Modeling of chromatin architecture: implications in health and disease

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Priyojit Das
December 2022

Copyright © 2022 by Priyojit Das.
All rights reserved.

DEDICATION

I would like to dedicate the dissertation work to my mom and dad, Purnima and Tapan Das for their heartfelt love and never ending support.

ACKNOWLEDGEMENTS

Joining a biological science graduate program without any prior biology experience was never meant to be easy. I was trained to work with computers, not with the cells. But still I took the challenge and decided to make my scientific career in biology. Almost four and half years are gone and today, I am writing this acknowledgement section of my doctoral thesis. The journey was hard, challenging and demanding, but every single day of this journey made me a better researcher and better person. And also, during this time, I came into contact with many kind and generous individuals, without whom I cannot even imagine my graduate life, both inside and outside the lab space. I am greatly indebted to all those people for making this journey a worthy and fulfilling period in my life.

First and foremost, I would like to thank my dearest parents for providing me with such a wonderful life. You both are always inspirational to me. Thank you for supporting and encouraging me to pursue my dreams all the time. Thank you for teaching me the importance of hard work and dedication. Thank you for everything.

The person who comes next is my supervisor and mentor, Dr. Rachel Patton McCord. I couldn't dream of any person better than you as my supervisor. I am grateful to you for believing in me and giving me the opportunity to do research in your lab. I am grateful for the amazing guidance and supervision you have provided throughout this period. You gave me enough freedom to design my own doctoral research projects, which has put me at least a step closer to becoming an independent researcher. You taught me how to think critically, which is one such skill every researchers must possess. Thank you for supporting my travels to workshops, conferences, and meetings, which allowed me to make connections with the people in the scientific community and to be more confident presenting my research. I am again grateful for everything you have done to make my research experience enriching and for allowing me to grow as a scientist.

Apart from Rachel, another person who has made significant contribution to my graduate research journey is Dr. Tongye Shen. I always considered you as my second supervisor. You were always there if I had any theoretical and computational problems related to projects. I genuinely appreciate your effort in challenging my models by finding their shortcomings and making me think more critically. Thank you for pointing out my weak areas and teaching me how to improve them. And thank you for being available anytime of the day to discuss science. I would also like to thank you and late professor Dr. Elizabeth Howell for instilling the passion for doing structural biology in me, which I really want to continue doing in future also. If you two had not been there, I would have been working on some other sections of biology.

I would also like to thank my other committee members Dr. Tian Hong, Dr. Mariano Labrador, and Dr. Jianbin Wang. Thank you all for dedicating your valuable time in providing constructive feedbacks on my research. Your ideas provided me with different perspectives when I am stuck focusing too much on a specific thing. I appreciate those all.

I would also like to take this opportunity to thank the members of McCord lab. I joined the lab without any prior knowledge in biology. But the lab members were more than welcoming to teach me whenever I have any questions regarding biology problems. I am greatly thankful to McCord lab member, Dr. Rebeca San Martin. You always helped me to think about and understand the biological system first before asking the research question and further designing methods to solve them. Moreover, you taught me how to think like an experimental biologist even though I am a computational researcher and have very minimal wet-lab experience. Another person in the lab who was invaluable to my research and development was Dr. Rosela Golloshi. I am thankful to her for being a such wonderful mentor to me. You were very welcoming from my first days in McCord lab. I greatly enjoyed discussing science and collaborating with you. I genuinely appreciate your significant contribution in shaping my early graduate research career and I owe you a lot for that. You have always been a role model for me. Also, the other past and present members of the McCord lab: Dr. Jacob Sanders, Dr. Yang Xu, Heng Li, Chris Playter, Nishita Gogia and Tianchun Xue, all your feedback and support helped me immensely in my research. I owe you all for helping me out with everything.

I am hugely thankful to my peers and friends from GST and BCMB for making the journey very enjoyable and fulfilling. I would especially like to thank Bright, Nadejda, Dhanya, Abhishek, Nishita, Ajay, Debabrata, Manjima, Dibyayan and many others for making a homely experience for me far from home.

Last but not least, I would like to thank my previous mentors, Drs. Sujay Saha, KA Abdul Nazeer and Ashish Misra for installing the passion for pursuing computational biology research in me. Without them, it would not be possible for me to reach this level.

ABSTRACT

In eukaryotic cells, the long linear sequence of DNA is packed tightly inside the nucleus in a non-random folded fashion. This organization is formed via the interactions of chromatin with itself and with other nuclear substructures. However, this spatial organization is not static: inside the nucleus, chromatin is subjected to various active forces and thermal noise that produce structural dynamics. As a result, an ensemble of three-dimensional structures are produced both temporally and spatially. Several studies have been performed to characterize the structural properties of the chromatin by analyzing the population-averaged mean structure, but the properties of the chromatin conformational ensemble have not received equal attention. For example, it is not clear how the structural ensemble varies from one cell type to another. Or, how the structural ensemble is transformed during different biological processes such as the cell cycle. Therefore, in this dissertation, I focused on characterizing the structural and dynamical properties of the chromatin structural ensemble at different resolutions ranging from individual chromosomes to local regions. I started my work by developing a computational method to infer the radial distribution of chromosome territories inside the nucleus based on chromosome contact data and genome intrinsic properties (Chapter 1). Next, I moved on to analyze single-cell chromatin conformational data to characterize the degree of variation present in chromosome structural ensemble (Chapter 2). In Chapter 3, I characterized the transition of the structural ensemble during different conditions through contact correlation analysis. After characterizing different properties of the structural ensemble, I focused on the role of the three-dimensional genome organization in different physiological and pathological conditions. In Chapter 4, I studied the relationship between the lamina-associated domains, genomic compartmentalization, and gene expression across different cell types by integrating different genomic data. And, in the final project, I characterized chromosome structural properties associated with different subtypes of breast cancer and breast cancer metastasis to distant organ (Chapter 5). Overall, findings through these projects help to characterize the structural and dynamical properties of chromosome ensemble and how these are altered during different conditions.

TABLE OF CONTENTS

INTRODUCTION 1
Chromosome territories organize in a non-random cell-type-specific fashion inside the nucleus 4
Single-cell chromosome conformations reveal heterogeneity present in chromosome structural ensemble 6
Properties of chromosome structural ensemble and their transition during different processes 7
Role of nuclear lamina in genome organization and gene regulation 9
Breast cancer metastasis organotropism and role of spatial genome organization in cancer 11
REFERENCES 14
CHAPTER I Inferring chromosome radial organization from Hi-C data 22
Abstract 23
Introduction 24
Results 27
Thresholding contacts extracts meaningful chromosomal interaction patterns from Hi-C data 27
Radial chromosome ordering can be inferred from PCA on inter-chromosomal strong interaction pattern matrix 29
Network modeling predicts probabilistic radial CT organization 31
Chromosome property-based tuning improves predicted consensus radial arrangement of CTs 36
Paternal and maternal homologs show mostly similar radial arrangement of CTs 37
Radial CT organization in epithelial cells changes depending on their nuclear shape 39
Variation between the radial arrangement of lymphoblastoid cells and neutrophils 39
Discussion 41
Conclusions 43
Methods 44
Pre-processing of Hi-C data 44
Simulating random ligation Hi-C from original Hi-C data 46
Identifying CT radial distribution patterns with PCA transformation on the pairwise strong inter-chromosomal interaction pattern matrix 46
Constructing 3D network graphs from Hi-C matrix 46
Fitting a geometrical structure to the 3D network graphs 47
Identifying the specific cluster of network models having meaningful CT organization 48
Gene density and chromosome length based tuning of consensus radial organization of CTs 48
Availability of data and materials 49
Abbreviations 49
REFERENCES 50

APPENDIX 57
CHAPTER II Characterizing the variation in chromosome structure ensembles in the context of the nuclear microenvironment 70
Abstract 71
Introduction 72
Material and methods 74
Construction of distance matrices from single-cell chromosome imaging data 76
Generating chromosome structural landscapes using a statistical analysis technique 77
Quantification of chromosome structural ensemble variation using a network-based similarity index 77
Calculating chromosome structure variability for each genomic bin 78
Characterization of chromatin states using epigenetic profiles 79
Polymer simulation of chromatin segments based on underlying chromatin states . 79
Loop extrusion simulation of chromatin segments based on underlying RAD21 binding sites 81
Analysis of nucleolus-associated domain (NAD) mapping data 82
Results 82
Chromosome structural landscape reveals major features of the structural ensemble 82
Ensemble structure similarity index captures cell-type- and chromosome-specific features of the chromosome structural ensemble 84
Contribution of epigenetic properties to the cell-type-specific variation of chromosome structural ensemble 87
Structure heterogeneity properties are associated with epigenetic state and boundary positions 89
Loop extrusion alone does not reproduce experimentally detected chromosome structure variation patterns 91
The effect of cohesin depletion on structural heterogeneity depends on the underlying epigenetic state of the chromosome region 92
Discussion 94
Availability 95
Supplementary data 95
Acknowledgements 96
Funding 96
Conflict of interest 96
REFERENCES 97
APPENDIX 102
CHAPTER III Using contact statistics to characterize structure transformation of biopolymer Ensembles 114
Abstract 115
Introduction 116
The generic polymer limit and the problem of removing distance decay in contact matrices 119

Contact analysis for partially structured proteins and chromosomes 123
Proteins approach the generic polymer limit with rising temperature 123
Mitotic chromosomes approach the effective generic polymer limit during the cell cycle 125
Chromosomes maintain detailed structure under salt and drug treatments 127
Cross comparison using score functions 129
Concluding remarks 133
Acknowledgement 133
REFERENCES 134
CHAPTER IV Differential contributions of nuclear lamina association and genome compartmentalization to gene regulation 139
Abstract 140
Introduction 141
Results 142
Chromosome compartmentalization and lamina association segregate cells by lineage 142
Lamina association adds another layer of gene regulation to compartment identity 144
Coupling and decoupling of lamina association and compartmentalization effects on gene expression 146
The strength of compartment identity correlates with lamina association and vice versa 146
Lamina-associated regions with similar compartment identity exhibit differential histone modifications between cell types 148
Lamina association reduces gene expression inducibility after physicochemical treatment 151
Discussion 151
Methods 154
Cell culture 154
Prostate cancer cell line Lamin B1 ChIP-seq 155
DamID-seq data processing 156
Hi-C data processing 156
RNA-seq data processing 157
Batch effect removal 157
Differential expression analysis 157
Histone ChIP-seq data processing 157
LaminB1 ChIP-seq data processing 158
Hierarchical clustering of genome-wide lamina association and compartmental organization data 158
Combine lamina association and compartmental organization data 158
Acknowledgements 159
Disclosure statement 159
Data availability 159
Author contributions 159

Supplementary material 159
Funding 159
REFERENCES 160
APPENDIX 166
CHAPTER V Rearrangement of 3D Genome Organization in Breast Cancer Metastasis Organotropism 174
Abstract 175
Introduction 176
Results 177
Modeling of breast cancer lung metastasis system 177
Breast cancer compartmental organization exhibits subtype and epithelial-mesenchymal state based segregation 178
Proportions of lung specific changes are higher in lung metastatic breast cancers compared to the localized cancers 181
Discussion 182
Methods 185
Hi-C data processing 185
RNA-seq data processing 185
Hierarchical clustering of genome-wide compartmental organization and transcriptome data 185
REFERENCES 186
APPENDIX 190
CONCLUSION 192
REFERENCES 195
VITA 196

LIST OF TABLES

Table 1 Dataset description. 57
Table 2 Correlation between microscopy data radial CT positions and those inferred by different sources of information. 58
Table 3 ENCODE accession IDs for the eleven epigenetic marks from three different cell-types. 102
Table 4 Parameters values for Lennard-Jones Potential $\boldsymbol{UPr}$ used in the polymer model. Each entry represents the interaction strength and distance cutoff (separated by |) between beads of that two specific types. Energy values and distance cutoff values are represented in terms of $\boldsymbol{kBT}$ and $\boldsymbol{\sigma}$ respectively. 103
Table 5 NCBI GEO, EMBL ENA, 4DN Data Portal and ENCODE accession IDs for the Hi-C, DamID-seq, RNA-seq and ChIP-seq data from thirteen different cell-types. RC = replicates combined for analysis, R1 = replicate 1 used for analysis. Specific conditions chosen from each accession number indicated where needed in parentheses. 166
Table 6 NCBI GEO, EMBL ENA, 4DN Data Portal and ENCODE accession IDs for the ChIP-seq data from thirteen different cell-types. RC = replicates combined for analysis, R1 = replicate 1 used for analysis. Specific conditions chosen from each accession number indicated where needed in parentheses. 167
Table 7 NCBI GEO or Encode accession codes for different breast and lung cell types. 190

LIST OF FIGURES

Figure 1 Chromatin organizes in a multi-layered fashion inside the nucleus of eukaryotic cell. 2
Figure 2 Structural dynamics exhibited by different layers of the multi-layered spatial genome organization. Adopted from Finn and Misteli (Finn & Misteli, 2019). 3
Figure 3 Determination of strong contact threshold and inference of radial CT distribution type. 28
Figure 4 Rearrangement of radial CT position in progeria and senescent cells compared to proliferating fibroblasts 32
Figure 5 Network modeling generated distributions of CT radial positions and comparison of the predicted tuned result with microscopy data. 34
Figure 6 Paternal and maternal homologs follow a similar CT distribution pattern. 38
Figure 7 Epithelial cells show different radial CT organization in a nuclear morphology dependent manner. 40
Figure 8 Radial arrangement of CTs in irregularly shaped neutrophil nuclei. 42
Figure 9 Overview of the computational analysis approaches. 45
Figure 10 Radial CT arrangement of GM12878 (FitHiC). 59
Figure 11 Radial CT arrangements inferred from GM12878 and BJ1hTERT simulated random ligation data. 60
Figure 12 Network modeling generated clusters for GM12878. 61
Figure 13 The radial distance distributions of 23 CTs for both GM12878 and BJ1-hTERT obtained from the respective selected model clusters based on inferred CT distribution types (# of models in the selected clusters- GM12878: 109, and BJ1-hTERT: 122). 62
Figure 14 Comparison of network modeling generated radial CT distance profiles with experimentally obtained distributions. 63
Figure 15 The radial distance distribution profiles of 23 CTs for GM12878, GM12878 simulated random ligation, BJ1-hTERT and BJ1-hTERT simulated random ligation obtained from the respective selected model clusters based on inferred CT distribution types (# of models in the selected clusters: for GM12878- 109, GM12878 simulated random ligation- 95, BJ1-hTERT- 122, and BJ1-hTERT simulated random ligation- 113). 64
Figure 16 Radial arrangement of BJ1-hTERT and BJ-5ta follow similar CT distribution patterns. 65
Figure 17 Radial arrangement of GM12878 replicates R1 and R2 show similar CT distribution pattern. 66
Figure 18 The radial distance profiles of 23 CTs for GM12878 R1, GM12878 R2, BJ1-hTERT and BJ-5ta obtained from the respective selected model clusters based on inferred CT distribution types (# of models in the selected clusters: for GM12878 R2- 109, GM12878 R1- 142, BJ1-hTERT- 122, and BJ-5ta- 115). 67
Figure 19 The radial distance profiles of 23 CTs for GM12878 standard, paternal, and maternal copies obtained from the respective selected model clusters based on

inferred CT distribution types (# of models in the selected clusters: for GM12878 standard- 109, paternal copy- 194, and maternal copy- 148). 68
Figure 20 Statistical comparisons of the CT distance profiles obtained from network modeling for GM12878 and BJ1-hTERT for different sample sizes ns (number of network structures generated from Hi-C data) using the two-sided Mann-Whitney U test. 69
Figure 21 Schematic representation of the statistical pipeline used to quantify conformational diversity present in single-cell chromosome distance matrices. 75
Figure 22 The chromosome structural landscape orders single-cell structures on a continuum based on their radius of gyration followed by domain reorganization. .. 83
Figure 23 Network based similarity score quantifies cell-type- and chromosome-specific variability present in chromosome structural ensemble at different resolutions. 85
Figure 24 Epigenetic properties (histone modifications) contribute to the cell-type-specific variation in the chromosome structural ensemble. 88
Figure 25 Evaluation of structural variation of regions with different epigenetic, genomic, and structural properties. 90
Figure 26 Cohesin depletion alters chromosome structural heterogeneity depending on underlying epigenetic state constraints. 93
Figure 27 Assignment of different chromatin states based on relative enrichment of different histone marks using ChromHMM tool for chr21:28-30 Mb genomic region from three different cell-types - IMR90, K562 and A549 at 30 Kb resolution. 104
Figure 28 a) Pearson's correlation between chromosome structural landscape PC1 values and radius of gyration ($\boldsymbol{Rg}$) of single-cell structures (calculated from 3D position traces) for IMR90 chr21:28-30 Mb genomic region. b) HiCRatio score of PC2 ordered structures belonging to different PC1 groups for IMR90 chr21:28-30 Mb genomic region, calculated by HiCRatio approach (with a 300 kb window size). Here, PC1G1 refers to the group1 (G1; smallest $\boldsymbol{Rg}$ structures) based on PC1. Within PC1G1, then PC2G1 means the group1 (G1) based on PC2. Successive higher numbered PC1 groups have higher $\boldsymbol{Rg}$ structures. c) For K562 chr21:28-30 Mb genomic region, structures are first divided into three groups based on the PC1 ordering. Within each group, structures are further divided into three subgroups based on their PC2 ordering. To display the domain reorganization along PC2 in K562, averaged distance maps from only groups 2 and 3 from PC1 are shown here. d) For the same groups as in c, HiCRatio (with a 300 kb window size) scores of structures are shown. Colored lines represent HiCRatio scores from PC2 subgroups within two different PC1 subgroups (top = smaller $\boldsymbol{Rg}$ and bottom = larger $\boldsymbol{Rg}$). 105
Figure 29 For IMR90 chr2 structure data, structures are first divided into ten groups based on the PC1 ordering. Within each group, structures are further divided into ten subgroups based on their PC2 ordering. To display the domain reorganization phenomenon along PC2 in IMR90 chr2, only PC1 subgroups 6 (lower $\boldsymbol{Rg}$) and 10 (higher $\boldsymbol{Rg}$) are shown here. Heatmaps display the average distances for structures within each PC2 subgroup in each PC1 subgroup. 106
Figure 30 For IMR90 chr21 entire q-arm structure data, structures are first divided into sixteen groups based on the PC1 ordering. Within each group, structures are further

divided into sixteen subgroups based on their PC2 ordering. To display the domain reorganization phenomenon along PC2 in IMR90 chr21, only PC1 subgroups 12 (lower $\boldsymbol{Rg}$) and 16 (higher $\boldsymbol{Rg}$) are shown here. Heatmaps display the average distances for structures within each PC2 subgroup in each PC1 subgroup. 107
Figure 31 a,c) Hierarchical clustering of IMR90 (a) and K562 (c) structures for chr21:28-29.5 Mb genomic segment, based on $\boldsymbol{Q}$ similarity metric as described in Cheng et al. (1). To assign structures to the different clusters, the trees are cut at 0.7. b,d) Proportion of structures in different clusters obtained from IMR90 (b) and K562 (d) structural ensembles. 108
Figure 32 a) Chromosome structural landscape of IMR90 and K562 chr21:28-30 Mb chromatin regions constructed for different numbers of cells. Points are colored according to the radius of gyration ($\boldsymbol{Rg}$) of the corresponding structures (dark blue – smaller $\boldsymbol{Rg}$ and dark red – larger $\boldsymbol{Rg}$). b) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb segment from IMR90 and K562 for different number of cells (shown as n). Here, for resampling 200 single-cell structures are selected randomly each time. c) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb segment from IMR90 and K562 for different correlation matrix cutoff. Here, for resampling 200 single-cell structures are selected randomly each time. 109
Figure 33 a) Chromosome structural landscape of IMR90 and K562 chr21:28-30 Mb chromatin regions constructed for different numbers of cells after removing extremely compact or large conformations. Points are colored according to the radius of gyration ($\boldsymbol{Rg}$) of the corresponding structures (dark blue – smaller $\boldsymbol{Rg}$ and dark red – larger $\boldsymbol{Rg}$). b) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb segment from IMR90 and K562 for different numbers of cells, before (green) or after (blue) removing extremely compact or large conformations ("outliers"). 200 single-cell structures are selected randomly each time for resampling. 110
Figure 34 a) Population-averaged distance maps of active and inactive chrX from IMR90 cell at TAD resolution. b) Jointly constructed chromosome structural landscape of active and inactive chrX from IMR90 cell. c) Bin Variability (BSI standard deviation) of active and inactive chrX, where a higher value represents a higher level of structure variation for that bin across the population and vice versa. d) Chromosome structural ensemble SSI distributions for chr2 segments of varying length. Here, for resampling 300 single-cell structures are selected randomly each time. 111
Figure 35 a) Representative polymer models of IMR90 and K562 for chr21:28-29.5 Mb genomic segment, obtained from different time points of the simulation. The color of the beads shows corresponding epigenetic states. b) Chromatin states of six different 2 Mb spanning genomic regions of IMR90 chr21 at 50 kb resolution. The chromatin states are inferred based on 11 different histone marks for each cell-type by ChromHMM tool. c) Epigenetic state similarity of different 2 Mb spanning genomic regions of IMR90 chr21 q-arm with the K562 28-30 Mb region using binary comparison. Here, a higher comparison value shows higher similarity with

the K562 epigenetic state. d) Chromosome structural ensemble SSI distributions for 31 different 2 Mb spanning genomic regions of IMR90 chr21. For resampling, 300 single-cell structures are selected randomly each time. Mean boundary strength of those regions inferred by HiCRatio have been shown in green color. e) Pearson's correlation between the mean boundary strength and mean ensemble SSI of the 31 different genomic regions obtained from IMR90 chr21. 112
Figure 36 a) Population-averaged distance maps of IMR90 and K562 chr21:28-30 Mb genomic region from loop extrusion polymer simulations (see Methods) at 1 kb resolution. b) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb from loop extrusion simulated IMR90 and K562. For resampling, 500 single-cell structures are selected randomly each time from a single run (left) or across multiple runs (right). c,d) Jointly constructed chromosome structural landscapes of chr21:28-30 Mb (c) or chr21:34-37 Mb (d) of HCT116 for WT and +Auxin conditions. Points are colored according to the radius of gyration ($\boldsymbol{Rg}$) of the corresponding structures (dark blue – smaller $\boldsymbol{Rg}$ and dark red – larger $\boldsymbol{Rg}$). e) Bin Variability of chr21:28-30 Mb genomic region for wild-type (WT) and cohesin-depleted (+Auxin) conditions. f) Bin Variability of chr21:34-37 Mb genomic region for wild-type (WT) and cohesin-depleted (+Auxin) conditions. 113
Figure 37 A cartoon illustration of three different analysis methods, (a) I-PCA, (b) MI-PCA, and (c) E-PCA, on contact matrices. 118
Figure 38 (a) The dominant mode of MI-PCA for a random polymer, MI-PC1, including power law form of mean matrix, exponential decay form of mean matrix, and a result of computer simulation of the polymer. (b) The top five eigenvectors of MI-PCA for the case of the exponential decay with α=1.5. Eigenvectors with smaller eigenvalues have increasing number of nodes. 121
Figure 39 (a) Structure and contact map of protein CI2. (b) MI-PC1 of protein CI2 at different temperatures with the mean contact matrix at highest and lowest temperatures shown. The dashed line is the reference system of backbone interaction only. (c) The corresponding E-PC1 of protein CI2 at different temperatures are shown using its significant elements (a cutoff of $|d_{ij}| > 0.025$). 124
Figure 40 MI-PC1 of chicken chromosome 21 (chr21) at timepoints progressing from G2 to late prometaphase. The x-axis represents the 40 kb bin index along the chromosome. 126
Figure 41 The comparison of drug (TSA) treated human chromosome 21 and the corresponding control using direct contact analysis, i.e., MI-PCA on u_{ij} (a) and the conventional approach with the distance-decay scaled (b), i.e., MI-PCA on v_{ij}..... 128
Figure 42 (a) Protein CI-2 at different temperatures characterized by two parameters R_{str} and R_{ip}, where the x-axis measures against a specific folded structure at the low temperature limit and the y-axis represents how well the structure resembles the features of the corresponding generic polymer. (b) Chicken chromosome 21 at different time points during the cell cycle characterized by two parameters R_{str} and R_{ip}. 131

Figure 43 (a) The structural features of chicken chromosomes at t = 60 min expressed by the MI-PC1s of the unpretreated contact matrices. (b) The R_{ip} versus the corresponding chromosome length. 132
Figure 44 Cell lineage relationships are detected in both compartmentalization and lamin association patterns. 143
Figure 45 Gene expression change between cell types follows alterations in lamina association for genomic regions with unchanged compartment identity. 145
Figure 46 Mean gene expression fold change for different categories of genomic regions based on compartment identity and lamina association. 147
Figure 47 Compartment identity strength shifts with changes in LAD status even when compartment identity remains unchanged. 149
Figure 48 Fold change of different histone marks for the similar compartment identity regions which exhibit differential lamina association between cell types. 150
Figure 49 Effect of compartment identity and lamina association on differential gene regulation after different treatments for various cell types. 152
Figure 50 Proportion of LAD and compartment type for each cell type. 168
Figure 51 Gene expression changes with alteration in lamina association for similar compartment identity genomic regions. 169
Figure 52 Mean gene expression fold change between cell types for different categories of genomic regions based on compartment identity and lamina association. 170
Figure 53 Compartment identity strength shifts with changes in LAD status even when compartment identity remains unchanged. 171
Figure 54 Lamina association strength shifts with changes in compartment status. 172
Figure 55 Fold change of different histone marks for the similar compartment identity regions which exhibit differential lamina association between cell types. 173
Figure 56 Modelling of breast cancer lung metastasis system. 179
Figure 57 Unsupervised analysis separates breast cancer cells depending on their molecular subtypes and epithelial-mesenchymal states. 180
Figure 58 Percentages of different lung permissive changes shown by localized and metastatic cancers from different breast cancer subtypes. 183
Figure 59 Breast cancer cell lines are arranged depending on their epithelial-mesenchymal status. 191

INTRODUCTION

Deoxyribonucleic acid (DNA), the material responsible for encoding the genetic information, is enclosed within the cell nucleus (Alberts, 2017). Though the nucleus is the largest organelle of the mammalian cell, when compared its size with the total length of the enclosed DNA sequence, it becomes geometrically equivalent to packaging miles of highway inside a tennis ball. Therefore, a compact folding of the linear DNA sequence must exist inside the nucleus. There are many biological processes which can only be explained via the spatial proximity of distant genomic regions (McCord & Balajee, 2018; Schoenfelder & Fraser, 2019). Additionally, deleterious mutations or translocations can affect gene regulation without even occurring near to the genes they influence. Intrigued by these phenomena, I set the overarching goal of my research to characterize the folding of the 3D genome in physiological conditions and how this folding changes during pathological conditions.

A part of this compact DNA structure has already been known for decades: the linear sequence of DNA wrap around histone proteins and forms nucleosomes, which then further coils into a chromatin fiber (Bednar et al., 1998). Recent advances in molecular biology and imaging techniques however have shown the existence of further higher order structure where the chromatin fiber folds in a multi-layered fashion (Fig. 1) (Lieberman-Aiden et al., 2009; Rachel Patton McCord et al., 2020). At the local scale, the interactions between distant chromatin regions which are generally hundreds of kilobases apart, lead to the formation of chromatin loops. As a result, domains are formed which are enriched for intra-domain contacts compared to inter-domain contacts. The domains further organize into separate active and inactive chromatin spatial regions and then these regions are organized into chromosome territories. These structural layers work independently as well as in concert to contribute to various essential biological mechanisms.

As chromatin experiences different active forces and thermal noise inside the nucleus, the higher order spatial organization also becomes dynamic (Hübner & Spector, 2010; Oliveira et al., 2021; Put et al., 2019). Due to these active forces and thermal noise, chromosome structures produce both fast and slow structural dynamics, which eventually lead to the emergence of the spatial and temporal structural heterogeneity (Fig. 2) (Zhang & Wolynes, 2017). The structural fluctuations also affect a number of genetic and biological functions such as DNA replication, DNA repair, gene replication, regulation, and transcription. Often the effects of such heterogeneity are ignored while characterizing the structural properties of the mean chromosome conformational data (Luppino & Joyce, 2020; Ulianov et al., 2017). That also obscures the mechanism involved with the chromatin folding at different structural organization levels. Furthermore, the existence of different sub-population structures which may associate with different biological mechanisms can also be masked only if bulk structures are analyzed (Cheng et al., 2020; Das et al., 2022; Xie et al., 2020). Characterizing single cell properties therefore can help understand the folding pathway of the chromosomes as well as the structural features associated with different subpopulations

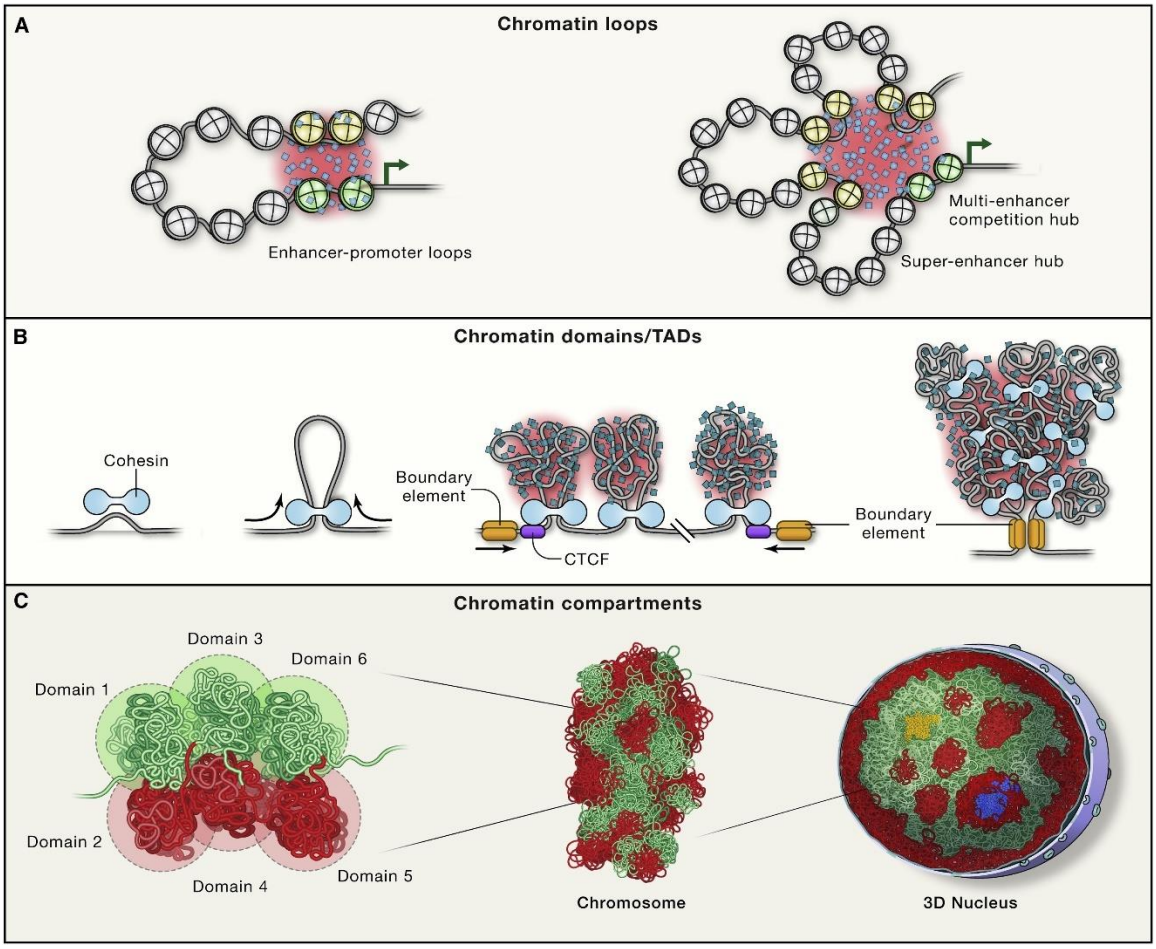


Figure 1 Chromatin organizes in a multi-layered fashion inside the nucleus of eukaryotic cell.

a) Chromatin loops form due to the spatial proximity of the distal genomic regions at the length scale of several hundred kilobases. These include enhancer promoter interactions as represented by yellow and green nucleosomes respectively in the figure. The interactions get stabilized further by phase separation of relevant components such as transcription factors. b) Loop extrusion process mediated by cohesin and CTCF leads to the formation of topologically associated domains (TADs) which are enriched for intra-domain contacts compared to inter-domain contacts. c) The TADs of similar chromatin type further associate together to from compartmental organization of the genome which leads to the segregation of the euchromatin and heterochromatin regions. Adopted from Misteli (Misteli, 2020).

