D² Plot, a Matrix of DNA Density and Distance to Periphery, Reveals Functional Genome Regions

Yizhuo Che, Xiaofei Yang, Peng Jia, Tingjie Wang, Dan Xu, Tianxue Guo, and Kai Ye*

The execution of biological activities inside space-limited cell nuclei requires sophisticated organization. Current studies on the 3D genome focus on chromatin interactions and local structures, e.g., topologically associating domains (TADs). In this study, two global physical properties: DNA density and distance to nuclear periphery (DisTP), are introduced and a 2D matrix, D² plot, is constructed for mapping genetic and epigenetic markers. Distinct patterns of functional markers on the D² plot, indicating its ability to compartmentalize functional genome regions, are observed. Furthermore, enrichments of transcription-related markers are concatenated into a cross-species transcriptional activation model, where the nucleus is divided into four areas: active, intermediate, repress and histone, and repress and repeat. Based on the trajectories of the genomic regions on D² plot, the constantly active and newly activated genes are successfully identified during olfactory sensory neuron maturation. The analysis reveals that the D² plot effectively categorizes functional regions and provides a universal and transcription-related measurement for the 3D genome.

1. Introduction

DNA molecules are highly condensed inside cell nuclei to compact into limited space, with the necessary loci exposed for active transcription.¹ The rapid development of 3C-based technologies offers insight into the sophisticated organization of the genome, allowing for the discovery of functional nuclear structures, for example, TADs, loops, and compartments.² Nevertheless, most of these interaction-derived structures focus on local spatial organizations, lacking a universal ability to compare distant regions. Although compartments are genome-wise structures, they fail to correlate a given region to a specific biological function without additional functional data.³ While some researchers use interdisciplinary approaches to delineate the chromatin structure,⁴-⁶ their methods continue to focus on interactions. Therefore, universal measurement for a 3D genome with functional correlation is required to elucidate the structural basis of genome function.

DNA density (hereafter referred to as density) and distance to nuclear periphery (DisTP) are two primary physical properties that are related to transcriptional activities in genome spatial organization. A negative correlation between density and transcriptional activity has been implied by several nuclear structures in previous studies.⁶-⁸ Specifically, the densely packed heterochromatin is less accessible and transcriptionally repressed,⁶-⁸ whereas euchromatin, which is less condensed, exhibits more active transcription.⁶

Y. Che, P. Jia, T. Wang, T. Guo, K. Ye
School of Automation Science and Engineering
Faculty of Electronic and Information Engineering
Xi’an Jiaotong University
Xi’an, Shaanxi 710049, China
E-mail: kaiye@xjtu.edu.cn

Y. Che, X. Yang, P. Jia, T. Wang, T. Guo, K. Ye
MOE Key Lab for Intelligent Networks and Networks Security
Faculty of Electronic and Information Engineering
Xi’an Jiaotong University
Xi’an, Shaanxi 710049, China

X. Yang
School of Computer Science and Technology
Faculty of Electronic and Information Engineering
Xi’an Jiaotong University
Xi’an, Shaanxi 710049, China

D. Xu
Key Laboratory of Biomedical Information Engineering of the Ministry of Education
School of Life Sciences and Technology
Xi’an Jiaotong University
Xi’an, Shaanxi 710049, China

K. Ye
School of Life Science and Technology
Xi’an Jiaotong University
Xi’an, Shaanxi 710049, China

K. Ye
Faculty of Science
Leiden University
Leiden 2300, The Netherlands

K. Ye
Genome Institute
The First Affiliated Hospital of Xi’an Jiaotong University
Xi’an, Shaanxi 710049, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202202149

© 2022 The Authors. Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202202149
Several studies have shown that DisTP is positively correlated with transcriptional activity. The lamina-associated domains (LADs), located at the nuclear periphery, are believed to repress gene expression. Experimental evidence has reinforced the idea that reduced proximity to the nuclear lamina is usually accompanied with higher transcriptional activity, while increased proximity to the nuclear lamina results in more repressed transcription. Previous works have clustered the genomic segments into groups with different DisTP (1 Mb resolution), either according to the results of experimentation or those of simulation. By computing the correlation with genetic markers, such as histone modifications, they found that the inner groups were more likely to enrich for active markers compared with the peripheral groups.

Owing to the lack of appropriate techniques for detecting the physical properties in genome-wide scale with high resolution, the extent to which density or DisTP are connected to transcriptional activities, and the analysis comprising the properties for the comprehensive description of genome structure and function are unclear. We developed an algorithm, D², to compute the two properties: density and DisTP, and integrated them as a 2D matrix, D² plot, for mapping functional markers at the genome-wide scale. We identified distinct enrichment patterns of functional markers on the D² plot and proposed a cross-species transcriptional activation model on D² plot. Our study revealed a novel analysis procedure for correlating nuclear organization with gene transcriptional activities and provided a new perspective for the study of the 3D genome structure.

