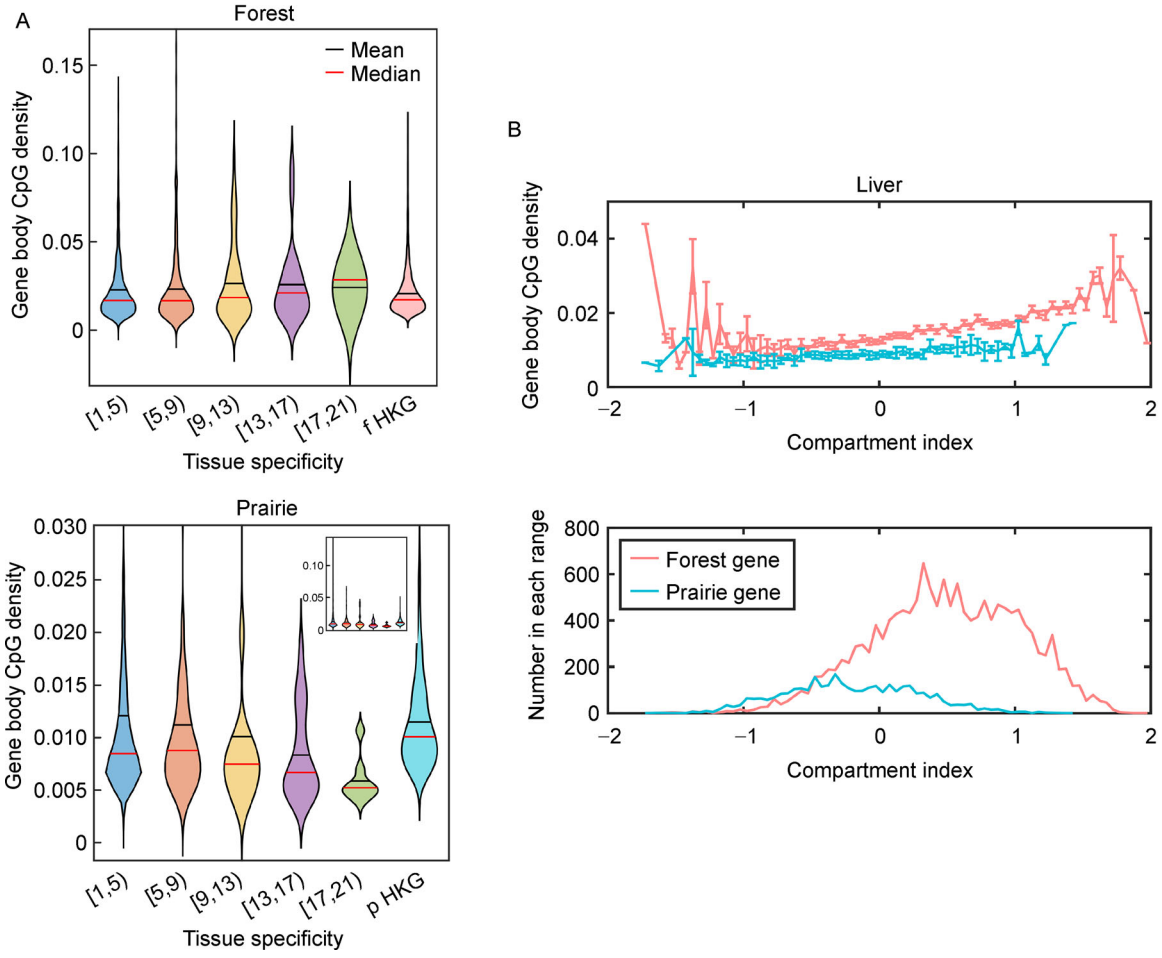

Figure 2. The sequence dependence of gene tissue specificity and 3D environment gene located in. (A) The relation between gene tissue specificity and gene body CpG density. The gene tissue specificity is its maximum value among all tissues (except for testis, which contains a very large number of TSGs, see Supplementary File 2). Inner: unamplified figure. (B) The relation between compartment index and gene body CpG density. Upper, the median value of gene body CpG density, the compartment indices of corresponding genes lie within one 0.05-long range, *e.g.*, [–2, –1.95). Error bar: Median±SE (standard error). Down, the number of genes in each range.

based on Hi-C matrix? Forest and prairie regions tend to spatially segregate [16] and the compartment assignment is similar between different cell types [15]. These two points, together with the large overlap between forest and A as well as prairie and B, indicate that the formation of compartment A/B is largely based on forest and prairie (sequence property), but also possesses tissue specificity (*e.g.*, A-B switch between different tissues). Many studies [14,15] have revealed that compartment A-B switch is related to specific gene expression and tissue-specific transcription factors, histone marks and other factors are suggested to play an essential role in the formation of tissue-specific compartment A/B. Akin to these, as shown above, prairie TSGs need to enter a more forest and compartment A environment for activation, under the help of transcription-related factors (which will be discussed later), letting these specific prairie segments be more compartment A in the corresponding tissue. Furthermore, we also noticed that the activation of many prairie TSGs does not require B (eigenvector < 0) to A (eigenvector > 0) switch (Fig. 3A), but a movement towards compartment A suffices. Therefore, the compartmentalization degree of prairie genes as well as their distribution in compartments in different tissues provide information on how tissue specificity is established given the invariant sequence and how small compartment assignment difference between different tissues is formed under the forest-prairie phase separation model as mentioned above.

The tissue-specific functional modules defined earlier also show tissue-specific enhanced contacts with forest domains. We calculated the segregation ratio R_s (see

methods) to quantify the segregated state (a prairie with other prairies) of these prairie domains. High value of R_s indicates that the prairie domain tends to spatially contact with the other prairie domains. The functional modules in liver, spleen and adrenal all showed the nearly smallest R_s values in the corresponding tissues (Fig. 3E and Supplementary Fig. S9), indicating the expected F-P intermingling of these domains is consistent with tissue-specific active transcription. Overall, the tissue-specific genome organization is consistent with expression level of genes residing in these domains. The functional module of pancreas was not included in this calculation as the pancreas TSG (amylase gene) cluster locates in a very small, variable and dynamic region [33], for which the Hi-C contact information is missing.