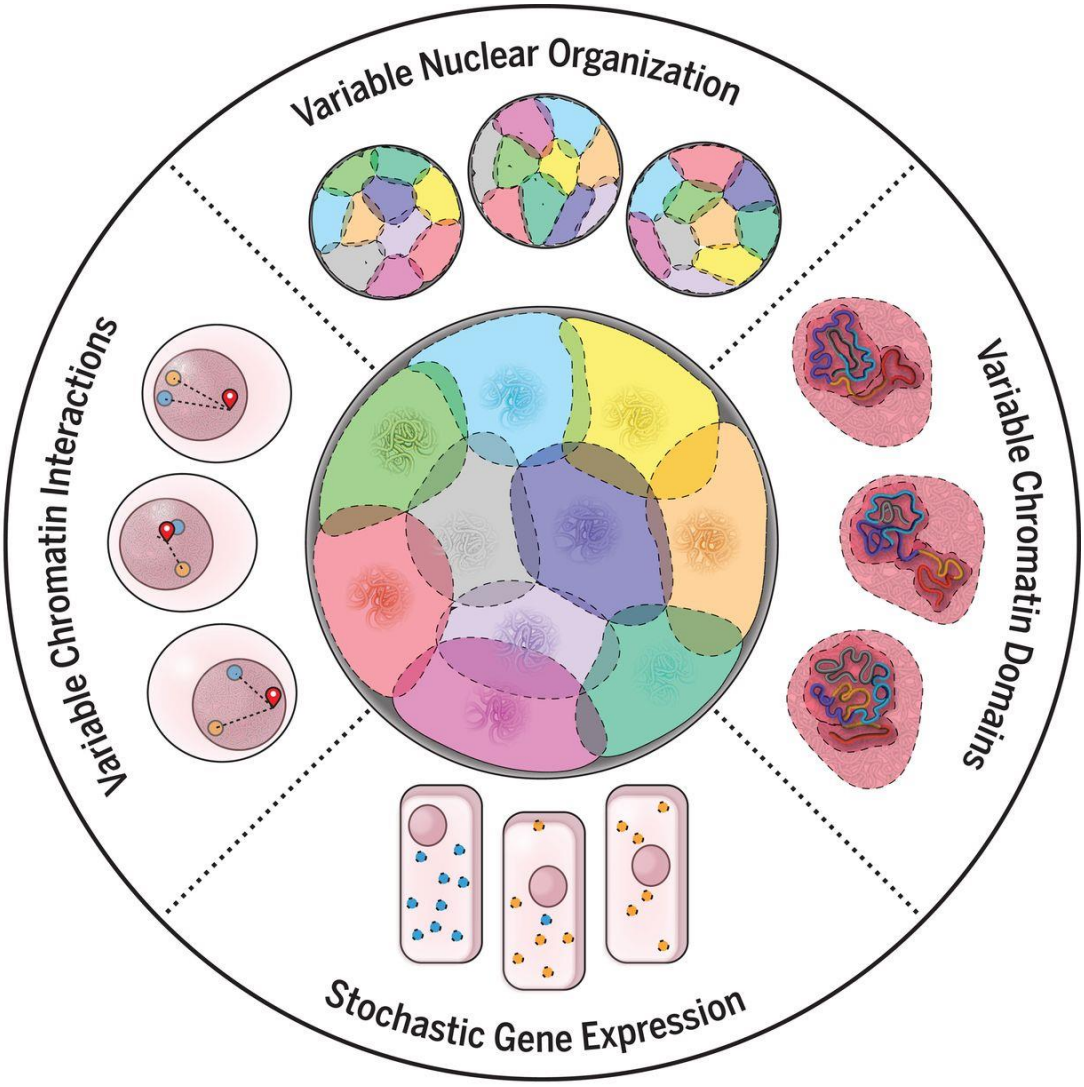


Figure 2 Structural dynamics exhibited by different layers of the multi-layered spatial genome organization. Adopted from Finn and Misteli (Finn & Misteli, 2019).

and their relationship with different biological processes. In addition to that, it is still not known how and to what extent these structural and dynamical properties of the chromosome structural ensemble change during different physiological and pathological conditions. Therefore, at the University of Tennessee Knoxville, under the supervision of Dr. Rachel Patton McCord, for my doctoral research, I have focused on analyzing both population-averaged and single cell chromosome conformation data in order to extract insight about folding mechanisms and impact on gene regulation. To solve these problems throughout my PhD research, I have developed and applied different computational, mathematical, and physics-based modeling and simulation techniques in combination with graph theory, machine learning, and soft computing algorithms.

Chromosome territories organize in a non-random cell-type-specific fashion inside the nucleus

Individual chromosomes inside the nucleus organize in discrete territorial domains (Cremer & Cremer, 2010). These domains, which are better known as chromosome territories (CTs) form the highest level of the microscopic spatial genome organization. Moreover, this spatial organization is nonrandom in nature and has been found to regulate different cellular mechanisms such as DNA replication, ribonucleic acid (RNA) translation and gene expression (Cremer et al., 1993). Apart from these essential processes, the nonrandom organization has recently also found to contribute to the protection of the genome from deleterious rearrangements that may happen due to DNA damage (Rosin et al., 2019). Although CT positions remain largely stable during interphase, they can get altered as the cell divides due to nuclear envelope breakdown and re-establishment. In addition to this, the radial organization of the CTs changes as the cells differentiate during healthy conditions and also as they become diseased (Kuroda et al., 2004; Mehta et al., 2011). For example, human chr18, which is a gene-poor chromosome, is more likely to be found internally in premature aging progeria syndrome cells compared to its position in normal quiescent fibroblast cells. In contrast, certain gene-rich chromosomes tend to exhibit peripheral localization, specifically within blebs (Bercht Pfleghaar et al., 2015). The likelihood of particular chromosome translocations also increases due to the radial CT rearrangements. One such example is the translocation between chr12 and chr16, which may occur during adipogenesis due to spatial proximity of these two chromosomes and may drive liposarcoma tumorigenesis (Kuroda et al., 2004). In addition, the reorganization also contributes to the gene expression regulation as shown in breast cancer, where due to radial rearrangements, both active and inactive chrX exhibits higher interchromosomal contacts and that correlates with the overall increase in the expression of chrX genes (Fritz et al., 2014).

The non-random radial organization of the interphase CTs are cell-type-specific as shown by microscopy imaging techniques (Boyle et al., 2001; Sun et al., 2000). Prior studies have revealed that, this cell-type-specific radial positioning can be broadly categorized into two groups based on the genome intrinsic properties: gene-density and chromosome length. In the gene-density based organization, the gene rich chromosome territories are more likely

to be found near nuclear center whereas chromosomes with lesser number of genes are preferentially located at the periphery of the nucleus (Boyle et al., 2001). On the other hand, in the case of length based organization, with the increase in the chromosome size the likelihood of peripheral localization also increases (Sun et al., 2000). It is likely that the competition between the gene activity and lamina association might drive these different distributions. Generally, lamina associated regions are heterochromatic repressive in nature and can be found frequently on the chromosomes having lower gene density which again are more likely to be lengthier compared to the chromosomes having higher gene-density (Guelen et al., 2008). In addition, the cell proliferation rate has also been found to be associated with the radial CT distribution: proliferating cells are likely to exhibit gene-density based distribution whereas the quiescent and senescent cells tend to follow length based one (Bridger et al., 2000). Furthermore, it has also been shown that the non-random radial organization of the CTs to some extent correlated with the shape of nuclei, where spherical nuclei are more likely to exhibit a distribution based on gene-density whereas the flat ellipsoidal nuclei exhibit chromosome size-based organization (Bolzer et al., 2005; Tanabe et al., 2002; Tanabe et al., 2005).

Although the CTs show a preference for non-random radial localization inside the nucleus depending different factors (e.g., gene density and lamina association) discussed in the previous paragraph, their positions can still vary between cells across a population, and this single-cell variability may originate due to alteration of CT positions during cell division (Lucas & Cervantes, 2002; Strickfaden et al., 2010). Therefore, analyzing only the mean position by considering their average position across the population of cells is not sufficient. The advanced microscopy imaging techniques have made significant contributions to capturing the variation present in spatial CT positions across the population, nonetheless their application has been limited to only few cell types possibly due to technical challenges and accessibility issues (Fritz et al., 2016; Gué et al., 2005; Iannuccelli et al., 2010).

Molecular biology techniques such as genome-wide chromosome conformation capture (Hi-C) on the other hand, capture the spatial organization of the genome at different resolutions ranging from enhancer-promoter to inter-chromosomal interactions, and are being applied to characterize the structural properties of the 3D genome in an ever-increasing manner across different cell-types and conditions (Rachel Patton McCord et al., 2020). Although the inter-chromosomal patterns do not capture the exact spatial position of the CTs inside the nucleus, the patterns can still infer the relative position of chromosomes with respect to one another. These relative positions can further be used to reconstruct the positioning of CTs inside the nucleus. Typically, inter-chromosomal interactions are more noisy and highly variable compared to intrachromosomal interactions. Recent improvements of Hi-C techniques, however, have been able to reduce the effect of noise on inter-chromosomal contacts, and as a result studies have started to discover relevant features of interchromosomal contacts (Golloshi et al., 2022; Horta et al., 2018; Xiong & Ma, 2019). For example, mouse olfactory receptor genes and enhancers form regulatory contacts between different chromosomes, and novel chromosomal

rearrangements in cancers can be detected by Hi-C interchromosomal contacts (Horta et al., 2018). Apart from these issues, the contact data also suffers from an inherent disadvantage which stems from the population-averaged nature of the experiments. The contacts measured by Hi-C between any two regions reflects the frequency of contacts occurring across a population of cells and as a result reconstructing the single-cell information becomes challenging (Ulianov et al., 2017).

In order to reconstruct the folding of individual chromosomes and the entire genome at different resolutions based on Hi-C contact and contact-derived data, a number of modeling techniques based on constraint and polymer physics have been devised (Di Pierro et al., 2017; Kalhor et al., 2012; Tjong et al., 2016; Zhang & Wolynes, 2017). Some models also consider external information in addition such as lamina association and histone marks, in order to make the model prediction more accurate, as multiple models can satisfy same set of contact constraints. Among those, however, only a few have explored the radial CT arrangements in their models (Agrawal et al., 2017; Paulsen et al., 2017; Tjong et al., 2016). For example, 3D genome models reconstructed by considering pairwise contacts obtained from the genome-wide conformation capture experiments as restraints have been able to capture the mean radial CT position consistent with the microscopy measurement (Tjong et al., 2016). This approach is further fine-tuned in another model, named Chrom3D, by considering the lamina association of chromatin region (Paulsen et al., 2017). Apart from these contact derived models, physics based approaches using different genome intrinsic and extrinsic properties (e.g., gene activity, chromatin looping and chromatin states) have been proposed to study the spatial and radial genome organization (Agrawal et al., 2020; Ganai et al., 2014). In Chapter 1, I present a computational approach to predict CT radial organization and heterogeneity from Hi-C contact data.

Single-cell chromosome conformations reveal heterogeneity present in chromosome structural ensemble

After focusing on the chromosome territorial organization and their cell-type-specific non-random distribution, now we will move on to the organization within individual chromosomes. As mentioned previously, chromosome structures exhibit both slow and fast dynamics due to the active force and thermal noise experienced inside the nucleus, which leads to chromosome structural heterogeneity (Hübner & Spector, 2010; Oliveira et al., 2021; Put et al., 2019). In the past two decades, numerous studies have been performed to study conformations of individual chromosomes at different resolutions ranging from euchromatin and heterochromatin separation to local topologically associated domain formation (Rachel Patton McCord et al., 2020). Among those studies, most of them focus on the population averaged structure of the chromosome which is obtained by considering structures across millions of cells in bulk. Analyzing the mean structural data can reveal the changes in chromosome conformation between different cell types and their association with biological functions like gene regulation. However, those bulk studies fail to capture the heterogeneity of spatial genome organization present in individual cells (Luppino & Joyce, 2020; Ulianov et al., 2017). In-silico modeling studies based on first principles and

experimental data have been proposed to characterize the heterogeneity present in chromosome structural ensemble (Di Pierro et al., 2017; Tjong et al., 2016). But, due to the problems associated with the experimental data as well as our incomplete understanding of chromatin folding mechanisms, it becomes challenging to know whether the structural variation captured by these models are indeed true for the biological systems.

To address these challenges, on the other hand, single-cell Hi-C and oligonucleotide-based fluorescence in situ hybridization (FISH) imaging techniques are being developed which have the ability to capture chromosome conformations at different resolution in single cells (Bintu et al., 2018; Su et al., 2020; Wang et al., 2016). By analyzing these single-cell data, researchers have started to uncover chromosome structural heterogeneity present in a population of cells. One such example is the structural heterogeneity of inactivated X-chromosome in human female cells (Cheng et al., 2021; Takei et al., 2021; Wang et al., 2016). It is known that in females, one copy of the chrX gets inactivated for the dosage compensation purpose. When the inactivated structure is characterized using bulk techniques, it shows the existence of two unstructured mega-domains separated by a satellite region (Nora et al., 2012). However, the analysis of single-cell data reveals that domain-like structures similar to the active chrX can still exist within the mega-domains of the inactive one at the single-cell level, which gets averaged out in bulk experiments. Furthermore, these studies have also found that the domains inter-mingle more among themselves compared to their active counterpart (Wang et al., 2016). In addition, studies have also started to delineate the folding principles of the chromatin regions (Cheng et al., 2020; Das et al., 2022; Xie et al., 2020). In one study, the structural transition of a short chromatin region between the open and closed conformations has been proposed to be linked with the transcriptional regulation of the genes present in that region (Cheng et al., 2020).

Though these studies focuses on fundamental principles behind the emergence of structural heterogeneity, they mostly ignore the extent of variation present in chromosome structural ensemble. It has been recently shown that during early embryonic development, *Caenorhabditis elegans* chromosomes exhibit a continuum of structures instead of a handful of discrete states (Sawh et al., 2020). Similar observations therefore can also be expected for individual chromosomes across a population of cells. But the question then becomes how that structural continuum differs from one cell type to another. Moreover, it is important to investigate what factors might contribute to cell type variations in structure heterogeneity. Therefore, in Chapter 2, I investigate the extent of variation present in chromosome structural ensemble and factors governing that.

Properties of chromosome structural ensemble and their transition during different processes

In addition to characterizing single-cell structural heterogeneity present in a population of cells, it is also important to have ways to quantify structural transitions within chromosomes between conditions at the ensemble level. During different physiological and

pathological conditions, the structural continuum gets altered and that eventually leads to the emergence of a new structural ensemble. Therefore, by analyzing the population level structure, the transition of structural ensembles can be characterized.

Biopolymers, such as chromosomes are semi-flexible and semi-structured in nature. This suggests that although the chromosomes have structural components, they do not have defined structures in the same way that proteins do. Analysis of bulk or population-averaged chromosome structure data at different resolutions reveals statistical properties which are found to be associated with different structural components of the chromosome (Lieberman-Aiden et al., 2009; Rachel Patton McCord et al., 2020). One such component is the compartmental organization of the individual chromosomes, where euchromatic A compartment regions segregates from the heterochromatic B compartment region. Studies have started to unravel the biological functionalities associated with chromatin compartmentalization. However, how the chromosomes fold into compartments and sustain the specific structure is still not well understood. Similar to other biopolymers, most likely, higher affinity of different types of chromatin regions for their like types give rise to the compartmental folding of the chromosome structures. For example, in the simplistic model, which considers only two types of chromatin: euchromatin and heterochromatin, the interaction between same types of chromatin regions leads to the separation of the B compartment regions from the other (Tiana & Giorgetti, 2019).

It is therefore important to characterize the properties of the chromosome structural ensemble for the understanding of the physical description of the chromosomes as well as their proper association with the biological functions. In order to analyze the semi-structured ensembles of biopolymers, researchers often use contacts (based on spatial proximity) between different parts of the polymer (Bryngelson & Wolynes, 1987; Go, 1983). Nevertheless, contact analysis is also useful for situations where biopolymer structure transforms from a structured to disordered state. In the case of chromosome structures, the contacts are obtained from chromosome conformation capture techniques (e.g., Hi-C) (Golloshi et al., 2018; R. P. McCord et al., 2020).

Statistical analysis of the contacts, specifically using principal component analysis (PCA) based methods has been extremely productive in characterizing the collective “modes” of the biopolymer ensembles (Johnson et al., 2018; Lindsay et al., 2018). One such technique is compartment analysis from the chromosome field, which has become a standard for characterizing the euchromatin and heterochromatin domains on chromosome structures (Lieberman-Aiden et al., 2009; Rachel Patton McCord et al., 2020). This compartment analysis performs distance decay normalization on the contact data to amplify the long range domain signatures present on a folded chromosome structure. As a result, the generic polymer aspect of the chromosome structure, such as the overall decaying of contacts with increasing sequence distance, is preemptively removed. However, chromosome structures are semi-structured and even contain disordered regions. Moreover, during the physicochemical perturbations to study chromosome structural and dynamical alterations, the biopolymers are often moved from a semi-structured state into unfolded state. Also, as

a cell progresses through the cell cycle, at certain points the chromosomes may behave as weakly interacting polymer (having low number of long-range interactions) approaching a generic polymer limit. Therefore, in Chapter 3, I present an approach to quantify statistical properties of ensemble chromosome structures and the degree to which they deviate from a generic polymer.

Role of nuclear lamina in genome organization and gene regulation

So far, we have focused on the contribution of within and between chromosomal interactions on the formation and dynamics of different large scale structural components. However, chromosomes do not exist in isolation inside the nucleus as there exist other nuclear substructures and bodies with which chromosomes interact (Quinodoz et al., 2018). Among those, the interactions with the nuclear lamina (NL), are of particular interest as these interactions help to properly organize the genome spatially and regulate the expression of the genes which are in contact or in the vicinity of lamina (Van Steensel & Belmont, 2017).

The NL beneath the inner nuclear membrane (INM) acts as a nucleoskeleton and primarily consists of type V intermediate filament proteins, called lamins (Shimi et al., 2008). These lamins are responsible for maintaining the nuclear structure-function relationship and non-random organization of genome inside the nucleus. The nuclear lamins are divided into two categories - type A and type B (Shimi et al., 2008). The type A lamins comprises of lamin A and C which are encoded by LMNA gene. On the other hand, LMNB1 and LMNB2 genes encode the type B lamins: lamin B1 and lamin B2 respectively. Each of these lamins forms their separate but interacting microdomains in the NL and show different mobilities in the nucleoplasm - nucleoplasmic lamin A/C is highly mobile whereas nucleoplasmic lamin B1/B2 are relatively immobile (Shimi et al., 2015; Shimi et al., 2008). Similar to their distinct localization, the expression of different types of lamins varies among cell types (Constantinescu et al., 2006; Kim et al., 2011; Melcer et al., 2012). While B-type lamins are constitutively expressed in all types of cells, A-type lamins show low abundance in stem cells and start to express as cells differentiate. Furthermore, lamins also regulate transcription, replication, DNA repair, and senescence by interacting directly or indirectly (through the chromatin binding proteins) with the chromatins (Mattout et al., 2006). Since lamin A/C has a direct association with chromatin, it is believed to be a contributor to the nuclear rigidity and chromatin stability and disruption of this A-type of lamina proteins make the nucleus more malleable (Lammerding et al., 2006).

The chromatin regions tethered to the nuclear envelope and lamina are known as lamina associated domains (LADs) (Van Steensel & Belmont, 2017). LADs are generally AT-rich, transcriptionally suppressed, mostly heterochromatic regions which correlate largely with Hi-C derived B compartment regions (Lochs et al., 2019; Rao et al., 2014; Robson et al., 2017). Although these LAD regions contain lower number of genes overall, they are specifically enriched for developmentally regulated genes (Guelen et al., 2008; Harr et al., 2015; Peric-Hupkes et al., 2010). As a cell commits to a specific lineage during

differentiation, the genes associated with other lineages get sequestered to the nuclear lamina for a stable repression. LADs are broadly categorized in two subgroups: genomic regions those remain lamina associated irrespective of cell-types (constitutive LAD; cLAD) and those which reorganize from one cell-type to another (facultative LAD; fLAD) (Lochs et al., 2019; Peric-Hupkes et al., 2010; Van Steensel & Belmont, 2017). The cLADs are mostly gene poor regions whereas the fLADs harbor genes which are active or poised to be activated. LADs are primarily enriched for different histone methylation marks, specifically H3K9me2 and H3K9me3 (Guelen et al., 2008; Harr et al., 2015; Poleshko et al., 2019). It has also been found that heterochromatin regions are sequestered to nuclear lamina by histone methyltransferases (See et al., 2020). Functionally, nuclear lamina is considered for their heterochromatic repressive nature. Generally ectopic tethering of a region which is actively transcribing to the lamina makes the region repressed and that results in the downregulation of expression values of the genes harbored there. For example, tethering of reporter gene to the nuclear lamina via lamin C leads to the repression of the gene (Dialynas et al., 2010). In another study, peripheral repositioning of *lac* operator array (*lacO*) via inner nuclear membrane protein exhibits ~2-3 fold changes in the reporter gene expression (Reddy et al., 2008). However, there exist evidence that shows local chromatin state and promoter activity can still decouple the gene regulation from peripheral positioning. In one report it is shown that peripheral tethering of a gene to lamina does not alter its kinetics of transcriptional induction (Kumaran & Spector, 2008). Other instances are also known where gene maintain their expression level irrespective of whether they are lamina associated or not (Forsberg et al., 2019). Similar findings have also been observed in the case of mouse liver gene where lamina association and dissociation based on circadian clock do not influence gene expression alteration (Brunet et al., 2019). In a recent study, upon systematic investigation it has been found that some promoters become repressed as they associate with the repressive nuclear lamina whereas some escapes that environment mostly harboring H3K27me3 facultative heterochromatin mark (Leemans et al., 2019).

Genomic regions that switch from A compartment to B often exhibit gene expression repression and the opposite trend can be seen for B-to-A switches (Bonev et al., 2017; Dixon et al., 2015; Miura et al., 2019; San Martin et al., 2021). However, upon systematic inspection of the genome, a significant number of genomic regions can be found where switch in genomic compartmentalization is incongruent to the lamina association change. And for those incongruent regions, how the gene expression is regulated depending on different active or repressive environments in a combinatorial fashion is not fully understood. It is still not known what effect would be observed for the genes which are part of the transcriptionally active euchromatin regions located near the nuclear periphery. In contrast, what effect will be observed by moving a heterochromatin region from the interior to the periphery, does it allow for more stable repression of the genes from that region? Furthermore, the studies to delineate the contribution of genome organization, lamina association on transcriptional regulation often focus on a specific type of cell. But, whether the contributions are cell-type-specific or cell-type-invariant, still needed to be investigated. I explore these questions through systematic analyses in Chapter 4.

Breast cancer metastasis organotropism and role of spatial genome organization in cancer

Till this point, we have primarily discussed structural properties and dynamics of the chromosomes in healthy states. Next, we will focus on how the reorganization of all these different structural components contribute to the dysregulation of genes during pathological conditions. Cancer is one such complex disease which involves the misregulation of homeostatic conditions at different cellular levels (Stratton et al., 2009). Several recent genome organization studies focusing on different types of cancers have revealed that the genesis of the gene signatures depends, in part, on the rearrangement of the 3D genome architecture at different hierarchical levels such as heterochromatin/euchromatin compartmentalization, topologically associated domains, and loops (Golloshi et al., 2022; Hnisz et al., 2016; Lebeau et al., 2022; San Martin et al., 2021). Therefore, it is important to understand alteration of the chromosome structural properties in the context of pathological conditions.

Metastasis is the major cause of death in all breast cancer-related diseases, and it results from the dissemination of cancer cells from the primary tumor site to a new site leading to the formation of the secondary tumor (Chaffer & Weinberg, 2011). However, the organs affected by breast cancer metastasis vary remarkably. This phenomenon is known as 'metastasis organotropism' (Chen et al., 2018). For instance, luminal subtypes of breast cancers show a higher percentage of bone metastasis whereas luminal B and basal subtype breast cancers preferentially metastasize to the lung (Savci-Heijink et al., 2016; Smid et al., 2008). The mechanism behind this organotropic behavior may be explained by the 'seed and soil' hypothesis, according to which a cancer cell (seed) having specific intrinsic abilities will only be able to survive in a local microenvironment (soil) that is favorable to it (Langley & Fidler, 2011). The cancer cells with these intrinsic abilities are known as 'pre-selected seeds'. This hypothesis is supported by the accumulating evidence suggesting that not only extrinsic factors such as circulation patterns, vascular wall accessibility, and organ-specific niche but also the intrinsic genotype and epigenotype of the cancer cells contribute to the organ tropism.

The process of organ colonization by primary tumor cells is a complex process and is achieved by the dynamic and evolving interactions between cancer cells and the surrounding microenvironment. In most situations, the microenvironment is not very welcoming, and the disseminated cells fail to survive in that soil. Despite the low survival rate of the cancer cells during the metastatic colonization process, the cells with the intrinsic capabilities to interact with the host microenvironment are able to survive and take over the host organ (Obenauf & Massagué, 2015). Based on recent studies, some relationship between the metastatic traits and gene signatures has been found in some cancers, but in most cases, the origin of these traits still poses as an outstanding question to the research community (Padua et al., 2008; Zhang et al., 2013). I investigate the hypothesis that the intrinsic properties that make a cancer cell suited for the metastatic microenvironment may be associated with the rearrangement of the 3D genome

organization at different hierarchical levels such that compartment switch, topologically associated domain fusion and aberrant looping.

Recent applications of the chromosome conformation capture techniques in cancer research have found that disruption of the genome organization has the potential to contribute to oncogenic transformation and progression (Achinger-Kawecka et al., 2020; Hnisz et al., 2016; San Martin et al., 2021). Among different layers of structural organization, disruption of topologically associated domains has primarily been found to be associated with gene dysregulation (Hnisz et al., 2016; Lebeau et al., 2022). For example, it has been shown that in acute lymphoblastic lymphoma, the elimination of boundary element can place oncogenes under the influence of regulatory element which otherwise are insulated in separate domains (Hnisz et al., 2016). In another study, it is reported that due to hypermethylation of the CTCF binding sites in gliomas, the overall CTCF binding gets reduced and which in turn leads to the loss of boundary insulation and aberrant oncogene activation (Flavahan et al., 2016). Similar results involving DNA hypermethylation and reduced CTCF binding have also been found in the case of gastrointestinal stromal cancers (Flavahan et al., 2019). Separately, working on a collaborative project from our lab, I found that genome-wide compartmental organization changes help explain prostate cancer progression (San Martin et al., 2021). In another collaborative project that characterized the genome organization changes associated with constricted migration of melanoma cells, I found that cells migrating through narrow constrictions for multiple times exhibit distinct compartmental organization compared to cells never migrated (Golloshi et al., 2022). Thus, it is important to examine the relationship between the 3D genome reorganization and the origin of the metastatic traits in addition to when and where the metastatic traits arise - whether at the primary tumor site or at the metastatic site.

Overall, this dissertation focuses on characterizing the structural and dynamical properties of the chromosome structures at different resolution and how these properties are getting altered due to different physiological and pathological conditions. In order to investigate these questions, I used different computational, mathematical, and physics-based modeling and simulation techniques in combination with graph theory, machine learning, and soft computing algorithms. In the first project, to study chromosome territorial organization and their non-random radial distribution, I developed a computational method process Hi-C contact data and genome intrinsic properties using principal component analysis and a network layout algorithm (Chapter 1). The following project then characterizes the degree of variation present in chromosome structural ensemble by analyzing the single-cell chromatin tracing data of different chromosomes for various cell types at different resolutions (Chapter 2). In the third project, the generic polymer aspect of the chromosome structures is explored to reveal the structural transition of chromosome structures under different physical perturbation or biological processes (Chapter 3). Next, Chapter 4 investigates the association between lamina association and genome compartmentalization and their contribution in gene expression regulation. And finally, Chapter 5 focuses on characterizing the relationship between three-dimensional genome reorganization and

different breast cancer subtypes organ-specific metastatic traits in breast cancer. In summary, this dissertation aims to address the properties of human chromosome structures and their alterations during different conditions.

REFERENCES

Achinger-Kawecka, J., Valdes-Mora, F., Luu, P.-L., Giles, K. A., Caldon, C. E., Qu, W., Nair, S., Soto, S., Locke, W. J., & Yeo-Teh, N. S. (2020). Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. *Nature communications*, *11*(1), 1-17.

Agrawal, A., Ganai, N., Sengupta, S., & Menon, G. I. (2017). Chromatin as active matter. *J Stat Mech Theory Exp*, *2017*. https://doi.org/10.1088/1742-5468/aa5287

Agrawal, A., Ganai, N., Sengupta, S., & Menon, G. I. (2020). Nonequilibrium biophysical processes influence the large-scale architecture of the cell nucleus. *Biophys J*, *118*. https://doi.org/10.1016/j.bpj.2019.11.017

Alberts, B. (2017). *Molecular biology of the cell*. WW Norton & Company.

Bednar, J., Horowitz, R. A., Grigoryev, S. A., Carruthers, L. M., Hansen, J. C., Koster, A. J., & Woodcock, C. L. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proceedings of the National Academy of Sciences*, *95*(24), 14173-14178.

Bercht Pfleghaar, K., Taimen, P., Butin-Israeli, V., Shimi, T., Langer-Freitag, S., Markaki, Y., Goldman, A. E., Wehnert, M., & Goldman, R. D. (2015). Gene-rich chromosomal regions are preferentially localized in the lamin b deficient nuclear blebs of atypical progeria cells. *Nucleus*, *6*. https://doi.org/10.1080/19491034.2015.1004256

Bintu, B., Mateo, L. J., Su, J.-H., Sinnott-Armstrong, N. A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A. N., & Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, *362*(6413), eaau1783. https://doi.org/10.1126/science.aau1783

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., & Speicher, M. R. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, *3*(5), e157.

Bonev, B., Cohen, N. M., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., & Tanay, A. (2017). Multiscale 3D genome rewiring during mouse neural development. *Cell*, *171*(3), 557-572. e524.

Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A., & Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, *10*(3), 211-220.

Bridger, J., Boyle, S., Kill, I., & Bickmore, W. (2000). Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts. *Curr Biol*, *10*. https://doi.org/10.1016/S0960-9822(00)00312-2

Brunet, A., Forsberg, F., Fan, Q., Sæther, T., & Collas, P. (2019). Nuclear lamin B1 interactions with chromatin during the circadian cycle are uncoupled from periodic gene expression. *Frontiers in genetics*, *10*, 917.

Bryngelson, J. D., & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, *84*(21), 7524-7528.

Chaffer, C. L., & Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science*, *331*(6024), 1559-1564.

Chen, W., Hoffmann, A. D., Liu, H., & Liu, X. (2018). Organotropism: new insights into molecular mechanisms of breast cancer metastasis. *NPJ precision oncology*, *2*(1), 1-12.

Cheng, R. R., Contessoto, V. G., Aiden, E. L., Wolynes, P. G., Di Pierro, M., & Onuchic, J. N. (2020). Exploring chromosomal structural heterogeneity across multiple cell lines. *ELife*, *9*, e60312.

Cheng, Y., Liu, M., Hu, M., & Wang, S. (2021). TAD-like single-cell domain structures exist on both active and inactive X chromosomes and persist under epigenetic perturbations. *bioRxiv*, 2021.2005.2012.443887. https://doi.org/10.1101/2021.05.12.443887

Constantinescu, D., Gray, H. L., Sammak, P. J., Schatten, G. P., & Csoka, A. B. (2006). Lamin A/C expression is a marker of mouse and human embryonic stem cell differentiation. *Stem cells*, *24*(1), 177-185.

Cremer, T., & Cremer, M. (2010). Chromosome territories. *Cold Spring Harbor Perspect Biol*, *2*. https://doi.org/10.1101/cshperspect.a003889

Cremer, T., Kurz, A., Zirbel, R., Dietzel, S., Rinke, B., Schröck, E., Speicher, M. R., Mathieu, U., Jauch, A., & Emmerich, P. (1993). Role of chromosome territories in the functional compartmentalization of the cell nucleus. Cold spring harbor symposia on quantitative biology,

Das, P., Shen, T., & McCord, R. P. (2022). Characterizing the variation in chromosome structure ensembles in the context of the nuclear microenvironment. *PLoS computational biology*, *18*(8), e1010392.

Di Pierro, M., Cheng, R. R., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, *114*(46), 12126-12131.

Dialynas, G., Speese, S., Budnik, V., Geyer, P. K., & Wallrath, L. L. (2010). The role of Drosophila Lamin C in muscle function and gene expression. *Development*, *137*(18), 3067-3077.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., & Xie, W. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331-336.

Finn, E. H., & Misteli, T. (2019). Molecular basis and biological function of variability in spatial genome organization. *Science*, *365*(6457), eaaw9498.

Flavahan, W. A., Drier, Y., Johnstone, S. E., Hemming, M. L., Tarjan, D. R., Hegazi, E., Shareef, S. J., Javed, N. M., Raut, C. P., & Eschle, B. K. (2019). Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature*, *575*(7781), 229-233.

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., & Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, *529*(7584), 110-114.

Forsberg, F., Brunet, A., Ali, T. M. L., & Collas, P. (2019). Interplay of lamin A and lamin B LADs on the radial positioning of chromatin. *Nucleus*, *10*(1), 7-20.

Fritz, A. J., Barutcu, A. R., Martin-Buley, L., Van Wijnen, A. J., Zaidi, S. K., Imbalzano, A. N., Lian, J. B., Stein, J. L., & Stein, G. S. (2016). Chromosomes at work: organization of chromosome territories in the interphase nucleus. *Journal of cellular biochemistry*, *117*(1), 9-19.

Fritz, A. J., Stojkovic, B., Ding, H., Xu, J., Bhattacharya, S., Gaile, D., & Berezney, R. (2014). Wide-scale alterations in interchromosomal organization in breast cancer cells: defining a network of interacting chromosomes. *Human Mol Genet*, *23*. https://doi.org/10.1093/hmg/ddu237

Ganai, N., Sengupta, S., & Menon, G. I. (2014). Chromosome positioning from activity-based segregation. *Nucl Acids Res*, *42*. https://doi.org/10.1093/nar/gkt1417

Go, N. (1983). Theoretical studies of protein folding. *Annual review of biophysics and bioengineering*, *12*(1), 183-210.

Das, P., Shen, T., & McCord, R. P. (2022). Characterizing the variation in chromosome structure ensembles in the context of the nuclear microenvironment. *PLoS computational biology*, *18*(8), e1010392.

Golloshi, R., Playter, C., Freeman, T. F., Das, P., Raines, T. I., Garretson, J. H., Thurston, D., & McCord, R. P. (2022). Constricted migration is associated with stable 3D genome structure differences in cancer cells. *EMBO reports*, e52149.

Golloshi, R., Sanders, J. T., & McCord, R. P. (2018). Iteratively improving Hi-C experiments one step at a time. *Methods*, *142*. https://doi.org/10.1016/j.ymeth.2018.04.033

Gué, M., Messaoudi, C., Sun, J. S., & Boudier, T. (2005). Smart 3D-FISH: automation of distance analysis in nuclei of interphase cells by image processing. *Cytom Part A J Int Soc Anal Cytol*, *67*. https://doi.org/10.1002/cyto.a.20170

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., De Klein, A., Wessels, L., & De Laat, W. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, *453*(7197), 948-951.

Harr, J. C., Luperchio, T. R., Wong, X., Cohen, E., Wheelan, S. J., & Reddy, K. L. (2015). Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins. *Journal of Cell Biology*, *208*(1), 33-52.

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. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, *351*(6280), 1454-1458.

Horta, A., Monahan, K., Bashkirova, E., & Lomvardas, S. (2018). Cell type-specific interchromosomal interactions as a mechanism for transcriptional diversity. *bioRxiv*, 287532.

Hübner, M. R., & Spector, D. L. (2010). Chromatin dynamics. *Annual review of biophysics*, *39*, 471.