2. Results

2.1. Detection of DNA Density and DisTP

D² algorithm computed Density and DisTP at the genome-wide scale in high resolution (Figure 1A; the Experimental Section). The input of the D² algorithm was a set of DNA particles, which were reconstructed from the single-cell Hi-C contact map by Dip-C or NucDynamics (Table S1, Supporting Information; the Experimental Section). Each particle had its own spatial coordinate and genomic position, thus linking spatial structure to genomic sequence. First, we introduced a 3D mesh segmentation to assign the input DNA particles into different cubes. The cube length was adjusted so that the average number of particles for each cube was approximately equal among cells, thereby allowing for intercell comparison. Density was calculated as the number of particles in one cube and then smoothed to reduce noise. To obtain DisTP, we first identified the membrane cubes (cubes located at the nuclear periphery), and then calculated DisTP for each cube as its distance to the nearest membrane cubes (Figure S1, Supporting Information). Both properties were consistent across different reconstructed replicates of a given individual cell (Figure S2, Supporting Information), indicating that D² was robust against the randomness introduced by the reconstruction algorithm. Therefore, we selected the first reconstruction of each cell for the following analysis. Serial sections of density and DisTP of a sample cell were illustrated in Figure 1B.
We then validated the results of D² with experimental evidence. LADs were the conspicuous nucleic structures that were mainly located at the nuclear periphery. We found a significantly high correlation between the probability of appearing at the nuclear periphery (low DisTP) and the experimental LAD data, thereby verifying the fitness of DisTP (Figure 2A; Pearson coefficient correlation (PCC): 72.87%, p-value: 4.8 × 10⁻²¹¹). Rods of nocturnal animals adopted an inverted radial organization, in which euchromatin resided at periphery and heterochromatin preferred the interior. Inverted distributions between the computed DisTP of rods and that of other cell types were observed (Figure 2B), indicating that our algorithm is able to capture diverse configurations of DisTP from dramatically different genome structures. Owing to the lack of experimental measurement for density, we attempted to validate this through highly compacted chromocenters located at the inner nuclei of mouse cells but not in human cells (Figure S3, Supporting Information). We found that high-density regions preferentially located at the inner nuclei of mouse cells but this was not the case in human cells (Figure 2C), indicating the feasibility of density detection.

We then investigated the correlation between density and DisTP. In human cells, density and DisTP were largely independent of each other (PCC: −0.019 and −0.103; Figure 3A,B; Table S2, Supporting Information). By contrast, they were positively correlated in mouse cells (PCC: 0.278 to 0.673, Figure 3A,B; Table S2, Supporting Information). This interspecies difference indicated that human and mouse cells adapted different strategies for nuclear spatial organization.

Next, we explored the stochasticity of density and DisTP by computing their standard deviations (SDs; Figure S4, Supporting Information; the Experimental Section). SDs of density (≈1.5) and DisTP (≈4.7) were comparable to their means (≈2.2 and ≈6.8, Table S3, Supporting Information), indicating a strong intrinsic stochasticity that was consistent with previous observations. In addition, certain regions with specific density or DisTP had lower standard deviation (SD). Specifically, for DisTP, the periphery regions showed lower SD for both human and mouse cells (Figure 3C; Figure S5A, Supporting Information), thereby coinciding with their roles as anchors for attachment to the nuclear membrane. For density, the lower SD regions exhibited noticeable cell-type specificity. For most cell types (11 out of 17), high-density regions were lower in SD, while for the other six mouse neuron cell types, low-density regions had lower SD (Figure 3; Figures S5B and S6, Supporting Information), suggesting a cell-type-specific anchoring mechanism for density.

2.2. Distinct Patterns of Functional Markers on D² Plot Revealed Functional Genomic Regions

The spatial organization of eukaryotic cell nuclei plays a pivotal role in regulating the biological activities of DNA molecules. However, a comprehensive and systematic analysis of the correlation of density and DisTP with chromatin activities was lacking. To achieve that, we collected different types of epigenomic and genomic data from human GM12878 cells and mouse neonatal forebrain neurons, including DNA methylations, histone modifications, transcription factor binding sites, genomic repeats, and chromatin states.

For joint analysis, we constructed a 2D matrix of density (x-axis) and DisTP (y-axis), named D² plot (the Experimental Section). D² plot was composed of physical states (the Experimental
Figure 3. Strong intrinsic stochasticity of density and DisTP. A) Scatter and kernel density estimate (KDE) plots of density and DisTP. The gray line delineates the fitting line of linear regression. The correlations are computed by Pearson correlation. Detailed information for the additional 15 cell types is shown in Table S2 of the Supporting Information. B) Boxplot of PCC between density and DisTP for human and mouse cells. n indicates the number of cell types. C) 2D histogram and KDE plots of mean and SD for DisTP. One human and two mouse cell lines are shown here, and additional 14 cell types are shown in Figure S5 of the Supporting Information. D) Similar to (C) but for density.

Section), the rectangular bins with specific values of density and DisTP (Figure 4A; Table S4, Supporting Information). We only kept the states with enough genomic segments for statistical significance. Genetic markers were then mapped to the D2 plot according to their enrichments on genomic segments. Enrichment scores of each marker at each physical state were calculated, compared to the whole genome as the background (the Experimental Section). The enrichment scores of all markers are shown in Figure 4 and the figures in the Supporting Information.

We first examined the distribution of active markers on D2 plot. We found that H3K4me3, a well-established marker for transcriptional activation, enriched at the center rather than at the periphery of the nuclei, which was consistent with previous microscopic observations (Figure 4B; Figure S7, Supporting Information). Regions with enriched H3K4me3 distribution showed lower density (Figure 4), which confirmed that transcriptional regulators around active genes lowered the local DNA density. Other active histone markers and chromatin states, and even the poised or weakly transcribed chromatin states exhibited similar patterns (Figure S8, Supporting Information). The coexistence of active and weak transcriptional markers implied that the inner low-density areas were not a determining factor for transcriptional activity, but rather provided a hospitable environment, allowing for the recruitment of regulating proteins for
Figure 4. Enrichment scores of genetic and epigenetic markers on D² plot. A) 2D histogram and KDE plots of density and DisTP in human GM12878 cells and mouse neonatal forebrain neurons. The top and right subpanels show the 1D histograms of density (top) and DisTP (right). Yellow lines mark the boundaries to exclude outliers from downstream analysis. The boundaries of 16 cell types are shown in Table S4 of the Supporting Information. B) Enrichment scores of H3K4me3 marker of human and mouse cells on D² plot. The enrichment score for the bottom right physical state in mouse
transcriptional activation, and thereby reinforcing the previously proposed “permissive role” of spatial organization.\cite{24}

