Genome by Multidimensional Scaling

Subjects: Mathematical & Computational Biology
Contributor: Ryo Ishibashi

The positions of enhancers and promoters on genomic DNA remain poorly understood. Chromosomes cannot be observed during the cell division cycle because the genome forms a chromatin structure and spreads within the nucleus. However, high-throughput chromosome conformation capture (Hi-C) measures the physical interactions of genomes. In previous studies, DNA extrusion loops were directly derived from Hi-C heat maps. By using Multidimensional Scaling (MDS), we can easily locate enhancers and promoters more precisely.

Keywords: multidimensional scaling ; high-throughput chromosome conformation capture ; enhancer ; promoter

1. Introduction

For cells to utilize genetic information, many genes must be expressed in a coordinated manner. The accessibility of genomic information depends on how DNA is packed into the chromatin. Chromatin is the basis of various biological processes, including cell cycle regulation and, DNA replication, repair, and maintenance [1]. Euchromatin is a genome region consisting of DNA with a relatively loose structure. The open structure allows RNA polymerase and other proteins to access the genome for DNA transcription. Enhancers and promoters also approach the euchromatin region to form DNA loops. Gene expression is controlled by promoters near the gene and by gene regulatory sites named as enhancers that are distant from the gene. However, how promoters and enhancers interact with each other to regulate gene expression is not well understood. High-throughput chromosome conformation capture (Hi-C) can be used to analyze the 3D structure of a genome by detecting genomic regions that are spatially close to each other using next-generation sequencing [2]. This conventional method led to an approximation of the genome structure from the Hi-C heat map [3]. We demonstrated the potential of using this method for identifying enhancers and promoters by applying multi-dimensional scaling (MDS).

2. Hypothetical Chromosomes

Figure 1 shows a heat map of Hi-C data after arranging these data as shown in Equation 1 below.

$$d_{ij}\mathrm{new} = \begin{cases} d_{ij} & , |i-j| \leq 5 \\ d_{ij} \times \log|i-j| & , |i-j| > 5 \end{cases}$$

where ***i*** and ***j*** are coordinates, and d_{ij} is the number of Hi-C detections. Pairs with large values in the matrix indicate region pairs with a high contact probability. The inverse of the Hi-C data was used as the distance data because MDS was used for similarity matrices. Then, MDS was applied to the Hi-C data, and the resulting hypothetical chromosomes are shown in **Figure 3**. The euchromatin region was identified (**Figure 2**).

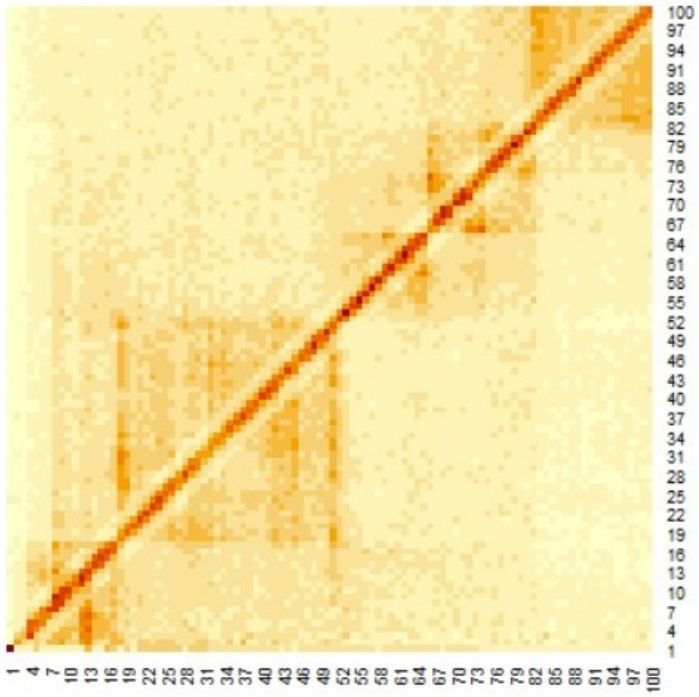

Figure 1. Heat map of Hi-C data after adding weighting.