The above phenomena, in which specific spatial interactions with forest domains are coupled with gene regulation, were also observed in forest (Supplementary Figs. S10 and S11). However, one can see that in the aspect of transcriptional activation, the compartment index changes of forest TSGs are generally less significant than prairie TSGs (Supplementary Figs. S12 and S13). To quantify this difference, we calculated the PCC between gene expression levels and forest/compartment indices. One can see from Fig. 3D that the values for forest genes tend to be more negative than those for prairie genes. Therefore, the activation of forest genes depends less on the specific spatial interactions (*i.e.,* movement to forest and compartment A environment) compared to prairie genes, suggesting the importance of other mechanisms other than compartmentalization in forest gene expression regulation. Interestingly, we found that forest genes showing strong negative correlations between their gene expression level and forest index tend to be brain-related (Supplementary Figs. S14 and S15). This observation, together with the broad and uniform distribution of brain prairie TSGs in the genome sequence that need to intermingle with forest domains for activation, is indeed accordant with the cortex BA9 chromatin structure derived from Hi-C data [16].

In addition, a positive (although weak) correlation was also observed between the tissue specificity of prairie genes and the mean tissue specificity of its highly contacted forest genes (Fig. 3F and Supplementary Figs. S16 and S17), how highly contacted gene pairs are identified can be found in "Construction of gene networks" of "Materials and methods". Briefly, two genes *i* and *j* are regarded in strong spatial contact if their normalized contact probability exceeds a cutoff derived from a vector composed of the contact probabilities, the genomic distance of which is equal to the linear distance between genes *i* and *j*. This result indicates that prairie genes of high tissue specificities tend to contact with the genes of related functions residing in forests. Besides, a positive correlation was indeed found to exist between gene expression levels of prairie genes and its highly contacted forest genes (Fig. 3F inner and Supplementary Figs. S18 and S19). We also performed a control analysis calculating the correlation coefficients between the expression level/tissue specificity of prairie genes and the average expression level/tissue specificity of its some no highly contacted forest genes (randomly selected, the number of which is equal to the number of highly contacted forest genes for the same prairie gene), which were found to be smaller and more negative than the correlation between prairie genes and its highly contacted forest genes (Fig. 3G and Supplementary Fig. S20), further emphasizing the role of tissue-specific forest-prairie gene contact in their regulation.

The interplay between gene pair expression coordination and genome organization

Given that the co-regulation of gene pairs is largely associated with genome organization [17–19], we next asked whether gene pair expression coordination is also related to domain segregation. To address this question, we first constructed gene networks based on their expression correlation coefficients (see "Materials and methods") following standard practice [34,35]. Briefly, two genes are considered to be co-regulated if their expression correlation coefficient exceeds the 99.5th percentile of all correlations. To quantify the effect of compartmentalization on gene expression correlation, we calculated the average number of compartment X(X = A or B) genes connected to genes in compartment Y(Y = A or B), $N_{Y\text{-}X}$, in the co-regulation networks. Interestingly, genes tend to correlate with other genes residing in the same other than different compartment type, since $N_{B\text{-}A}$ and $N_{A\text{-}B}$ are always lower than $N_{A\text{-}A}$ and $N_{B\text{-}B}$, respectively, with the exception of cortex BA9 (Fig. 4A and Supplementary Fig. S21). Such a result strongly suggests a non-negligible association between compartmentalization and gene pair co-regulation and is accordant with previous findings revealing that paralogs tend to locate in the same compartment, and many of which are expressed with high coordination [18]. Our results presented here were derived from the analysis of all genes (18413) of autosomes. For the cortex, the $N_{B\text{-}A}$ is larger than $N_{A\text{-}A}$, implying the prominent behaviors of compartment B genes in mediating brain cell activity, and we further selected the top10 compartment B genes, which are correlated with the largest number of genes in compartment A. These ten genes are *SYNGR1, MYH10, TMEM246, FBXL2, ATP8A2, ELMOD1, PTPN3, FAXC, HECW1* and *ARHGAP44*. Among those, *SYNGR1* is associated with presynaptic vesicles in neuronal cells, *ATP8A2* is involved in the regulation of neurite out-