Next, we focused on the facultative and constitutive heterochromatin with repressed transcription, as marked by the trimethylation of lysine 27 on histone H3 (H3K27me3) and trimethylation of lysine 9 on histone H3 (H3K9me3), respectively.\cite{25} On the D\textsuperscript{2} plot, H3K27me3 was more likely to occupy the condensed regions inside the nuclei, while H3K9me3 showed an obvious preference for nuclear periphery (Figure 4C), as confirmed by the immunofluorescence experiments (Figures S9 and S10, Supporting Information).\cite{26, 27} We also found that H3K9me3 enriched at inner high-density regions in mouse cells but not in human cells, suggesting the species-specific features of heterochromatin arrangements.\cite{18, 19}

We then investigated the repetitive elements on the D\textsuperscript{2} plot. We found that long repeats, including long terminal repeats and long interspersed nuclear elements (for example, L1), preferred the inner high-density regions or periphery regions (Figure 4D; Figure S11, Supporting Information). Contrarily, the short interspersed nuclear elements (including Alu, B2, and ID) preferably enriched at inner low-density regions (Figure 4; Figure S10, Supporting Information), which was consistent with the FISH analysis performed by Lu et al., who found that Alu clustered at the nuclear interior, while L1 clustered at the nuclear periphery.\cite{28} The distributions of Alu and L1 were significantly negatively correlated in both species (Figure S12, Supporting Information), indicating a competitive relationship between the two transposons in spatial organization.

Next, we examined the distribution of the architectural proteins, including RAD21, SMC3 (both subunits of cohesin), CTCF, and YY1. They are believed to cooperate in the construction of genomic structures via the use of a previously proposed loop extrusion model.\cite{29} The patterns of these proteins largely overlapped on the D\textsuperscript{2} plot at the inner region of the nuclei (Figure 4E; Figure S13, Supporting Information). Nonetheless, YY1 and CTCF were more likely to enrich at the lower-density regions than cohesin, indicating that YY1 and CTCF preferred to locate at the active compartment. Consistently, the fold changes of YY1 (1.71) and CTCF (1.23) enrichments in compartment A to B were higher than those of RAD21 (1.13) and SMC3 (1.15) (Figure S14, Supporting Information). The differences in enrichments implied that, in addition to working together, these proteins might function individually, as previously observed by Wei et al.\cite{30}

2.3. Cross-Species Transcriptional Activation Model on D\textsuperscript{2} Plot

To link transcription activity to the D\textsuperscript{2} plot, we selected 16 transcription-related markers and performed the hierarchy cluster on physical states (the Experimental Section). As shown in Figure 5A, we successfully distinguished the active states from the repressed ones. However, there were still intermediate states marked by both active markers and repressive histone modifications, corresponding to the “bivalent” genomic regions where promoters were prepared for rapid activation during differentiation.\cite{31}

Based on the hierarchy clusters for both human and mouse cells, we divided the physical states into four groups: active, intermediate, repress and histone (repressed states marked by repressed histone modifications), and repress and repeat (repressed states marked by long repeats) areas. Then we assigned these four areas back to the D\textsuperscript{2} plot to reveal their transcriptional patterns (Figure 5B). The patterns between human GM12878 cells and mouse neonatal forebrain neurons showed slight differences. For example, by computing the property strength (Figure S15, Supporting Information; the Experimental Section), we found that DisTP had less impact on the nuclei interior in human cells compared to mouse cells, probably owing to the different strategies of heterochromatin organization between the two species.\cite{18} Other than the marginal difference, the patterns were highly similar, out of which we proposed a cross-species transcriptional activation model on the D\textsuperscript{2} plot (Figure 5C). The model showed that the most inactive area, repress and repetitive, located at the periphery regardless of density. At the inner nucleus, however, density and DisTP jointly affected gene regulation. From inactive to active areas (repress and histone, intermediate to active area), density decreased while DisTP increased. This model is largely consistent with the current understanding of density and DisTP, but provides additional functional annotation.

To verify the proposed model, we projected the lineage-specific active genes of mouse brain cells onto the D\textsuperscript{2} plot and found that those active genes located at more interior regions with lower density in the corresponding cell types compared with other cell types (Figure S16, Supporting Information). Next, we focused on the genes that were previously defined as either newly activated or repressed during maturation from oligodendrocyte progenitors to mature oligodendrocytes.\cite{32} As expected, the newly activated genes distributed almost randomly on the D\textsuperscript{2} plot of progenitors, but enriched at inner low-density regions after the maturation (Figure 5D), corresponding to the active and intermediate areas. Conversely, the newly repressed genes located at inner low-density regions in progenitors but distributed randomly in mature cells (Figure 5E).

Tan et al. found no explanation for the adult-neuron-specific genes maintaining the same radial position during differentiation, despite the newly activation of these genes.\cite{31} We revealed that DisTP values of these genes remained unchanged (Fisher Test p-value: 0.68), but density values declined significantly (Fisher Test p-value: $3.34 \times 10^{-18}$; Figure S17, Supporting Information), demonstrating that although DisTP alone was not sufficient for predicting gene regulation, its combination with DNA density was instructive.
Figure 5. Cross-species transcriptional activation model on D² plot. A) Hierarchy clustering of D² plot physical states (n = 225 for human, n = 224 for mouse due to one state with fewer than 200 genomic segments) based on the 16 transcription related markers. Using the optimal ranking hierarchy cluster, the states are ranked automatically according to their activation levels. The number of clusters is set as 4 for both species. The black dashed lines mark the boundaries of the clusters. The abbreviations have been spelled out in the legends of Figures S7 and S10 of the Supporting Information. B) The four areas on each D² plot for either human GM12878 cells or mouse neonatal neurons, corresponding to (A). C) Illustration of the transcriptional activation model on the D² plot. The arrows point to more active regions. D) The enrichment scores of newly activated genes on the D² plots of oligodendrocyte progenitors and mature oligodendrocytes. The enrichment scores of blank physical states are not shown owing to the small number (n < 200) of genomic bins. E) Similar to (D) but with newly repressed genes.
2.4. Trajectories on D² Plots Revealed Transcriptional Activation Modes

After establishing the correlation between the transcriptional level and the D² plot, we aimed to identify the transcriptional patterns of genomic regions merely by their trajectories on the D² plot. To do this, the activation index was introduced as the probability of a given genomic region appearing at the active or intermediate areas, and changes in its value reflected the trajectories on the D² plot during cell maturation (Figure S18, Supporting Information; the Experimental Section). We examined our approach on olfactory sensory neurons (OSNs), which included three cell types at different differentiating stages (newborn, developing, and mature OSNs). By investigating the trajectories of the whole genome on the D² plot, we identified three transcriptional modes: constantly active (CA), repressed to active (RA), and constantly repressed (CR), which revealed the dynamic regulation of gene expression during OSN development.