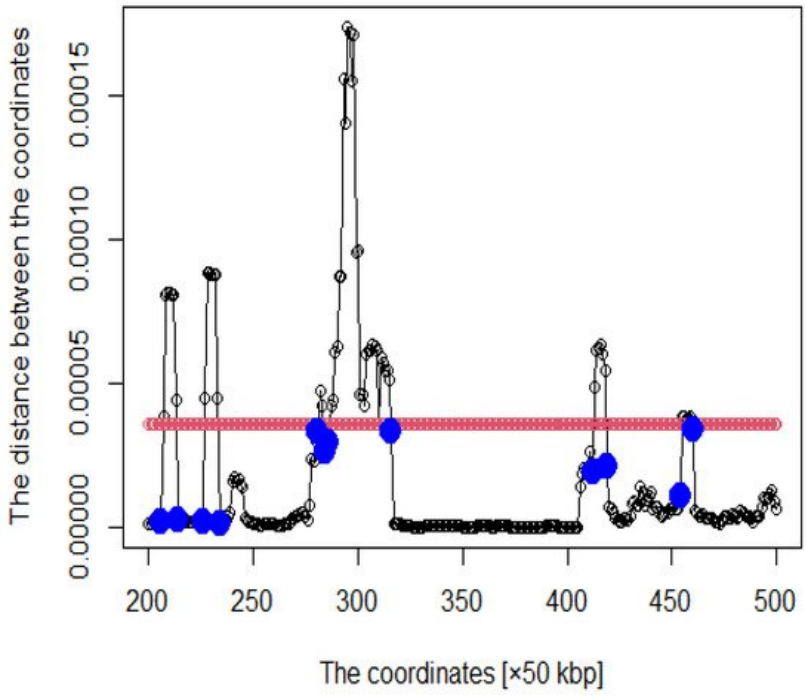

Figure 2. Distance plot between coordinates (The red line is the threshold and blue points are roots).

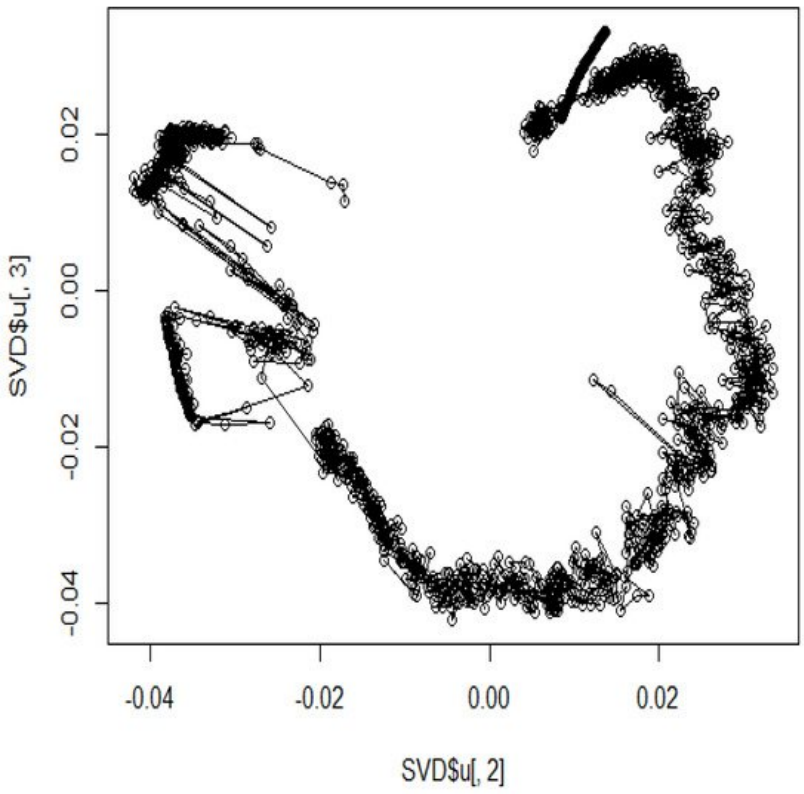

Figure 3. The hypothetical chromosomes 18 (0 bp–86,000 kbp).