Iannuccelli, E., Mompart, F., Gellin, J., Lahbib-Mansais, Y., Yerle, M., & Boudier, T. (2010). NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-fish experiments. *Bioinformatics*, *26*. https://doi.org/10.1093/bioinformatics/btq013

Johnson, Q. R., Lindsay, R. J., & Shen, T. (2018). CAMERRA: an analysis tool for the computation of conformational dynamics by evaluating residue–residue associations. In: Wiley Online Library.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., & Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, *30*. https://doi.org/10.1038/nbt.2057

Kim, Y., Sharov, A. A., McDole, K., Cheng, M., Hao, H., Fan, C.-M., Gaiano, N., Ko, M. S., & Zheng, Y. (2011). Mouse B-type lamins are required for proper organogenesis but not by embryonic stem cells. *Science*, *334*(6063), 1706-1710.

Kumaran, R. I., & Spector, D. L. (2008). A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *The Journal of cell biology*, *180*(1), 51-65.

Kuroda, M., Tanabe, H., Yoshida, K., Oikawa, K., Saito, A., Kiyuna, T., Mizusawa, H., & Mukai, K. (2004). Alteration of chromosome positioning during adipocyte differentiation. *J Cell Sci*, *117*. https://doi.org/10.1242/jcs.01508

Lammerding, J., Fong, L. G., Ji, J. Y., Reue, K., Stewart, C. L., Young, S. G., & Lee, R. T. (2006). Lamins A and C but not lamin B1 regulate nuclear mechanics. *Journal of Biological Chemistry*, *281*(35), 25768-25780.

Langley, R. R., & Fidler, I. J. (2011). The seed and soil hypothesis revisited—The role of tumor-stroma interactions in metastasis to different organs. *International journal of cancer*, *128*(11), 2527-2535.

Lebeau, B., Jangal, M., Zhao, T., Wong, C. K., Wong, N., Cañedo, E. C., Hébert, S., Aguilar-Mahecha, A., Chabot, C., & Buchanan, M. (2022). 3D chromatin remodeling potentiates transcriptional programs driving cell invasion. *Proceedings of the National Academy of Sciences*, *119*(36), e2203452119.

Leemans, C., van der Zwalm, M. C., Brueckner, L., Comoglio, F., van Schaik, T., Pagie, L., van Arensbergen, J., & van Steensel, B. (2019). Promoter-intrinsic and local chromatin features determine gene repression in LADs. *Cell*, *177*(4), 852-864. e814.

Lieberman-Aiden, E., Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., & Dorschner, M. O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*. https://doi.org/10.1126/science.1181369

Lindsay, R. J., Pham, B., Shen, T., & McCord, R. P. (2018). Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic acids research*, *46*(16), 8143-8152.

Lochs, S. J., Kefalopoulou, S., & Kind, J. (2019). Lamina associated domains and gene regulation in development and cancer. *Cells*, *8*(3), 271.

Lucas, J., & Cervantes, E. (2002). Significant large-scale chromosome territory movement occurs as a result of mitosis, but not during interphase. *Int J Radiat Biol*, *78*. https://doi.org/10.1080/09553000110097190

Luppino, J. M., & Joyce, E. F. (2020). Single cell analysis pushes the boundaries of TAD formation and function. *Current opinion in genetics & development*, *61*, 25-31.

Mattout, A., Dechat, T., Adam, S. A., Goldman, R. D., & Gruenbaum, Y. (2006). Nuclear lamins, diseases and aging. *Current opinion in cell biology*, *18*(3), 335-341.

McCord, R. P., & Balajee, A. (2018). 3D genome organization influences the chromosome translocation pattern. *Chromosome Translocation*, 113-133.

McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Molecular cell*, *77*(4), 688-708.

McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Mol Cell*, *77*. https://doi.org/10.1016/j.molcel.2019.12.021

Mehta, I. S., Eskiw, C. H., Arican, H. D., Kill, I. R., & Bridger, J. M. (2011). Farnesyltransferase inhibitor treatment restores chromosome territory positions and active chromosome dynamics in Hutchinson-Gilford progeria syndrome cells. *Genome biology*, *12*(8), 1-14.

Melcer, S., Hezroni, H., Rand, E., Nissim-Rafinia, M., Skoultchi, A., Stewart, C. L., Bustin, M., & Meshorer, E. (2012). Histone modifications and lamin A regulate chromatin protein dynamics in early embryonic stem cell differentiation. *Nature communications*, *3*(1), 1-12.

Misteli, T. (2020). The self-organizing genome: principles of genome architecture and function. *Cell*, *183*(1), 28-45.

Miura, H., Takahashi, S., Poonperm, R., Tanigawa, A., Takebayashi, S.-i., & Hiratani, I. (2019). Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nature genetics*, *51*(9), 1356-1368.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., & Sedat, J. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381-385.

Obenauf, A. C., & Massagué, J. (2015). Surviving at a distance: organ-specific metastasis. *Trends in cancer*, *1*(1), 76-91.

Oliveira, G. M., Oravecz, A., Kobi, D., Maroquenne, M., Bystricky, K., Sexton, T., & Molina, N. (2021). Precise measurements of chromatin diffusion dynamics by modeling using Gaussian processes. *Nature communications*, *12*(1), 1-11.

Padua, D., Zhang, X. H.-F., Wang, Q., Nadal, C., Gerald, W. L., Gomis, R. R., & Massagué, J. (2008). TGFβ primes breast tumors for lung metastasis seeding through angiopoietin-like 4. *Cell*, *133*(1), 66-77.

Paulsen, J., Sekelja, M., Oldenburg, A. R., Barateau, A., Briand, N., Delbarre, E., Shah, A., Sørensen, A. L., Vigouroux, C., & Buendia, B. (2017). Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol*, *18*. https://doi.org/10.1186/s13059-016-1146-2

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R. M., & van Lohuizen, M. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, *38*(4), 603-613.

Poleshko, A., Smith, C. L., Nguyen, S. C., Sivaramakrishnan, P., Wong, K. G., Murray, J. I., Lakadamyali, M., Joyce, E. F., Jain, R., & Epstein, J. A. (2019). H3K9me2 orchestrates inheritance of spatial positioning of peripheral heterochromatin through mitosis. *ELife*, *8*, e49278.

Put, S., Sakaue, T., & Vanderzande, C. (2019). Active dynamics and spatially coherent motion in chromosomes subject to enzymatic force dipoles. *Physical Review E*, *99*(3), 032421.

Quinodoz, S. A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J. M., Detmar, E., Lai, M. M., Shishkin, A. A., Bhat, P., & Takei, Y. (2018). Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, *174*(3), 744-757. e724.

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. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*. https://doi.org/10.1016/j.cell.2014.11.021

Reddy, K., Zullo, J., Bertolino, E., & Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, *452*(7184), 243-247.

Robson, M. I., Jose, I., Czapiewski, R., Sivakumar, A., Kerr, A. R., & Schirmer, E. C. (2017). Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome research*, *27*(7), 1126-1138.

Rosin, L. F., Crocker, O., Isenhart, R. L., Nguyen, S. C., Xu, Z., & Joyce, E. F. (2019). Chromosome territory formation attenuates the translocation potential of cells. *ELife*, *8*. https://doi.org/10.7554/eLife.49553

San Martin, R., Das, P., Dos Reis Marques, R., Xu, Y., Roberts, J. M., Sanders, J. T., Golloshi, R., & McCord, R. P. (2021). Chromosome compartmentalization alterations in prostate cancer cell lines model disease progression. *Journal of Cell Biology*, *221*(2), e202104108.

Savci-Heijink, C. D., Halfwerk, H., Koster, J., & van de Vijver, M. J. (2016). A novel gene expression signature for bone metastasis in breast carcinomas. *Breast cancer research and treatment*, *156*(2), 249-259.

Sawh, A. N., Shafer, M. E., Su, J.-H., Zhuang, X., Wang, S., & Mango, S. E. (2020). Lamina-dependent stretching and unconventional chromosome compartments in early C. elegans embryos. *Molecular cell*, *78*(1), 96-111. e116.

Schoenfelder, S., & Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, *20*(8), 437-455.

See, K., Kiseleva, A. A., Smith, C. L., Liu, F., Li, J., Poleshko, A., & Epstein, J. A. (2020). Histone methyltransferase activity programs nuclear peripheral genome positioning. *Developmental biology*, *466*(1-2), 90-98.

Shimi, T., Kittisopikul, M., Tran, J., Goldman, A. E., Adam, S. A., Zheng, Y., Jaqaman, K., & Goldman, R. D. (2015). Structural organization of nuclear lamins A, C, B1,

and B2 revealed by superresolution microscopy. *Molecular biology of the cell*, *26*(22), 4075-4086.

Shimi, T., Pfleghaar, K., Kojima, S.-i., Pack, C.-G., Solovei, I., Goldman, A. E., Adam, S. A., Shumaker, D. K., Kinjo, M., & Cremer, T. (2008). The A-and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. *Genes & development*, *22*(24), 3409-3421.

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G., Foekens, J. A., & Martens, J. W. (2008). Subtypes of breast cancer show preferential site of relapse. *Cancer research*, *68*(9), 3108-3114.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719-724.

Strickfaden, H., Zunhammer, A., Koningsbruggen, S., Köhler, D., & Cremer, T. (2010). 4d chromatin dynamics in cycling cells: Theodor Boveri's hypotheses revisited. *Nucleus*, *1*.

Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B., & Zhuang, X. (2020). Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, *182*(6), 1641-1659. e1626.

Sun, H. B., Shen, J., & Yokota, H. (2000). Size-dependent positioning of human chromosomes in interphase nuclei. *Biophysical journal*, *79*(1), 184-190.

Takei, Y., Zheng, S., Yun, J., Shah, S., Pierson, N., White, J., Schindler, S., Tischbirek, C. H., Yuan, G.-C., & Cai, L. (2021). Single-cell nuclear architecture across cell types in the mouse brain. *Science*, eabj1966.

Tanabe, H., Habermann, F. A., Solovei, I., Cremer, M., & Cremer, T. (2002). Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutat Res Fundam Mol Mech Mutagenes*, *504*. https://doi.org/10.1016/S0027-5107(02)00077-5

Tanabe, H., Küpper, K., Ishida, T., Neusser, M., & Mizusawa, H. (2005). Inter-and intra-specific gene-density-correlated radial chromosome territory arrangements are conserved in old world monkeys. *Cytogenet Genome Res*, *108*. https://doi.org/10.1159/000080824

Tiana, G., & Giorgetti, L. (2019). *Modeling the 3D conformation of genomes*. CRC Press.

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., & Gros, M. A. (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci*, *113*. https://doi.org/10.1073/pnas.1512577113

Ulianov, S. V., Tachibana-Konwalski, K., & Razin, S. V. (2017). Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. *BioEssays*, *39*(10), 1700104.

Van Steensel, B., & Belmont, A. S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, *169*(5), 780-791.

Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C.-t., & Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, *353*(6299), 598-602.

Xie, W. J., Qi, Y., & Zhang, B. (2020). Characterizing chromatin folding coordinate and landscape with deep learning. *PLoS computational biology*, *16*(9), e1008262.

Xiong, K., & Ma, J. (2019). Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat Commun*, *10*. https://doi.org/10.1038/s41467-018-07882-8

Zhang, B., & Wolynes, P. G. (2017). Genomic energy landscapes. *Biophys J*, *112*. https://doi.org/10.1016/j.bpj.2016.08.046

Zhang, X. H.-F., Jin, X., Malladi, S., Zou, Y., Wen, Y. H., Brogi, E., Smid, M., Foekens, J. A., & Massagué, J. (2013). Selection of bone metastasis seeds by mesenchymal signals in the primary tumor stroma. *Cell*, *154*(5), 1060-1073.

CHAPTER I
INFERRING CHROMOSOME RADIAL ORGANIZATION FROM HI-C DATA

A version of this chapter was originally published by Priyojit Das, Tongye Shen and Rachel Patton McCord:

Das, P., Shen, T., & McCord, R. P. (2020). Inferring chromosome radial organization from Hi-C data. BMC bioinformatics, 21(1), 1-28.

My contributions to this work include: (1) conceived the project, (2) performed computational modeling and analysis, (3) wrote the manuscript, and (4) prepared the figures.

Abstract

The nonrandom radial organization of eukaryotic chromosome territories (CTs) inside the nucleus plays an important role in nuclear functional compartmentalization. Increasingly, chromosome conformation capture (Hi-C) based approaches are being used to characterize the genome structure of many cell types and conditions. Computational methods to extract 3D arrangements of CTs from this type of pairwise contact data will thus increase our ability to analyze CT organization in a wider variety of biological situations. A number of full-scale polymer models have successfully reconstructed the 3D structure of chromosome territories from Hi-C. To supplement such methods, we explore alternative, direct, and less computationally intensive approaches to capture radial CT organization from Hi-C data. We show that we can infer relative chromosome ordering using PCA on a thresholded inter-chromosomal contact matrix. We simulate an ensemble of possible CT arrangements using a force-directed network layout algorithm and propose an approach to integrate additional chromosome properties into our predictions. Our CT radial organization predictions have a high correlation with microscopy imaging data for various cell nucleus geometries (lymphoblastoid, skin fibroblast, and breast epithelial cells), and we can capture previously documented changes in senescent and progeria cells. Our analysis approaches provide rapid and modular approaches to screen for alterations in CT organization across widely available Hi-C data. We demonstrate which stages of the approach can extract meaningful information, and also describe limitations of pairwise contacts alone to predict absolute 3D positions.

Introduction

The three dimensional (3D) structure of the human genome is composed of different structures at different length scales. At smaller length scales, nucleosome positions, loops, and topologically associating domains are the most salient features, followed by compartmentalization at a longer length scale (McCord et al., 2020). At the largest scale of this genome organization, the 3D bodies of individual chromosomes arrange mostly as discrete entities, known as chromosome territories (CTs) (Cremer & Cremer, 2010). The arrangement of all the CTs inside the nucleus with respect to nuclear center and periphery forms the higher-order genome architecture. This CT organization is nonrandom with respect to the nucleus periphery and can play important roles in different nuclear mechanisms ranging from DNA replication and gene expression to the processing of RNA (Cremer et al., 1993). Recently, it has been shown that chromosome territorial organization can also protect genome from deleterious rearrangements during DNA damage (Rosin et al., 2019). Alterations in CT organization can also be important during cell differentiation (Kuroda et al., 2004) and in different disease conditions. For example, in Hutchinson–Gilford progeria syndrome cells, chr18 shifts toward the nucleus interior as compared to its position in normal proliferating fibroblast cells (Mehta et al., 2011), while certain gene-rich chromosomes localize near the periphery in blebs in progeria cells (Bercht Pfleghaar et al., 2015). Alterations in CTs can influence the likelihood of chromosomal translocations. For example, during adipogenesis, chr12 and chr16 become spatially proximal, increasing the chance of translocation between those chromosomes, which is the driving event of liposarcoma tumorigenesis (Kuroda et al., 2004). Recently, researchers have shown that in breast cancer, a gain in inter-chromosomal interactions for chrX correlates with its gene expression changes (Fritz et al., 2014). Approaches to characterize CT positions can thus further our understanding of the implications of CT arrangements in health and disease.

From careful microscopic measurements over the past several decades, largely using sequence specific probes in fluorescence in situ hybridization (FISH), principles of CT organization in certain cell types have been identified. It has been observed that CTs are often organized according to one of two different distributions: either a gene density or chromosome length based pattern (Boyle et al., 2001; Sun et al., 2000). Opposing forces of gene activity and lamina associated domain (LAD) density on different chromosomes likely contribute to these different distributions (Amendola & Steensel, 2014; Van Steensel & Belmont, 2017). In general, LADs are likely to be repressive to gene activity and occur more frequently on gene poor chromosomes, which also tend to be longer than gene rich chromosomes (Guelen et al., 2008). The proliferation rate of cells and nuclear shape have also been implicated as factors influencing CT organization. For example, proliferating cells tend to follow a gene density-based organization compared to the length-based distribution in quiescent or senescent cells (Bridger et al., 2000). Further, the spherical human lymphocyte nucleus follows a gene density based organization, which is conserved across several related species (Tanabe et al., 2002; Tanabe et al., 2005), while the chromosome size based distribution is more prevalent in ellipsoidal fibroblast nuclei

(Bolzer et al., 2005). Our understanding of the relationships between factors influencing CT distribution are limited, however, by the fact that relatively few different cell types have been characterized in depth by this type of microscopic analysis.

Thorough analysis of CT positions requires not only observing their average, but also the distribution of possible positions across the cell population. The position of each chromosome varies between cells in the population, even while following certain tendencies (Kind et al., 2015). Throughout interphase the CT positions remain stable, but change from one generation to another during mitosis as the nuclear envelope is broken down and re-established (Kind et al., 2013; Lucas & Cervantes, 2002; Strickfaden et al., 2010; Walter et al., 2003). Improvements to microscopy experiments (Shachar et al., 2015; Solovei & Cremer, 2010) and image analysis methods make sampling this variation in CT position across the population increasingly feasible (Gué et al., 2005; Iannuccelli et al., 2010; Ollion et al., 2013), yet such analyses of all CT positions in a large number of cells exist for only few cell types (Fritz et al., 2016).

In contrast, genome wide chromosome conformation capture (Hi-C) approaches are being applied to characterize the 3D genome structure of rapidly increasing numbers of cell types and conditions for the past decade (Dekker et al., 2017; Golloshi et al., 2018; Lieberman-Aiden et al., 2009; McCord et al., 2020). This Hi-C technique and its variants capture a snapshot of pairwise chromosomal interactions ranging from specific enhancer-promoter interactions (Schoenfelder et al., 2018) to large scale inter-chromosomal interactions. Due to the effect of noise and high variability, the inter-chromosomal interactions have received less attention compared to their intra-chromosomal counterparts. In a past few years, with the improvement of experimental protocols, the effect of noise on inter-chromosomal contacts has been reduced and studies based on relevant inter-chromosomal interactions have started to emerge (Maass et al., 2019; Pal et al., 2019; Xiong & Ma, 2019; Zhang et al., 2018). For example, genes corresponding to olfactory receptors from different chromosomes form specific inter-chromosomal contacts in mouse olfactory sensory neurons which strengthen upon differentiation from a progenitor cell (Horta et al., 2018). In an another set of studies, by analyzing Hi-C inter-chromosomal contacts obtained from different malignant diseases, researchers have identified several novel chromosomal rearrangements (Golloshi et al., 2022; Steininger et al., 2018). Though Hi-C does not directly capture the radial position of the CTs, approaches that infer 3D chromosome positioning information from Hi-C contact data provide a valuable supplement to microscopic data, greatly increasing the number of cell types for which CT positions can be analyzed. In this study, we explore a set of Hi-C analysis approaches focused on rapid and efficient prediction of CT radial organization, ranging from very simple direct calculations on the Hi-C contact matrix to a network model tuned by additional chromosome property information.

Many approaches have been developed to reconstruct the 3D folding of individual chromosomes and the genome from Hi-C data at different resolutions using restraint and polymer physics based approaches (Meluzzi & Arya, 2020; Oluwadare et al., 2019). Some

models focus on detailed structures of local regions of chromatin rather than the whole genome, and for others, the primary focus is often to understand the mechanistic principles underlying the organization of the interphase and metaphase genome rather than a prediction of CT arrangement (Chiariello et al., 2016; Gibcus et al., 2018; Kalhor et al., 2012; Pierro et al., 2016; Serra et al., 2017; Wettermann et al., 2020; Zhang & Wolynes, 2017). But, some of the approaches have also explored the radial arrangement of CTs in their 3D models. An early restraint-based model of the 3D genome based on tethered chromosome conformation capture (TCC) data predicted CT positions that agreed with the major principles of lymphoblast genome organization characterized by microscopy (Kalhor et al., 2012). In an another work, Hi-C maps were probabilistically deconvolutated into a population of single cell structures, and the averaged radial position of the CTs predicted from those structures matched fairly well with microscopic measurements from a single cell type (Tjong et al., 2016). A series of coarse-grained polymer simulation based studies have been performed to characterize non-random organization of the chromosomes using gene activity and random and biological looping constraints (Agrawal et al., 2017, 2020; Ganai et al., 2014). In a recent polymer modeling based study, the researchers used Hi-C derived properties and a chromatin state based energy function to study the principles of the spatial as well as radial genome organization (Qi et al., 2020).

Because numerous arrangements can potentially be consistent with a set of Hi-C contacts, in many cases, whole genome 3D models generated from Hi-C also have to take into account external information in order to increase the accuracy of CT positioning predictions. For example, Stevens et al. combine imaging and Hi-C contacts on single cells to orient their 3D chromosome models (Stevens et al., 2017). Similarly, the Chrom3D algorithm combines LAD data along with Hi-C data to capture spatial and radial organization of the chromosomes (Paulsen et al., 2017). Di Stefano et al. used a steered molecular dynamics based approach to reconstruct diploid genome organization from Hi-C data (Stefano et al., 2016). Using significant Hi-C interactions as the constraints, the modeling technique was able to capture the preferential nuclear position of different genomic regions based on their gene density, lamina association and epigenetic marks.

To supplement this landscape of approaches to predict 3D genome conformations from Hi-C data, we have several specific goals in this study. We describe and test direct, computationally non-intensive analysis approaches that have the focused aim of inferring CT radial positions from Hi-C data rather than a full model of 3D chromosome folding. Specifically, we demonstrate that radial organization patterns can be inferred from PCA analysis of thresholded inter-chromosomal contact matrices and show the utility of a force-directed graph layout algorithm to infer the average and variation around the average CT positions. These approaches thus do not require the computational resources necessary to calculate high resolution polymer models, and could be used to screen for potentially important differences in CT organization across a wide range of cell types and conditions without creating detailed 3D models in each case. We further evaluate the strengths and limitations of Hi-C contact data when it is used toward the goal of inferring large scale 3D positions of chromosomes, and where additional reference information needs to be added

to the contacts to generate reliable radial positioning information. We finally evaluate the different stages of our approach on a variety of cell types and conditions, comparing to a variety of published microscopic data and predicting additional details of CT organization where limited microscopy data exists.

Results

Thresholding contacts extracts meaningful chromosomal interaction patterns from Hi-C data

Hi-C experiments capture spatial genome organization by measuring the interaction frequency between different genomic fragments, yielding information about both intra-chromosomal and inter-chromosomal interactions. Inter-chromosomal interactions generally occur much less frequently than intra-chromosomal interactions, and true interactions are mixed in with noise that arises from random background ligation (Golloshi et al., 2018; Lajoie et al., 2015). Despite this inherent noise and sometimes low signal, the interactions between chromosomes also contain information that reflects the radial organization of the chromosome territories inside the nucleus. But, excluding contacts that may primarily reflect the background is important to prevent those contacts from masking the true signal. In order to extract strong interactions that are more likely to distinguish radial chromosome positions between cell types, we applied a thresholding technique to the genome-wide contact matrix (see "Methods" section). The thresholding cutoff h_{cut} value was calculated by taking a certain percentile of all the genome-wide Hi-C interactions and then the interactions greater than this cutoff limit are considered as strong interactions. To determine the desired cutoff value, we compared a Hi-C matrix from a specific cell type with a corresponding simulated random ligation matrix (see "Methods" section) for different values of h_{cut} ranging from the 5th to 95th percentile. We examined two different cell types with different nuclear shapes (Appendix: Table 1), since the nuclear shape has been observed to correlate to some extent with the non-random radial organization of the CTs. The human blood lymphoblastoid cell, GM12878, has a spherical nucleus and follows a gene density based radial CT organization (Boyle et al., 2001). On the other hand, BJ1-hTERT human skin fibroblast cell has an ellipsoidal nucleus, which shows a chromosome length based CT organization (Bolzer et al., 2005). The whole genome Hi-C contact matrices were obtained from Sanders et al. (Sanders et al., 2019).

For each h_{cut} value, the number of chromosomal contact bin pairs passing the threshold were summed between each pair of chromosomes to a single bin (Eqn. 4).

$$CHR_{ij}^{STRONG} = total\ number\ of\ strong\ interaction\ bins\ between\ chr\ i\ and\ j \quad (1)$$

Then, Pearson's correlation was applied to compare the whole chromosome pair interaction sums obtained from the original and corresponding random ligation Hi-C data. From Fig. 3a, it can be seen that with the increase in h_{cut} values from the 45th to the 85th percentile, the pairwise strong chromosomal interaction similarity between random and real data

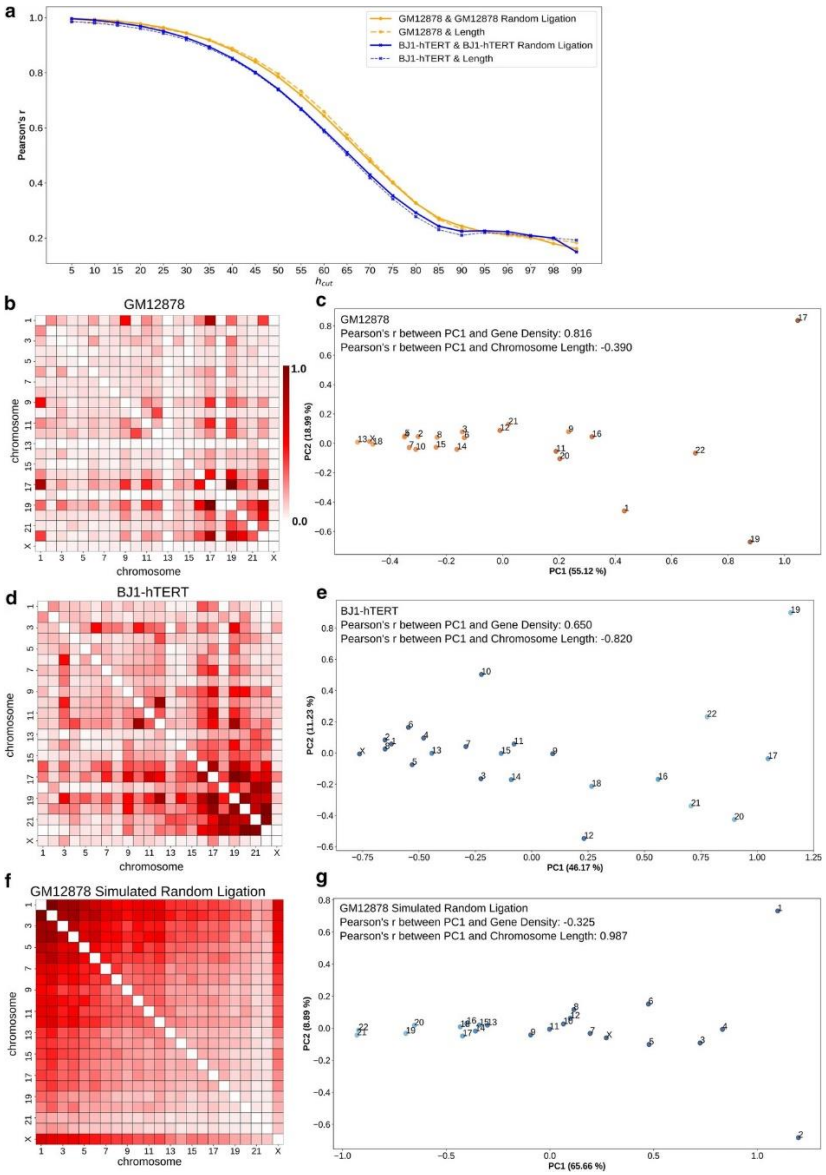


Figure 3 Determination of strong contact threshold and inference of radial CT distribution type.

a Correlation of pairwise chromosomal strong interaction matrix obtained from original Hi-C with the matrix from corresponding randomly ligated Hi-C (solid lines) and chromosome length (dotted lines) at different values of $\boldsymbol{h_{cut}}$. b, d, f Pairwise inter-chromosomal strong interaction pattern matrix for GM12878 (b), BJ1-hTERT (d) and GM12878 simulated random ligation (f) Hi-C data respectively. Each entry of the matrix represents the total number of strong interacting bins between a pair of chromosomes. c, e, g 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrices obtained from GM12878 (c), BJ1-hTERT (e) and GM12878 simulated random ligation (g) Hi-C data respectively. Correlations of the PC1 projection with gene density and chromosome length (in bp) are shown. Chromosomes are color coded based on their gene density (orange) or length (blue). Dark color = high gene density/large chromosome size and light color = low gene density/short chromosome length.

decreased rapidly and reached a stable value around the 90th percentile. Similarly, we compared the strong chromosomal interaction sums with the pairwise product of the chromosome lengths to measure how much the number of interactions was primarily driven by chromosome length (i.e. two large chromosomes will have more interactions at random overall than two small chromosomes). This analysis produced a similar correlation trend as the random ligation effect comparison: strong interaction sums are no longer primarily explained by chromosome length at 90th–95th percentile h_{cut} (Fig. 3a). Based on these two comparison results for both GM12878 and BJ1-hTERT, we chose the 95th percentile as the final value of h_{cut} that leads to a minimized effect of chromosome length and random ligation for both cell types. While this optimal value was similar for two different datasets we considered, we note that Hi-C library complexity, read depth, and cis/trans ratio could affect the most appropriate h_{cut} value. As an alternative to this thresholding approach, we also explored the FitHiC (Ay et al., 2014) algorithm to extract significant chromosomal interactions from the Hi-C data. The analysis results obtained using those interactions are discussed in the following subsection.

Radial chromosome ordering can be inferred from PCA on inter-chromosomal strong interaction pattern matrix

The pairwise chromosomal strong interaction pattern is not only able to distinguish the true biological interaction pattern from the random ligation, but also reveals distinct patterns specific to each cell type. By looking at the inter-chromosomal component of the pairwise strong interaction patterns (Eq. 5) for GM12878 and

$$CHR_{ij}^{TRANS} = CHR_{ij}^{STRONG} \text{ if } i \neq j \text{ otherwise } 0 \quad (2)$$

BJ1-hTERT represented in Fig. 3b, d respectively, we can clearly see that these two different cell types have distinct patterns—in BJ1-hTERT, the smaller chromosomes have higher strong inter-chromosomal interactions among themselves, whereas in GM12878 that pattern is dispersed. In order to capture the major interaction trends between chromosomes from those distinct patterns, we applied principal component analysis (PCA) to these pairwise strong interaction matrices. PCA has been frequently applied to contacts within or between chromosomes to detect the spatial segregation of A/B compartments (Das et al., 2020; Imakaev et al., 2012; Lieberman-Aiden et al., 2009; Lindsay et al., 2018). Further, the A/B compartment status of bins within a chromosome has been found to correlate with their lamina association and radial positioning, and therefore has been used in models predicting the lamin associations of domains within a chromosome (Luperchio et al., 2018). Here, by contrast, we are not detecting the compartment status of regions within a chromosome but instead applying PCA to the pattern of thresholded pairwise interactions between whole chromosome territories and then examining the projection of chromosomes onto the first two principal components. We find that PCA on the pattern of strong contacts between all pairs of chromosomes can detect the spatial segregation and relative ordering of chromosome territories. (Here we note that since the Hi-C matrix represents the average of the two copies of each chromosome, we use chromosome (chr)

and chromosome territory (CT) interchangeably, noting that the CT positions will represent the average of the chromosome locations of each homolog across the cell population). Figure 3c shows the 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrix for the GM12878 cell. From this figure, it can be seen that chromosome 17, 19 and 22 and chromosome 13, 18 and X are on the opposite ends along the PC1 axis due to the high dissimilarity in their inter-chromosomal interaction patterns. In addition to that, this separation also correlates highly with the gene density based distribution of the chromosomes, as gene-rich chromosomes (e.g., chr19, chr17) are on the right extreme and gene-poor chromosomes (e.g., chr18, chr13) on the other extreme. This ordering thus corresponds to previous reports showing that lymphoblast cells with spherical nuclei tend to position their chromosomes in the nucleus in a gene density associated pattern (Boyle et al., 2001). Indeed, the PC1 values of the chromosomes show a higher absolute value correlation with their gene density than with their chromosome length. Next, we analyzed the pairwise inter-chromosomal significant interaction pattern matrix obtained using FitHiC for the GM12878 Hi-C data. We again see a higher PC1 correlation with gene density than chromosome length, but the correlation is much weaker (Appendix: Fig. 10b). This is not surprising given that the interaction pattern in the FitHiC chromosome pair interaction map is less distinct and more uniform across chromosomes (Appendix: Fig. 10a). Therefore, we choose to proceed with the strong interaction thresholding approach for our remaining analyses to detect the inherent radial arrangement of the CTs.

When we applied PCA to the BJ1-hTERT pairwise inter-chromosomal strong interaction pattern matrix, PC1 showed a different separation, where chromosomes are ordered from left to right roughly according to decreasing length (Fig. 3e). The strong correlation between this ordering and chromosome length matches previous observations from fibroblast nuclei (Bolzer et al., 2005). There is still some gene density correlation with this pattern, likely reflecting the complex picture of the radial CT organization in fibroblast nuclei. It has indeed been reported that while chromosomes are generally radially positioned by length in fibroblast nuclei, CT18 (gene poor and short) is still nearer to the nuclear envelope than CT19 (gene rich and short). When we applied PCA to the simulated random-ligation matrices generated from the GM12878 and BJ1-hTERT, we find that chromosomes are ordered along PC1 in a strong length-based distribution (Fig. 3g and Appendix: Fig. 10d). This is expected, since by chance longer chromosomes will have more random strong interactions than shorter chromosomes, as is evident in Fig. 3f.

The above results suggest that a simple application of PCA to the pairwise inter-chromosomal strong pattern matrix can reveal chromosome spatial distribution type, while also showing the limitations of this direct use of Hi-C contacts alone. Though PC1 can infer the radial CT distribution type, the direction of the ordering (whether inwards to outwards or outwards to inwards) is arbitrary and cannot be inferred based only this PCA result. Further, the random ligation result provides a caution that a length-based radial distribution would be detected even when no distinct interaction patterns are present.