Next, we investigated the enrichments of functional elements in the three modes. Compared to the background, the percentages of exons in CA and RA were remarkably higher, while that of CR was lower (Figure 6A), in accordance with their expression levels. In stark contrast, the percentages of L1 in CA and RA were lower while that of CR was higher. As expected, CA was enriched with foundational gene pathways (ribosome, translation, and lipid transport), whereas RA was enriched with olfactory receptor (OR)-related pathways (Figure 6B). No significantly enriched pathways (FDR < 0.01) were identified for CR.

We then spotted an ≈3 Mb genomic segment (chromosome 10: 126.16–129.28 Mb), harboring 19 housekeeping (HK) genes (mainly located at CA), 10 ORs (at RA), and 129 L1 repeats (at CR) (Figure 6C). The activation indexes of HK genes stabilized at high levels across the three cell types, while those of L1 repeats remained at low levels (Figure 6D). On the D² plot, the housekeeping genes persistently enriched at the inner low-density regions, while the L1 repeats enriched at the periphery (Figure 6E; Figure S19, Supporting Information), corresponding to their constitutive activation or repression, respectively.

ORs were a large gene family that expressed at mature OSNs for sensing odors.[33] Correspondingly, their activation indexes increased during development (Figure 6D), and they moved on the D² plot from the periphery (newborn OSNs), to inner high-density (developing OSNs), and then toward lower density (mature OSNs), in consonance with their activation process.
(Figure 6E; Figure S18, Supporting Information). Additionally, for each individual OSN, only one specific OR expressed while all the others were repressed, which would leave featured distributions on D2 plot. Our D2 plot demonstrated that ORs collectively enriched inside nuclei with varying densities, corresponding to the active, intermediate, and repress and histone areas (Figure S20, Supporting Information), and reflecting their diverse levels of activation at the bulk level.

3. Conclusion

Inside the cell nuclei, the genome is compartmentalized into regions with different physical properties, to effectively execute their biological functions. Based on the reconstructed genome structures from single-cell Hi-C, we developed the D2 algorithm to define density and DisTP at a genome-wide scale and integrated them into a two-dimensional D2 plot. Different genetic and epigenetic markers exhibited distinct enrichment patterns on the D2 plot, indicating that these two physical properties, density and DisTP, were two pivotal parameters for identifying the functional compartmentalization of the genome structure. For instance, transcription-related enzymes and active genomic sites altogether clustered at inner low-density regions, whereas repressed heterochromatin markers enriched at inner high-density or periphery regions.

In this study, we investigated the enrichments of architectural proteins (CTCF, YY1, and cohesin) on the D2 plot and revealed their largely overlapping enrichment zones, correlating to their cooperative functions. In addition, we spotted cohesin-independent enrichments of YY1 and CTCF at inner low-density regions, hinting their independent functions. Furthermore, all the architectural proteins located inside the nuclei, suggesting alternative mechanisms for genome construction at the nuclear periphery. L1 repeats, which were previously reported to promote heterochromatin compartmentalization, were found to mainly locate at the nucleic periphery, implying their possible involvement in maintaining genome construction.

In this study, we proposed a cross-species transcriptional activation model on the D2 plot among different human and mouse cells. This model revealed that the distance to the periphery was negatively correlated with the transcriptional activities of genes, which seemingly contrasted to the center-positioned nucleoli that were mainly composed of heterochromatin. The significantly low genome coverage (<5%) of nucleolus-associated domains might render these regions neglected in our model. To examine the fitness of our model to reflect gene activation, mouse forebrain neurons, is capable of linking the structure with gene expression during OSN development, verifying its broad fitness and feasibility. Nonetheless, further verification and application of this model in other species is required, but these are currently unavailable owing to the lack of high-quality single-cell Hi-C data. Our model would be applicable to most species, where sophisticated Mb-scale spatial organization is essential for regulating gene expression. It might be helpful to elucidate human tumorigenesis analysis, which usually involves dramatic alteration of the genome structure.

Our study provides a new perspective for 3D genome analysis. Previous studies mainly focused on well-known structures such as loops or TADs, which only revealed the spatial adjacency of focal structures. Contrarily, our D2 model allows for universal measurements of the genome structure and correlates the genomic structure with its function. Nonetheless, we speculate that physical properties of DNA act more as microenvironments to facilitate or repress gene transcription activity rather than direct and decisive factors, as protein structure to its function. In summary, our study paves the way for further methodological exploration to link genome structure with its function.

4. Experimental Section

D2 for Detecting Density and DisTP: Python script D2.py implemented the D2 algorithm, which rapidly and accurately computed the DNA density and DisTP values based on the reconstructed genome structure. D2 algorithm can be found at https://github.com/xjtu-omics/D2.