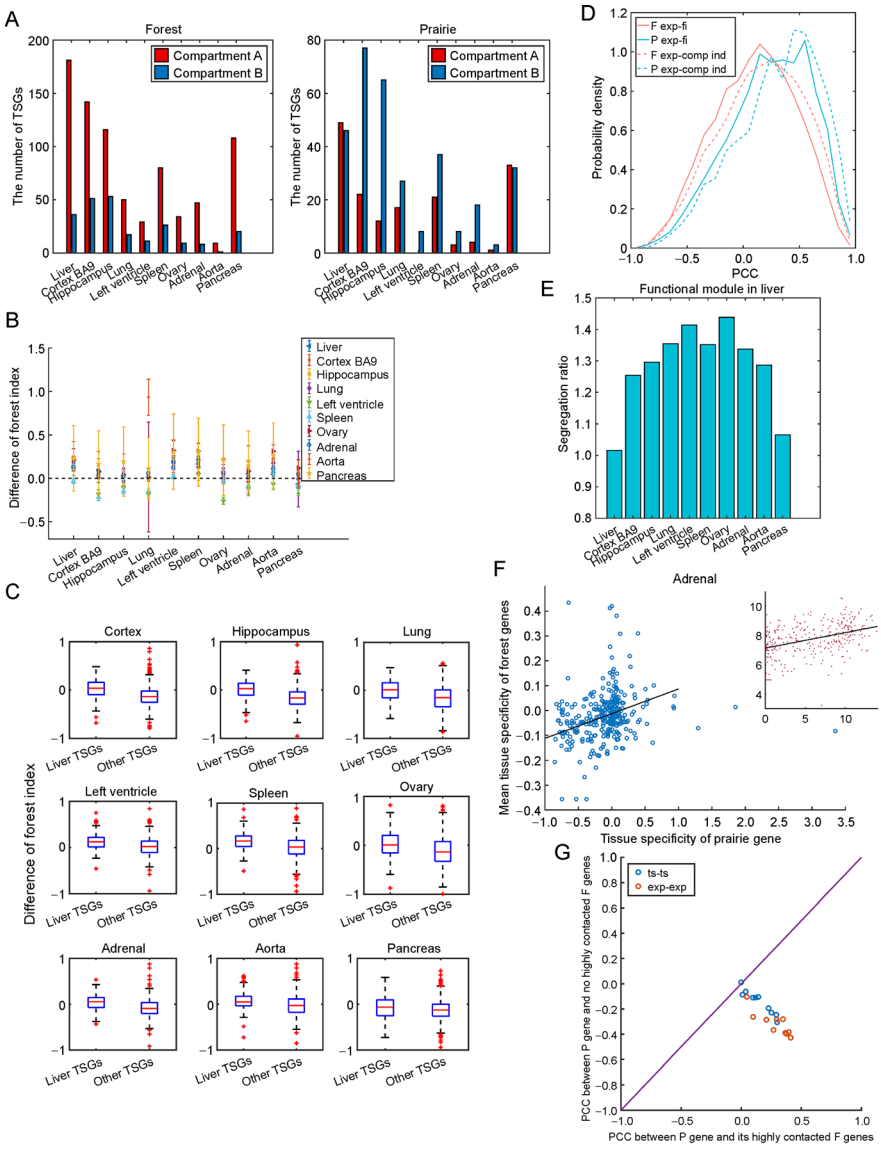

Figure 3. Prairie gene expression regulation and genome organization. (A) The number of TSGs residing in compartment A (eigenvector > 0, see "Materials and methods") or B (eigenvector < 0). (B) Forest index changes of relevant prairie TSGs (TSGs belonging to one tissue, *e.g.*, liver prairie TSGs) from other nine tissues (control tissues, *e.g.,* spleen) to the relevant tissue (*e.g.,* liver). *Y*-axis represents the median value of forest index difference calculated between relevant tissue (illustrated in legend) and control tissue (labeled in *x*-axis) for relevant tissue's prairie TSGs. For instance, pancreas (five-pointed star) has nine values calculated between pancreas and the remaining 9 tissues, for the prairie TSGs of pancreas, respectively. Almost all values are positive, indicating that when tissue changes from T1 to T2, at least half of prairie TSGs belonging to T2 enter a more forest environment for activation. Error bar: Median±SE (standard error). (C) Comparison of forest index changes between relevant prairie TSGs (prairie TSGs of one certain tissue, *e.g.*, liver TSGs) and prairie TSGs specific to other tissues (complementary TSGs, *e.g.*, non-liver TSGs) when tissue type changes from one other tissue (control tissue, *e.g.*, non-liver tissue) to the relevant tissue (*e.g.*, liver). Liver was used as illustration (relevant tissue), title of each boxplot represents the control tissue. *p*-value < 0.001 (Welch's unequal variance t-test) in most cases, except for left ventricle and pancreas. For clarity, these nine figures were amplified and the corresponding intact figures can be found in Supplementary File 3. (D) Pearson correlation coefficient (PCC) between expression level and forest/compartment index for forest and prairie genes. *p*-value = 4.6e–75 (P exp-fi vs F exp-fi), 9.9e–23 (P exp-fi vs F exp-comp ind), 4.0e–170 (P exp-comp ind vs F exp-fi) and 4.9e–87 (P exp-comp ind vs F exp-comp ind) by Welch's unequal variance t-test. (E) The segregation ratio of liver functional module in ten tissues. (F) Scatter plot for tissue specificity (expression level, inner figure) of prairie genes and the averaged tissue specificity (expression level) of forest genes in strong contact (see methods). (G) Comparisons of Pearson correlation coefficients (PCC) calculated from several conditions. Condition 1/2: prairie gene expression level/tissue specificity-average gene expression level/tissue specificity of its highly contacted forest genes. Condition 3/4: prairie gene expression level/tissue specificity-average gene expression level/tissue specificity of its some no highly contacted forest genes. *p*-value < 0.05 in eight (except for cortex and lung) and six (except for liver, hippocampus, lung and left ventricle) tissues for conditions 1 and 2, respectively. As a control, conditions 3 and 4 only have one correlation level, the *p*-value of which is smaller than 0.05 (ovary for condition 3).

growth, and *FAXC* may play a role in axonal development. Besides, a cortex prairie TSG, *NEUROD6*, was also located in compartment B and showed high expression correlations with many compartment A genes (the number is 408). Intriguingly, previous work has revealed that downregulation of *NEUROD6* could be a possible biomarker for Alzheimer's disease brains [36]. Therefore, perturbation of *NEUROD6*-related/mediated pathways may be associated with the disease occurrence or development.

To further explore the spatial interactions of co-regulated gene pairs, we next reconstructed a different type of gene network. The edges of this new network connect genes that are not only highly correlated but also are in strong spatial contact (see "Materials and methods"). Two tissues, liver and cortex BA9, were selected to illustrate the results as their F-P domain segregation patterns are significantly different as shown in our previous work [16]. The former displays a strong separation between forest and prairie domains and an aggregation between prairie domains distantly distributed along the linear genome, whereas the latter shows a strong intermingling between forest and prairie domains. All genes of chr1 were included in constructing the gene network. The results revealed that such networks in liver and cortex BA9 are roughly composed of two independent sub-networks (Supplementary Figs. S22 and S23), the gene tissue specificity distributions of which are distinctly different (Fig. 4C and Supplementary Fig. S24). One network is directly associated with cell identity and the other is of more generic functions, indicating that the 3D chromatin structure and spatial contacts among genes may play an important role in the formation of tissue-specific gene network and the establishment of tissue specificity. To further validate the above phenomena, we constructed such gene networks in the remaining eight tissues (hippocampus, lung, left ventricle, spleen, ovary, adrenal, aorta and pancreas, see Supplementary File 4) and found that six tissues (except for spleen and aorta) display similar patterns (Supplementary Fig. S25, the network is roughly composed of two independent sub-networks, the gene tissue specificity distribution of which is significantly different) compared to liver and cortex. In addition, the pattern of interactions between forest and prairie genes in the tissue-specific sub-network does reflect the overall segregation of forest and prairie domains in that particular tissue (Fig. 4B): In liver, the network core is mainly composed of forest genes and prairie genes occupy only the marginal positions, whereas the cortex BA9 network displays a strong intermingling of forest and prairie genes (the non-tissue-specific sub-network in liver also displays F-P separated state, whereas in cortex BA9 the scale of such network is very small (Supplementary Figs. S22 and S23)). Overall, the proportion of F-P edges in the whole network in liver and cortex is 6.04% and 17.78% (p-value = 1.2e–7 by Fisher's exact test).

We next compared the expression-only (see Supplementary File 5) and expression + Hi-C gene networks and found the former already showed a tendency to contain two sub-networks, which are connected by only small number of edges, indicating the separated correlation modules characterized by the distinct tissue specificity distribution. Intriguingly, the number of highly correlated F-P gene pairs in the forest-prairie intermingling tissue, cortex, is significantly larger than that in liver, a forest-prairie segregated tissue (3088 vs 1934, p-value = 5.82e–80), again implying the association between forest-prairie domain segregation and forest-prairie gene co-regulation.

Since unlike the standard gene expression correlation network, the expression + Hi-C network presented here also includes spatial information and can be used to infer information on the co-regulatory mechanism (such as the binding of TFs and spatial clustering) of the genes involved. For example, in the liver network, *CA14* (a forest gene) is found to connect with six genes, of which five (*HFE2*, *ANXA9*, *SELENBP1*, *PKLR*, *RHBG*) reside in forest domains and one (*FMO3*) in prairie. In liver, *HFE2*, *PKLR*, *RHBG*, *ANXA9* and *FMO3* are highly expressed and the promoter regions of all these seven genes bind HNF4A, a liver-specific TF, RXRA and YY1 [37], strongly suggesting that the co-localization of these genes, with the help of TF binding, plays a role in the formation of tissue-specific gene regulation network and their co-activation in liver. Hence, like in the phase separation model, the tissue–specific formation of such clusters as well as spatial contacts between forest and prairie genes may render an efficient regulatory mechanism in the gene transcription, especially for prairie genes, and the subsequent execution of hepatic functions. These interactions are worthy of experimental validation.