To further explore the utility of such PCA ordering of CT positions, we next applied this analysis technique to Hi-C data from conditions which have been previously shown to exhibit large-scale rearrangement of chromosome territories. Specifically, in the premature aging disease Hutchinson–Gilford Progeria syndrome, it has been shown that chr18 moves to the nuclear interior compared to its peripheral location in normal proliferating fibroblasts, while chr10 shows the opposite trend (Meaburn, 2016; Mehta et al., 2011). To check whether our analysis technique can also detect these CT reorganizations, we analyzed WI38-hTERT proliferating fibroblast Hi-C data (Chandra et al., 2015) and Hi-C data from progeria patient fibroblasts at passage 19, when the cells are approaching senescence (McCord et al., 2013). Figure 4a, b show the 2D PCA projections of the pairwise inter-chromosomal strong interaction pattern matrices obtained from the proliferating fibroblast and the progeria cells respectively. As mentioned above, while these projections reflect the underlying CT organization, inferring the directionality requires additional experimental information. There is evidence that chrX does not change its peripheral position in progeria cells compared to normal proliferating fibroblasts (Mehta et al., 2011). Indeed, in our analyses, chrX is positioned at a far extreme of PC1 in both progeria and proliferating fibroblasts, so we assigned the end of the PC1 axis near chrX as the periphery and the opposite end as the nucleus center, as seen in Fig. 4a, b. Now, if we examine the positions of chr18 in these projections, we can see that chr18 occupies a more internal position in the progeria cell (Fig. 4a, b). We can visualize the relative changes in chromosome positions by ordering the chromosomes based on the PC1 value in an increasing fashion from center to periphery in a radar plot (Fig. 4d). Here, we can see the internal shift of chr18 in progeria compared to the proliferating fibroblast. We did not find any significant change in the chr10 position in our analysis. However, we noticed that chr13 in progeria also moves to interior similar to chr18, which can be found in proliferating laminopathy fibroblasts (Meaburn, 2016). Progeria cells approaching senescence have a CT arrangement that to some extent matches other quiescent and senescent cells (Mehta et al., 2011). Therefore, next, we performed PCA ordering analysis on WI38-hTERT oncogene induced senescent cell Hi-C data (Chandra et al., 2015). When we compared the result with normal proliferating cells, we found chr10, chr13, chr18 and chrX radial rearrangements similar to the progeria cells (Fig. 4c, d bottom). Overall, this analysis shows that the strong inter-chromosomal interaction pattern from Hi-C has the potential to infer changes in the underlying CT distribution within the nucleus.

Network modeling predicts probabilistic radial CT organization

Though applying a simple PCA approach to a pairwise inter-chromosomal strong interaction pattern matrix is surprisingly powerful at detecting relative chromosome ordering in different cell types, as shown above, it is limited to predicting a single position for each chromosome, and will predict ordering patterns even for data that is actually random noise. In reality, chromosome territory positions vary substantially within individual cells, even while following certain general trends. The positioning of CTs in individual cells changes from mother cell to daughter cell during mitosis (Lucas & Cervantes, 2002; Walter et al., 2003). Thus, we sought an approach to model not only the

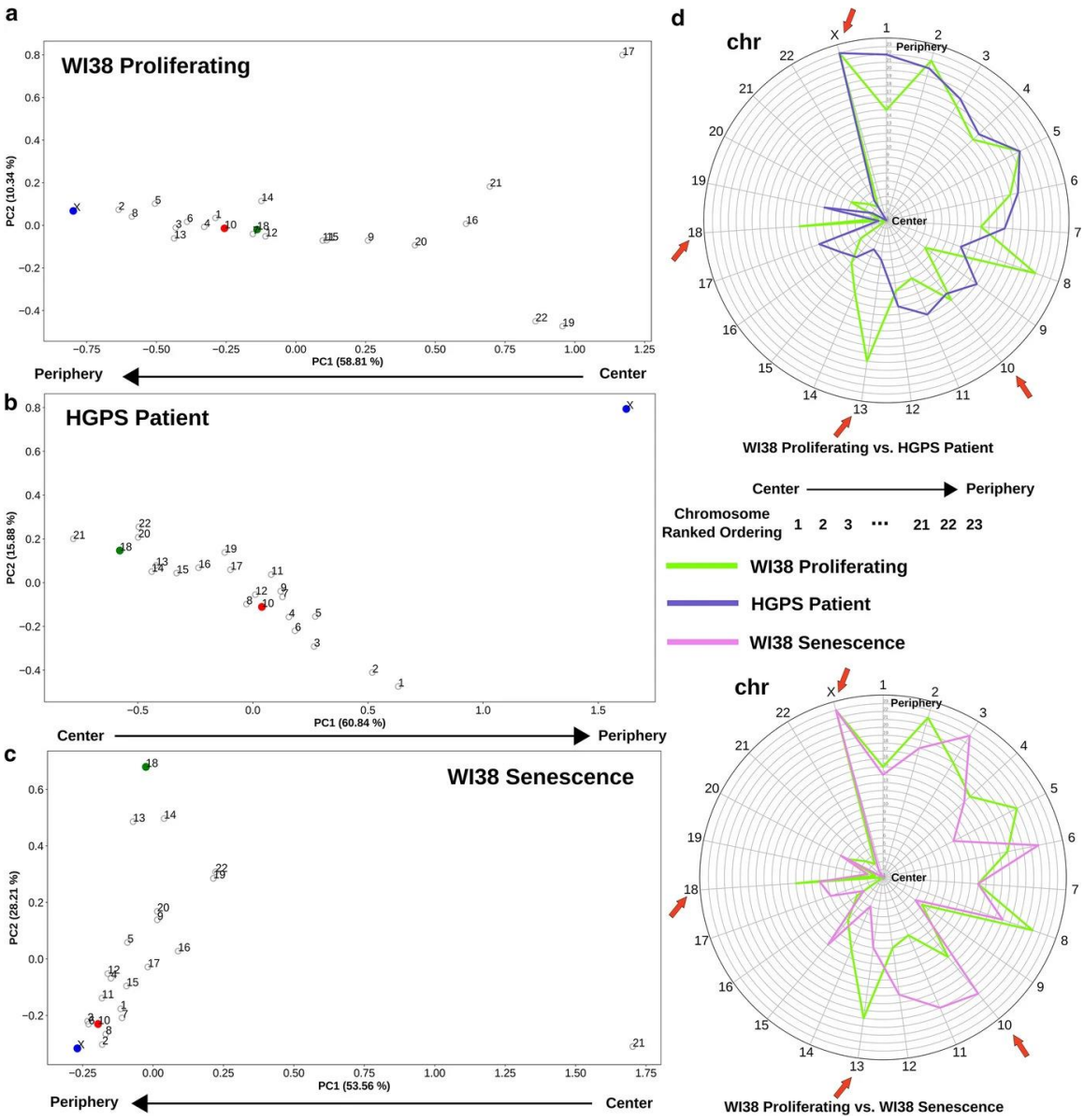


Figure 4 Rearrangement of radial CT position in progeria and senescent cells compared to proliferating fibroblasts.

a–c 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrices obtained from WI38 proliferating (a), Hutchinson–Gilford progeria syndrome patient passage 19 (b) and WI38 senescent fibroblast (c) Hi-C data respectively. In progeria and senescent cells, chr10 (red color) and chr18 (green color) change their radial positions compared to the proliferating cells (chr10 change is not significant in progeria), while chrX maintains its peripheral position in all of the cases. d Radar plot representing the change in the radial order of the CTs in progeria and senescence compared to the proliferating condition. Chromosome radial positions are shown as ordered (ranked) in an increasing fashion from the center to periphery. Salient chromosomes discussed in the text are highlighted with red arrows.

average position, but the variability of chromosome positions in an ensemble of arrangements derived from Hi-C data. For this, we tested a network modeling approach. We used the strong interaction matrix obtained from the thresholding step to construct an undirected weighted graph in three-dimensional space using the 3D Fruchterman-Reingold (FR) force-based layout (Fruchterman & Reingold, 1991) algorithm. Each node of the graph represents a genomic region of fixed size and the weights are the contact frequencies. In such a graph layout, the nodes within each chromosome become clustered together due to high intra-chromosomal contacts while chromosomes that have higher inter-chromosomal contacts among themselves will end up close together in 3D space compared to the other chromosomes with less inter-chromosomal contacts. To model the variable CT arrangements possible within the average Hi-C data, 1000 independent runs of the FR algorithm were performed with different random initial configurations, which led to the formation of 1000 different 3D network graphs (see "Methods" section). Next, a minimum volume ellipsoid algorithm was used to fit a geometrical object around the models to serve as the nuclear periphery. From each of those models, we calculated the distance of the center of mass of each chromosome from the nucleus center (center of the geometrical fitted object).

When we examined the models, we found that they cluster into several different possible chromosome organization patterns (Appendix: Fig. 12). As we discovered with the PCA analysis, the Hi-C data alone cannot determine the absolute ordering from interior to periphery, so some clusters of models represent inverted patterns of organization. We also find some clusters of models that do not coincide with the length or gene density based distributions inferred from PCA CT ordering. These may represent local minima reached from certain initial conditions of the network model which do not reflect the true chromosome arrangements. Given these observations, we combine information from the previous PCA analysis with these network models to select the cluster of models for further analysis. After observing that the PCA analysis reports whether the cell type in general follows a length or gene density-based radial distribution, we use this information to select the cluster of models which has the highest absolute correlation between its mean CT distances and gene density or chromosome length (based on the inferred CT distribution type from the pairwise inter-chromosomal PCA transformation).

With the selected cluster of models (Fig. 5a, b), we next examined the predicted heterogeneity in chromosome territory positioning within each cell type (Fig. 5 and Appendix: Fig. 13). Figure 5c shows the radial distance profiles of four example CTs - CT1, CT18, CT19 and CT20 obtained from GM1878 and BJ1-hTERT network model clusters. Among these CTs, CT18 (length 78.1 Mbp and gene density 3.4 genes/sequenced Mbp) and CT19 (length 59.1 Mbp and gene density 23.9 genes / sequenced Mbp) have comparable DNA content but have drastically different gene densities. The network modeling result shows a trend in which, in GM12878, CT18 has a peak near the periphery and CT19 has a more internal peak. On the other hand, in BJ1-hTERT, both CT18 and CT19 peaks are located toward the nuclear center with a similar distribution. The next contrasting pair is CT1 (length 249.3 Mb and gene density 8.70

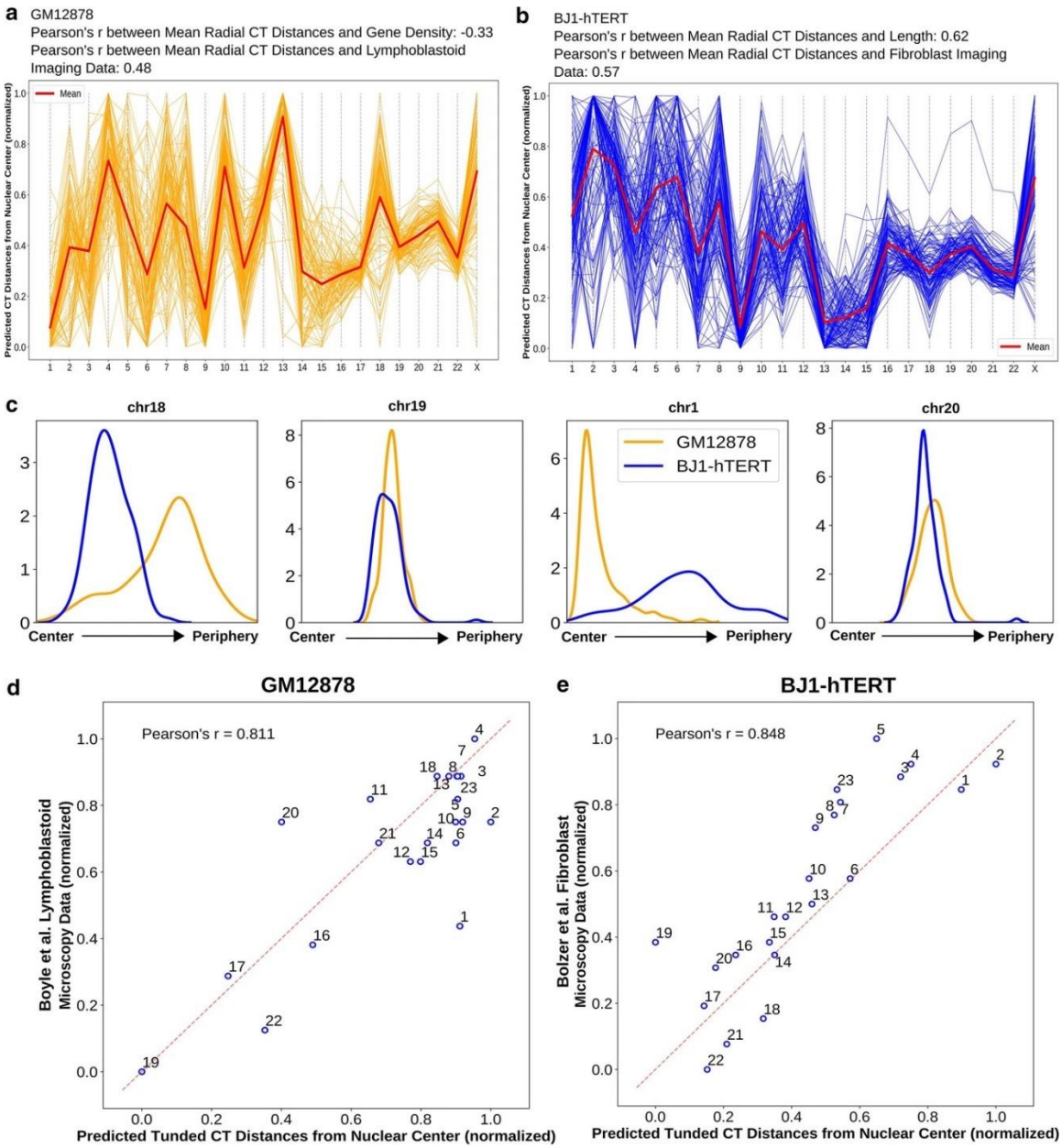


Figure 5 Network modeling generated distributions of CT radial positions and comparison of the predicted tuned result with microscopy data.

a, b Network modeling generated model clusters for both GM12878 (a) and BJ1-hTERT (b), selected based on respective inferred CT distribution types. c The radial distance profiles of 4 CTs (chr18, chr19, chr1 and chr20) for both GM12878 and BJ1-hTERT obtained from the respective selected model clusters (# of models in the selected clusters: for GM12878—109, and BJ1-hTERT—122). d Correlation between predicted tuned CT distance from GM12878 Hi-C data and lymphoblastoid microscopy imaging data. e Correlation between predicted tuned CT distance from BJ1-hTERT Hi-C data and fibroblast microscopy imaging data.

genes/sequenced Mb) and CT20 (length 63.0 Mb and gene density 8.71 genes/sequenced Mb). They both have comparable gene densities but strikingly different lengths. Again based on modeling results, we can see that CT1 and CT20 have internal locations in GM12878, but in BJ1-hTERT CT1 occupies a much peripheral location. As expected given the overall gene density or length correlations of the model cluster, these mean positions are consistent with microscopy data regarding the different positioning of chr18, chr19, and chr1 in different cell types. Beyond mean position shifts, we also observe that the distributions of possible chromosome positions also match microscopic evidence for chromosome position variation in some respects. For example, in the models, CT4 in lymphoblasts is highly skewed toward peripheral positions and almost never observed near the center of the nucleus. This matches the distribution of positions of a gene located on chr4 measured by FISH (Gué et al., 2005) in which the gene was almost never observed in the interior 30 percent of the nuclear radius in any individual cell (Appendix: Fig. 14d). In contrast, the models predict that CT10 in fibroblasts can be found throughout the middle of the nucleus, but almost never at the extreme center or periphery (Appendix: Fig. 14c). This matches the measured chr10 distribution by FISH (Mehta et al., 2010). In some cases, however, the model-inferred CT distribution matches the mean position observed in microscopy, but not the distribution of positions. For example, the models locate chr19 near the nucleus center on average in both cell types we considered, as in microscopy data (Appendix: Fig. 14a, b). However, in single cell images, chr19 is found over a broader range of positions (Cremer et al., 2001), while our models predict a more tightly focused consistent positioning of chr19. This narrow predicted distribution likely stems from the highly specific interaction pattern for chr19, visible in the strong contact matrix and distinct PCA position of this chromosome in Fig. 3. Additionally, while chr13 has been observed by microscopy to be strongly skewed toward the nuclear periphery (Mehta et al., 2010; Mora et al., 2006), our model sometimes predicts a strongly peripheral distribution (GM12878) and sometimes a strongly internal distribution (BJ1-hTERT) (Appendix: Fig. 13). This occurs because there are relatively few strong inter-chromosomal interactions detected by chr13 in the Hi-C data, and the model cannot distinguish whether this means the chromosome is located on its own far to the center or far to the periphery. So, while our network model approach shows a strong ability to predict chromosome position variability for some chromosomes, in other cases, there are inherent limitations of what can be predicted from Hi-C contacts alone.

In contrast to the original Hi-C contact maps, simulated random ligation maps produced network models that were much more variable and did not form tight clusters (Appendix: Fig. 11a, e). This primarily happens due to non-specific pairwise inter-chromosomal interaction pattern driven by length which ultimately leads to the generation of wide variety of random network model configurations. Although these simulated random ligation datasets show a strong length based distribution of CTs from the PCA ordering, the pairwise inter-chromosomal strong interaction pattern is quite different from true length based distribution as shown by ellipsoidal cells. When we examine the distributions of chromosome positions predicted for these random models, most chromosomes showed broad indistinct distributions that did not vary based on the cell type the random matrix

was derived from (Appendix: Fig. 15). This demonstrates that real Hi-C contacts contribute important information to our network modeling predicted radial distance distributions.

Chromosome property-based tuning improves predicted consensus radial arrangement of CTs

Although the network modeling generated radial distribution of the CTs can reveal several interesting features of the CT organization that match with prior observations, the direct correlation with microscopy measured mean positions is only modest (Fig. 5a, b). Meanwhile, the direct correlation between gene density or chromosome length and microscopy measured positions may be quite high (Appendix: Table 2), but these values are cell type invariant and can never be used alone to infer different CT positions between different cell types. Thus, we next tested an approach of using chromosomal properties, informed by PC1 ordering, to further tune the network model inferred mean radial positions. We add the effects of gene density and chromosome length to the averaged positions using weighted averaging and loess (locally estimated scatterplot smoothing) (Cleveland, 1979) techniques. This tuning procedure is described in detail in the "Methods" section, but the basic idea behind it involves several concepts. First, we incorporate information about the length and gene density of each chromosome explicitly. Second, going back to our earlier observation that the PC1 of the pairwise inter-chromosomal strong interaction pattern matrix can provide meaningful ordering information, we combine the network model output with PC1 ordering result obtained from selected model cluster through weighted averaging. Finally, in calculating weights for our averaging calculation, we take into account how much of the variance in the chromosome positioning pattern in the selected model cluster is captured by PC1 and PC2. We find this metric captures a major distinction between random and real Hi-C data. In the previous steps, random data will give some pattern, often highly chromosome length related, but we find that PC1 and PC2 explain a much lower percentage of the overall variance in the chromosome contact pattern for random data, while for real data these first two PCs capture most of the variance, indicating that the orderings along these PCs are highly meaningful. Thus, weighting by the variance explained by PC1 and PC2 will emphasize meaningful patterns over random patterns.

In order to evaluate the accuracy of the tuning procedure, we compared the tuned radial CT positions of GM12878 and BJ1-hTERT with corresponding microscopy imaging data, as shown in Fig. 5d, e. From Fig. 5d, it can be seen that, in GM12878, most of the CT positions correlate well with the experimentally obtained position with slight displacements, apart from the CT1 position, which moves outwards compared to the experimental data. Similarly, predicted CT distances for the BJ1-hTERT cell shows high similarity with the corresponding imaging data. For this particular cell, out of all the CTs, CT19 showed a higher amount of displacement towards the center in the predicted result. When the tuning technique was tested on simulated random ligation Hi-C data generated from GM12878 and BJ1-hTERT, it produced a far weaker correlation with the fibroblast imaging result in both cases (length based inferred CT distribution type) compared to the

original Hi-C analysis results (Appendix: Fig. 11b, f). Furthermore, we checked the consistency of the predicted tuned results between two Hi-C replicates of the BJ1-hTERT cell and found very little difference in chromosome positioning (Appendix: Fig. 16f). In case of GM12878, variation between replicates was higher (Appendix: Figs. 17f and 18), perhaps reflecting the larger variation in quality metrics between these two datasets. However, a strength of employing the PCA ordering is revealed in comparing these replicates: the PCA ordering of the strong interaction matrix is robust to such variations in dataset quality (Appendix: Fig. 17d).

Paternal and maternal homologs show mostly similar radial arrangement of CTs

Our modeling approach places only one copy of each chromosome in the radial organization network, and we assume that this represents the average of the two homolog positions. In any given cell, it is known from microscopy that the CT positions of homologs can be quite different (Cremer et al., 2001), but we assume that both homologs would follow the same overall distribution of positions, and thus averaging their positions is reasonable. To test this assumption, we applied our approach to the deeply sequenced GM12878 Hi-C data from Rao et al. that can be mapped to maternal and paternal homologs based on allele-specific single nucleotide polymorphisms (SNPs) (Rao et al., 2014). Once we had the genome wide Hi-C contact matrices for the paternal and maternal copies, thresholding was applied and pairwise inter-chromosomal strong interaction pattern matrix was analyzed using PCA transformation for each of them. Figure 6c shows the 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrix obtained from the paternal copy. By looking at this figure, we can clearly see that the paternal homologs are following a gene density driven distribution along the PC1 axis, which we also found for the maternal homologs (Fig. 6d y-axis). In addition, the PC1 values from the paternal and maternal copies are highly correlated to each other and also with the PC1 values obtained from the previous GM12878 data (Fig. 6d, e). This suggests that the paternal and maternal homologs follow a similar radial distribution inside the nucleus, and that it is fair to infer a single average position for both homologs. We also applied the network modeling approach to the paternal and maternal copies individually and selected the model cluster for each of them whose mean CT distances has highest absolute correlation with the inferred CT distribution type, as above. By looking at the mean CT distances from Fig. 6f, g, we can see that the mean radial CT distribution pattern between two copies are highly comparable with a single drastic exception in case of chr2, which shows an internal position in maternal copy and occupies a peripheral position in the paternal one. Again this can be explained by distinct patterns of inter-chromosomal strong interaction of chr2 paternal and maternal homologs. Finally, when we compared the radial distance profiles of the 4 CTs - CT1, CT18, CT19 and CT20 among the paternal, maternal and the standard GM12878, we found the density peaks in similar radial positions (Fig. 6h and Appendix: Fig. 19). However, in case of chr1, we found a strong narrow peak from the standard GM12878 data compared to the short broad peak obtained from both the paternal and maternal copies, which might arise due to the less specific inter-

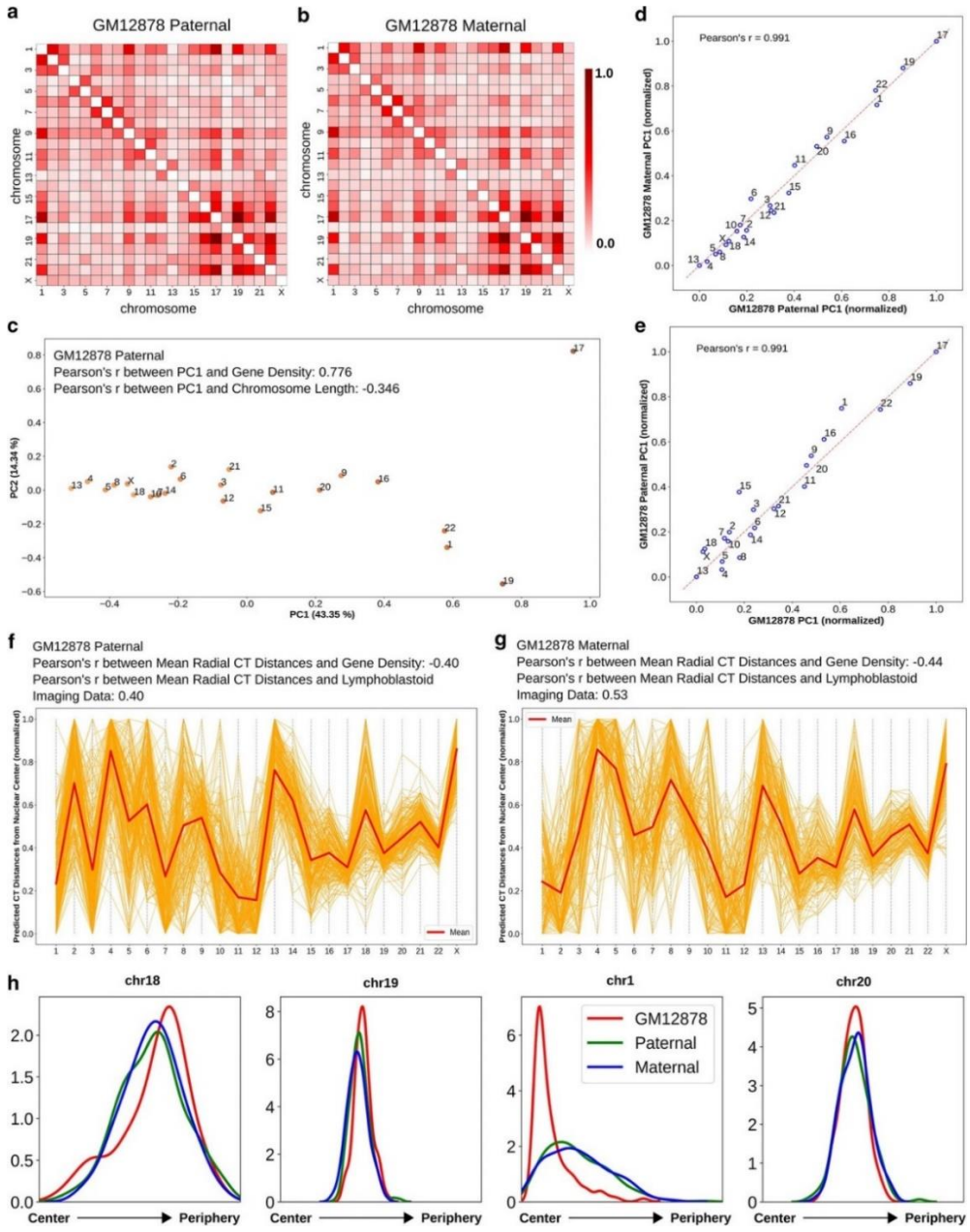


Figure 6 Paternal and maternal homologs follow a similar CT distribution pattern.

a, b Pairwise inter-chromosomal strong interaction pattern matrices for GM12878 paternal (a) and maternal (b) copies. c 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrix for GM12878 paternal copy. d, e Correlation of GM12878 paternal copy PC1 values with maternal copy (d) and diploid averaged GM12878 (e) PC1 values obtained from the PCA transformation of the respective pairwise inter-chromosomal strong interaction pattern matrices. f, g Network modeling generated model clusters for both GM12878 paternal (f) and maternal (g) copies, selected based on respective inferred CT distribution types. h Predicted radial distance distributions of 4 CTs (chr18, chr19, chr1 and chr20) for GM12878 diploid averaged, paternal and maternal copies, obtained from the respective selected model clusters (# of models in the selected clusters: GM12878 diploid average—109, paternal—194, and maternal—148).

chromosomal interaction pattern of chr1 in both maternal and paternal copies compared to the interaction patterns of rest of the chromosomes in those copies. Overall, we observe that representing both homologs with an average predicted position is valid.

Radial CT organization in epithelial cells changes depending on their nuclear shape

After testing the performance of the analysis technique on the GM12878 and BJ1-hTERT cells, we applied our analysis to MCF10A non-tumorigenic breast epithelial cells. The Hi-C contact data of this cell was downloaded from Barutcu et al. (Barutcu et al., 2015) and binned at 2.5 Mb resolution. We obtained corresponding imaging data for 8 CTs (CT1, CT4, CT11, CT12, CT15, CT16, CT18 and CT21) of the MCF10A from Fritz et al. (Fritz et al., 2014). When we applied PCA to the pairwise inter-chromosomal strong interaction matrix, PC1 showed higher absolute correlation with chromosome length compared to gene density, as expected from experimental data (Fig. 7b), and corroborating the association of the length based distribution with ellipsoidal nucleus shape. Next, we used the network modeling approach to generate radial distance profiles of the 8 CTs and found the distributions match the ordering of CT positions reported in imaging data (Fig. 7c). We note that we predict some chromosomes (chr4, chr15) have a much broader distribution than others (chr16, chr21), but this distribution is not reported for comparison in microscopy data. Finally, when the tuned CT positions were compared with the experimental data, a high correlation was obtained (Fig. 7d).

Since epithelial cells can take different nuclear shapes depending on their properties in culture, we next applied our analysis to Hi-C data from another human normal mammary epithelial cell (HMEC) (Rao et al., 2014). In this cell type, the PCA analysis of the pairwise inter-chromosomal strong interaction pattern inferred a gene density based distribution as represented in Fig. 7e. We hypothesize that this is related to the fact that these cells are closer to normal epithelium than MCF10A, which harbor chromosomal translocations, and that classical epithelial patterns would lead to a more spherical nucleus shape even in 2D culture, rather the flat, spreading growth pattern of MCF10A cells (Nandakumar et al., 2012). When the network modeling inferred radial distance profiles of four CTs - CT1, CT18, CT19 and CT20 were compared between MCF10A and HMEC (Fig. 7f), CT19 and CT20 showed density peaks at a similar radial position in both cell types. On the other hand, CT18 showed a preference for interior radial positions in MCF10A (length based) and for peripheral positions in HMEC (gene density based) and in case of CT1 that trend was the opposite. These results suggest that similar cell types can show different CT organization that correlates with their cell morphology, and Hi-C contact patterns can capture these differences.

Variation between the radial arrangement of lymphoblastoid cells and neutrophils

To test the utility of our analyses on a cell having an irregular nuclear shape, we focused on neutrophils and obtained the corresponding Hi-C data from Javierre et al. (Javierre et al., 2016). Neutrophils have a multi-lobed nucleus with a toroid shaped genome where

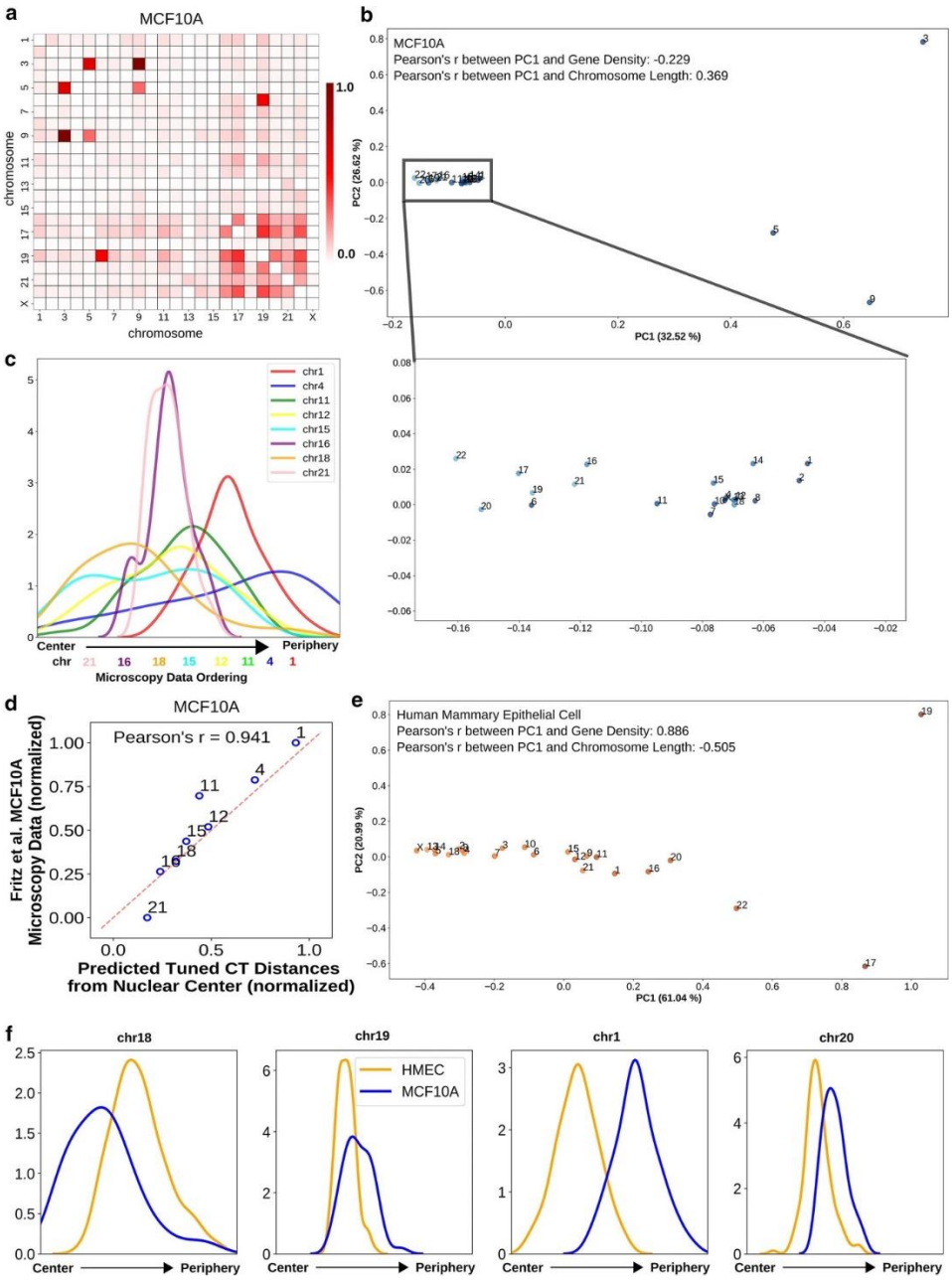


Figure 7 Epithelial cells show different radial CT organization in a nuclear morphology dependent manner.

a Pairwise inter-chromosomal strong interaction pattern matrix for MCF10A Hi-C data. b, e 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrices obtained from MCF10A (b) and HMEC (e) Hi-C data respectively. c The radial distance profiles of 8 CTs (chr1, chr4, chr11, chr12, chr15, chr16, chr18 and chr21) for MCF10A obtained from the selected model cluster (# of models in the selected cluster: 48). Relative mean chromosome positions measured by microscopy shown below graph. d Correlation between predicted tuned CT distance from Hi-C data and microscopy imaging data for MCF10A cells. f The radial distance profiles of 4 CTs (chr18, chr19, chr1 and chr20) for both MCF10A and HMEC obtained from the respective selected model clusters (# of models in the selected clusters: for HMEC: 165).

lobes are connected by thin filaments (Zhu et al., 2017). When we analyzed the pairwise inter-chromosomal strong interaction pattern matrix using PCA, we found that the PC1 values have a higher absolute correlation with gene density (Fig. 8a, b), though the effect is much weaker compared to GM12878 case, which is consistent with the divergence of a neutrophil from its precursor cell's round nucleus shape. This is another indication that the Hi-C contact data provides more information about chromosome positioning than just the underlying length based or gene density based distribution. Based on microscopy imaging data, it has been reported that chr2 and chr18, which have drastic differences in their lengths, both occupy a position near nuclear envelope in neutrophil cells (Sanchez et al., 1997). Upon inspecting the positions of those chromosomes along the PC1 axis in Fig. 8b, we found the two chromosomes in very close proximity near the left extreme of the PC1 axis. Although projection data does not infer the directionality of the ordering, based on its higher similarity with gene density, we can assume the left extreme as the periphery which contains mostly gene poor chromosomes. Furthermore, for those two chromosomes, we also observed similar trend in the network model-derived radial CT distance profiles as represented in Fig. 8c. The network model predicts a highly variable positioning of chr2, likely related to the overall dearth of strong contacts between chr2 and other chromosomes in the initial thresholded contact map. We observe that the predicted tuned CT positions for neutrophils are only weakly correlated with lymphoblastoid imaging data (Fig. 8d), revealing the ability of the model to predict different chromosome positions from different initial Hi-C contact data.