Inputted Reconstructed Structures: Reconstructed genome structures, which were generated from single-cell Hi-C, were the main input for the D2 algorithm. Genome structures from two research groups were collected in this study (Table S1, Supporting Information). Stevens et al. processed single-cell Hi-C on eight individual haploid mouse ES cells and then reconstructed the nuclear structures, using the NucDynamics algorithm. The algorithm was designed based on molecular dynamic simulation. First, the genome was divided into 100 kb genomic bins, each of which represents a DNA particle in 3D structure. Then, a structure was constructed using the physical force field, including attraction, repel, and restriction. Using the simulated annealing algorithm, the structures were constructed to optimally simulate actual structures. The other group, Tan et al. developed a novel chromatin conformation capture method, Dip-C, which can detect more contacts than single-cell Hi-C. They omitted biotin pull down and conducted high-coverage whole-genome amplification (META). Then, they used a genome reconstruction algorithm, hickit, which was similar to NucDynamics and enabled the construction of diploid cells at a relatively higher resolution (20 kb).

Partition Particles by 3D Segmentation: Owing to the heterogeneity of nuclei sizes among cells, the reconstructed structures were not directly comparable among cells. This batch effect resulted in different nucleic radii and DNA densities. To address this, the DNA particles were placed inside a 3D mesh segmentation with equally shaped cubes, with cell-specific cube length. The length was set as the average distance to the nearest ith particle of all particles from a given cell (i is the predetermined parameter, default as 7), to ensure the average number of DNA particles within a cube was approximately the same across cells. This predetermined parameter was not sensitive for further analysis, because the following analysis concentrated on the differences between the density and DisTP of genomic regions (e.g., lower-density regions), while the change in this parameter increased or decreased density or DisTP universally. The segmentation was constructed according to the cube length and covered the particles. Then the spatial coordinate of particles was replaced using the cube coordinate. This normalization was based purely on DNA density, but the nucleic radius (half of the maximum distance among filled cubes) was adequately normalized (Figure S21, Supporting Information). Furthermore, this partition reduced the number of tasks from particles to cubes, thus efficiently speeding up and saving memory during calculation.
Computing DNA Density: Once segmentation was performed, DNA density could simply be defined as the number of particles per cube. However, the arbitrary partition above could result in an inaccurate density. For example, if a DNA particle resided in the transiting regions between compacted and loose regions, it should be labeled as middle density. However, it may be partitioned into high-density or low-density cubes, instead of a middle one. To solve this problem, the density of a given cube was smoothed by its nearby cubes through the following equation:

\[ D_i = \frac{\sum_{s=5}^{15} C_j / \| x - y \|}{\text{num}(S)} \]  

(1)

where \( D_i \) is smoothed DNA density of the spatial cube \( i \), \( C_j \) is the DNA particle number within the cube \( i \), and \( S \) is the set of nearby cubes. The nearby range was set as 3. A longer range would be more optimized for calculation, but it was unnecessary as distant cubes contributed little.

Computing Distance to Periphery: Previous methods for detecting DisTP were simple and, to a certain extent, inaccurate (Note S2, Supporting Information). One of them was developed in 2017 by Stevens et al. [16]. It first identified the empty cubes around the nonempty cubes as the surface. Then, they calculated the DisTP as the nearest distance to the surface. However, imaging experiments showed that chromocenters would cause they were discrete values and highly sensitive. For example, if one distance to membrance cubes. The minimum distances were not used because they were discrete values and highly sensitive. For example, if one inner cube was incorrectly defined as a membrane cube, the cubes around it would be defined as periphery.

D^2 Plot: Construction of D^2 Plot: To combine density and DisTP for further systematic analysis, a 2D density–DisTP matrix, the D^2 plot was constructed. Outlier bins were eliminated during plotting. For the density axis, the distribution resembled the normal distribution. Therefore, the 1.96 \( \times SD \) was selected from the center as the separating line, to retain 95% of the center bins. For the DisTP axis, all the bins were maintained near the periphery, but eliminated the top 5% bins further away from the periphery. The ranges were similar among the diploid cell types, even among the species (Table S4, Supporting Information). Therefore, a standard range was set for all diploid cells (1 to 3.2 for density, 1.21 to 16 for DisTP). Each axis was then divided into 15 equally shaped bins. The bin number was set arbitrarily, but it was shown that different bin numbers did not affect the enrichment patterns (Figure S22, Supporting Information). Each 2D bin at the D^2 plot represented a specific physical state with a unique combination of density and DisTP. Notably, states with fewer than 200 particles were removed from subsequent analysis to ensure statistical significance (marked as empty).

Computing Marker Enrichment Scores: To calculate the marker enrichment scores on the D^2 plot, genetic data were weighted and averaged by the probability of the genomic segment appearing at the physical state, as shown below:

\[ E_a = \frac{\sum_{i=5}^{15} P_{ga} \cdot NV_a}{\sum_{i=5}^{15} P_{ga}} \]  

(2)

where \( E_a \) is the computed enrichment score of the genomic marker in the physical state, \( a \). NV_a is the z-scored normalized enrichment score of a genetic marker at the genome segment, \( g \). P_{ga} is the probability of the genome segment \( g \) appearing in the state \( a \), defined as the ratio of the number of cells whose genomic segment \( g \) located at the state \( a \) with respect to the total cell number. Afterward, the enrichment scores were normalized by z-score normalization.

Through this procedure, the density and DisTP of single cells were determined. There was another way of computing the enrichment scores, where the density and DisTP were averaged first

\[ E'_a = \frac{\sum_{i=5}^{15} NV_a}{N(G_a)} \]  

(3)

where \( E'_a \) is the alternative enrichment score derived from the average density and DisTP at physical state, \( a \). G_a is the collection of genomic segments whose average density and DisTP were located at state \( a \). N(G_a) is the number of collections. The average values delineated the bulk activities of a cell type more optimally, but were unable to represent the raw density and DisTP detected from the single cell. For example, the average values could not reflect the sparse distribution, for example, the multimodal distribution.