Next, we calculated the Hi-C rank between pairs of genes (see "Materials and methods") to quantify their relative Hi-C contact strength. Zero value of Hi-C rank indicates that there is no contact between two genes. The distribution of Hi-C rank of highly correlated gene pairs (see "Materials and methods") within a tissue at different Hi-C contact levels was calculated and compared with all gene pairs from the gene list. The results given in Fig. 4D show that in all tissues examined here, the ratio calculated between the Hi-C rank distributions of highly correlated gene pairs and all gene pairs is always less than 1 for pairs of zero Hi-C rank and greater than 1 for pairs of non-zero Hi-C rank. Such a result shows that highly correlated gene pairs are more likely to form Hi-C contact. Finally, the gene expression correlation coefficients of gene pairs that

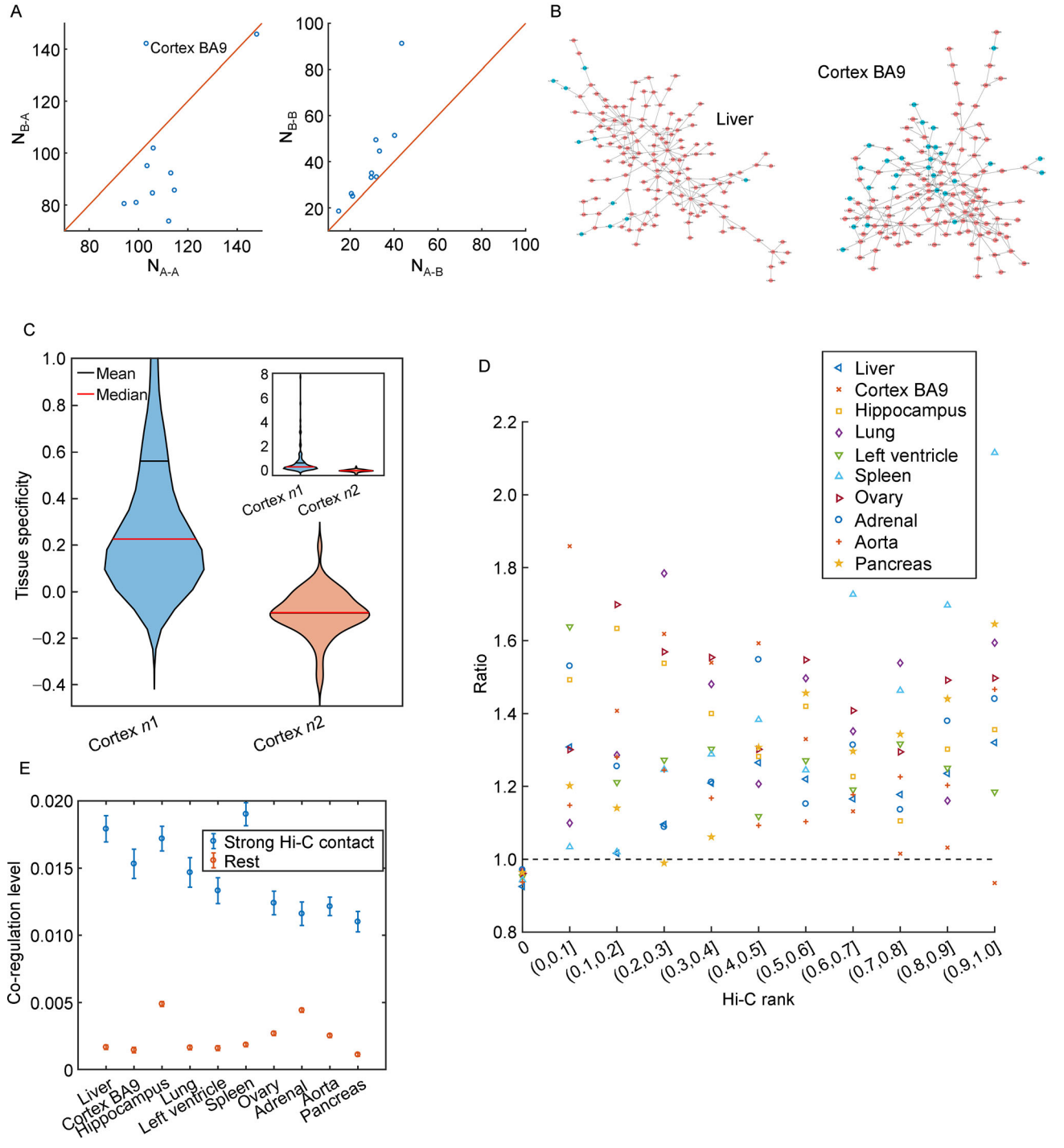

Figure 4. The interplay between gene pair expression coordination and chromatin structure. (A) Comparisons between N_{A-A} and N_{B-A} as well as N_{A-B} and N_{B-B} in the co-regulation networks of the ten tissues. The only outlier is cortex BA9. (B) The two tissue-specific sub-networks for liver and cortex BA9 (Cytoscape software was used). (C) Comparison of the gene tissue specificity distribution between the two sub-networks in cortex BA9. *n*1, tissue-specific sub-network; *n*2, the generic sub-network. *p*-value = 3.8e–12 by Welch's unequal variance t-test. Inner: unamplified figure. (D) The ratio calculated between the Hi-C rank distributions of highly correlated gene pairs within chr1 and all gene pairs ($\binom{n}{2}$, n is the number of genes of chr1 (1979)). (E) The average co-regulation levels of highly contacted gene pairs compared to those that are not. *p*-value < 1e–15 by Welch's unequal variance t-test in all tissues. All genes of chr1 (1979) were included in this analysis. Error bar: mean±SE (standard error).

are in high contact (see methods) are larger than those of low contact probability (Fig. 4E and Supplementary Fig. S26). Therefore, gene pairs forming stronger spatial contacts are also more likely to be highly correlated in expression, accordant with previous studies [20]. Together, these results show the existence of an association between chromatin structure, especially domain segregation, and gene pair expression correlation, in addition to the expression levels of individual genes.