Discussion

In this study, our goal has not been only to predict one final set of CT positions from Hi-C data. Instead, we have explored what aspects of 3D chromosome radial positioning can and cannot be inferred with a series of direct and non-computationally intensive Hi-C contact analysis approaches that have not been previously used for this purpose.

Our results demonstrate that a straightforward statistical calculation (PCA) on the pattern of strong inter-chromosomal contacts can capture important biological features of CT radial positioning. With this approach, we not only can capture important patterns of gene density or chromosome length based ordering of chromosomes previously observed for very different cell types, but also can detect meaningful shifts of individual chromosomes in related cell types. We were able to infer changes in CT ordering in premature aging and senescence directly from contact data, and these changes are supported by previous microscopy results. We also note that this PCA ordering approach is actually quite robust to differences in Hi-C data quality and depth. While our subsequent network graph layout approach was somewhat sensitive to different quality metrics of different Hi-C replicates, the PCA ordering of chromosomes was robust to different levels of noise in these datasets.

We have also demonstrated that a network graph layout algorithm approach can generate an ensemble of models that capture the experimentally validated mean and variation of CT positions in a cell type. We demonstrate that these approaches can be applied to a variety of cell types and can even detect differences in underlying CT radial distributions between

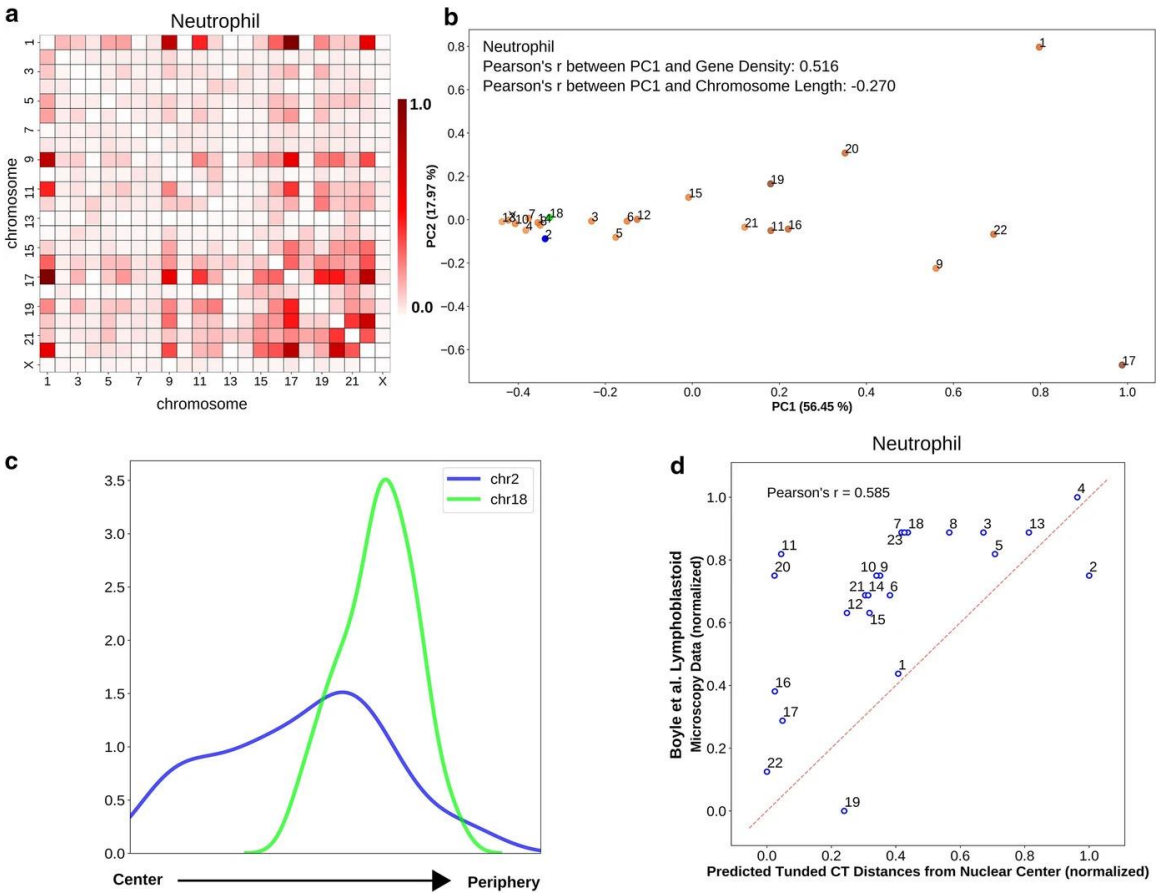


Figure 8 Radial arrangement of CTs in irregularly shaped neutrophil nuclei.

a Pairwise inter-chromosomal strong interaction pattern matrix for neutrophil Hi-C data. b 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrix obtained from neutrophil Hi-C data. In neutrophil cells, chr2 (blue color) and chr18 (green color) occupy radial positions near nuclear envelope. c The radial distance profiles of 2 CTs (chr2, and chr18) in neutrophils obtained from the selected model cluster (# of models in the selected clusters: 100). d Comparison of predicted tuned CT distances from neutrophil Hi-C data and lymphoblastoid microscopy imaging data shows important differences between these cell types.

highly related cell types (two different breast epithelial cell types). Interestingly, for these cell types, the difference in CT distribution corresponds to documented differences in nucleus geometry of these cell types when grown in culture. This suggests that not only do spherical and elliptical nuclei exhibit different CT organization in completely different cell lineages (lymphoblast vs. fibroblast), as previously documented, but that even cells within the same overall type may have different CT positioning associated with their nucleus shape. This result emphasizes the importance of approaches to detect changes in CT ordering directly from a given cell type in its particular condition, rather than assuming that a measurement in one circumstance will define CT organization principles across, for example, all epithelial cells. Our rapid approaches to inferring CT distribution makes screening across such varieties of datasets more feasible, as a complement to more intensive computational models or microscopic measurements.

In addition to the validation of these approaches that makes them useful for future applications, our results also show that while Hi-C contacts are useful for inferring a relative ordering of radial chromosome positions, inferring absolute ordering often requires an external reference point. We observe this in that the PCA based ordering of chromosomes can represent either direction (interior-exterior or vice versa) and that some clusters of network models display the same chromosome relative ordering in reverse orientation. This is an important factor to consider in any model that attempts to use Hi-C contacts in isolation to generate 3D structure models.

The analysis approach described in this paper is highly flexible due to its modular nature and can be integrated with different kinds of genome analysis applications. For example, both PCA ordering and inferred tuned positions can be used to characterize the changes in the radial CT organization in a perturbed (e.g. relocation of CTs during DNA damage response (Fatakia et al., 2017)) or diseased cell (e.g. mislocalization of CT18 and CT19 in lamin B2 depleted colorectal cancer cells (Ranade et al., 2017)) compared to a control cell, which in turn allows us to study how these changes affect higher-order chromatin structure. The method also generates a population of 3D network structures, which can further be used to characterize inter-chromosomal dynamics and can be compared to single-cell Hi-C results.

Conclusions

We have described a set of approaches that can be used sequentially or as separate modules to predict chromosome territory radial positioning in the nucleus from Hi-C data. We find that analyzing only the strongest interchromosomal contacts emphasizes differences in CT arrangement between cell types, and that PCA on this strong contact matrix can be used as a simple, fast approach to detect relative CT radial positions. We describe a network modeling approach that builds on this overall pattern detected by PCA to simulate the variability in CT positioning across the cell population. Finally, we demonstrate both the strengths and limitations of what Hi-C data alone can predict about radial CT organization and show that intrinsic chromosome properties can be added to tune CT organization predictions. These methods provide researchers with additional tools to infer the properties

of radial CT organization for the growing numbers of cell types that have available Hi-C data but not detailed microscopic measurements of chromosome positions.

Methods

The schematic representation of our whole analysis approach is given in Fig. 9.

Pre-processing of Hi-C data

All original Hi-C data sources for this study are listed in Appendix: Table 1. We mapped the data to hg19 to obtain a set of unique valid interacting fragment pairs and then binned the data at a resolution of 2.5 Mb, where each bin of the contact matrix represents the interaction frequency between a pair of 2.5 Mb genomic regions. We then perform two pre-processing steps on the Hi-C matrix. First, to remove biases related to GC content and cut site frequency, the raw contact matrix is normalized using the ICE technique (Imakaev et al., 2012). Next, the unmappable (repetitive) genomic regions are removed from the normalized Hi-C matrix. Also, we do not consider chrY in our analysis since this chromosome is not present in both male and female cell types, and we remove those corresponding bins from the Hi-C matrix. This modified Hi-C matrix is used for further downstream processing.

Let $H_{m\times m}$ represent a square symmetric Hi-C contact matrix where each row and column correspond to genomic regions (bins) of specific size, with each element h_{ij} representing the normalized number of contacts between i^{th} and j^{th} bins.

Determining a threshold to capture strong interactions in Hi-C matrix
After prepossessing the Hi-C contact data, we proceed to identify the strong interactions which have the best potential to infer cell-type specific radial organization of the CTs, distinct from the levels of average background noise. To do this, we define a cutoff limit and apply that to $H_{m\times m}$, to obtain a matrix of only filtered strong interactions $S_{m\times m}$.

$$s_{ij} = h_{ij} \text{ if } h_{ij} > h_{cut} \text{ otherwise } s_{ij} = 0 \quad (3)$$

where s_{ij} is the element of matrix $S_{m\times m}$, which represents the normalized number of contacts between i^{th} and j^{th} bins and h_{cut} is the cutoff limit. For 2.5 Mb genomic resolution, the value of cutoff limit is set to: h_{cut} = 95th percentile of all the genome-wide interactions from $H_{m\times m}$. The significance of and reason for choosing this 95th percentile cutoff is discussed in the "Results" section. Due to the application of this cutoff limit h_{cut}, the resultant matrix $S_{m\times m}$ contains mostly the intra-chromosomal interactions and a few inter-chromosomal interactions. In addition, to detect significant chromosomal interactions, we apply the FitHiC tool (Ay et al., 2014) with default parameters to the genome-wide Hi-C contact data, binned at 2.5 Mb resolution. From those significant intra- and inter-chromosomal interactions detected, we further select highly significant interactions for analysis by applying a q-value cutoff – 10^{-2} (for intra-chromosomal) and 10^{-12} (for inter-

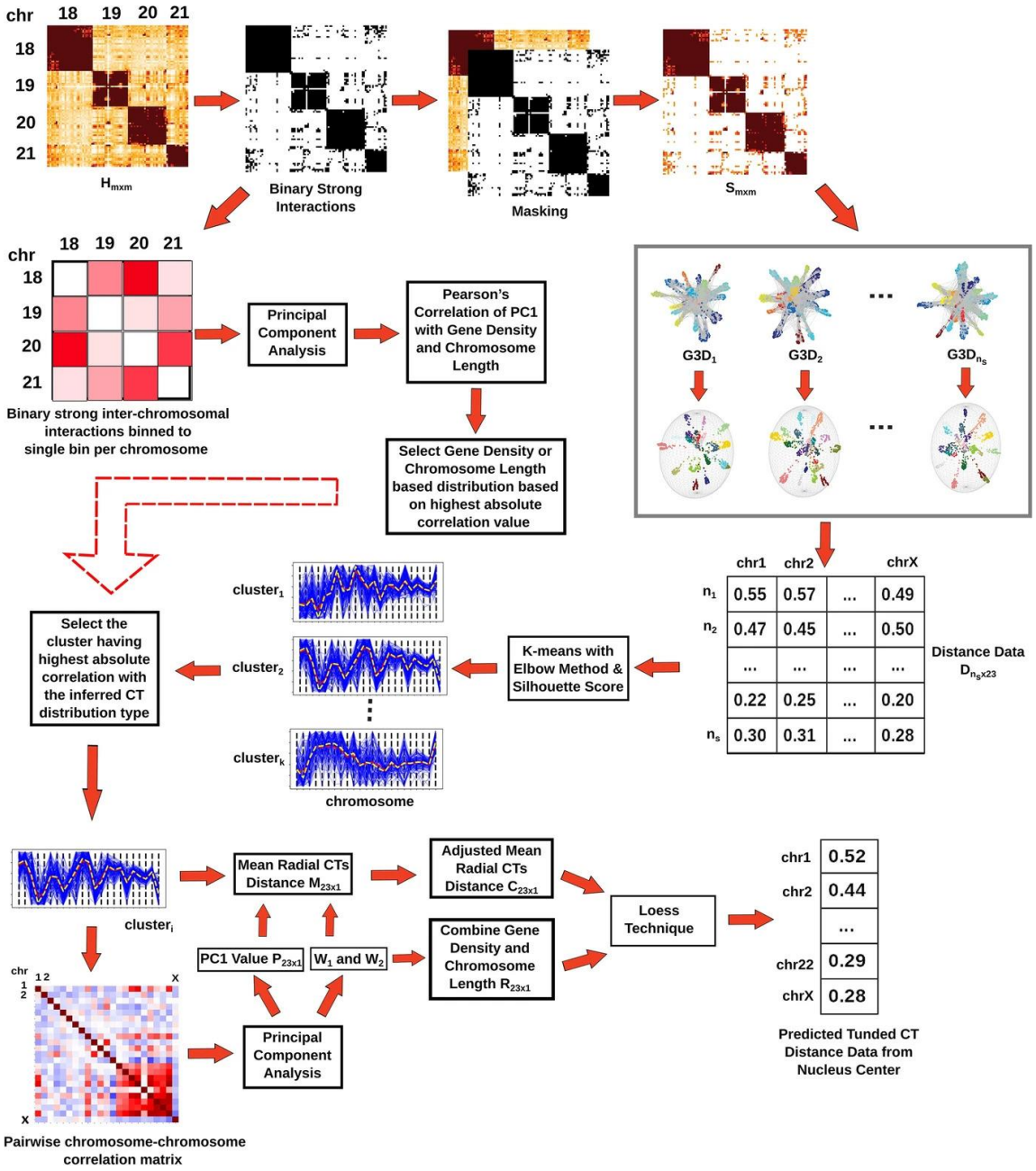


Figure 9 Overview of the computational analysis approaches.

At the first step, the whole genome Hi-C matrix is thresholded in order to extract strong interactions. Next, the pairwise inter-chromosomal strong interaction pattern matrix obtained from the thresholded Hi-C data is analyzed using PCA to determine the underlying CT distribution type. The radial CT organization is also modeled from the thresholded interactions by generating multiple 3D network model configurations. Out of those configurations, clusters of models having similar radial arrangements are identified and the particular cluster matching the inferred CT distribution type from PCA is selected for further downstream analysis. Finally, the network modeling inferred radial organization can be tuned by incorporating chromosome gene density and length properties.

chromosomal). The reason behind selecting a very stringent q-value cutoff for inter-chromosomal interactions is to make the number of significant inter-chromosomal interactions from FitHiC comparable to the strong interactions from our method.

Simulating random ligation Hi-C from original Hi-C data

Random ligation Hi-C data are simulated by taking an original raw/non-normalized Hi-C contact map, binned at 2.5 Mb, and shuffling the bins (including the diagonal) of this matrix five times. Then, the shuffled matrix is passed through the ICE normalization step (Imakaev et al., 2012). The reason for generating the random ligation matrix from the original matrix by random shuffling is to ensure that both the contact matrices have an equal number of total reads and a similar dynamic range of values, ensuring a matched comparison of the results from real and random Hi-C data.

Identifying CT radial distribution patterns with PCA transformation on the pairwise strong inter-chromosomal interaction pattern matrix

For our most direct approach to infer a radial organization pattern from Hi-C data, we begin by creating a pairwise strong inter-chromosomal interaction pattern matrix. For the chosen h_{cut} value, the number of bin pairs passing the threshold are summed between each pair of chromosomes to a single bin (Eqn. 4). This sum thus includes a count of the number of bin pairs that passed a threshold, rather than the total number of interactions in each included bin.

CHR_{ij}^{STRONG} = total number of strong interaction bins between chr i and j (4)

Then, we exclude cis contacts and look only at the inter-chromosomal component of this pairwise strong interaction pattern (Eqn. 5).

$CHR_{ij}^{TRANS} = CHR_{ij}^{STRONG}$ if $i \neq j$ otherwise 0 (5)

To capture the major interaction trends from this matrix, we apply principal component analysis (PCA) to this pairwise strong inter-chromosomal interaction pattern. We then calculate the projection of all chromosomes onto PC1, and we find that the ordering of chromosomes along this PC can capture the radial organization ordering of the chromosomes.

Constructing 3D network graphs from Hi-C matrix

The thresholded strong interaction matrix $S_{m\times m}$ is treated as a weighted adjacency matrix to generate an undirected weighted graph G, where nodes represent genomic bins and the number of interactions between each pair of bins is used as the edge weight. Next, we apply a 3D Fruchterman-Reingold (FR) force based layout to draw the undirected Hi-C graph in 3D space. This graph drawing layout adds an attractive force between the connected nodes

and creates repulsion between the nodes that are not connected. Along with this, gravitational force is used in this layout to pull the nodes towards the center. The rationale behind using the FR layout is that it uncovers the intrinsic structure of the network, as the strong interactions only matrix $S_{m\times m}$ has a large number of intra-chromosomal interactions and fewer inter-chromosomal interactions. The resulting 3D network graph infers the radial organization of the CTs inside the nucleus. Next, ns number of 3D network graphs – $G3D_1, G3D_2, \dots, G3D_{n_s}$ are generated by performing independent runs of the FR algorithm with different random initial configurations to model the variability in the radial CT organization.

Fitting a geometrical structure to the 3D network graphs

For each 3D network graph, the next objective is to find the distance of each CT from the center of that network graph. This step is performed in two parts.

In the first part, given a 3D network graph, the 3D Cartesian coordinates of the nodes are extracted and then Khachiyan's algorithm is used to find a minimum volume ellipsoid enclosing the set of nodes (Todd & Yıldırım, 2007). The center of the fitted object is calculated and assigned as the nucleus center. Along with this, the center of mass of each of the individual CTs is calculated from the coordinates of the nodes of the network graph.

$$COM_j^i = \frac{\sum coordinates\ of\ nodes\ of\ CTj\ in\ G3D_i}{\#\ of\ nodes\ of\ CTj\ in\ G3D_i} \tag{6}$$

where COM_j^i represents the center of mass of a particular CTj in the structure $G3D_i$ and $j \in \{1,2,\dots,22,X\}$.

After obtaining the center of mass of each CT and nucleus center, the Euclidean distance is calculated between the center of mass of each CT and the nucleus center. Also, to remove the heterogeneity that arises from different minimum volume ellipsoid fits of different 3D network structures, for each structure the distance values are normalized in the [0,1] range using Min-max normalization. In this way, by iterating over all of the 3D network graphs, for each CT, ns distances from the center of the nucleus are obtained. This distance information is represented in a matrix $D_{n_s\times 23}$ of size $n_s \times 23$, where rows represent n_s 3D network graphs and columns correspond to 23 CTs (22 autosomes and one X chromosome), and each matrix entry d_{ij} is the distance of the CT j from the nucleus center of network graph $G3D_i$.

The parameter n_s represents the number of configurations required to mimic the heterogeneity in CT arrangement originating during cell division within the same cell type. To estimate this parameter, for different values of n_s we compare the fitted distance profiles of all CTs between GM12878 and BJ1-hTERT cells using non-parametric hypothesis testing - a Two-sided Mann-Whitney U test. From the statistical test results, it can be observed that with increasing n_s value, more chromosomes show significant differences in

their CT distance profiles between lymphoblastoid and fibroblast cells, reaching a maximum at n_s = 1000 (Appendix: Fig. 20). Hence, we set the n_s parameter to 1000 in our analysis.

Identifying the specific cluster of network models having meaningful CT organization

The network modeling approach produces 3D graph models with heterogeneous CT organization, as intended, but we find that some groups of models may capture an inverted ordering of some groups of CTs or all CTs (reversing central to peripheral distances). Thus, rather than blindly averaging all models together, we first identify these different clusters of organization patterns and then choose for further consideration the cluster of models that follows the radial CT organization distribution captured by the PCA analysis described above. We perform K-means clustering (Lloyd, 1982) with deterministic initialization on the distance matrix $D_{n_s \times 23}$ by treating the models (rows) as samples and chromosomes (columns) as features. The initial centroid positions for the clustering technique are calculated using the algorithm from Nazeer et al. (Nazeer & Sebastian, 2009). In addition to that, to estimate optimal number clusters for the K-means, we use a modified elbow method approach. Here, first we calculate inertia (Pedregosa et al., 2011) which represents within cluster sum of squares for different increasing number of clusters and detect the elbow of the inertia curve with the help of the algorithm from Satopaa et al. (Satopaa et al., 2011). After detecting the elbow, again we calculate the average silhouette score (Pedregosa et al., 2011) for each of the different number of clusters and select the point as optimal number of clusters which will have the highest average silhouette score in the vicinity of the elbow point (two points upstream of elbow, two points downstream and the elbow itself). As our set of predictions for further analysis, we take the cluster whose mean radial CT positions shows the highest absolute correlation with the CT distribution type obtained by the PCA transformation of the pairwise inter-chromosomal strong interaction pattern matrix.

Gene density and chromosome length based tuning of consensus radial organization of CTs

The last aspect of our approach considers how the predicted averaged radial CT distances can be further tuned using two chromosomal properties - gene density and chromosome length. Human chromosome gene density (genes/ sequenced Mb) is obtained from "Short guide to the human genome" (Scherer, 2008) and length information from UCSC Genome Browser hg19 (Haeussler et al., 2018). But before discussing the steps involved with this tuning procedure, let us assume, $D'_{t \times 23}$ represents the selected model cluster based on inferred CT distribution type. This cluster contains t number of network models and the average CT distance of each chromosome across all of the t models is denoted by the column vector $M_{23 \times 1}$. After having the selected model cluster $D'_{t \times 23}$, we calculate a pairwise chromosome correlation matrix of size 23×23 from that and perform PCA transformation on that correlation matrix. The PC1 value obtained from this transformation represents the major separation of the chromosomes based on pairwise interactions and is

denoted by the column vector $P_{23\times1}$. Once we have the two column vectors - $M_{23\times1}$ and $P_{23\times1}$ representing the average radial position of the CTs and their separation respectively, we combine them using a weighted averaging technique as per Eqn. 7.

$$C_{23\times1} = W_1 \times M_{23\times1} + W_2 \times P_{23\times1}$$
$$W_1 = \frac{\%\text{ of variance explained by PC1}}{\%\text{ of variance explained by PC1}+\%\text{ of variance explained by PC2}}$$
$$W_2 = \frac{\%\text{ of variance explained by PC2}}{\%\text{ of variance explained by PC1}+\%\text{ of variance explained by PC2}} \quad (7)$$

where $C_{23\times1}$ represents the resultant combined column vector and W_1 and W_2 denote the weights calculated from the percentage of the the variance explained by PC1 and PC2.

Next, in order to model the effect of both chromosome length and gene density, we combine these two properties using weighted averaging as described in the Eqn. 8. The first and second components on the right hand side of each of the equations are related to the normalized gene density and normalized chromosome length respectively. Also, we use the same set of weights W_1 and W_2 in this equation but the weights are ordered based on the inferred CT distribution type. For example, if the inferred CT distribution follows a gene density based radial positioning, in that case the component having gene density will have the higher weight W_1 and vice verse.

$$R_{23\times1} = W_1 \times e^{(1-GD_{23\times1})} + W_2 \times LN_{23\times1} \text{ if inferred – gene density}$$
$$W_2 \times e^{(1-GD_{23\times1})} + W_1 \times LN_{23\times1} \text{ if inferred – length} \quad (8)$$

Here, GD and LN are column vectors of size 23×1 and contain the gene density and length of each chromosomes respectively.

Following the chromosomal properties modeling procedure, in the final step of the tuning, the modified consensus radial distance of the CTs $C_{23\times1}$ is locally smoothed based on $R_{23\times1}$ using the Loess technique. The radial distance of the CTs obtained from this step represents the tuned arrangements of the CTs inside the nucleus.

Availability of data and materials

The source code of the analysis technique is available at https://github.com/rpmccordlab/Radial-CT-Analysis-HiC. Source Hi-C data are publicly available from GEO and EGA.

Abbreviations

CT: Chromosome territory
PCA: Principal component analysis
FISH: Fluorescence in situ hybridization

REFERENCES

Agrawal, A., Ganai, N., Sengupta, S., & Menon, G. I. (2017). Chromatin as active matter. *J Stat Mech Theory Exp*, *2017*. https://doi.org/10.1088/1742-5468/aa5287

Agrawal, A., Ganai, N., Sengupta, S., & Menon, G. I. (2020). Nonequilibrium biophysical processes influence the large-scale architecture of the cell nucleus. *Biophys J*, *118*. https://doi.org/10.1016/j.bpj.2019.11.017

Amendola, M., & Steensel, B. (2014). Mechanisms and dynamics of nuclear lamina–genome interactions. *Curr Opin Cell Biol*, *28*. https://doi.org/10.1016/j.ceb.2014.03.003

Ay, F., Bailey, T. L., & Noble, W. S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res*, *24*. https://doi.org/10.1101/gr.160374.113

Barutcu, A. R., Lajoie, B. R., McCord, R. P., Tye, C. E., Hong, D., Messier, T. L., Browne, G., Wijnen, A. J., Lian, J. B., & Stein, J. L. (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol*, *16*. https://doi.org/10.1186/s13059-015-0768-0

Bercht Pfleghaar, K., Taimen, P., Butin-Israeli, V., Shimi, T., Langer-Freitag, S., Markaki, Y., Goldman, A. E., Wehnert, M., & Goldman, R. D. (2015). Gene-rich chromosomal regions are preferentially localized in the lamin b deficient nuclear blebs of atypical progeria cells. *Nucleus*, *6*. https://doi.org/10.1080/19491034.2015.1004256

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., & Speicher, M. R. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, *3*(5), e157.

Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A., & Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, *10*(3), 211-220.

Bridger, J., Boyle, S., Kill, I., & Bickmore, W. (2000). Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts. *Curr Biol*, *10*. https://doi.org/10.1016/S0960-9822(00)00312-2

Chandra, T., Ewels, P. A., Schoenfelder, S., Furlan-Magaril, M., Wingett, S. W., Kirschner, K., Thuret, J. Y., Andrews, S., Fraser, P., & Reik, W. (2015). Global reorganization of the nuclear landscape in senescent cells. *Cell Rep*, *10*. https://doi.org/10.1016/j.celrep.2014.12.055

Chiariello, A. M., Annunziatella, C., Bianco, S., Esposito, A., & Nicodemi, M. (2016). Polymer physics of chromosome large-scale 3d organisation. *Sci Rep*, *6*. https://doi.org/10.1038/srep29775

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*, *74*. https://doi.org/10.1080/01621459.1979.10481038

Cremer, M., Hase, J., Volm, T., Brero, A., Kreth, G., Walter, J., Fischer, C., Solovei, I., Cremer, C., & Cremer, T. (2001). Non-random radial higher-order chromatin

arrangements in nuclei of diploid human cells. *Chromosome Res*, *9*. https://doi.org/10.1023/A:1012495201697

Cremer, T., & Cremer, M. (2010). Chromosome territories. *Cold Spring Harbor Perspect Biol*, *2*. https://doi.org/10.1101/cshperspect.a003889

Cremer, T., Kurz, A., Zirbel, R., Dietzel, S., Rinke, B., Schröck, E., Speicher, M. R., Mathieu, U., Jauch, A., & Emmerich, P. (1993). Role of chromosome territories in the functional compartmentalization of the cell nucleus. Cold spring harbor symposia on quantitative biology,

Das, P., Golloshi, R., McCord, R. P., & Shen, T. (2020). Using contact statistics to characterize structure transformation of biopolymer ensembles. *Physical Review E*, *101*(1), 012419.

Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'shea, C. C., Park, P. J., & Ren, B. (2017). The 4D nucleome project. *Nature*, *549*. https://doi.org/10.1038/nature23884

Fatakia, S. N., Kulashreshtha, M., Mehta, I. S., & Rao, B. J. (2017). Chromosome territory relocation paradigm during dna damage response: some insights from molecular biology to physics. *Nucleus*, *8*. https://doi.org/10.1080/19491034.2017.1313938

Fritz, A. J., Barutcu, A. R., Martin-Buley, L., Wijnen, A. J., Zaidi, S. K., Imbalzano, A. N., Lian, J. B., Stein, J. L., & Stein, G. S. (2016). Chromosomes at work: organization of chromosome territories in the interphase nucleus. *J Cell Biochem*, *117*. https://doi.org/10.1002/jcb.25280

Fritz, A. J., Stojkovic, B., Ding, H., Xu, J., Bhattacharya, S., Gaile, D., & Berezney, R. (2014). Wide-scale alterations in interchromosomal organization in breast cancer cells: defining a network of interacting chromosomes. *Human Mol Genet*, *23*. https://doi.org/10.1093/hmg/ddu237

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Softw Pract Exp*, *21*. https://doi.org/10.1002/spe.4380211102

Ganai, N., Sengupta, S., & Menon, G. I. (2014). Chromosome positioning from activity-based segregation. *Nucl Acids Res*, *42*. https://doi.org/10.1093/nar/gkt1417

Gibcus, J. H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M. T., Xie, L., Paulson, J. R., & Earnshaw, W. C. (2018). A pathway for mitotic chromosome formation. *Science*, *359*. https://doi.org/10.1126/science.aao6135

Das, P., Shen, T., & McCord, R. P. (2022). Characterizing the variation in chromosome structure ensembles in the context of the nuclear microenvironment. *PLoS computational biology*, *18*(8), e1010392.

Golloshi, R., Playter, C., Freeman, T. F., Das, P., Raines, T. I., Garretson, J. H., Thurston, D., & McCord, R. P. (2022). Constricted migration is associated with stable 3D genome structure differences in cancer cells. *EMBO reports*, e52149.

Golloshi, R., Sanders, J. T., & McCord, R. P. (2018). Iteratively improving Hi-C experiments one step at a time. *Methods*, *142*. https://doi.org/10.1016/j.ymeth.2018.04.033

Gué, M., Messaoudi, C., Sun, J. S., & Boudier, T. (2005). Smart 3D-FISH: automation of distance analysis in nuclei of interphase cells by image processing. *Cytom Part A J Int Soc Anal Cytol*, *67*. https://doi.org/10.1002/cyto.a.20170

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., De Klein, A., Wessels, L., & De Laat, W. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, *453*(7197), 948-951.

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., & Gonzalez, J. N. (2018). The ucsc genome browser database: 2019 update. *Nucl Acids Res*, *47*. https://doi.org/10.1093/nar/gky1095

Horta, A., Monahan, K., Bashkirova, E., & Lomvardas, S. (2018). Cell type-specific interchromosomal interactions as a mechanism for transcriptional diversity. *bioRxiv*, 287532.

Iannuccelli, E., Mompart, F., Gellin, J., Lahbib-Mansais, Y., Yerle, M., & Boudier, T. (2010). NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-fish experiments. *Bioinformatics*, *26*. https://doi.org/10.1093/bioinformatics/btq013

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, *9*. https://doi.org/10.1038/nmeth.2148

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., & Thiecke, M. J. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, *167*. https://doi.org/10.1016/j.cell.2016.09.037

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., & Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, *30*. https://doi.org/10.1038/nbt.2057

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., Vries, S. S., Janssen, H., Amendola, M., Nolen, L. D., Bickmore, W. A., & Steensel, B. (2013). Single-cell dynamics of genome–nuclear lamina interactions. *Cell*, *153*. https://doi.org/10.1016/j.cell.2013.02.028

Kind, J., Pagie, L., Vries, S. S., Nahidiazar, L., Dey, S. S., Bienko, M., Zhan, Y., Lajoie, B., Graaf, C. A., & Amendola, M. (2015). Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, *163*. https://doi.org/10.1016/j.cell.2015.08.040

Kuroda, M., Tanabe, H., Yoshida, K., Oikawa, K., Saito, A., Kiyuna, T., Mizusawa, H., & Mukai, K. (2004). Alteration of chromosome positioning during adipocyte differentiation. *J Cell Sci*, *117*. https://doi.org/10.1242/jcs.01508

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, *72*. https://doi.org/10.1016/j.ymeth.2014.10.031

Lieberman-Aiden, E., Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., & Dorschner, M. O. (2009). Comprehensive

mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*. https://doi.org/10.1126/science.1181369

Lindsay, R. J., Pham, B., Shen, T., & McCord, R. P. (2018). Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic acids research*, *46*(16), 8143-8152.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans Inform Theory*, *28*. https://doi.org/10.1109/TIT.1982.1056489

Lucas, J., & Cervantes, E. (2002). Significant large-scale chromosome territory movement occurs as a result of mitosis, but not during interphase. *Int J Radiat Biol*, *78*. https://doi.org/10.1080/09553000110097190

Luperchio, T., Sauria, M., Hoskins, V., Wong, X., DeBoy, E., Gaillard, M. C., Tsang, P., Pekrun, K., Ach, R., Yamada, N., Taylor, J., & Reddy, K. (2018). The repressive genome compartment is established early in the cell cycle before forming the lamina associated domains. *bioRxiv*. https://doi.org/10.1101/481598

Maass, P. G., Barutcu, A. R., & Rinn, J. L. (2019). Interchromosomal interactions: a genomic love story of kissing chromosomes. *J Cell Biol*, *218*. https://doi.org/10.1083/jcb.201806052

McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Molecular cell*, *77*(4), 688-708.