To test if the average density and DisTP would give similar enrichment patterns, the dip test was first carried out for unimodal distribution. All genomic bins were unimodal for both density and DisTP (Table S3, Supporting Information, percentage of both p-value < 0.01: 98.11–99.88%). Then, the enrichment score analysis with average density and DisTP was repeated and it was found that it exhibited similar patterns to Figure 4 (Figure S23, Supporting Information), indicating that the average density and DisTP equivalently reflect biological activity.

Stochasticity of Marker Enrichment Score: Considering the high cell stochasticity of density and DisTP, the stochasticity of marker enrichment scores was examined by computing their SD using the following equation:

\[ SD_a = \sqrt{\frac{\sum_{i=5}^{15} P_{ga} \cdot (NV_a - E_a)^2}{\sum_{i=5}^{15} P_{ga}}} \]  

(4)

where \( SD_a \) is the calculated SD of the marker enrichment score at physical state \( a \). NV_a, P_{ga}, and G are defined as above. To compare SD with the mean enrichment scores, they were presented together at Figure S24 of the Supporting Information. The max–min normalization on mean values was utilized for comparison between markers. This normalization was also used on SD values. Therefore, markers with SD less than 0 denoted low stochasticity, while others with greater than 1 denoted high stochasticity. Certain markers (e.g., H3K27me3 and H3K9me3) showed lower SD than the mean value, indicating their consistent enrichment among cells. Nonetheless, most markers showed a comparable or larger SD, indicating high stochasticity. To test the robustness of the enrichment score analysis under such stochasticity, the analysis was repeated using only one cell (Figure S25, Supporting Information). The enrichment patterns, even from one cell, were highly similar to Figure 3, suggesting a variable despite the robust correlation between physical properties and genetic markers.

Hierarchical Cluster and Property Strength: To obtain a transcriptional pattern on the D^2 plot, 16 transcription-related markers (Tsxn-St, Ttxn-W, PrAc, Pr-W, En-St, H3K27ac, H3K36me3, H3K4me3, H3K4me2, Alu, H3K9me3, H3K27me3, L1, Ht-L; the abbreviations are spelled out in the legend of Figures S7 and S10, Supporting Information) were selected and the hierarchy cluster (scipy.cluster.hierarchy with parameter optimal_ordering = True, method = “ward”) was performed on physical states from the D^2 plot. The hierarchy cluster number was set as 4 for both species.

The property strength reflected the intensity of impact or DisTP on the transcription level. The active markers were selected and the property strength was computed at a given DisTP as the slope between property changes (\( x \)) and marker enrichment score (\( y \)), as follows

\[ S_{d,p} = \text{slope of linear fitting} (x = MFG_{d,p} \ y = P_{d,p}) \]  

(5)
where \( S_{d,p} \) is the computed property strength of property \( p \) at DisTP \( d \). For density, the \( x \) input for linear fitting is the collection of all densities at a given DisTP and \( y \) is the corresponding marker enrichment score. For DisTP, \( x \) is the collection of the given DisTP and its nearby DisTP, and \( y \) is the corresponding enrichment score. The property strength reflected the association between the physical properties and the transcription level.

**Distribution of Lineage-Specific Active Genes**: The distribution of active genes was examined on the \( D^2 \) plot, to validate the transcriptional activation model. Considering the nonuniform distribution of whole genome bins on the \( D^2 \) plot, the enrichment scores of active genes over the background were computed. Specifically, the distribution of active genes was first calculated, and then it was divided by the distribution of all genomic bins as the background.

**Trajectories on \( D^2 \) Plots**: The activation index and its change were introduced to reveal the trajectories of a given genomic region on the \( D^2 \) plot, which reflected its transcriptional activation modes (Figure S18, Supporting Information). Specifically, the activation index was defined as the sum of the enrichment scores of a given genomic bin on active or intermediate regions. Based on the activation index, the top 10% genomic bins were marked as the active bins, the bottom 10% as the repressed bins, and the others as medium. Three transcriptional modes: constantly active (CA), constantly repressed (CR), repressed to active (RA), were identified. During OSN maturation, the bins, which were active at all OSN cell stages, were marked as the CA bins. Conversely, the bins repressed at all stages were marked as CR bins. RA bins comprised bins that were not active at newborn OSNs, but active at mature OSNs.

The genomic positions of exons were obtained from the gff file, and those of L1s were obtained from RepeatMasker. Housekeeping genes were obtained from the HRT Altas v1.1 database.\(^{[35]}\) Their percentage of length at the given genomic bins was calculated.

The gene pathway enrichment analysis was performed by DAVID.\(^{[36]}\)

**Only KEGG pathways and GO terms were chosen for enrichment.**

**Reference Genome**: The mouse reference genome (GRCm38) and gene annotations (ALL) were downloaded from the GENCODE (https://www.gencodegenes.org/mouse/) M19 release.

The human reference genome (GRCh37) and gene annotations (ALL) were downloaded from the GENCODE (https://www.gencodegenes.org/human/) 19 release.

**Statistical Analysis**: All of the Pearson correlations were two-sided and were computed using scipy.stats.pearsonr. The linear fitting was conducted using scipy.stats.linregress, with parameter deg = 1. The sample size for each statistical analysis is available in the figure legends or in the Supporting Information. Fisher tests were two-sided and computed by scipy.stats.fisher_exact.

**Preprocess of Genetic Markers**: To integrate density and DisTP with other genomic data, the other data were transferred into the same indexed data. All data (except DamID-LmnB1 data of mESC) were indexed by 20 kb genomic bins. For unphased data, the data were copied to two haplotypes.

**ChiP-seq Data**: ChiP-seq sequenced protein binding sites. ChiP-seq data of transcription factors (CTCF, YY1, RAD21, SMC3, POLR2A, POLR2Ap5 for human GM12878 cells) and histone modifications (H3K4me2, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3 for both human GM12878 cells and mouse neonatal neurons) were collected.