3. Enrichment Analysis

We used BiomaRt [4] in R to retrieve genes from the obtained coordinates. Finally, the obtained euchromatin regions were subjected to enrichment analysis using g: Profiler [5]. The results are presented in **Table 1** and **Table 2**. The functions and processes involved in transcription were also determined such as the pre-transcriptional initiation complex and RNA polymerase II initiation complex, and transcription factors involved in cancer, such as CAMP responsive element binding protein 3 (CREB3) [6] and forkhead box M1 (FOXM1) [7].

Table 1. Results of enrichment analysis of 90 min Hi-C data by g:Profiler.

Term_Name	Term_ID	Adjusted_*p*_Value
transcription preinitiation complex assembly	GO:0070897	9.67×10^{-5}
RNA polymerase II preinitiation complex assembly	GO:0051123	5.90×10^{-3}
immunoglobulin complex	GO:0019814	6.72×10^{-49}
immunoglobulin complex, circulating	GO:0042571	3.97×10^{-8}
DNA packaging complex	GO:0044815	3.75×10^{-7}
protein-DNA complex	GO:0032993	5.36×10^{-4}
transcription factor TFIID complex	GO:0005669	9.22×10^{-3}
RNA Polymerase I Promoter Opening	REAC:R-HSA-73728	2.04×10^{-8}
Transcriptional regulation by small RNAs	REAC:R-HSA-5578749	5.98×10^{-7}
Factor: ER-alpha; motif: TGACCYN	TF:M03547	5.59×10^{-4}
Factor: Foxm1; motif: NTGTTTRT	TF:M07255	5.79×10^{-3}
Factor: MafG; motif: CMATGACTCAGCAGA; match class: 1	TF:M07048_1	1.04×10^{-2}
Factor: AML2; motif: TGTGGTNNN	TF:M07372	1.39×10^{-2}
Factor: PR; motif: NNNNNNRGNACNNKNTGTTCTNNNNNN	TF:M00957_1	2.66×10^{-2}

Table 2. Results of enrichment analysis of 120 min Hi-C data by g:Profiler.

Term_Name	Term_ID	Adjusted_*p*_Value
transcription preinitiation complex assembly	GO:0070897	2.93×10^{-4}
RNA polymerase II preinitiation complex assembly	GO:0051123	3.00×10^{-3}
immunoglobulin complex	GO:0019814	1.91×10^{-41}
immunoglobulin complex, circulating	GO:0042571	1.32×10^{-8}
transcription factor TFIID complex	GO:0005669	5.16×10^{-3}
Factor: ER-alpha; motif: TGACCYN; match class: 1	TF:M03547_1	1.15×10^{-4}
Factor: RARA; motif: GAGGTCAAAAGGTCAAKK	TF:M08018	2.84×10^{-3}
Factor: AML2; motif: TGTGGTNNN	TF:M07372	5.68×10^{-3}
Factor: MafG; motif: CMATGACTCAGCAGA; match class: 1	TF:M07048_1	8.21×10^{-3}
Factor: CREB3; motif: NTGCCACGTCAYCN	TF:M04207	4.57×10^{-2}

4. Comparison with Previous Studies

In addition, several studies have used MDS to analyze Hi-C data for accurately reproducing 3D genome structures. The framework for predicting 3D genomic structures using t-distributed stochastic neighbor embedding (t-SNE) is named as StoHi-C [8]. MDS has inherent problems with very sparse high-dimensional Hi-C datasets, whereas tSNE overcomes these limitations. This method can reproduce the characteristics of chromosome 3D structures more clearly than MDS in yeast Hi-C data. The distances between the coordinates obtained from the 3D structure reproduced by the StoHi-C method are shown in Figure 4. As shown in Figure 4, attempts to precisely reproduce the 3D structure resulted in no significant difference in the distance between coordinates, even when acquiring DNA loops with a threshold value. Therefore, the enhancers and promoters cannot be precisely identified. We focused on the ones with a large number of Hi-C detections, although the distance between coordinates is large because the goal of this study was to identify enhancers and promoters. Therefore, we added weights as shown in Equation (1).