McCord, R. P., Nazario-Toole, A., Zhang, H., Chines, P. S., Zhan, Y., Erdos, M. R., Collins, F. S., Dekker, J., & Cao, K. (2013). Correlated alterations in genome organization, histone methylation, and DNA–lamin A/C interactions in Hutchinson–Gilford progeria syndrome. *Genome Res*, *23*. https://doi.org/10.1101/gr.138032.112

Meaburn, K. J. (2016). Spatial genome organization and its emerging role as a potential diagnosis tool. *Frontiers in genetics*, *7*, 134.

Mehta, I. S., Amira, M., Harvey, A. J., & Bridger, J. M. (2010). Rapid chromosome territory relocation by nuclear motor activity in response to serum removal in primary human fibroblasts. *Genome Biol*, *11*. https://doi.org/10.1186/gb-2010-11-1-r5

Mehta, I. S., Eskiw, C. H., Arican, H. D., Kill, I. R., & Bridger, J. M. (2011). Farnesyltransferase inhibitor treatment restores chromosome territory positions and active chromosome dynamics in Hutchinson-Gilford progeria syndrome cells. *Genome biology*, *12*(8), 1-14.

Meluzzi, D., & Arya, G. (2020). Computational approaches for inferring 3D conformations of chromatin from chromosome conformation capture data. *Methods*, *181*, 24-34.

Mora, L., Sánchez, I., Garcia, M., & Ponsà, M. (2006). Chromosome territory positioning of conserved homologous chromosomes in different primate species. *Chromosoma*, *115*. https://doi.org/10.1007/s00412-006-0064-6

Nandakumar, V., Kelbauskas, L., Hernandez, K. F., Lintecum, K. M., Senechal, P., Bussey, K. J., Davies, P. C., Johnson, R. H., & Meldrum, D. R. (2012). Isotropic 3d nuclear morphometry of normal, fibrocystic and malignant breast epithelial cells reveals

new structural alterations. *PLoS One*, *7*. https://doi.org/10.1371/journal.pone.0029230

Nazeer, K. A., & Sebastian, M. (2009). Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. Proceedings of the world congress on engineering,

Ollion, J., Cochennec, J., Loll, F., Escudé, C., & Boudier, T. (2013). TANGO: a generic tool for high-throughput 3d image analysis for studying nuclear organization. *Bioinformatics*, *29*. https://doi.org/10.1093/bioinformatics/btt276

Oluwadare, O., Highsmith, M., & Cheng, J. (2019). An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biol Proced Online*, *21*. https://doi.org/10.1186/s12575-019-0094-0

Pal, K., Forcato, M., Jost, D., Sexton, T., Vaillant, C., Salviato, E., Mazza, E. M. C., Lugli, E., Cavalli, G., & Ferrari, F. (2019). Global chromatin conformation differences in the drosophila dosage compensated chromosome x. *Nat Commun*, *10*. https://doi.org/10.1038/s41467-018-07882-8

Paulsen, J., Sekelja, M., Oldenburg, A. R., Barateau, A., Briand, N., Delbarre, E., Shah, A., Sørensen, A. L., Vigouroux, C., & Buendia, B. (2017). Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol*, *18*. https://doi.org/10.1186/s13059-016-1146-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *J Mach Learn Res*, *12*.

Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2016). Transferable model for chromosome architecture. *Proc Natl Acad Sci*, *113*. https://doi.org/10.1073/pnas.1613607113

Qi, Y., Reyes, A., Johnstone, S. E., Aryee, M. J., Bernstein, B. E., & Zhang, B. (2020). Data-driven polymer model for mechanistic exploration of diploid genome organization. *bioRxiv*. https://doi.org/10.1101/2020.02.27.968735

Ranade, D., Koul, S., Thompson, J., Prasad, K. B., & Sengupta, K. (2017). Chromosomal aneuploidies induced upon lamin b2 depletion are mislocalized in the interphase nucleus. *Chromosoma*, *126*. https://doi.org/10.1007/s00412-016-0580-y

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. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665-1680.

Rosin, L. F., Crocker, O., Isenhart, R. L., Nguyen, S. C., Xu, Z., & Joyce, E. F. (2019). Chromosome territory formation attenuates the translocation potential of cells. *ELife*, *8*. https://doi.org/10.7554/eLife.49553

Sanchez, J. A., Karni, R. J., & Wangh, L. J. (1997). Fluorescent in situ hybridization (fish) analysis of the relationship between chromosome location and nuclear morphology in human neutrophils. *Chromosoma*, *106*. https://doi.org/10.1007/s004120050236

Sanders, J. T., Freeman, T. F., Xu, Y., Golloshi, R., Stallard, M. A., Martin, R. S., Balajee, A. S., & McCord, R. P. (2019). Radiation-induced DNA damage and repair effects on 3D genome organization. *bioRxiv*. https://doi.org/10.1101/740704

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a" kneedle" in a haystack: Detecting knee points in system behavior. 2011 31st international conference on distributed computing systems workshops,

Scherer, S. (2008). *Short guide to the human genome*. Cold Spring Harbor Laboratory Press.

Schoenfelder, S., Javierre, B. M., Furlan-Magaril, M., Wingett, S. W., & Fraser, P. (2018). Promoter capture HI-C: high-resolution, genome-wide profiling of promoter interactions. *J Vis Exp*, *136*.

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G. J., & Marti-Renom, M. A. (2017). Automatic analysis and 3D-modelling of Hi-C data using tadbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol*, *13*. https://doi.org/10.1371/journal.pcbi.1005665

Shachar, S., Voss, T. C., Pegoraro, G., Sciascia, N., & Misteli, T. (2015). A high-throughput imaging-based mapping platform for the systematic identification of gene positioning factors. *Cell*, *162*. https://doi.org/10.1016/j.cell.2015.07.035

Solovei, I., & Cremer, M. (2010). 3D-FISH on cultured cells combined with immunostaining. In J. Bridger & E. Volpi (Eds.), *Fluorescence in situ hybridization (FISH)*. Springer. https://doi.org/10.1007/978-1-60761-789-1_8

Stefano, M., Paulsen, J., Lien, T. G., Hovig, E., & Micheletti, C. (2016). Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci Rep*, *6*. https://doi.org/10.1038/srep35985

Steininger, A., Ebert, G., Becker, B. V., Assaf, C., Möbs, M., Schmidt, C. A., Grabarczyk, P., Jensen, L. R., Przybylski, G. K., & Port, M. (2018). Genome-wide analysis of interchromosomal interaction probabilities reveals chained translocations and overrepresentation of translocation breakpoints in genes in a cutaneous t-cell lymphoma cell line. *Front Oncol*, *8*. https://doi.org/10.3389/fonc.2018.00183

Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., Leeb, M., Wohlfahrt, K. J., Boucher, W., & O'Shaughnessy-Kirwan, A. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, *544*. https://doi.org/10.1038/nature21429

Strickfaden, H., Zunhammer, A., Koningsbruggen, S., Köhler, D., & Cremer, T. (2010). 4d chromatin dynamics in cycling cells: Theodor Boveri's hypotheses revisited. *Nucleus*, *1*.

Sun, H. B., Shen, J., & Yokota, H. (2000). Size-dependent positioning of human chromosomes in interphase nuclei. *Biophysical journal*, *79*(1), 184-190.

Tanabe, H., Habermann, F. A., Solovei, I., Cremer, M., & Cremer, T. (2002). Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutat Res Fundam Mol Mech Mutagenes*, *504*. https://doi.org/10.1016/S0027-5107(02)00077-5

Tanabe, H., Küpper, K., Ishida, T., Neusser, M., & Mizusawa, H. (2005). Inter-and intra-specific gene-density-correlated radial chromosome territory arrangements are conserved in old world monkeys. *Cytogenet Genome Res*, *108*. https://doi.org/10.1159/000080824

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., & Gros, M. A. (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci*, *113*. https://doi.org/10.1073/pnas.1512577113

Todd, M. J., & Yıldırım, E. A. (2007). On Khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. *Discrete Appl Math*, *155*. https://doi.org/10.1016/j.dam.2007.02.013

Van Steensel, B., & Belmont, A. S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, *169*(5), 780-791.

Walter, J., Schermelleh, L., Cremer, M., Tashiro, S., & Cremer, T. (2003). Chromosome order in hela cells changes during mitosis and early g1, but is stably maintained during subsequent interphase stages. *J Cell Biol*, *160*. https://doi.org/10.1083/jcb.200211103

Wettermann, S., Brems, M., Siebert, J., Vu, G., Stevens, T., & Virnau, P. (2020). A minimal gō-model for rebuilding whole genome structures from haploid single-cell Hi-C data. *Comput Mater Sci*, *173*. https://doi.org/10.1016/j.commatsci.2019.109178

Xiong, K., & Ma, J. (2019). Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat Commun*, *10*. https://doi.org/10.1038/s41467-018-07882-8

Zhang, B., & Wolynes, P. G. (2017). Genomic energy landscapes. *Biophys J*, *112*. https://doi.org/10.1016/j.bpj.2016.08.046

Zhang, X., Zhang, Y., Zhu, X., Purmann, C., Haney, M. S., Ward, T., Khechaduri, A., Yao, J., Weissman, S. M., & Urban, A. E. (2018). Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. *Nat Commun*, *9*. https://doi.org/10.1038/s41467-017-02088-w

Zhu, Y., Gong, K., Denholtz, M., Chandra, V., Kamps, M. P., Alber, F., & Murre, C. (2017). Comprehensive characterization of neutrophil genome topology. *Genes Dev*, *31*. https://doi.org/10.1101/gad.293910.116

APPENDIX

Table 1 Dataset description.

Cell Line	Type	Nuclear Shape	Hi-C Data	GEO or EGA Accession Number	Microscopy Data
GM12878	Lymphoblastoid	Spherical	Sanders et al., 2019	GSE136899	Boyle et al., 2001
GM12878 Rao	Lymphoblastoid	Spherical	Rao et al., 2014	GSE63525	Boyle et al., 2001
BJ1-hTERT	Skin Fibroblast	Ellipsoidal	Sanders et al., 2019	GSE136899	Bolzer et al., 2005
BJ-5ta	Skin Fibroblast	Ellipsoidal	Sanders et al., 2019	GSE136899	Bolzer et al., 2005
MCF10A	Breast Epithelium	Ellipsoidal	Barutcu et al., 2015	GSE66733	Fritz et al., 2014
HMEC	Breast Epithelium	Spherical	Rao et al., 2014	GSE63525	X
Neutrophil	Neutrophil	Multi Lobed	Javierre et al., 2016	EGAS00001000327	X
WI38-hTERT	Lung Fibroblast	Ellipsoidal	Chandra et al., 2015	ENAPRJEB8073	X
WI38-hTERT Senescent	Lung Fibroblast	Ellipsoidal	Chandra et al., 2015	ENAPRJEB8073	X
HGPS Patient p19	Lymphoblastoid	Ellipsoidal	McCord et al., 2013	GSE41763	X

Table 2 Correlation between microscopy data radial CT positions and those inferred by different sources of information.

	Correlation with microscopy data from:		
	GM12878	BJ1-hTERT	MCF10A
Pairwise strong contact PC1 for given cell type	-0.873	-0.791	0.763
Gene density	-0.8	-0.3	0.312
Chromosome length	0.43	0.89	0.959
Tuned network model for given cell type	0.811	0.848	0.941

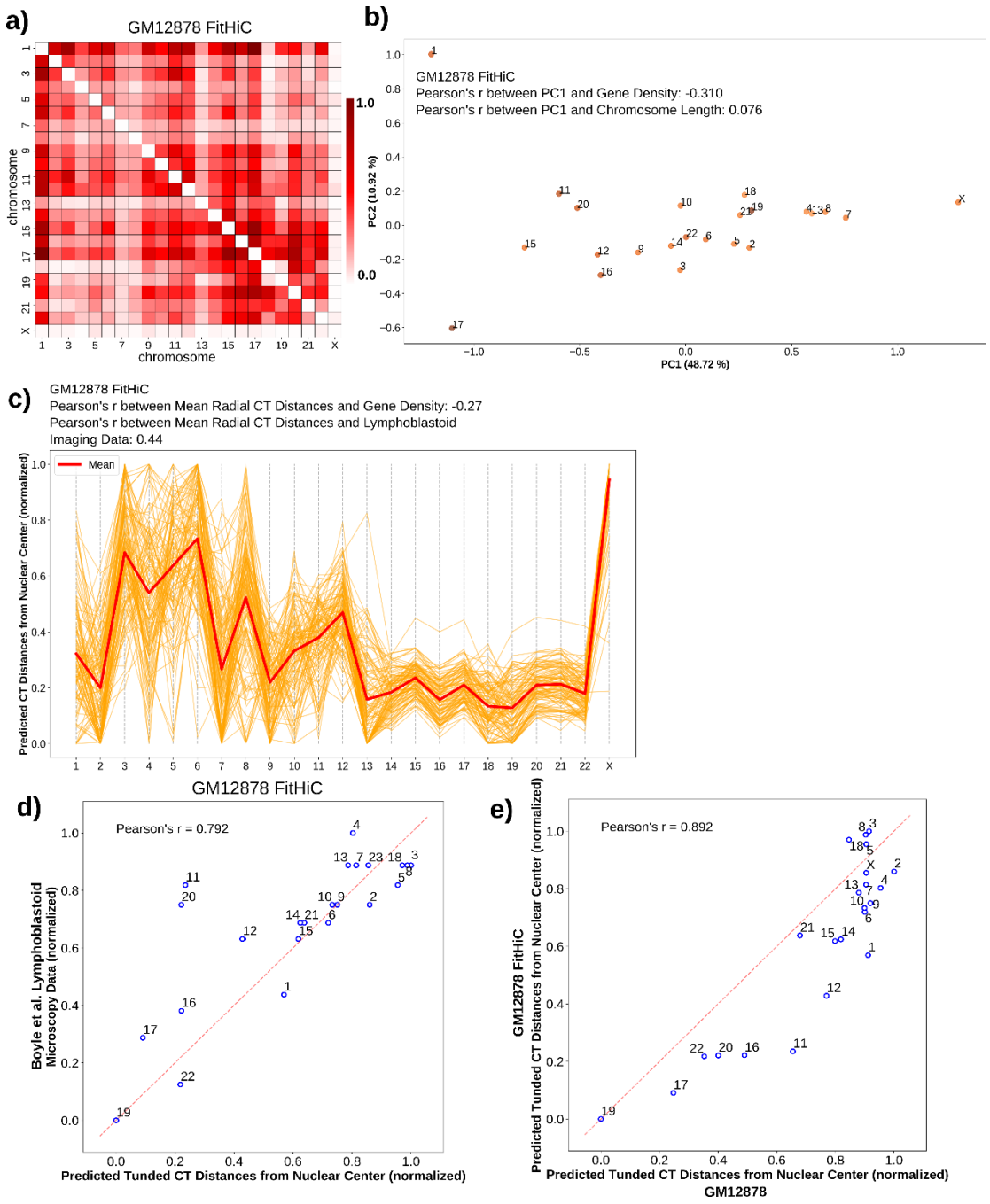


Figure 10 Radial CT arrangement of GM12878 (FitHiC).

a) Pairwise inter-chromosomal significant interaction pattern matrix derived from FitHiC for GM12878 Hi-C data. b) 2D PCA projection of the FitHiC pairwise inter-chromosomal significant interaction pattern matrix obtained from GM12878 Hi-C data. c) Network modeling generated model cluster for FitHiC GM12878 contacts, selected based on inferred CT distribution type. d) Correlation between predicted tuned CT distance from GM12878 (FitHiC) and lymphoblastoid microscopy imaging data. f) Correlation between predicted tuned CT distances obtained from GM12878 (FitHiC) and GM12878 (strong interactions).

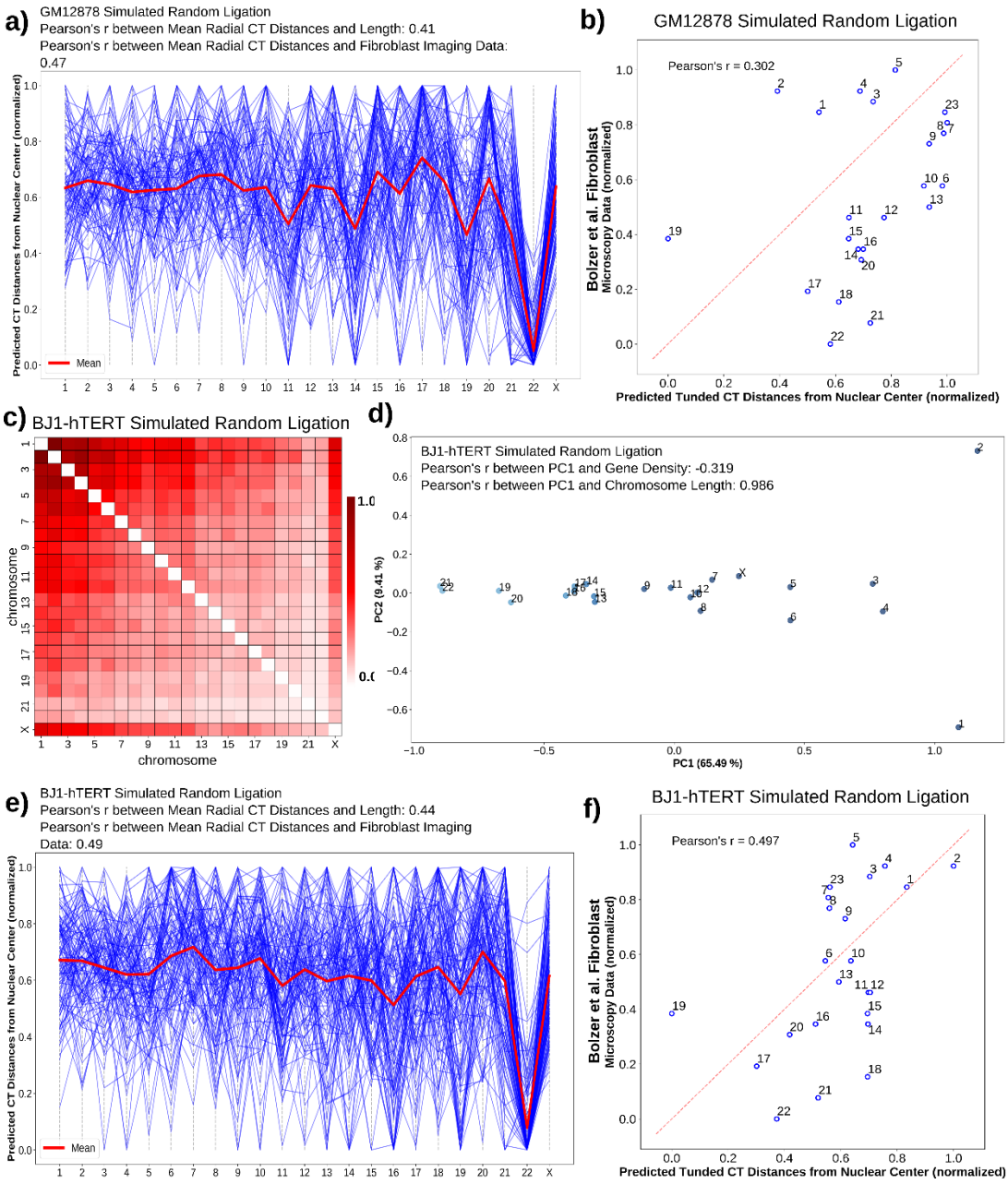


Figure 11 Radial CT arrangements inferred from GM12878 and BJ1hTERT simulated random ligation data.

a) Network modeling cluster for GM12878 simulated random ligation, selected based on inferred CT distribution type. b) Correlation between predicted tuned CT distance from GM12878 simulated random ligation Hi-C data and lymphoblastoid microscopy imaging data. c) Pairwise inter-chromosomal strong interaction pattern matrix for BJ1-hTERT simulated random ligation Hi-C data. d) 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrix obtained from BJ1-hTERT simulated random ligation Hi-C data. e) Network modeling cluster for BJ1-hTERT simulated random ligation, selected based on inferred CT distribution type. f) Correlation between predicted tuned CT distance from BJ1-hTERT simulated random ligation Hi-C data and fibroblast microscopy imaging data.

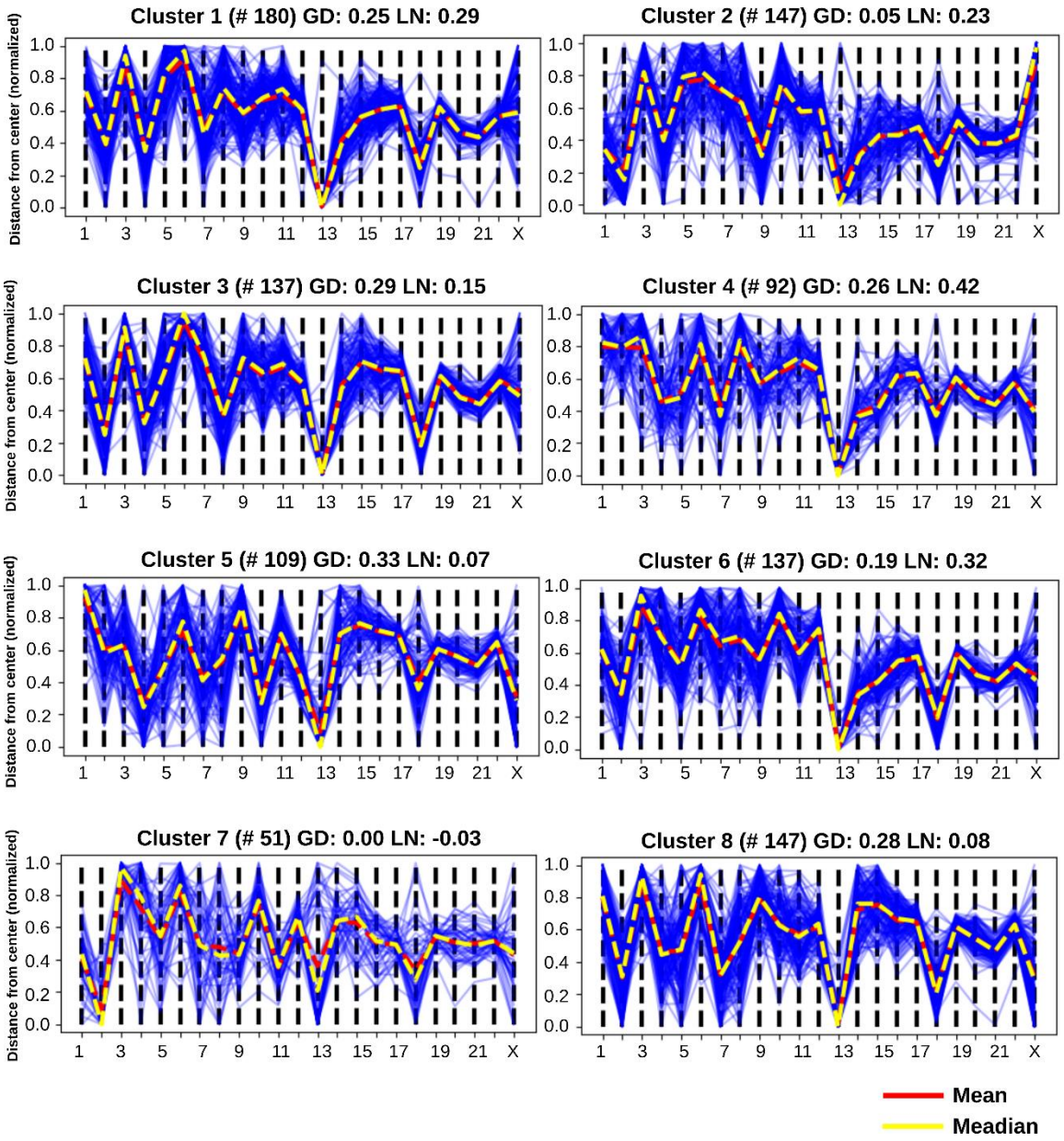


Figure 12 Network modeling generated clusters for GM12878.

For each cluster, the Pearson's correlation of mean radial CT distances with gene density (GD) and chromosome length (LN) are shown on the top of the cluster. Number of models in each cluster is indicated in parentheses above each cluster.

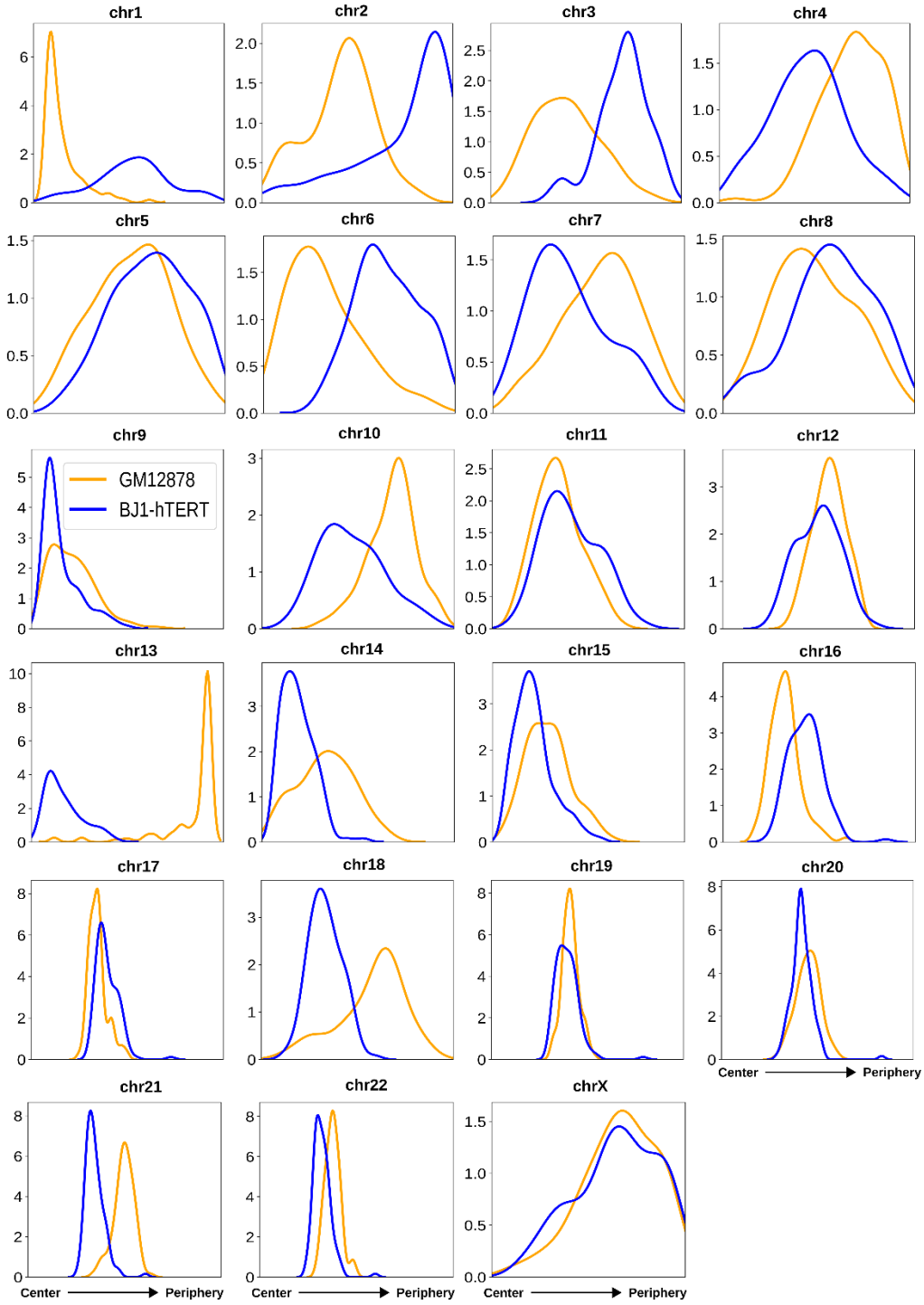


Figure 13 The radial distance distributions of 23 CTs for both GM12878 and BJ1-hTERT obtained from the respective selected model clusters based on inferred CT distribution types (# of models in the selected clusters- GM12878: 109, and BJ1-hTERT: 122).

All distributions are ordered left to right from center to periphery.

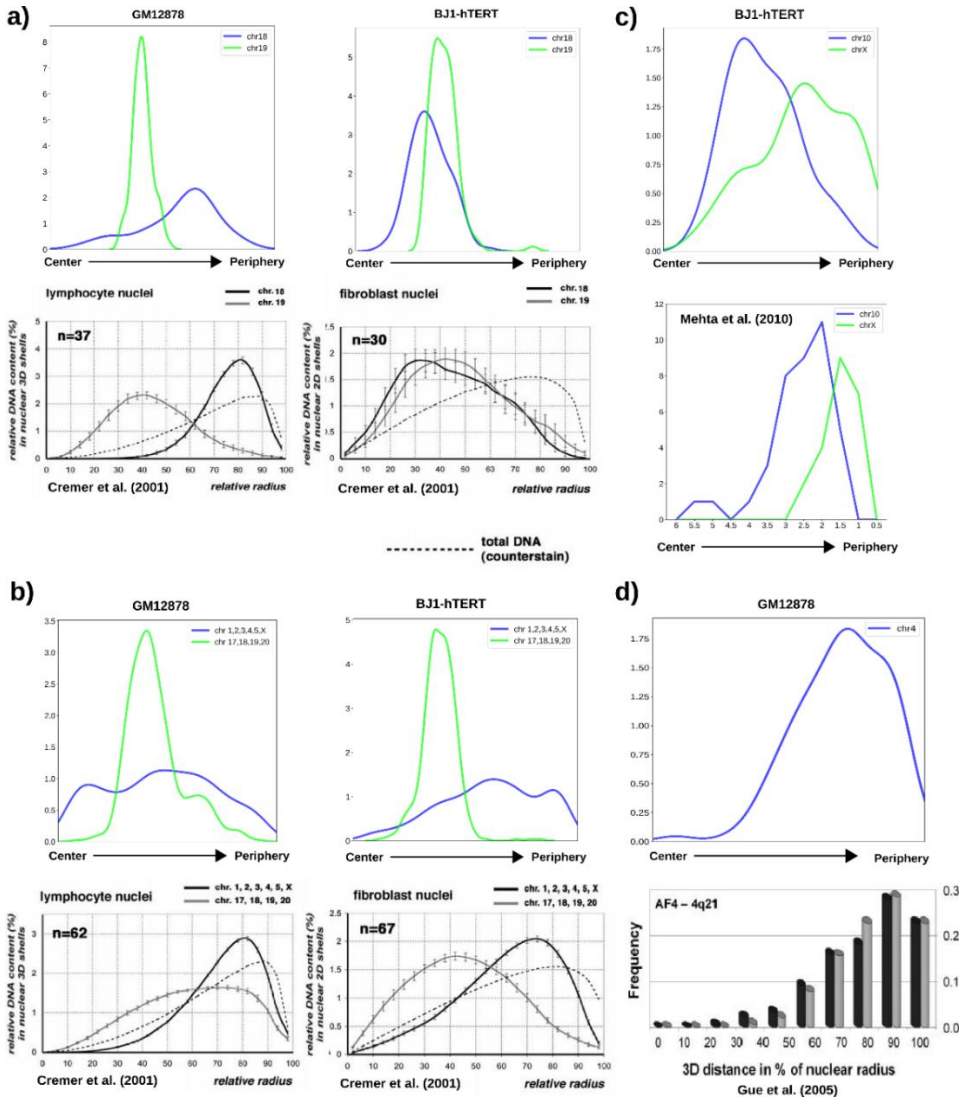


Figure 14 Comparison of network modeling generated radial CT distance profiles with experimentally obtained distributions.

a-b) The radial distance profiles of CT18 and CT19 (a top) and short and long chromosomes (b top) for GM12878 and BJ1-hTERT obtained from the network model clusters in this current work. These distributions are similar to the radial arrangement previously measured by microscopy for CT18 and CT19 (a bottom) and short and long chromosomes (b bottom) in lymphocyte (2D arrangement) and fibroblast (3D arrangement) nuclei. Panel (a-b) bottom parts are adapted from Cremer et al. (Cremer et al., 2001). Copyright Springer Nature. Used with permission. c) Top- The radial distance profiles of CT10 and CTX for BJ1-hTERT obtained from our network model cluster. Bottom- The distance distribution of CT10 and CTX from the nuclear periphery in human dermal fibroblasts obtained using 3D FISH. Panel (c) bottom part is created with data obtained from Mehta et al. (Mehta et al., 2011). d) Top- The radial distance profiles of CT4 for GM12878 obtained from our network modeling cluster. Bottom- Radial distance profiles of AF4 gene, which is located on chr4, from the nucleus center in two lymphoblastic cell lines- NALM-6 (black) and IL-9 (grey). Panel (d) bottom part is adapted from Gué et al. (Gué et al., 2005), copyright John Wiley and Sons, used with permission.