DISCUSSION

In the present study, we try to connect the gene distribution in the linear and 3D genome to the establishment of tissue specificity. We integrated structural and genetic data to systematically study the relationship between gene expression and chromatin organization in tissues and observed a common role that 3D genome structure plays in gene regulation. The CGI-deficient prairies significantly enrich genes of high tissue specificity compared with CGI forests, whereas the latter is abundant in non-tissue-specific genes and their overall tissue specificity is close to zero [16]. Furthermore, we found that the distribution of genes in prairie regions shows significant function-related mosaicity. Prairie and, to a lesser extent, forest domains, tend to cluster genes of similar expression variations among different tissues (therefore probably similar biological functions). Accordingly, a small number of prairies were found to significantly enrich TSGs of similar functions. The assembly of genes of particular functions in linearly adjacent regions of low CpG density is likely to correlate with their co-activation in corresponding tissues, possibly under the assist of specific TFs, and, equally important, co-repression in other tissues.

Besides the enrichment of TSGs, the relation between gene body CpG density and gene tissue specificity is also distinctly different between forest and prairie domains. For prairie but not forest TSGs, the tissue specificity is negatively correlated to gene body CpG density. In addition, a positive correlation was found between prairie gene body CpG density and compartment index. Such an observation is in line with earlier studies [26,27] which showed a strong negative correlation between the repressive mark H3K9me3 and CpG density (or GC content). Interestingly, many highly transcribed prairie TSGs remain in compartment B but in stronger contact with forest domains in tissues where they are highly expressed than in other tissue types. Therefore, prairie gene activation appears to depend on tissue-specific interactions with the forest domains, resulting in a shift towards compartment A. The functional modules together also display the strongest domain-mixing (between forests and prairies) in their highly-expressed tissues. Such a trend is also observed for the activation of forest TSGs, but to a lesser extent. Thus, the mechanism of forest gene regulation needs to be further investigated. We also found that prairie genes of higher tissue specificities or expression levels tend to spatially contact with forest genes harboring relatively higher tissue specificity or expression level, which is reflected by the positive correlation between the tissue specificity (expression level) of prairie genes and the mean tissue specificity (expression level) of its highly contacted forest genes.

TFs have been widely reported to dynamically mediate chromatin structure. For example, TF binding sites can form clusters over long genomic distances [24,25]. CTCF was proposed to participate in the TAD formation through a loop extrusion model [38,39] and was involved in tethering genome regions to nucleolus [40], indicating its capacity in chromatin repositioning. YY1, similar to CTCF, can form dimers to enhance interactions between enhancers and promoters [41]. C/EBPα and OSKM were suggested to drive A-B compartment switch and TAD border insulation change during B-cell reprogramming [14]. These observations suggest to us that TFs may also play a non-negligible role in mediating tissue specific F-P interactions.

The current analyses yielded many F-P gene pairs (see Supplementary File 6), characterized by strong spatial contacts and high expression correlation levels for each tissue studied. Furthermore, the expression levels of a number of TFs tend to correlate with not only the expression level of these individual forest and prairie genes in that tissue but also their correlation coefficients in different tissues (Supplementary Table S3), strongly suggesting a co-regulation of the forest and prairies genes induced by these TFs. For example, in liver, *CYP4A11* (forest gene) and *CYP2J2* (prairie gene) are highly correlated in expression and in strong spatial contact. In addition, the expression level of one liver-specific TF, NR1I3, correlates not only with both *CYP4A11* and *CYP2J2* genes in liver, but also to the expression correlation coefficients between the latter two. Together, these results are consistent with TFs interacting with both forest and prairie genes and especially facilitating the movement of the latter to a more transcriptionally active (forest) environment for regulation. This mechanism is consistent with the phase separation model in which the multivalent interactions of specific proteins can induce their condensation and subsequently form liquid-like droplets, resulting in the co-localization of genes. A recent *in vitro* and *in silico* study has revealed that the DNA sequence features, including TF target site valence, density and affinity, play a vital role for driving condensation and both TF-DNA and TF-Mediator interactions determine the formation of condensates [42]. One should note that besides TFs, other molecules such as

non-coding RNA may also play important roles in the spatial co-localization of genes.

The analyses presented here on the gene co-regulation network showed that in general genes are more likely to correlate in expression with genes in the same compartments, when compared to those in different compartments, indicating a non-negligible connection between compartmentalization and gene co-regulation. We then constructed gene networks, the edges of which represent not only high co-regulation level but strong spatial proximity, to investigate the spatial interactions of co-regulated gene pairs. Compared to the standard gene network constructed by considering only gene expression data [34,35], this new network also contains information on the spatial interaction patterns of co-regulated genes and may be used in identifying tissue-specific gene clusters in the nuclear. We found that the interactions between forest and prairie genes in the tissue-specific subnetwork in liver and cortex BA9 do reflect the overall chromatin organization features, indicating the association between gene co-regulation and forest-prairie domain segregation, which was further validated by the expression-only network. Next, we quantitatively evaluated the relationship between gene co-regulation and spatial interaction, and found that the co-regulated gene pairs are more likely to form Hi-C contact and the expression correlation coefficients of gene pairs in strong spatial contact are on average higher than others, consistent with previous studies showing a significant overlap between Hi-C and connection tendency matrices [20]. These results illustrate the important roles chromatin organization plays not only in gene expression but also in gene regulatory network formation.

One intriguing discovery from the analysis of the gene network for cortex BA9 sample is that many genes associated with neuro or metabolism functions are tightly coupled (co-regulated) and co-localized in space with each other. For example, *FABP3*, playing a role in the metabolism of long-chain fatty acids and lipoprotein, directly connects two genes, *HPCA* and *SH2D5*. *HPCA* encodes a neuron-specific calcium-binding protein, which is thought to contribute to the function of neurons in the central nervous system (CNS), and the protein encoded by *SH2D5* is known to participate in the regulation of synaptic plasticity. *ST3GAL3*, which is associated with the metabolism of glycosaminoglycan and glycosphingolipid, connects two genes *EPHA10* and *SNAP47*. The protein encoded by *EPHA10* acts as a mediator of axon guidance and cell-cell communication in neuronal and epithelial cells. *SNAP47* is related to syntaxin binding. As another example, *SLC45A1*, which encodes glucose transporter in the brain, is correlated to and in strong spatial contact with *HTR6* that encodes a G protein-coupled protein. The latter regulates cholinergic neuronal transmission in brain and spatial memory. *SLC44A5*, similar to *SLC45A1*, is connected to *RGS7* which regulates G protein signaling and synaptic vesicle. *PRKCZ*, related to glucose transport upon insulin treatment, connects *GABRD* that mediates neuronal inhibition. These and other examples show that the execution of normal brain functions significantly depends on these cross-function (and inter-forests-prairies) interactions (both in space and in expression regulation). Interestingly, it is known that the pathway of glucose metabolism is highly relevant to neurodegenerative disorders such as Alzheimer disease [43]. Previous studies also showed that a possible connection does exist between neurodegenerative diseases, *e.g.*, Alzheimer disease, and the digestive system, such as the metabolism of lipid [44,45] and bile acids [46].