These data with different types of processed files, were collected from ENCODE. Here the fold change was utilized over the control file, because it provides whole genome protein binding information. This file was formatted as BigWig and was transformed to a bedgraph file using the UCSC script bigWigToBedGraph. Then bedtools intersect were used to find the intersected regions between the index file and the marker bedgraph file. Next, the fold change of all 20 kb bins was weight-averaged and they were indexed according to the index file.

**Genomic Repeat**: The repeat data were obtained from RepeatMasker. Owing to the low resolution of the 3D genome, only broadly distributed repeats would be statistically significant for enrichment. Therefore, repeats with less occurrences were filtered out. Only repeats with more than 20,000 occurrences were left. The remaining repeats with the index file were intersected and then their percentage at a given genomic bin was calculated, and finally they were indexed in the aforementioned procedure.

**Chromatin State**: Chromatin states described the transcriptional states of genomic loci. They were calculated using HMM and based on chip-seq data.\(^{[37,38]}\) Because chromatin states were discrete states, intersections, calculated percentages, and indexed were performed through processes similar to those of the repeats. Each genomic bin had a value for each chromatin state, representing the length of the state within that bin.

**CpG**: The CpG data were downloaded from ENCODE, formatted as bedMethyl. Sites were filtered out without reads. The remaining sites were intersected with the index file, the median CpG percentages were calculated in a given bin, and finally they were indexed using the aforementioned procedure.

**DamID-Lmn for Detecting LADs**: DamID-seq identified protein binding sites by expressing the proposed DNA-binding protein as a fusion protein with DNA methyltransferase. Using DamID-seq to detect the binding sites of Lmn-B1 (Lamin-B1), the main component of nuclear lamina, is a popular method of detecting LADs. The mESC LAD experimental data were utilized.\(^{[39]}\) Because these data were discrete like genomic repeats, the data were intersected, calculated, and indexed in the same way as the repeats. Notably, the bin size was 100 kb for the used mES cells.

To validate DisTP, experimental LAD data were compared to the probability of appearing at the nuclear periphery (percentages of cells whose given genomic bin had DisTP below 2). As they were both indexed by the same index file, they were associated with their index. The outliers were eliminated using three-sigma limits.

Noteworthy, due to the small number of available cells (\( n = 8 \)), the possible values for simulated LAD data are too sparse at 100 kb resolution. Therefore, the nearby bins were concatenated to 2 Mb resolution, in order to provide continuous percentages. The number of 2 Mb genomic bins is 1103. The two-sided Pearson correlation was computed using scipy.stats.pearsonr.

**Stochasticity of Density and DisTP**: To explore the stochasticity of DisTP, its standard deviations were computed (SDs; Figure S4, Supporting Information). SD was computed among cells of given genomic bins, thereby reflecting the cell heterogeneity. The lower SD meant a more stable DisTP value among cells.

The stochasticity of density was computed similarly to DisTP. In contrast with DisTP, the lower SD regions of density showed obvious cell-type specificity. Therefore, the Pearson correlation between density mean and density SD (scipy.stats.pearsonr) was conducted, to identify the values of density that were more stable.

**Supporting Information**

Supporting Information is available from the Wiley Online Library or from the author.

**Acknowledgements**

This work is supported by National Science Foundation of China (32105009, 32070663, 62172325, 32070134), by the Fundamental Research Funds for the Central Universities, by the National Science Basic Research Program of Shanxi (20211LB062), by the National Fundamental Research Project of Xi’an (No. 21RSGF0013), by Key Construction Program of the National “985” Project. The authors would like to thank members of Kai Ye’s laboratory for helpful comments to improve the manuscript. Special thanks to Cheng Li and Hebing Chen for helpful discussion.

**Conflict of Interest**

The authors declare no conflict of interest.

**Authors Contribution**

Y.C. developed the \( D^2 \) algorithm with the help of X.Y. Y.C. downloaded, analyzed, and interpreted the data in current studies. P.J. and T.W. provided help for analyzing data. Y.C. was the major contributor in writing the
manuscript. T.G. helped with the benchmark. X.Y., D.X., and K.Y. revised the manuscript. K.Y. supervised the whole program. All authors read and approved the final manuscript.

Data Availability Statement

The reconstructed genome structures analyzed during the current study are available in the GEO repository: GSE117876, GSE94449, GSE121791 and GSE162511. The ChIP-seq data analyzed during the human GM12878 study are available in the ENCODE repository: ENCF364OXN (CTCF), ENCF36592JH (POLR2A), ENCF002UPS (POLR2AP5), ENCF6631VY (Y1Y), ENCF2336X8 (SMC3), ENCF0567EGK (RAD21), ENCF828CQV (H3K4me2), ENCF674QZB (H3K4me3), ENCF776OVW (H3K9me3), ENCF180LKW (H3K27ac), ENCF594HSG (H3K27me3), ENCF662QFK (H3K36me3) and ENCF279HCL (CpG). The ChIP-seq data analyzed during the mouse forebrain tissue study are available in the ENCODE repository: ENCF879HIK (H3K4me2), ENCF0504HA (H3K4me3), ENCF9651FF (H3K9me3), ENCF422JMO (H3K27ac), ENCF742UNH (H3K27me3), ENCF3636EM (H3K36me3). The chromatin states data of human GM12878 are available in GEO GSE26386, while that of mouse (H3K4me2), ENCFF674QZB (H3K4me3), ENCFF776OVW (H3K9me3), ENCFF180LKW (H3K27ac), ENCF594HSG (H3K27me3), ENCF662QFK (H3K36me3) and ENCF279HCL (CpG). The ChIP-seq data analyzed during the mouse forebrain tissue study are available in the ENCODE repository: ENCF879HIK (H3K4me2), ENCF0504HA (H3K4me3), ENCF9651FF (H3K9me3), ENCF422JMO (H3K27ac), ENCF742UNH (H3K27me3), ENCF3636EM (H3K36me3). The chromatin states data of human GM12878 are available in GEO GSE26386, while that of mouse forebrain tissue are available at UCSC Table Browser.