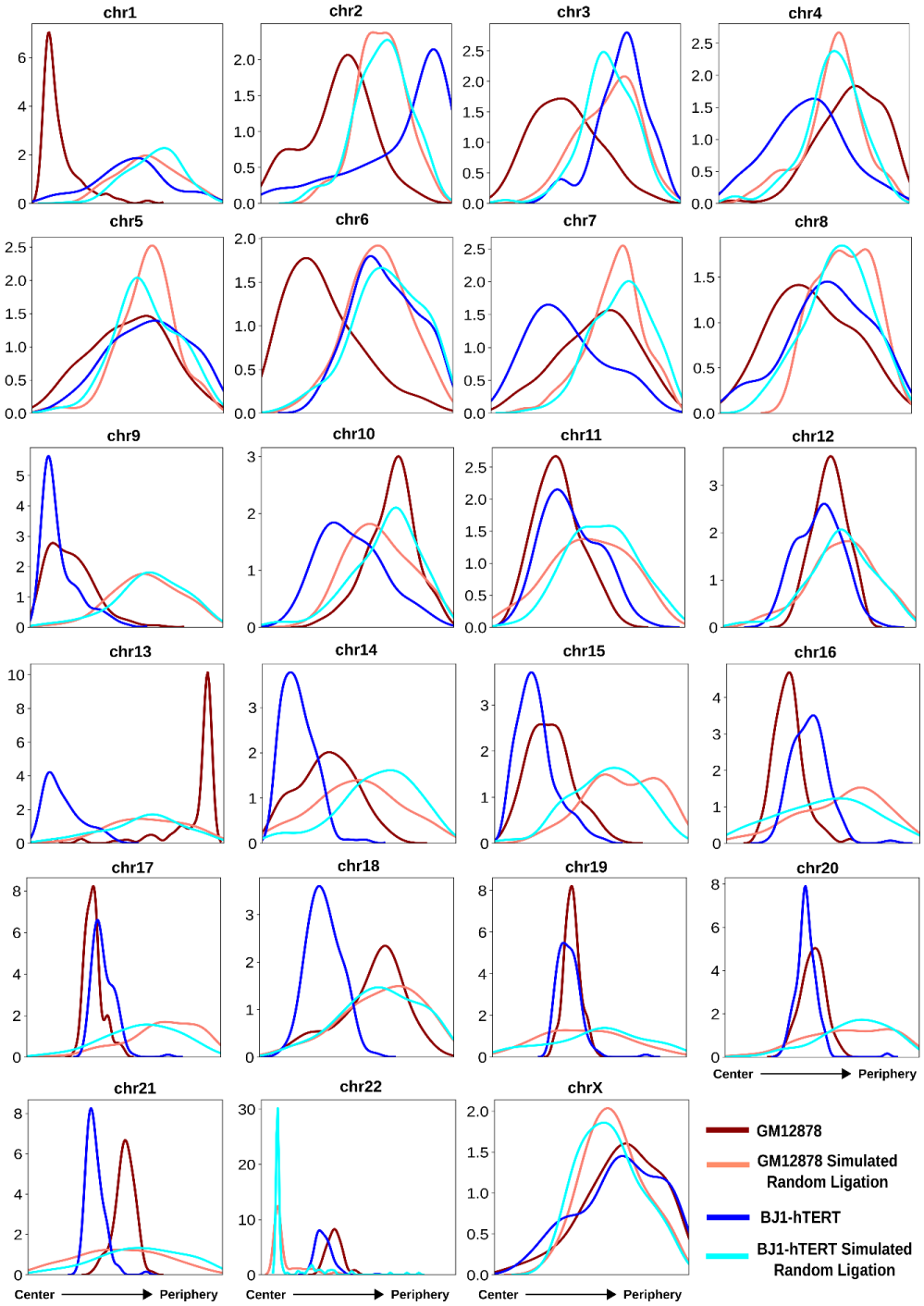


Figure 15 The radial distance distribution profiles of 23 CTs for GM12878, GM12878 simulated random ligation, BJ1-hTERT and BJ1-hTERT simulated random ligation obtained from the respective selected model clusters based on inferred CT distribution types (# of models in the selected clusters: for GM12878- 109, GM12878 simulated random ligation- 95, BJ1-hTERT- 122, and BJ1-hTERT simulated random ligation- 113).

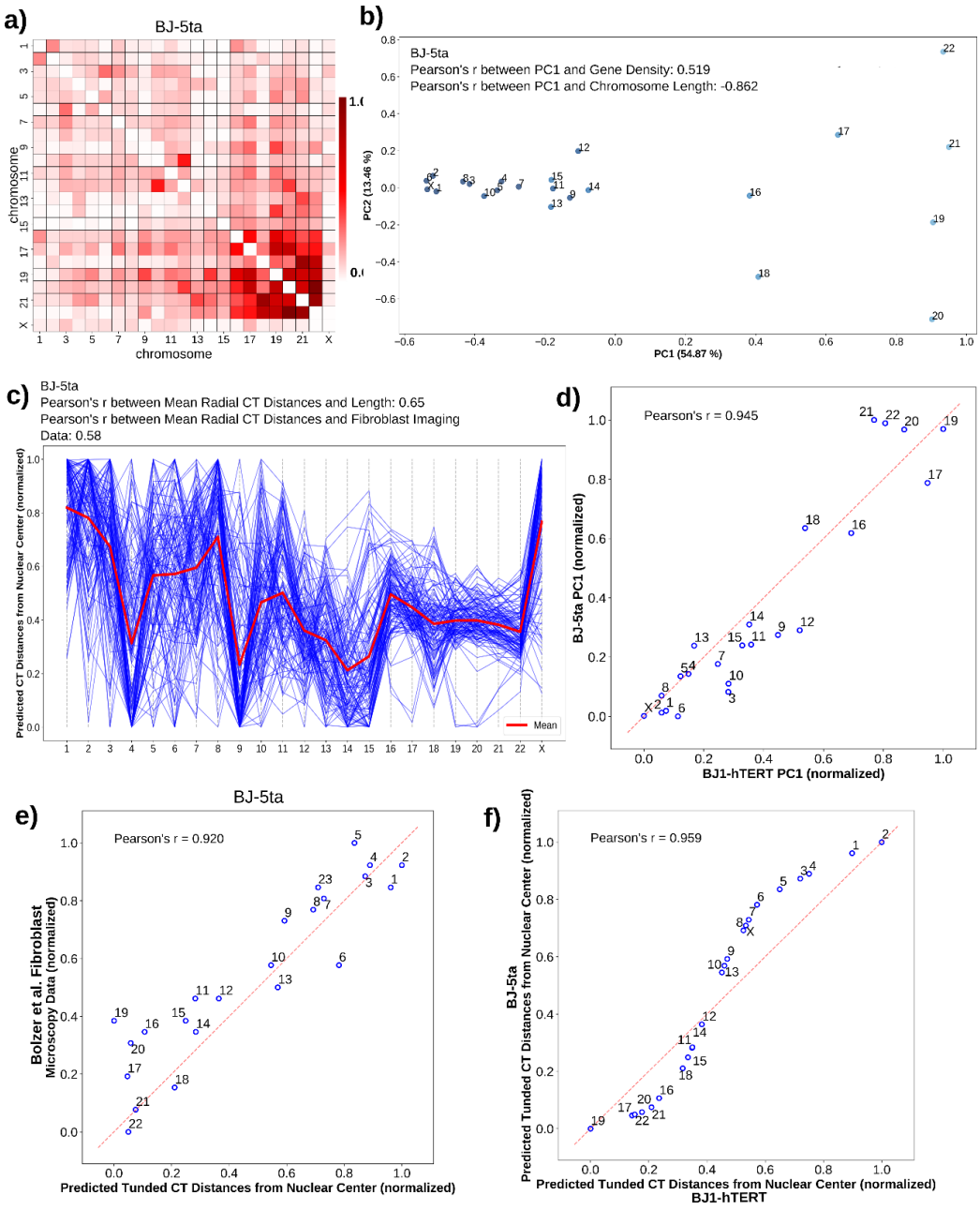


Figure 16 Radial arrangement of BJ1-hTERT and BJ-5ta follow similar CT distribution patterns.

a) Pairwise inter-chromosomal strong interaction pattern matrix for BJ-5ta Hi-C data. b) 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrix obtained from BJ-5ta Hi-C data. c) Network modeling generated model cluster for BJ-5ta, selected based on respective inferred CT distribution type (# of models in the selected clusters: 115). d) Correlation of BJ1-hTERT PC1 values with BJ-5ta PC1 values obtained from the PCA transformation of the respective pairwise interchromosomal strong interaction pattern matrices. e) Correlation between predicted tuned CT distance from BJ-5ta Hi-C data and fibroblast microscopy imaging data. f) Correlation between predicted tuned CT distances obtained from BJ1-hTERT and BJ-5ta.

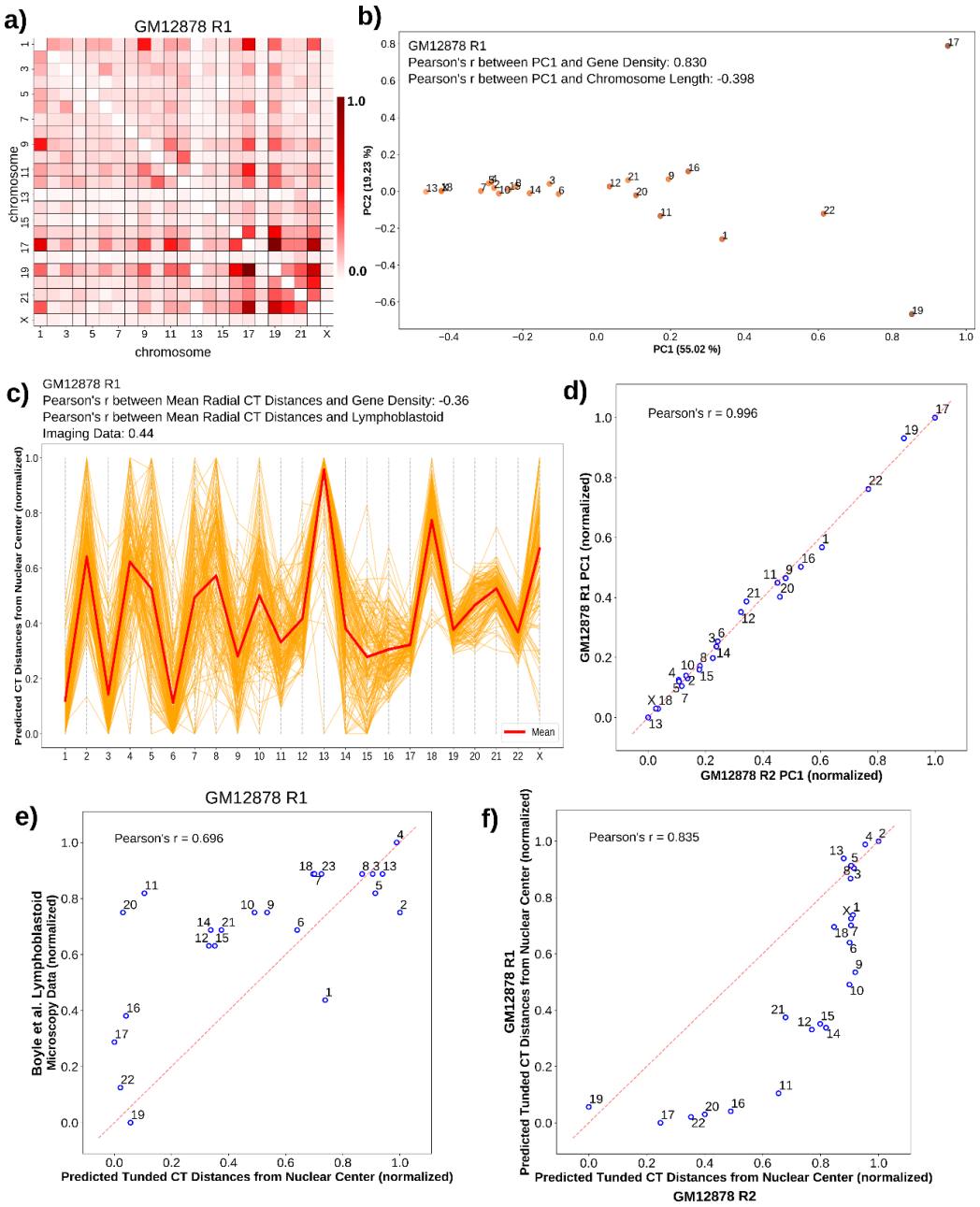


Figure 17 Radial arrangement of GM12878 replicates R1 and R2 show similar CT distribution pattern.

a) Pairwise inter-chromosomal strong interaction pattern matrix for GM12878 R1 Hi-C data (R2 used in main figures). b) 2D PCA projection of the pairwise inter-chromosomal strong interaction pattern matrix obtained from GM12878 R1 Hi-C data. c) Network modeling generated model cluster for GM12878 R1, selected based on respective inferred CT distribution type (# of models in the selected clusters: 142). d) Correlation of GM12878 R2 PC1 values with GM12878 R1 PC1 values obtained from the PCA transformation of the respective pairwise inter-chromosomal strong interaction pattern matrices. e) Correlation between predicted tuned CT distance from GM12878 R1 Hi-C data and lymphoblastoid microscopy imaging data. f) Correlation between predicted tuned CT distances obtained from GM12878 R2 and GM12878 R1.

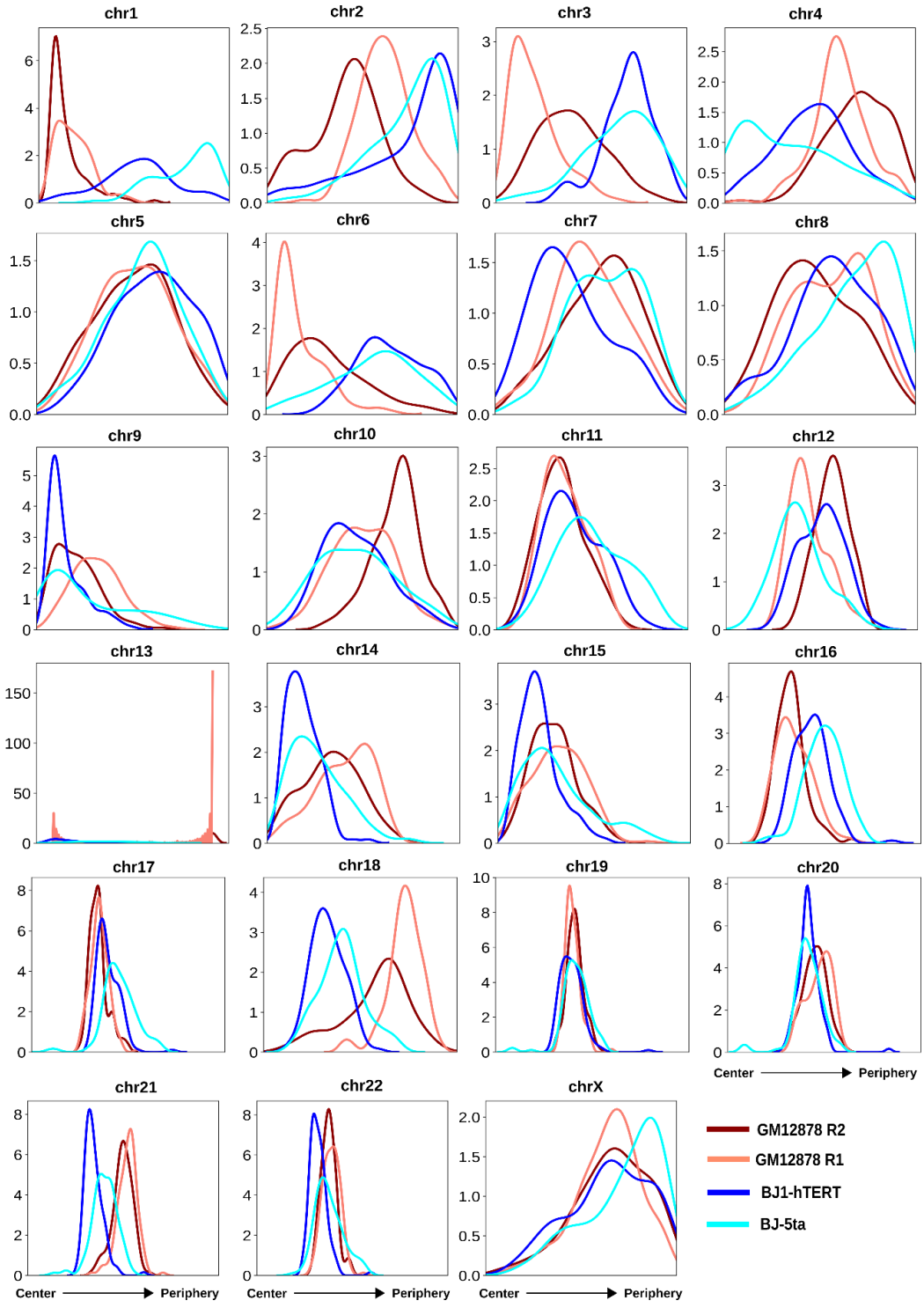


Figure 18 The radial distance profiles of 23 CTs for GM12878 R1, GM12878 R2, BJ1-hTERT and BJ-5ta obtained from the respective selected model clusters based on inferred CT distribution types (# of models in the selected clusters: for GM12878 R2- 109, GM12878 R1- 142, BJ1-hTERT- 122, and BJ-5ta- 115).

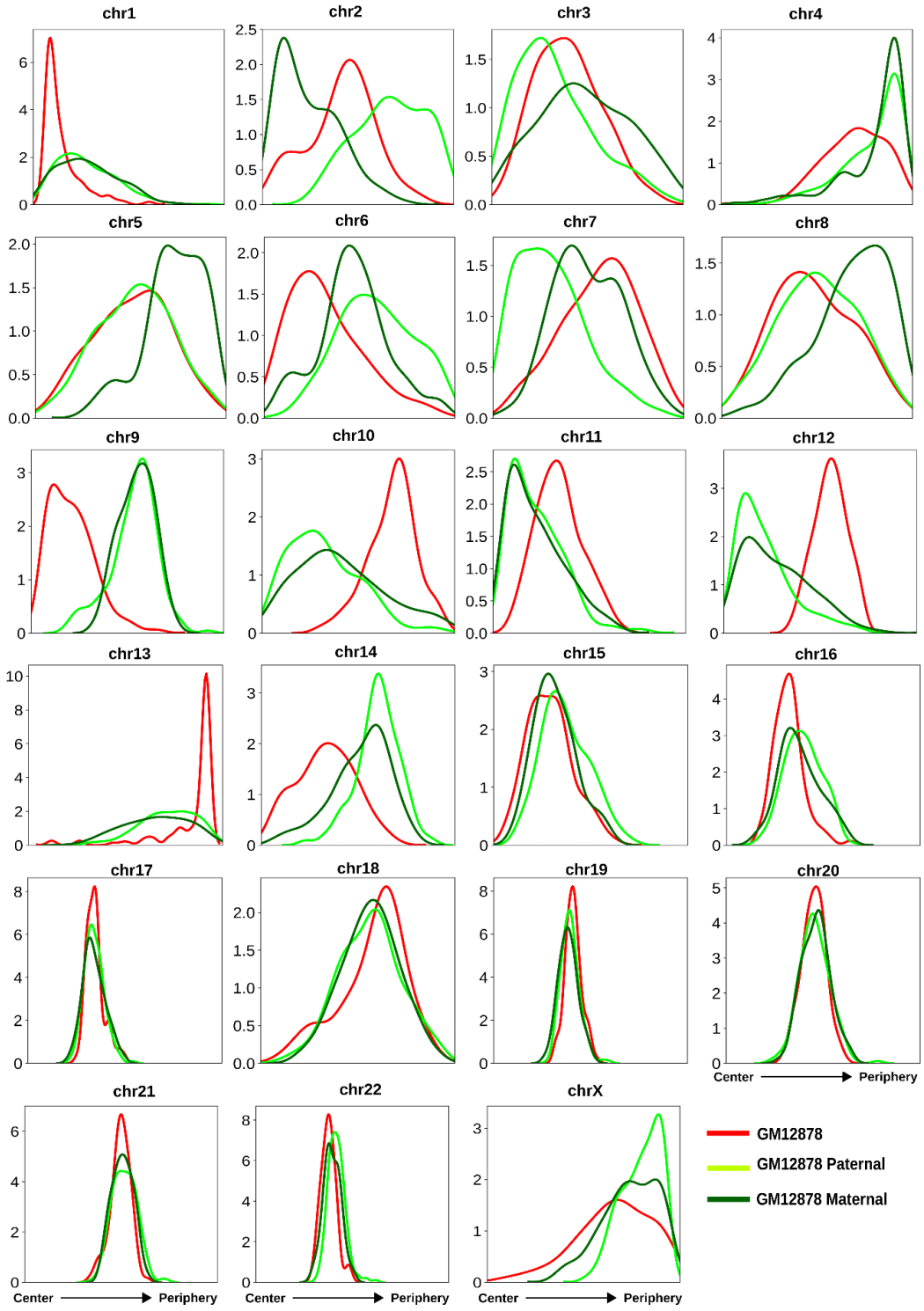


Figure 19 The radial distance profiles of 23 CTs for GM12878 standard, paternal, and maternal copies obtained from the respective selected model clusters based on inferred CT distribution types (# of models in the selected clusters: for GM12878 standard- 109, paternal copy- 194, and maternal copy- 148).

chr	n_s = 100	n_s = 250	n_s = 500	n_s = 750	n_s = 1000
1	****	****	****	****	****
2	****	****	****	****	****
3	****	****	****	****	****
4	ns	**	****	****	****
5	ns	****	***	****	****
6	****	****	****	****	****
7	*	****	****	****	****
8	ns	ns	ns	ns	ns
9	**	****	****	****	****
10	ns	ns	**	****	****
11	ns	ns	ns	*	**
12	ns	ns	ns	*	ns
13	****	****	****	****	****
14	****	****	****	****	****
15	***	****	****	****	****
16	****	****	****	****	****
17	**	***	****	****	****
18	****	****	****	****	****
19	ns	*	***	****	****
20	**	****	****	****	****
21	****	****	****	****	****
22	*	****	****	****	****
X	*	**	***	****	****

Figure 20 Statistical comparisons of the CT distance profiles obtained from network modeling for GM12878 and BJ1-hTERT for different sample sizes ns (number of network structures generated from Hi-C data) using the two-sided Mann-Whitney U test.

The color values and the asterisk marks represent the level of significance of the BJ1-hTERT vs. GM12878 distribution difference. Green- P-value ≤ 0.0001- ****; Blue- P-value ≤ 0.001- ***; Yellow- P-value ≤ 0.01- **; Red- P-value ≤ 0.05- *; White- P-value > 0.05- non significant (ns).

CHAPTER II
CHARACTERIZING THE VARIATION IN CHROMOSOME STRUCTURE ENSEMBLES IN THE CONTEXT OF THE NUCLEAR MICROENVIRONMENT

A version of this chapter was originally published by Priyojit Das, Tongye Shen and Rachel Patton McCord:

Das, P., Shen, T., & McCord, R. P. (2022). Characterizing the variation in chromosome structure ensembles in the context of the nuclear microenvironment. PLoS computational biology, 18(8), e1010392.

My contributions to this work include: (1) conceived the project, (2) performed computational modeling and analysis, (3) wrote the manuscript, and (4) prepared the figures.

Abstract

Inside the nucleus, chromosomes are subjected to direct physical interaction between different components, active forces, and thermal noise, leading to the formation of an ensemble of three-dimensional structures. However, it is still not well understood to what extent and how the structural ensemble varies from one chromosome region or cell-type to another. We designed a statistical analysis technique and applied it to single-cell chromosome imaging data to reveal the heterogeneity of individual chromosome structures. By analyzing the resulting structural landscape, we find that the largest dynamic variation is the overall radius of gyration of the chromatin region, followed by domain reorganization within the region. By comparing different human cell-lines and experimental perturbation data using this statistical analysis technique and a network-based similarity quantification approach, we identify both cell-type and condition-specific features of the structural landscapes. We identify a relationship between epigenetic state and the properties of chromosome structure fluctuation and validate this relationship through polymer simulations. Overall, our study suggests that the types of variation in a chromosome structure ensemble are cell-type as well as region-specific and can be attributed to constraints placed on the structure by factors such as variation in epigenetic state.

Introduction

In eukaryotic cells, long genomic DNA is packed within a relatively small nucleus, requiring chromosomes to fold into a multi-scale 3D spatial organization (McCord et al., 2020). At the kilobase length scale, loops form between different chromatin regions (Fudenberg et al., 2016; Nichols & Corces, 2015; Sanborn et al., 2015). Some loops are formed by the interaction of cohesin-mediated loop extrusion and boundary factors, a process which can create domains of enriched contacts called Topologically Associated Domains (TADs) at the scale of 100s of kilobases (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). Through processes such as phase separation and tethering to nuclear structures the chromatin is further organized at the megabase scale into separate euchromatin and heterochromatin spatial regions, known as A/B compartments (Falk et al., 2019; Lieberman-Aiden et al., 2009; Sanulli et al., 2015; Strom et al., 2017; Wang et al., 2019). Finally at the scale of the whole nucleus, chromosomes tend to form their own territories (Bolzer et al., 2005; Cremer et al., 2001; Stack et al., 1977). Each of these structural features act independently as well as synergistically to contribute to different biological processes, as reported by an ever-increasing number of studies. For example, chromatin loops and TAD boundary positioning can influence enhancer-promoter interactions and gene regulation (Braccioli & de Wit, 2019; Kadauke & Blobel, 2009; Zuin et al., 2021). This multi-layered organization of 3D genome is not static: the chromatin fiber is semiflexible in nature and is subjected to forces generated by ATP-driven molecular machines along with thermal fluctuations which generate chromosome structure heterogeneity on a temporal scale (Clapier et al., 2017; Kim et al., 2021; Kim & Yu, 2020). Some chromosome structures (such as the positioning of whole chromosomes or nuclear lamina associations) are stable on a timescale of hours (Kind et al., 2013; Strickfaden et al., 2010), but heterogeneity in these structures can arise, for example, due to stochastic variations that occur as each cell rebuilds its chromosome structure after mitosis (Kind et al., 2013; Nagano et al., 2013). Overall, therefore, variations observed in a chromosome structure ensemble reflects a combination of both cell-to-cell and temporal structural heterogeneity. Measuring the heterogeneity of the structural ensemble may reveal detailed folding mechanisms and subpopulations of cells with different chromosome structures that mediate different biological outcomes (Takei et al., 2021).

Several key molecular and imaging techniques to characterize chromosome structure have been developed in recent years (Jerkovic & Cavalli, 2021; Kempfer & Pombo, 2020). Chromosome conformation capture (3C) based techniques, including Hi-C, can map the spatial organization of the whole genome in terms of contacts between genomic regions. However, when Hi-C and its derivatives are applied to millions of cells in bulk, it is challenging to study the heterogeneity of chromosome structures from the resulting population averaged measurements. Several polymer- and restraint- based models have been developed to reconstruct ensembles of structures from such population averaged Hi-C data (Conte et al., 2020; Di Pierro et al., 2017; Di Pierro et al., 2018; Tjong et al., 2016; Wlasnowolski et al., 2020). However, there are usually multiple possible theoretical ensembles that could fit the population average, so without single cell measurements for

comparison, it is difficult to know whether models derived from bulk data capture the variation present in true biological systems. To resolve this issue, researchers have started to develop single-cell counterparts of Hi-C and related techniques (Arrastia et al., 2021; Ramani et al., 2017; Tan et al., 2021). However, these single cell contact capture techniques often suffer from low coverage, which makes it difficult to characterize the structural heterogeneity within a particular cell-type.

In contrast, oligonucleotide-based FISH imaging techniques have enabled the tracing of chromatin regions at high resolution in thousands of single cells (Bintu et al., 2018; Mateo et al., 2019; Sawh et al., 2020; Su et al., 2020; Takei et al., 2021; Wang et al., 2016). These techniques can generate distance distribution maps between imaged points along a chromosome that resemble Hi-C contact maps. By analyzing these single-cell structures at different resolutions, recent studies have started to characterize the heterogeneity associated with the structural ensemble. For example, it has been revealed that the inactive chrX exhibits not only canonically expected compact structures but also more chromosome domain inter-mingling as compared to the active chrX (Wang et al., 2016). Another study revealed that domain-like structure also exists on the inactive chrX at the single-cell level, but such structures are averaged out and not visible in bulk experiments (Takei et al., 2021). Bintu and co-authors reported the existence of TAD-like structures and sharp domain boundaries at a single cell level (Bintu et al., 2018). Though the boundary positions vary from one cell to another, they show higher preference for structural protein binding sites. By analyzing the single-cell whole chromosome tracing data at a higher resolution, researchers identified a substantial number of TAD-like domains in single cells that cross compartment boundaries observed in bulk data (Su et al., 2020).

While initial studies focused on showing that variability in chromosome structures exists, recent work has begun to turn toward the question of how and why this variability arises (Cheng et al., 2020; Conte et al., 2020; Xie et al., 2020). For example, transcriptional regulation has been proposed to influence an observed two state transition between closed and open dumbbell-like conformations of a short chromatin segment spanning a TAD boundary (Cheng et al., 2020). Other groups have analyzed the chromatin tracing data to evaluate the effects of the structural protein cohesin on the dynamics of TAD-like structures. Another study used a folding coordinate to represent the progression of TAD formation. These analyses showed that in the absence of structural protein cohesin, a local chromatin region may still exhibit TAD-like structures in individual snapshots, but these structures are less stable due to the entropy gain of the unfolded conformations (Xie et al., 2020). This has the effect of converting phase separated globules into more random coil-like state (Conte et al., 2020).

Although these recent studies have started to shed light on the fundamental mechanisms of local chromatin folding at the single-cell level, it is unclear how the chromosome structure variation differs between cell types or chromosome regions. Here, we focus on quantifying the full structure variation landscape rather than identifying multiple discrete structural states. Support for this view comes from a variety of recent studies. A FISH imaging study

in Caenorhabditis elegans showed that chrV exhibits four major conformational clusters during early embryonic development, but that the structure ensembles corresponding to different clusters exhibit a broad and continuous distribution rather than a handful of discrete conformational states (Sawh et al., 2020). Similarly, within a particular cell-type, the single-cell structure ensemble can be described as a distribution of conformations around the bulk average structure (mean). We do not know how this fluctuation around an average structure (variance) varies in a cell-type specific manner. This question becomes particularly interesting when the average structural organization looks similar across different cell types or conditions, but there may be different underlying structural ensembles that lead to similar averages. A logical follow up question is what factors might contribute to the extent of structural variation. There are various essential factors such as epigenetic marks, associated structural proteins, and nuclear lamina associations that vary in a cell-type specific manner and contribute to the organization of 3D genome (Bersaglieri et al., 2020; Chen et al., 2012; Zhu et al., 2021). Therefore, investigating the role of these different factors in the variation of chromosome structural ensemble is of high importance.

To answer these questions about chromosome structural ensembles, we designed a statistical pipeline to analyze structural data obtained from single-cell chromatin tracing. Using this pipeline, we first characterized the landscape of single-cell structures across different cell-types. We then identified cell-type specific differences in the structural ensembles through joint analysis of structural landscapes. To quantify these differences, we designed two different network-based measurements – one to quantify the degree of continuity across the chromosome structural ensemble (ensemble structure similarity index) and another to quantify the degree of structural variation of local genomic regions across the population (bin similarity index). Using these two measurements, we quantified the ensemble variation and local structural variation across cell-types and chromosome regions at different genomic resolutions. We identify overall locus compaction and domain boundary shifting as two primary but independent modes of structural fluctuation that are observed across all length scales. We observe differences in chromosome ensemble structural similarity between regions that have different epigenetic state profiles, and we are able to demonstrate that interactions governed by epigenetic state can help explain these differences in the type of structure variation exhibited by the chromosome region.

Material and methods

The statistical chromosome conformation analysis pipeline used in this work can be broadly divided into two parts. In the first part, the single-cell chromosome imaging data (of different genomic resolutions) is analyzed statistically to compare the structures in a pairwise fashion. Next, a structural similarity measurement is applied to quantify the degree of continuity present in the chromosome structural ensemble. In addition, the pairwise similarity data is used to construct a structural landscape to arrange the single-cell chromosome structures according to their individual and bulk structural properties (Fig. 21).

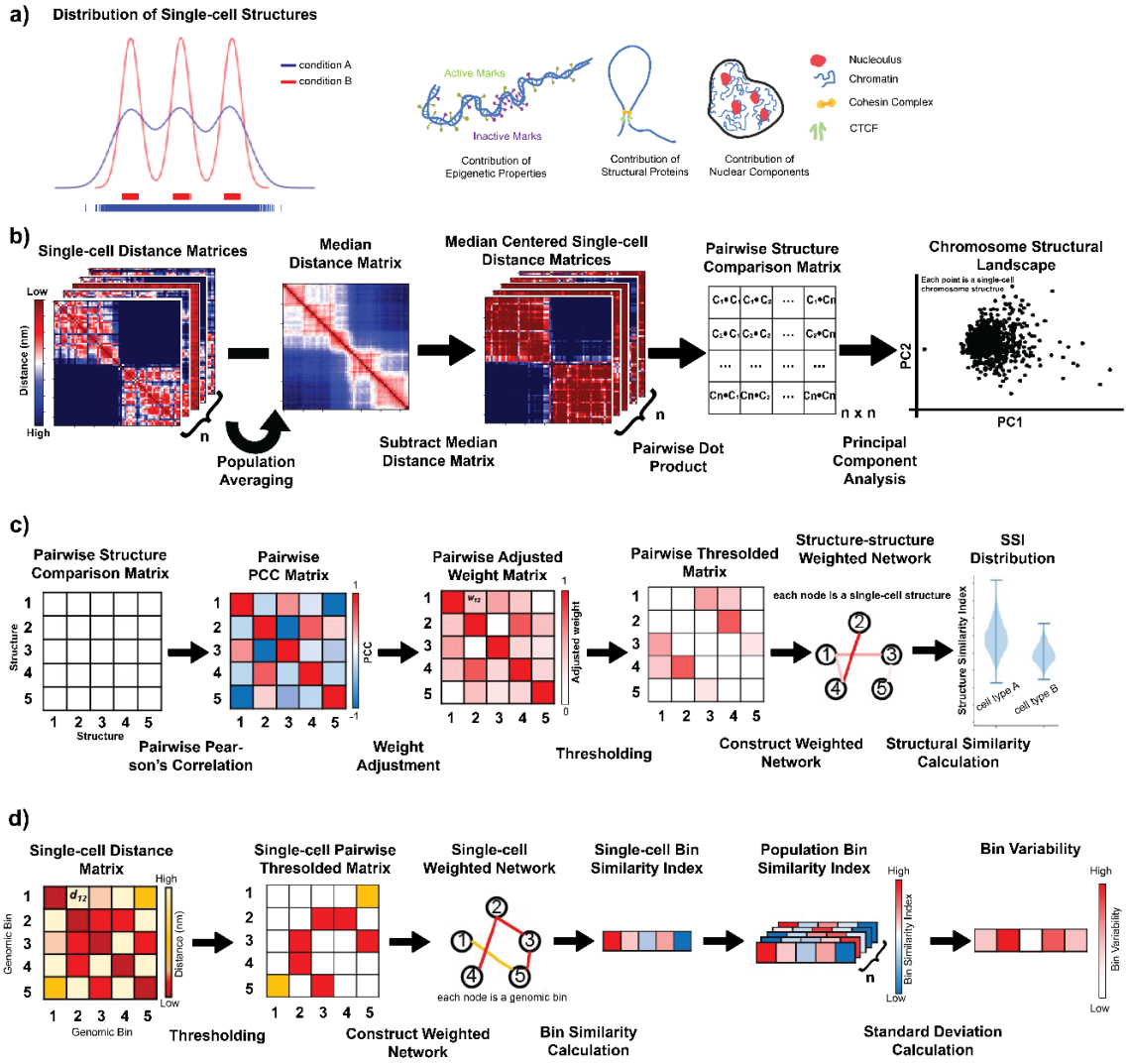


Figure 21 Schematic representation of the statistical pipeline used to quantify conformational diversity present in single-cell chromosome distance matrices.

a) Cartoon representation of potential differences in chromosome structure landscapes (left). Condition A has a continuous landscape including all structures along a trajectory, while in Condition B, chromosomes fall into 3 major structural clusters. (Right) Conditions and factors that may influence this chromosome structure landscape. b) Each pair of median-centered chromosome distance matrices are compared in a pairwise fashion. PCA is then performed on the pairwise structure comparison matrix and transformed data is projected in 2D space using PC1 and PC2 to create the "chromosome structural landscape". c) To calculate chromosome ensemble structure similarity index (SSI), we construct a Pearson's correlation matrix from the pairwise structure comparison matrix. This correlation matrix scaled, thresholded, and then converted into a weighted network from which SSI values can be calculated from each node. The median SSI value is plotted for many resamplings of nodes. d) To calculate bin variability, we convert each single-cell distance matrix into a thresholded matrix using a cutoff and then construct a network from each of those matrices with bins as nodes and contacts between bins as edges. We then calculate the bin similarity index for each bin in each single-cell and then take the standard deviation of the bin similarity index across the population as the bin variability.