Finally, as we mentioned above, the activation of many prairie TSGs does not require compartment switch (B to A), but spatial movements towards A potentially work, raising one question about the molecular mechanisms behind such genome reorganization. In fact, loop extrusion model [39], mediated by CTCF and cohesin, contributes to the TAD formation and has non-negligible influences on gene regulation through linking enhancer and promoter. For example, a recent study [47] performed enhancer-promoter analysis during neural development and identified a region, which we found to overlap with both forest and prairie domains and display stronger promoter-enhancer interactions in neurons than in induced pluripotent stem cells (iPSCs). Seven out of eight genes in this region are up-regulated. Therefore, enhancer-promoter interactions, possibly mediated by loop extrusion, may result in gene activation through forest-prairie specific spatial interactions. In addition, epigenetic information is also closely associated with 3D genome organization. For instance, Qi and Zhang [48] used one-dimensional data, including histone marks and CTCF binding sites that were found to correlate with compartment types [4], to predict high-resolution three-dimensional chromatin structures. They were able to reproduce experimental results and uncover long-range enhancer-promoter interactions. DNA methylation canyons [49], characterized by low DNA methylation and high H3K27me3 levels, can form cell-type specific and long-range chromatin interactions that are independent of CTCF and cohesin, contributing to the formation of sub-compartments. Similarly, specific spatial interactions between *HOXA* genes and DNA methylation canyon in acute myeloid leukemia promote *HOXA* genes expression [50], again emphasizing the vital role of epigenetic information in genome architecture. Cai *et al.* [51] identified H3K27me3-rich regions, termed as super-silencers, and found that these regions (similar chromatin state) prefer to spatially interact with each other, in

association with gene regulation (*e.g.*, the expression level of genes associated with super-silencers are more susceptible upon EZH2 inhibition). Therefore, different molecular mechanisms, including loop extrusion, specific spatial interactions between loci that are characterized by epigenetic information, may all contribute to the compartmentalization degree change of specific genes, which in turn influence the gene expression.

In summary, we performed here a quantitative analysis of distributions of genes with different tissue specificities in the linear and 3D genome, and their effects on the level and correlation of gene expression. It was found that (1) Consistent with earlier studies [16], tissue-specific genes are enriched in the prairie domains. More interestingly, genes of related functions tend to accumulate on individual prairie domains, and therefore are close in the linear genome of low CpG density sequences. (2) The expression of tissue-specific genes on prairies correlate more significantly with the large-scale chromatin structure formation, such as compartmentalization, compared to genes on forests. (3) The tissue specificity of prairie TSGs has a tendency to be negatively correlated with its gene body CpG density. In contrast, no significant correlation was observed in forests. (4) Gene co-regulation patterns in different tissues correlate with the tissue-specific domain segregation of the chromatin, that is, forest-prairie mixing tissue, cortex, tend to possess more forest-prairie gene interactions, compared to the forest-prairie separated tissue, liver. (5) The prairie TSGs interact spatially with gene-rich forest domains in a tissue-specific manner for activation. The forest genes in tissue-specific spatial contact with these prairie TSGs also tend to possess relatively high tissue specificity. On the contrary, the lowly expressed prairies genes tend to move towards compartment B, lose contact with the forest domains, and reside in a more repressive spatial environment. Since forest and prairie are identified solely based on sequence properties and are nearly constant in different biological stages, such as early embryo development, cell differentiation, cell senescence and tumorgenesis, the simple and uniform framework (forest-prairie domain segregation) provided here can therefore be used to further explore the biological mechanisms behind these stages.

MATERIALS AND METHODS

Definition of gene tissue specificity

The tissue specificity of gene i in tissue t is defined as

$$s_i^t = \frac{\varepsilon_i^t - \mu_i^{all}}{\mu_i^{all}},$$

where ε_i^t and μ_i^{all} are the mean expression level of gene i in tissue t and all tissue types included in the calculation (the normalized and comparable RNA-seq data was downloaded from https://zenodo.org/record/838734 [30], in which 38 tissues were provided, see Supplementary File 1), respectively. We consider gene i to be specific to tissue t if s_i^t exceeds a cutoff, ts. Different values of the cutoff ts were tested and the results are robust to this parameter. Without specific notation, the results we presented were obtained using a $ts = 2$.

Distribution of TSGs at individual forest/prairie domain level (saturation curve)

To better describe the distribution of relevant TSGs at the individual domain level, the vector composed of the number (non-zero) of forest/prairie TSGs related to one tissue (*e.g.*, liver) in individual domains was first sorted in a descending order and then each element, *e.g.*, the i-th, was converted to the ratio between the summation of the first i numbers and the total number of forest/prairie TSGs belonging to one tissue (*e.g.*, liver). For a given value of y, the corresponding x value was normalized to the ratio between i and the total number of forest (766)/prairie (801) domains.

Definition of segregation ratio

Segregation ratio of a prairie domain i, R_s^i, is defined as [16]

$$R_s^i = \frac{\sum_{n=1, n\neq i}^{N} D_{in}}{\sum_{m=1}^{M} D_{im}},$$

where N (n$\in$P) and M (m$\in$F) are the collection of all prairie and forest domains, respectively. D_{ij} is the summation of normalized Hi-C contact probability between domains i and j. Without specific notation, the Hi-C contact matrices were normalized using ICE (iterative correction and eigenvector decomposition) method [52]. A high value of R_s^i implies a high probability of the prairie domain i to be spatially in contact with prairie regions.

Calculation of compartment index

The compartment index of bin i, CI_i, is calculated following our previous work [53] as

$$CI_i = \ln\left(\frac{M_{iA}}{M_{iB}}\right),$$

where M_{iA} and M_{iB} are the mean spatial contact probabilities of bin i with compartments A and B, respectively. The identification of compartment A/B

follows our previous work with slight modifications [16]: the entire Hi-C matrix was disassembled into two parts, corresponding to p and q arms, and the eigenvalue decomposition was done within these two arms separately. A high and positive value of CI_i suggests that bin i locates in a more compartment A (interior of compartment A) environment. In contrast, lower (more negative) value indicates that bin i resides in the inner of compartment B.

Calculation of forest index

Forest index of a 40-kb bin i is defined as [16],

$$f_i = \ln\left(\frac{\sum_{j,j\neq i} C_{ij}\delta_j}{\sum_{j,j\neq i} C_{ij}(1-\delta_j)}\right),$$

where C_{ij} is the contact probability between bins i and j. δ_j equals to 1 if the 40-kb bin j locates in a forest, and 0 if it is in a prairie domain. Therefore, a larger value of f_i indicates that the 40-kb bin is spatially surrounded by a higher population of forests.