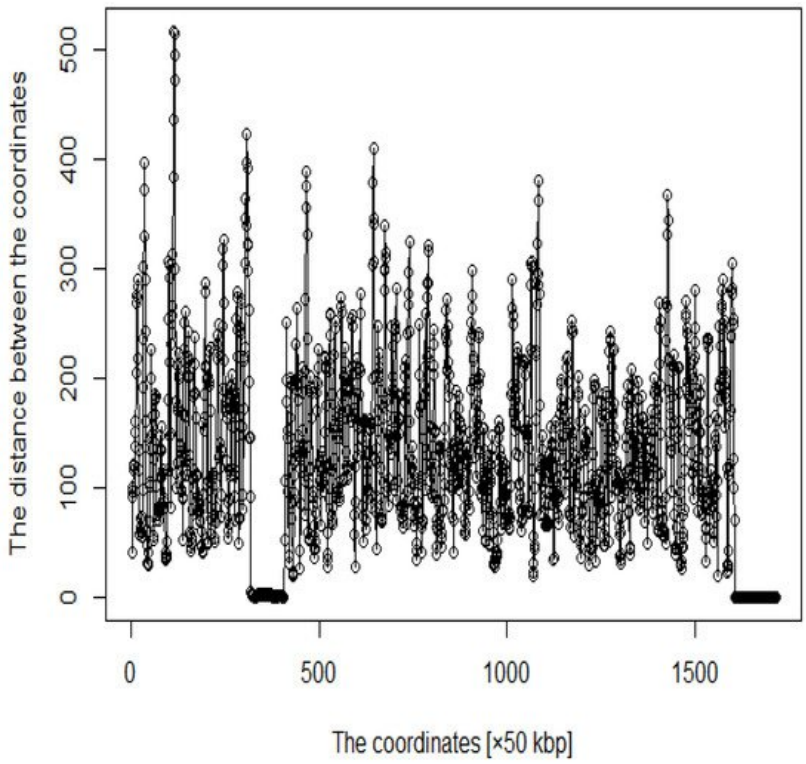

Figure 4. Distance plot between coordinates by StoHi-C.

Based on our results, it is useful to obtain DNA loops by automatically visualizing the chromosome structure using MDS, as performed in this study.

References

1. Job Dekker; Leonid Mirny; The 3D Genome as Moderator of Chromosomal Communication. *Cell* **2016**, *164*, 1110-1121, 10.1016/j.cell.2016.02.007.
2. Hyeseon Kang; Maxim N. Shokhirev; Zhichao Xu; Sahaana Chandran; Jesse R. Dixon; Martin W. Hetzer; Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation. *Genes & Development* **2020**, *34*, 913-930, 10.1101/gad.335794.119.
3. Irene Mota-Gómez; Darío G. Lupiáñez; A (3D-Nuclear) Space Odyssey: Making Sense of Hi-C Maps.. *Genes* **2019**, *10*, 415, 10.3390/genes10060415.
4. Steffen Durinck; Paul T Spellman; Ewan Birney; Wolfgang Huber; Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **2009**, *4*, 1184-1191, 10.1038/nprot.2009.97.
5. Uku Raudvere; Liis Kolberg; Ivan Kuzmin; Tambet Arak; Priit Adler; Hedi Peterson; Jaak Vilo; g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **2019**, *47*,

W191-W198, 10.1093/nar/gkz369.

6. Yizheng Wu; Ziang Xie; Junxin Chen; Jiaxin Chen; Weiyu Ni; Yan Ma; Kangmao Huang; Gangliang Wang; Jiying Wang; Jianjun Ma; et al. Circular RNA circTADA2A promotes osteosarcoma progression and metastasis by sponging miR-203a-3p and regulating CREB3 expression. *Molecular Cancer* **2019**, *18*, 1-20, 10.1186/s12943-019-1007-1.
7. Inken Wierstra; Jürgen Alves; FOXM1, a typical proliferation-associated transcription factor. *Biological Chemistry* **2007**, *388*, 1257-74, 10.1515/bc.2007.159.
8. Mackay, K.; Kusalik, A. StoHi-C: Using t-distributed stochastic neighbor embedding (t-SNE) to predict 3D genome structure from Hi-C Data. bioRxiv 2020.