Keywords

3D genome, DNA density, nuclear lamina, single-cell Hi-C

Received: April 13, 2022
Revised: August 9, 2022
Published online: August 30, 2022

[1] B. Bonev, G. Cavalli, Nat. Rev. Genet. 2016, 17, 661.
[2] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, Cell 2014, 159, 1665.
[3] V. Pancaldi, E. Carrillo-de-Santa-Pau, B. M. Javierre, D. Juan, P. Fraser, M. Spivakov, A. Valencia, D. Rico, Genome Biol. 2016, 17, 152.
[4] M. Madrid-Mencia, E. Raineri, T. B. N. Cao, V. Pancaldi, Nucleic Acids Res. 2020, 48, 4066.
[5] V. Pancaldi, Front. Bioeng. Biotechnol. 2021, 1, 742216.
[6] S. I. S. Grewal, S. Jia, Nat. Rev. Genet. 2007, 8, 35.
[7] I. Solovei, K. Thanisch, Y. Feodorova, Curr. Opin. Cell Biol. 2016, 40, 47.
[8] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Lund, A. R. Oldenburg, E. Delbarre, C. T. Freberg, I. Duband-Goulet, R. Eskeland, B. Buendia, P. Collas, Genome Res. 2013, 23, 1580.
[9] L. Tan, D. Xing, C. H. Chang, H. Li, X. S. Xie, Science 2018, 361, 924.
[10] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, O. O’Shaughnessy-Kirwan, J. Cramard, A. D. Mauer, F. M. Beh, E. Blanco, E. Morey, M. Sansó, M. G. S. Palayret, B. Lehner, L. di Croce, A. Wutz, B. Hendrich, D. Kleinermann, E. D. Laue, Nature 2017, 544, 59.
[11] I. Solovei, M. Kreysing, C. Lacntzt, S. Kösem, L. Peichl, T. Cremer, J. Guck, B. Jouffe, Cell 2009, 137, 356.
[12] K. L. Jost, B. Bertulat, M. C. Cardoso, Chromosoma 2012, 121, 555.
[13] Y. Wang, Z. Huang, K. Hu, J. Peng, W. Yao, W. Deng, J. Zuo, Y. Zhang, D. Yin, Acta Biochim. Biophys. Sin. 2020, 53, 124.
[14] H. E. Finn, G. Pegoraro, H. B. Brandão, A. L. Valton, M. E. Oomen, J. Dekker, L. Mirny, T. Misteli, Cell 2019, 176, 1502.
[15] J. Zheng, W. Xie, Nat. Rev. Mol. Cell Biol. 2019, 20, 535.
[16] A. M. Oudelaar, D. R. Higgs, Nat. Rev. Genet. 2021, 22, 154.
[17] A. J. Bannister, T. Kouzarides, Cell Res. 2011, 21, 381.
[18] E. Ing-Simmons, R. Vaid, X. Y. Bingham, M. Levine, M. Mannervik, J. M. Vaquerizas, Nat. Genet. 2021, 53, 487.
[19] J. C. R. Politz, D. Scalzo, M. Groudine, Annu. Rev. Cell Dev. Biol. 2013, 29, 241.
[20] E. J. Platt, L. Smith, M. J. Thayer, J. Cell Biol. 2018, 217, 541.
[21] A. Polishko, C. L. Smith, S. C. Nguyen, P. Sivaramakrishnan, K. G. Wong, J. I. Murray, M. Lakadamally, E. F. Joyce, R. Jain, J. A. Epstein, eLife 2019, 8.
[22] J. Y. Wu, L. Chang, T. Li, T. Wang, Y. Yin, G. Zhan, X. Han, K. Zhang, Y. Tao, M. Perchardé, L. Wang, Q. Peng, P. Yan, H. Zhang, X. Bi, W. Shao, W. Hong, Z. Wu, R. Ma, P. Wang, W. Li, J. Zhang, Z. Chang, Y. Hou, B. Zhu, M. Ramalho-Santos, P. Li, W. Xie, J. Na, Y. Sun, et al., Cell Res. 2021, 31, 613.
[23] Y. Li, J. H. I. Haahrhuis, A. S. Cacciatore, R. Oldenkamp, M. S. van Ruiten, L. Willems, H. Heunissen, K. W. Muir, E. de Wit, B. D. Rowland, D. Panne, Nature 2020, 578, 472.
[24] C. Wei, L. Jia, X. Huang, J. Tan, M. Wang, J. Niu, Y. Hou, J. Sun, P. Zeng, J. Wang, L. Qing, L. Ma, X. Liu, X. Tang, F. Li, S. Jiang, J. Liu, T. Li, L. Fan, Y. Sun, J. Gao, C. Li, J. Ding, Cell Res. 2022, 32, 744.
[25] E. Blanco, M. González-Ramírez, A. Alcaine-Colet, S. Aranda, L. di Croce, Trends Genet. 2020, 36, 118.
[26] L. Tan, W. Ma, H. Wu, Y. Zheng, D. Xing, R. Chen, X. Li, N. Daley, K. Deisseroth, X. S. Xie, Cell 2021, 184, 741.
[27] N. K. Hanchate, K. Kondoh, Z. Lu, D. Kuang, X. Ye, X. Qiu, L. Pachter, C. Trapnell, L. B. Buck, Science 2015, 350, 1251.
[28] A. van Koningsbruggen, M. Gierliński, P. Schofield, D. Martin, G. J. Barton, Y. Aniyurek, J. T. den Dunnen, A. I. Lamond, Mol. Biol. Cell 2010, 21, 3735.
[29] B. W. Hounkpe, F. Chenou, F. de Lima, E. V. de Paula, Nucleic Acids Res. 2021, 49, D947.
[30] D. W. Huang, B. T. Sherman, R. A. Lempicki, Nucleic Acids Res. 2009, 37, 1.