Construction of gene networks

Gene networks displayed in this study were constructed based on a combined usage of Hi-C [29] and RNA-seq [30] data (Hi-C data source, GEO with accession number GSE87112; RNA-seq data source, https://zenodo.org/record/838734). Two genes connected by an edge in the network not only possess a high co-regulation level, but are also in strong spatial proximity in terms of Hi-C contact.

Two genes i and j are considered to be under co-regulation, if the PCC calculated from $\vec{\varepsilon}_i^t$ and $\vec{\varepsilon}_j^t$ exceeds the 99.5th percentile of all correlations $\{P_{ij}\}$ ($i,j\in\{1,2,...,\mathrm{N}\}$ and $i<j$, N is the number of genes in the dataset we considered, based on such criterion we constructed gene co-regulation network. Another criterion, 99.9th was used for robustness and similar results were obtained). $\vec{\varepsilon}_i^t$ and $\vec{\varepsilon}_j^t$ are the expression vectors for genes i and j in tissue t, respectively. Each element of the expression vector represents the expression level of one sample belonging to tissue t. For instance, if tissue t has n_t samples in the corresponding RNA-seq data, the size of $\vec{\varepsilon}_i^t$ is $1\times n_t$.

The Hi-C contact matrices are of a 40-kb resolution and the co-regulated gene pairs i and j are therefore projected to 40-kb bins, denoted as a and b, respectively. The genomic distance between a and b is denoted as d_{ab}. Given that the contact probability decays quickly with the genomic separation, we use relative values to evaluate the strength of gene pair spatial interaction. Namely, genes i and j are regarded in strong spatial contact in tissue t if the normalized Hi-C contact probability of bins a and b, C_{ab}^t, exceeds the 75th percentile (90th percentile was used for robustness test and similar results were acquired) of $I_{d_{ab}}^{t'}$ that was defined as

$$I_{d_{ab}}^{t'} = \{C_{mn}^t | C_{mn}^t \in I_{d_{ab}}^t,\ C_{mn}^t > \alpha^t\}$$

where $I_{d_{ab}}^t$ is the vector composed of spatial contacts between two bins, the genomic distance of which is equal to d_{ab}. α^t is defined as

$$\alpha^t = s^t \times 0.05\%$$

where s^t is the summation of one row in tissue t's normalized Hi-C contact matrix. Such value was used to align the Hi-C data of different tissues to similar scale (The concrete values and results can be found in Supplementary Table S4).

Identification of gene functions

The website https://www.genecards.org was used for the analysis of gene functions.

Definition of Hi-C rank

The Hi-C rank $r_{i,j}$ between two genes i and j, is defined as

$$r_{i,j} = \frac{\sum_m \Pi(0 < I_{d_{a,b}}^t < C_{a,b}^t)}{\sum_m \Pi(I_{d_{a,b}}^t > 0)},$$

with $C_{a,b}^t$ and $I_{d_{a,b}}^t$ defined earlier. As the genomic distance increases, N_n/Z_n decays exponentially (Supplementary Fig. S27), where N_n and Z_n represent the number of non-zero and zero elements in the Hi-C matrix at the given genomic distance. Thus we retained only non-zero elements and calculated the rank within them.

Data availability

Hi-C and RNA-seq data analyzed during this study were downloaded from Gene Expression Omnibus Database (GEO) under the accession number GSE87112 and https://zenodo.org/record/838734, respectively.

ABBREVIATIONS

HKG	housekeeping gene
TSG	tissue-specific gene
TF	transcription factor
CGI	CpG island
F	CGI forest
P	CGI prairie
TAD	topologically associated domain

Hi-C	high throughput chromosome conformation capture
ChIA-PET	chromatin interaction analysis by paired-end tag sequencing
H3K9me3	histone H3 lysine 9 trimethylation
PCC	Pearson correlation coefficient

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.1007/s40484-020-0221-6.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Nos. 21927901, 21821004 and 21873007) and the National Key R&D Program of China (No. 2017YFA0204702).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Hao Tian, Ying Yang, Sirui Liu, Hui Quan and Yi Qin Gao declare that they have no competing interests.

The article does not contain any human or animal subjects performed by any of the authors.

REFERENCES

1. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326, 289–293
2. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462, 58–64
3. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485, 376–380
4. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159, 1665–1680
5. Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., Weintraub, A. S., Schuijers, J., Lee, T. I., Zhao, K., *et al.* (2014) Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell, 159, 374–387
6. Hnisz, D., Day, D. S. and Young, R. A. (2016) Insulated neighborhoods: structural and functional units of mammalian gene control. Cell, 167, 1188–1200
7. Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science, 351, 1454–1458
8. Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., Weintraub, A. S., Hnisz, D., Pegoraro, G., Lee, T. I., *et al.* (2016) 3D chromosome regulatory landscape of human pluripotent cells. Cell Stem Cell, 18, 262–275
9. Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell, 161, 1012–1025
10. Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., Mirny, L., *et al.* (2017) Two independent modes of chromatin organization revealed by cohesin removal. Nature, 551, 51–56
11. Smol, T., Sigé, J., Thuillier, C., Frénois, F., Brunelle, P., Rama, M., Roche-Lestienne, C., Manouvrier-Hanu, S., Petit, F. and Ghoumid, J. (2020) Lessons from the analysis of TAD boundary deletions in normal population. bioRxiv, 021188
12. Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell, 148, 84–98
13. Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C. T. and Zhuang, X. (2016) Spatial organization of chromatin domains and compartments in single chromosomes. Science, 353, 598–602
14. 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, 238–249
15. Bertero, A., Fields, P. A., Ramani, V., Bonora, G., Yardimci, G. G., Reinecke, H., Pabon, L., Noble, W. S., Shendure, J. and Murry, C. E. (2019) Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. Nat. Commun., 10, 1538
16. Liu, S., Zhang, L., Quan, H., Tian, H., Meng, L., Yang, L., Feng, H. and Gao, Y. Q. (2018) From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective. Nucleic Acids Res., 46, 9367–9383
17. Zhan, Y., Mariani, L., Barozzi, I., Schulz, E. G., Blüthgen, N., Stadler, M., Tiana, G. and Giorgetti, L. (2017) Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Res., 27, 479–490
18. Ibn-Salem, J., Muro, E. M. and Andrade-Navarro, M. A. (2017) Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res., 45, 81–91
19. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V. and Reyes, J. C. (2017) Analysis of the relationship between coexpression domains and chromatin 3D organization. PLOS Comput. Biol., 13, e1005708
20. Belcastro, V., Siciliano, V., Gregoretti, F., Mithbaokar, P.,

Dharmalingam, G., Berlingieri, S., Iorio, F., Oliva, G., Polishchuck, R., Brunetti-Pierri, N., *et al.* (2011) Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. Nucleic Acids Res., 39, 8677–8688

21. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. and Sharp, P. A. (2017) A phase separation model for transcriptional control. Cell, 169, 13–23
22. Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., *et al.* (2018) Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell, 175, 1842–1855.e16
23. Hnisz, D. and Young, R. A. (2017) New insights into genome structure: genes of a feather stick together. Mol. Cell, 67, 730–731
24. de Wit, E., Bouwman, B. A. M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M. J. A. M., Krijger, P. H. L., Festuccia, N., Nora, E. P., Welling, M., *et al.* (2013) The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature, 501, 227–231
25. Monahan, K., Horta, A. and Lomvardas, S. (2019) LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature, 565, 448–453
26. Hahn, M. A., Wu, X., Li, A. X., Hahn, T. and Pfeifer, G. P. (2011) Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. PLoS One, 6, e18844
27. Wang, Z. and Willard, H. F. (2012) Evidence for sequence biases associated with patterns of histone methylation. BMC Genomics, 13, 367–379
28. Kustatscher, G., Grabowski, P. and Rappsilber, J. (2017) Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol. Syst. Biol., 13, 937–950
29. Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., *et al.* (2016) A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep., 17, 2042–2059
30. Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., DeMeo, D. L., Quackenbush, J., Glass, K. and Kuijjer, M. L. (2017) Understanding tissue-specific gene regulation. Cell Rep., 21, 1077–1088
31. Hurst, L. D., Pál, C. and Lercher, M. J. (2004) The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet., 5, 299–310
32. Xu, H., Liu, J.-J., Liu, Z., Li, Y., Jin, Y.-S. and Zhang, J. (2019) Synchronization of stochastic expressions drives the clustering of functionally related genes. Sci. Adv., 5, eaax6525
33. Shwan, N. A. A., Louzada, S., Yang, F. and Armour, J. A. L. (2017) Recurrent Rearrangements of Human Amylase Genes Create Multiple Independent CNV Series. Hum. Mutat., 38, 532–539
34. Ponomarev, I., Wang, S., Zhang, L., Harris, R. A. and Mayfield, R. D. (2012) Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. J. Neurosci., 32, 1884–1897
35. Rosa, B. A., Jasmer, D. P. and Mitreva, M. (2014) Genome-wide tissue-specific gene expression, co-expression and regulation of co-expressed genes in adult nematode Ascaris suum. PLoS Negl. Trop. Dis., 8, e2678
36. Satoh, J., Yamamoto, Y., Asahina, N., Kitano, S. and Kino, Y. (2014) RNA-Seq data mining: downregulation of NeuroD6 serves as a possible biomarker for alzheimer's disease brains. Dis. Markers, 2014, 123165
37. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. and Kolpakov, F. (2019) GTRD: a database on gene transcription regulation-2019 update. Nucleic Acids Res., 47, D100–D105
38. Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. USA, 112, E6456–E6465
39. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L. A. (2016) Formation of chromosomal domains by loop extrusion. Cell Rep., 15, 2038–2049
40. Yusufzai, T. M., Tagami, H., Nakatani, Y. and Felsenfeld, G. (2004) CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. Mol. Cell, 13, 291–298
41. Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., *et al.* (2017) YY1 is a structural regulator of enhancer-promoter loops. Cell, 171, 1573–1588.e28
42. Shrinivas, K., Sabari, B. R., Coffey, E. L., Klein, I. A., Boija, A., Zamudio, A. V., Schuijers, J., Hannett, N. M., Sharp, P. A., Young, R. A., *et al.* (2019) Enhancer features that drive formation of transcriptional condensates. Mol. Cell, 75, 549–561.e7
43. Duran-Aniotz, C. and Hetz, C. (2016) Glucose metabolism: A sweet relief of Alzheimer's disease. Curr. Biol., 26, R806–R809
44. Di Paolo, G. and Kim, T.-W. (2011) Linking lipids to Alzheimer's disease: cholesterol and beyond. Nat. Rev. Neurosci., 12, 284–296
45. Tynkkynen, J., Chouraki, V., van der Lee, S. J., Hernesniemi, J., Yang, Q., Li, S., Beiser, A., Larson, M. G., Sääksjärvi, K., Shipley, M. J., *et al.* (2018) Association of branched-chain amino acids and other circulating metabolites with risk of incident dementia and Alzheimer's disease: A prospective study in eight cohorts. Alzheimers Dement., 14, 723–733
46. MahmoudianDehkordi, S., Arnold, M., Nho, K., Ahmad, S., Jia, W., Xie, G., Louie, G., Kueider-Paisley, A., Moseley, M. A., Thompson, J. W., *et al.* (2019) Altered bile acid profile associates with cognitive impairment in Alzheimer's disease—An emerging role for gut microbiome. Alzheimers Dement., 15, 76–92
47. Lu, L., Liu, X., Huang, W.-K., Giusti-Rodríguez, P., Cui, J., Zhang, S., Xu, W., Wen, Z., Ma, S., Rosen, J. D., *et al.* (2020) Robust Hi-C maps of enhancer-promoter interactions reveal the function of non-coding genome in neural development and diseases. Mol. Cell, 79, 521–534.e15
48. Qi, Y. and Zhang, B. (2019) Predicting three-dimensional genome organization with chromatin states. PLOS Comput. Biol., 15, e1007024
49. Zhang, X., Jeong, M., Huang, X., Wang, X. Q., Wang, X., Zhou,

W., Shamim, M. S., Gore, H., Himadewi, P., Liu, Y., *et al.* (2020) Large DNA methylation nadirs anchor chromatin loops maintaining hematopoietic stem cell identity. Mol. Cell, 78, 506–521.e6

50. David, Wang, X.Q., Gore, H., Himadewi, P., Feng, F., Yang, L., Zhou, W., Liu, Y., Wang, X., Chen, C-w., Su, J., *et al.* (2020) Three-dimensional regulation of *HOXA* cluster genes by a *cis*-element in hematopoietic stem cell and leukemia. bioRxiv, 017533
51. Cai, Y., Zhang, Y., Loh, Y. P., Tng, J. Q., Lim, M. C., Cao, Z., Raju, A., Li, S., Manikandan, L., Tergaonkar, V., *et al.* (2020) H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. bioRxiv, 684712
52. Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J. and Barillot, E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol., 16, 259–269
53. Xie, W. J., Meng, L., Liu, S., Zhang, L., Cai, X. and Gao, Y. Q. (2017) Structural modeling of chromatin integrates genome features and reveals chromosome folding principle. Sci. Rep., 7, 2818–2828