Construction of distance matrices from single-cell chromosome imaging data

For this work, we have examined experimental single-cell chromosome imaging data at three different resolutions, each of which reflects different levels of structural organization of the 3D genome at different genomic length scales. The single-cell chromosome imaging data obtained from Bintu *et al*. (Bintu et al., 2018). has the highest genomic resolution of 30 kb. Here, the authors traced the conformations of multiple ~ 2 – 2.5 Mb spanning chromatin regions of chr21, partitioned into consecutive 30 kb segments in single-cells across different cell-types (e.g., IMR90 – human fetal lung fibroblast cells, K562 – human erythroleukemic cells, A549 – human lung adenocarcinoma cells, and HCT116 – human colon cancer cells) during different physiological and perturbed conditions. The second set of IMR90 chromatin tracing data analyzed here is obtained from Su *et al*. (Su et al., 2020), where the single-cell conformations of chr21 non-repetitive region were traced using consecutive 50 kb bins. In addition, that dataset also provides chr2 conformations captured at the same 50 kb resolution, however with an interval of 250 kb between two genomic regions, due to very large size of human chr2. Both the Bintu *et al*. and Su *et al*. data capture chromatin conformations at a comparable resolution (30 and 50 kb respectively), differing only in the size of the genomic region analyzed. Whereas the Bintu *et al*. data primarily represents the structural heterogeneity at the level of specific TADs and sub-TADs, the Su *et al*. data captures the heterogeneity of almost the whole chromosome. Finally, we also analyzed chromosome structural heterogeneity of the whole active and inactive chrX of IMR90, obtained from Wang *et al*. (Wang et al., 2016). In this dataset, TADs are imaged as the basic subunits along the chromosome, so the resolution of this data is lower (ranging from 440 kb to 7.88 Mb) and reports on the conformations of compartments and the whole chromosome rather than the within-TAD level.

For all these different datasets, the authors reported the single-cell conformations of the chromosomes in a 3D Cartesian coordinate system where each basic genomic region (e.g., 30 or 50 kb genomic region or TAD; from here termed as bin) has X-, Y-, Z- coordinate values. However, in some single cell structures the 3D positions of some of the bins are missing due to technical difficulties. Therefore, to decrease the effect of missing data, we consider only structures with more than ~ 90% measured bins, and the values for missing bins in those structures are assigned by nearest-neighbor interpolation. The 3D positions of chromosomes are also associated with the global rotational and translational variations which might affect underlying statistical analysis. To resolve this issue, we convert each 3D structure into a 2D symmetric distance matrix by calculating the pairwise Euclidean distance between all the genomic bins of that chromosome structure. For example, if a particular chromosome has m genomic bins and a total of N copies originating from N single cells, then we will have N symmetric pairwise distance matrices each of size $m \times m$, where d_{ij} represents the Euclidean distance between i^{th} and j^{th} genomic bins. Further, the spatial distance can be converted into contacts by applying a genomic resolution specific thresholding cutoff. However, for our analyses, we use distance matrices instead of contact matrices, unless otherwise explicitly mentioned.

Generating chromosome structural landscapes using a statistical analysis technique

We next center all the single-cell matrices around zero by subtracting the median distance matrix from each of them, $h_{ij} = d_{ij} - median(d_{ij})$. Once we obtain the median centered single-cell distance matrices, we calculate the pairwise dot product between each of those single-cell matrices, such as, $s_{\alpha\beta} = \sum_{i,j} h_{ij}^{\alpha}.h_{ij}^{\beta}$, where h^{α} and h^{β} are median centered distance matrices obtained from single cells ($\alpha, \beta \in \{1, \dots, N\}$), respectively. For each pair of single-cells, this step compares the deviation from the median between the two structures and yields a single number. In the case of two structures having similar deviations, this dot product would yield a strong positive value and vice versa. The pairwise dot product comparison information between all possible pairs of the single cells is then represented in a symmetric matrix format, termed as pairwise structure comparison matrix, where each element of the matrix represents the pairwise comparison information between a specific pair of chromosome structures. Next, we perform PCA on the pairwise structure comparison matrix and project the single-cell conformations in 2D space using the first two principal components (PCs). This PC projection is termed the chromosome structural landscape.

Quantification of chromosome structural ensemble variation using a network-based similarity index

To quantify the variation associated with the chromosome structural ensemble, we perform a network-based similarity calculation on the pairwise comparison matrix (Fig. 21c). Similar approaches have been used to study cellular network heterogeneity associated with cancer and cellular differentiation (Banerji et al., 2013; Cheng et al., 2016; Teschendorff & Severini, 2010). In this approach, we first calculate Pearson's correlation coefficient (PCC) between each pair of rows of the comparison matrix, where c_{ij} denotes PCC between i^{th} and j^{th} rows of the pairwise similarity matrix. As PCC values range between -1 and 1, next we scale the PCC values in the range of 0 and 1, $w_{ij} = 0.5 * (1 + c_{ij})$. Once we have the rescaled pairwise PCC value matrix, we construct a weighted network from that where each node represents a single chromosome structure or conformation and the edges between them denotes rescaled PCC values. However, we only consider the edges having weights above a certain threshold, representing only connections between highly similar structures. This threshold is calculated by taking 50^{th} percentile of all the edge weights of the network (see Results for an analysis of the impact of this threshold choice). With the constructed network, we calculate the structure similarity index (SSI) of each node S_i, using a formula based on the Shannon-Jayne network entropy (Cheng et al., 2016).

$$S_i = -\frac{1}{\log k_i} \sum_{j \in N(i)} p_{ij} \log p_{ij}$$

where $N(i)$ represents the neighborhood set of the node i and k_i denotes the cardinality of that set. p_{ij} represents the proportionality of the edge weight between nodes i and j compared to all the edges associated with node i.

$$p_{ij} = \frac{w_{ij}}{\sum_{k \in N(i)} w_{ik}}$$

Network entropy is a commonly applied concept in the analysis of heterogeneity vs. continuity in biological networks (Chanda et al., 2020). Conceptually, entropy measures the degree of randomness in a system, or the extent to which all possible outcomes are sampled. Entropy in DNA sequence analysis measures whether all 4 DNA bases are represented uniformly in a motif position (high entropy) or some bases are strongly preferred in certain positions (low entropy) (Durbin et al., 1998; Saha & Sarkar, 2020). Network entropy in signaling or gene expression networks captures whether the network is more continuous or modular (Cheng et al., 2016; West et al., 2012). Similarly, here this metric will reveal whether the landscape of structures is highly connected with all intermediates sampled vs. more heterogeneous with disconnected structures. This index S_i is bound between 0 and 1. In our system, the resulting values are compressed near a value of 1, so for easier visual presentation of values, we transform them by subtracting 0.99 from each SSI and then multiplying by 1000. For our presented analyses, we calculate the median SSI across all nodes in the landscape and call this the ensemble SSI. Specifically, to normalize sample size between conditions, we iteratively resample a subset of nodes from each condition, then plot the distribution of median SSI values from the resampled node sets. Based on the distribution of the median SSI values, we characterize the heterogeneity of the chromosome structural landscape. An SSI distribution with higher mean is considered as having a smooth continuous structural landscape while a distribution with relatively lower mean would have more discrete, disconnected structures. Interpretation of the SSI distribution trends have been discussed in detail in the **Results** section.

Calculating chromosome structure variability for each genomic bin

To characterize the variability of structure at each genomic bin along the chromosome, we use a modified version of the network-based structural similarity calculation approach, discussed in the previous subsection (Fig. 21d). In this step, for each single-cell structure, we first construct a network based on its distance matrix, where each node is a single genomic bin and the edges between them represent the spatial distances. We then threshold the network using an experimentally-derived distance cutoff (330 nm for 30 kb resolution (Cheng et al., 2020) and 500 nm for 50 kb resolution (Su et al., 2020)), where edges with spatial distance higher than the cutoff are discarded from the analysis. Next, we calculate the bin similarity index of each of the nodes (bins) in the network using the same formula discussed in the previous section, except here using genomic bins as nodes rather than whole single-cell structures and spatial distances as edge weights rather than structure

correlation coefficients. We now have a bin similarity index value for each genomic bin in each single cell structure. To quantify the variability of structural positions occupied by a given bin across all structures, we then calculate the standard deviation of bin similarity index values across all single cell structures for a given bin. We call this measure the bin variability. A higher standard deviation corresponds to a more dynamic or variable genomic bin.

Characterization of chromatin states using epigenetic profiles

In order to characterize chromatin states of the IMR90, K562 and A549 cells, we downloaded a total of eleven histone marks – H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K27Ac, H4K20Me1, H3K36Me3, H3K27Me3, H3K9Me3, H2A.Z and DNase for all the three cell types from ENCODE repository (Appendix: Table 3) (Davis et al., 2018). Next, we used chromHMM software (Ernst & Kellis, 2012) to annotate genomic regions of desired resolutions into five chromatin states using a combinatorial pattern of the eleven histone marks. Based on the combinatorial histone marks, we defined chromatin states as – Active state 1, Active state 2, Inactive Poised, Repressed Polycomb and Heterochromatin (Appendix: Fig. 27). Briefly, if an inferred state shows strong enrichments for active marks such as H3K4Me1/2/3, H3K27Ac, H3K9Ac and DNase, we term that state as an Active state. Further, depending on the level of enrichment of those different active marks, we characterize them into Active state 1 and Active state 2. We assign the Heterochromatin label to the state for which we observe strong enrichment of H3K9Me3 and for the Repressive Polycomb assignment, we look for strong H3K27Me3 and a mix of H3K9Me3 and H4K20Me1 marks. And the rest of the states are assigned the Inactive Poised level.

Polymer simulation of chromatin segments based on underlying chromatin states

To test the contribution of epigenetic states to chromosome structure heterogeneity, we perform polymer simulations of a specific chromatin segment with the interactions defined based on their epigenetic states. In recent years, polymer simulation-based studies have shown how the compartmental profile or the underlying epigenetic state of the genomic regions can give rise to the population-averaged structural organization of chromatin at different resolutions (Chiang et al., 2019; Falk et al., 2019; Shi et al., 2018). Here, we represent a ~ 1.95 Mb chromatin region as a chain of beads using a bead-spring polymer model. The model system consists of 1950 beads where beads represent genomic regions of 1 kb. Each bead is assigned a specific chromatin state as described above. In addition to that, we also model loops within the specific chromatin regions based on publicly available high-resolution Hi-C derived chromatin loop sets from Rao *et al*. (Rao et al., 2014). For each loop, we connect the loop anchors using a harmonic constraint. The potential energy function of the polymer model is described below,

$$\mathrm{U}_{\mathrm{ref}}(r) = U_B(r) + U_P(r) + U_L(r) + U_W(r)$$

where $U_B(r)$ corresponds to bond stretching potential, modeled by finite extensible nonlinear elastic potential given by,

$$U_B(r) = -0.5\mathrm{K_S}R_0^2 \ln\left[1 - \left(\frac{r}{R_0}\right)^2\right]$$

where $\mathrm{K_S} = 30.0\frac{\epsilon}{\sigma^2}$ is spring constant and $\mathrm{R_0} = 1.6\sigma$ is the equilibrium bond length.

We use Lennard-Jones potential to model steric repulsion and attraction between the beads of different chromatin state with different parameters. For example, the potential between a bead of chromatin state A and that of B is represented below,

$$U_P(r) = 4\epsilon_{\mathrm{AB}}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where ϵ_{AB} denotes interaction strength between bead types A and B (Appendix: Table 4) and σ refers to the size of each bead.

In case of loops, we use harmonic potential between each of the loop anchors $U_L(r)$, represented by,

$$U_L(r) = K_L(r - R_0^L)^2$$

where $\mathrm{K_L} = 300\frac{\epsilon}{\sigma^2}$ is spring constant and $\mathrm{R_0^L} = 1.2\sigma$ is the equilibrium loop bond length.

Finally, in order to model the repulsive interaction between the chromatin beads and the simulation box boundaries, we use Lennard-Jones potential

$$U_W(r) = 4\epsilon_{\mathrm{W}}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where $\epsilon_{\mathrm{W}} = 1.0$ denotes the interaction strength between boundary walls and chromatin beads irrespective of their types.

The LAMMPS molecular dynamics package is used to perform polymer dynamics simulations (Plimpton, 1995). All the simulations are performed in terms of Lennard-Jones reduced units with τ for time, σ for length and ϵ for energy. To mimic the viscoelastic nature of the nucleoplasm, Langevin dynamics with a damping coefficient of $\gamma = 1.0\tau$ is used to maintain the temperature at $T = 1.0$. We then run the simulation for a total of 1,000,000 steps with a timestep of $dt = 0.000001\tau$. Followed by that, the timestep dt is increased to 0.01τ and a production run of a total 2,000,000 steps is performed and the

conformations are saved at intervals of 1,000 steps. We repeated this process for 20 times, each time with different initial conformation and random seeds.

Initial conditions and Equilibration The chromatin fiber is initialized as a self-avoiding worm like chain using the python package PolymerCpp (Stefko & Douglass, 2017) and is placed in a box of size $L = 35\sigma$ with fixed repulsive boundaries. We then allow the polymer to equilibrate inside the box using four consecutive phases. In the first phase, a soft potential is used to remove overlap in the polymer for 400,000 steps with a $dt = 0.01\tau$.

$$U_S(r) = \mathrm{A}\left[1 + cos\left(\frac{\pi \mathrm{r}}{\mathrm{R}_0^S}\right)^6\right]$$

where A is pre-factor, varied gradually from 0 to 100ϵ and $\mathrm{R}_0^S = 2^{1/6}\sigma$.

For the rest three subsequent phases, we switch to potential to Lennard-Jones potential and run the simulation for 400,000, 500,000 and 500,000 steps with timesteps 0.000001τ, 0.0001τ, and 0.01τ respectively.

$$U_G(r) = 4\epsilon_{\mathrm{G}}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where $\epsilon_{\mathrm{G}} = 1.0$ denotes the interaction strength between chromatin beads irrespective of their types.

Loop extrusion simulation of chromatin segments based on underlying RAD21 binding sites

To test the contribution of loop extrusion dynamics to the variability present in chromosome structural ensemble, here we perform loop extrusion polymer simulations of a specific chromatin segment according to the method described in Fudenberg et al. (Fudenberg et al., 2016). For these simulations, we define the domain boundaries based on RAD21 peaks present in the desired chromatin segment. Here, we model a 2 Mb spanning chromatin region as a 2000 beads monomer chain where beads represent genomic regions of 1 kb. All the simulation parameters are kept same between different cell types except the position of the boundary elements (here RAD21 peaks) and their strengths. The strength of boundary elements is defined in terms of boundary permeability parameter in the models which represents the boundary insulation. For each cell-type, these parameter values are obtained by calculating the relative strength of the RAD21 peaks. The RAD21 peaks for the IMR90 and K562 were downloaded from the ENCODE repository (Davis et al., 2018).

Analysis of nucleolus-associated domain (NAD) mapping data

In order to characterize the nucleolar association of the IMR90 chromosomal regions, we downloaded the NAD mapping data of the proliferating IMR90 cells from Dillinger et al. (Dillinger et al., 2017). This microarray data contains log2 normalized ratio of nucleolus-associated and genomic DNA at specific probe locations. For our analysis, we bin that log2 normalized ratio data at 10 kb resolution using multiBigwigSummary functionality of the deepTools software (Ramírez et al., 2014).

Results

Chromosome structural landscape reveals major features of the structural ensemble

We first apply our statistical analysis technique to the single-cell chromosome structures measured for a 2 Mb region of IMR90 chr21 (28-30 Mb) (Fig. 22a) and produce a chromosome structural landscape (Fig. 22b). We observe that the first major source of variation corresponds well to the compactness of the chromosome structures, which can be quantified using the radius of gyration, R_g (Appendix: Fig. 28a). We calculate the R_g of the single-cell structures from their experimentally obtained 3D conformations. When we compare the R_g values of the structures with their PC1 values (from here on we refer to the PC1 projection as "PC1"), we observe that the structures with more negative PC1 projections are highly compact with smaller R_g values while structures with more positive PC1 values are less compact with higher R_g scores. With increasing R_g, the structural divergence also increases, with individual structure points becoming spread further along PC2 and also more spread out from each other.

Along PC2, we observe that the structure ordering follows a progressive reorganization of the domains within the region and captures several subclusters of structures (Fig. 22c). As we move from negative to positive PC2, we observe a shifting of a boundary (weakening of one boundary and gaining of another) that leads to the change in the structure of the domain (Appendix: Fig. 28b). This phenomenon can be seen in different regimes of R_g, showing the existence of domain reorganization within both compact and open structures (Appendix: Fig. 28b).
When the same approach is applied to different chromosomes and cell-types at different resolutions, we found a consistent ordering of the dynamic modes of the structural landscape (Appendix: Fig. 28c, Appendix: Fig. 28d, Appendix: Fig. 29 and Appendix: Fig. 30). Although the major dynamic variation remains the R_g ordering for all the different situations, the reorganization of domains along PC2 is dependent on the length of the genomic region as well as the local structural properties.

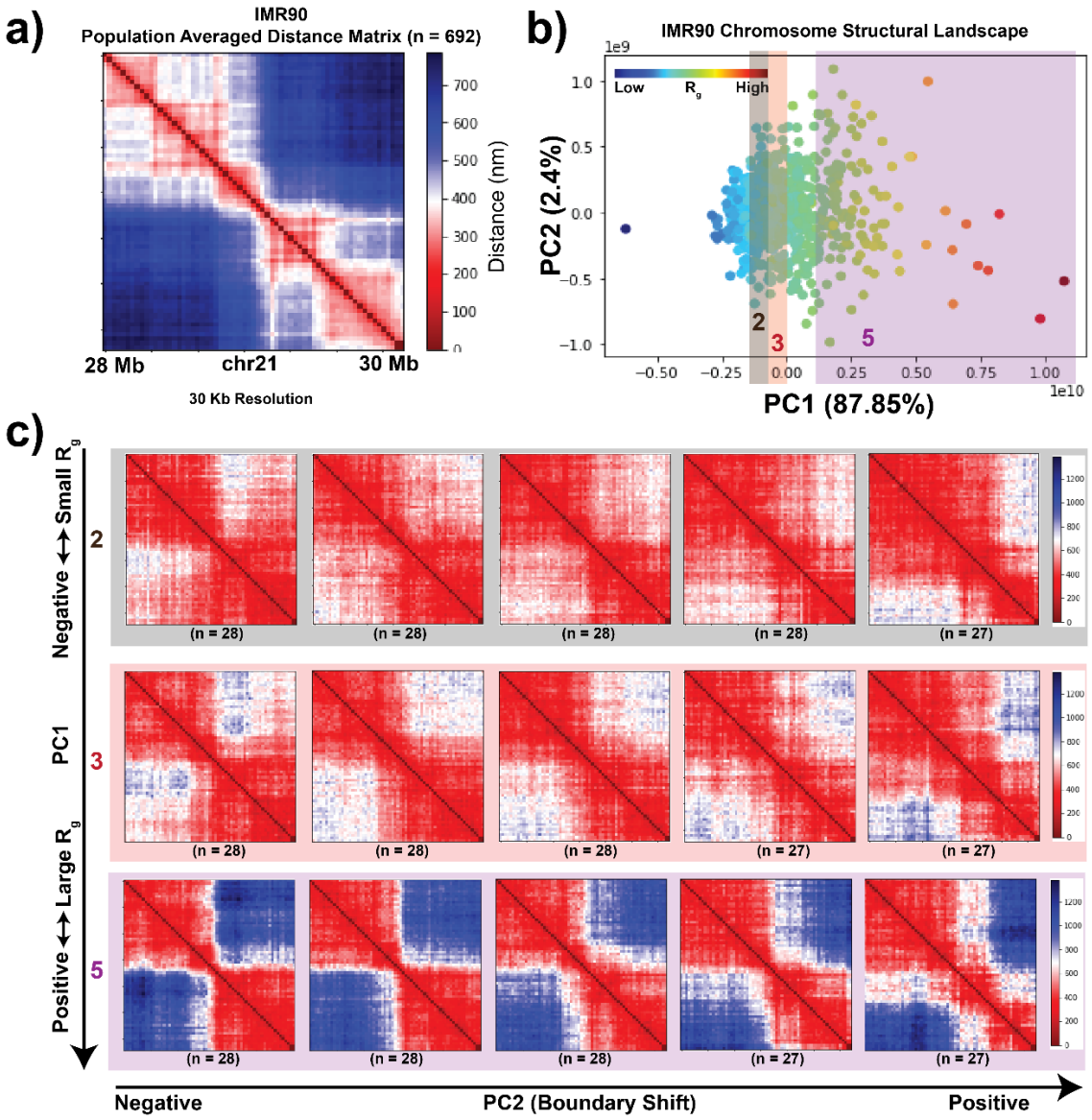


Figure 22 The chromosome structural landscape orders single-cell structures on a continuum based on their radius of gyration followed by domain reorganization.

a) Population-averaged distance map of IMR90 chr21:28-30 Mb at 30 kb resolution. b) Chromosome structural landscape of IMR90 chr21:28-30 Mb chromatin region, generated as described in the Methods section. Each point represents a single-cell structure. Points are colored according to the radius of gyration ($\boldsymbol{R_g}$) of the corresponding structures (dark blue – smaller $\boldsymbol{R_g}$ and dark red – larger $\boldsymbol{R_g}$). The percent of variation explained by each PC is shown in parentheses. c) Structures are first divided into five groups based on the PC1 ordering (which correlates highly with $\boldsymbol{R_g}$). Structures belonging to groups 2, 3 and 5 are shown. Within each group structures are further divided into five subgroups based on their PC2 ordering. Average structures within each subgroup are shown.

Ensemble structure similarity index captures cell-type- and chromosome-specific features of the chromosome structural ensemble

To compare the single cell structural heterogeneity between different cell types, we constructed a joint structural landscape using the 28-30 Mb chr21 region data from three cell types – IMR90 (lung fibroblast), K562 (leukemia) and A549 (lung adenocarcinoma) collectively. In this joint structural landscape we can visualize the cell-type-specific variation present in structural ensemble (Fig. 23a). To quantify the differences between cell types, we divide the landscape in 10 equally sized regimes along PC1 (since PC1 correlates with R_g, different PC1 regimes correspond to different R_g values) and calculated the proportion of different cell types within each regime. We observe that K562 shows a higher proportion of chromosome structures in lower R_g regimes and has an overall more compact structural ensemble compared to A549 and IMR90 (Fig. 23a and Fig. 23b).

While the relative R_g shift between cell types is straightforward to observe and quantify, other features of the chromosome structural landscape are more difficult to compare visually. Therefore, we devised a network-based measurement to quantitatively compare the structural landscapes, as discussed in the Methods section. According to this approach, a network with similar edge weights will exhibit a higher ensemble SSI value and vice-versa. In terms of the chromosome structural ensemble, a set of structures that smoothly transition across the entire conformational space will have a highly connected structural landscape and thus high ensemble SSI (Fig. 1a, blue line). On the other hand, if structures coalesce around discrete states (Fig. 21a, red line), the discontinuity in the structural ensemble will be higher, and the resulting network will have a larger variation in of edge weights, and a low ensemble SSI score will be produced.

We first applied the network-based similarity quantification approach to structures of 2 Mb regions of chr21 (30 Kb resolution) in different cell-types and conditions obtained from Bintu et al. (Bintu et al., 2018). There is variation in the sample size (i.e., the number of conformations) obtained for different experiments, so we relied on a resampling approach to make the comparison consistent. Based on this approach, when we compared the 28-30 Mb genomic region from three different cell-types analyzed above, we find that IMR90 has the highest ensemble SSI, followed by K562 and then A549 with the lowest ensemble SSI (Fig. 23c). Thus, this region in IMR90 cells has the highest continuity across different structures while the corresponding distribution in A549 cells exists as a relatively discrete set of structures.

To evaluate our interpretation of the ensemble SSI result in an orthogonal fashion, we performed hierarchical clustering of the 3D structures obtained from IMR90 and K562 cells based on a structure similarity metric, Q, as represented in Cheng et al. (Appendix: Fig. 31). At the same threshold, both cell types showed the same number of structure clusters. However, in IMR90, the vast majority of structures fall into one of these clusters, corresponding to our high ensemble SSI value (Appendix: Fig. 31a and b). In K562, the structures are more evenly split into the different clusters, corresponding to our low

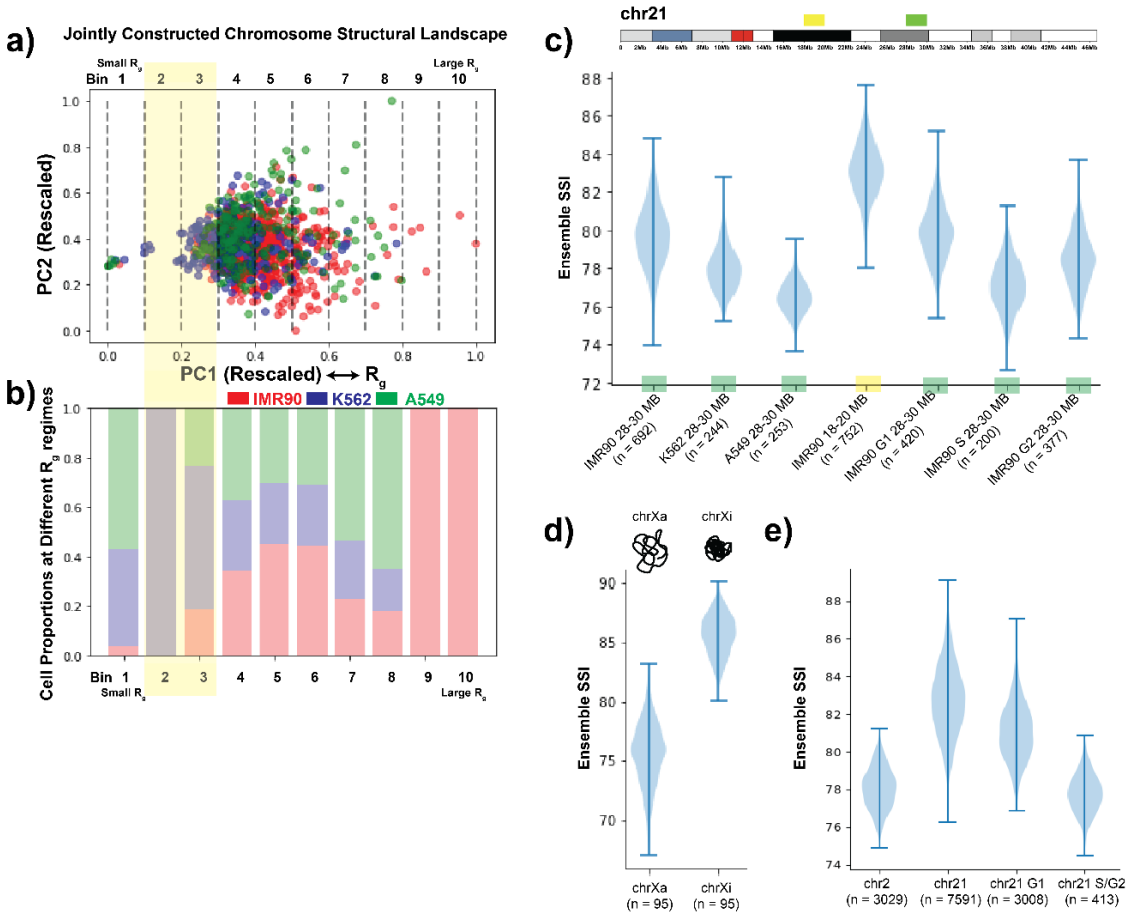


Figure 23 Network based similarity score quantifies cell-type- and chromosome-specific variability present in chromosome structural ensemble at different resolutions.

a) Jointly constructed chromosome structural landscape of genomic region chr21:28-30 Mb for cell types IMR90, K562 and A549. The only difference between individual and joint landscape is that all the different cell-type structures are considered together while calculating pairwise structure comparison matrix. Points are colored based on the cell types. Further, the landscape is divided into ten equal sized bins (or regions) based on PC1 values, where the bins on the left represent lower $\boldsymbol{R_g}$ regions and bins on the right are with higher $\boldsymbol{R_g}$. b) Relative proportion of IMR90, K562 and A549 single-cell structures present in each bin represented in a). c) Chromosome structural ensemble SSI distributions for different 2 Mb chromosome regions from different cell-types and conditions, obtained from Bintu et al. (Bintu et al., 2018). The SSI value of the structural ensemble is calculated using a network entropy based approach. For each structural ensemble (size of ensemble indicated as n), 200 single-cell structures are selected randomly (except IMR90 S 28-30 Mb where we select 150 structures for resampling), and the median of the SSI values of the nodes are calculated. This step is repeated 1000 times and values are plotted as a distribution. d) Chromosome structural ensemble SSI distributions for active and inactive IMR90 chrX at TAD resolution, obtained from Wang et al. (Wang et al., 2016). 50 single-cell structures are selected randomly each time for resampling. e) Chromosome structural ensemble SSI distributions for different IMR90 chromosomes and conditions at higher resolution, obtained from Su et al. (Su et al., 2020). Here, for resampling 300 single-cell structures are selected randomly each time.

ensemble SSI value (Appendix: Fig. 31c and d). Thus, our ensemble SSI metric indeed captures the degree of continuity within the structural ensemble. By using ensemble SSI rather than a hierarchical clustering approach, we avoid the need to arbitrarily select a cutoff or number of clusters. We further assessed the dependence of the ensemble SSI score on the number of cells measured (Appendix: Fig. 32). We observed that with an increase in the number of cells, the difference in ensemble SSI between cell-types follows the same trend, but the absolute difference becomes smaller (Appendix: Fig. 32b). It is important to note that as we increase the number of cells, we are including more low-quality measurements (cells with more missing data) and this may also contribute to a broader ensemble SSI distribution. We then examined how robust the ensemble SSI distribution is to changes in the correlation matrix cutoff chosen (Appendix: Fig 32c). Using a cutoff lower than our initially chosen 50th percentile produces similar results, but increases variability as more non-significant connections between structures are included. A high cutoff, on the other hand, can obscure the results by breaking a heterogeneous structure landscape into too many disconnected networks. Within each sub-network, there is high continuity, so this results in an artificially high SSI being reported for landscapes that are in fact heterogeneous. We conclude that a 50th percentile correlation matrix cutoff excludes noisy connections without creating disconnected subnetworks (see Appendix: Fig. 32c for extended discussion). As a final technical test of our metric, we repeated the ensemble SSI analysis after removing extremely compact or large structures from each cell type to see whether those extreme conformations are disproportionately affecting the results (Appendix: Fig. 33). We found that those extreme structures actually have little influence on the ensemble SSI calculation. Overall, we conclude that our ensemble SSI calculation is robust.

We next compared two different genomic regions of IMR90 chr21 using the ensemble SSI distribution. We found that although both are from same cell-type and same chromosome, the 28-30 Mb region has lower ensemble SSI values compared to 18-20 Mb region (Fig. 23c). This difference indicates that the latter region has a higher degree of continuity in its structural ensemble. This 18-20 Mb region lies entirely in the B compartment (putative inactive, gene poor regions) while the 28-30 Mb region contains a majority of A compartment regions mixed with B compartment regions. To further investigate the potential connection between compact, inactive states and the continuity of the structural landscape, we compared the ensemble structural similarity of the active vs. inactive chrX structural ensembles (440 Kb to 7.8 Mb) (Fig. 23d). We found that chrXi has higher ensemble SSI values compared to its active counterpart. This suggests that the chromosome with a highly compact global structure has a more continuous structural landscape compared to the more specific and discrete states of the open and active chromosome (Appendix: Fig. 34a, Appendix: Fig. 34b and Appendix: Fig. 34c) (Cheng et al., 2021).

When we analyzed the structural variation of the entire chr2 and chr21 q-arm at high resolution (Su et al., 2020), we observed that chr21 has a higher ensemble SSI compared to chr2 (Fig. 23e). To test whether the difference in length between chr2 and chr21 contributed to differences in the ensemble SSI values, we calculated the ensemble SSI of

regions of different lengths selected from chr2. We observed that as the length increases the ensemble similarity decreases (Appendix: Fig. 34d). This effect likely stems from the increased number of possible conformations that a longer polymer can adopt, which corresponds to a sparser coverage of conformational space. We also found that ensemble similarity is lower in S/G2 phase compared to G1 phase for chr21. This result is consistent with the ensemble SSI values of 28-30 Mb IMR90 chr21 region structural ensembles from different cell-cycle stages, where the G1 phase has highest ensemble SSI followed by S and G2 phases, as shown in Fig. 23a. One major difference between the chromosomes in the phases of the cell cycle is the addition of tethers between sister chromatids in S and G2 compared to G1 (Mitter et al., 2020). Increased constraints on the chromosome structure could then lead to more distinct structures and lower structural ensemble similarity.

Contribution of epigenetic properties to the cell-type-specific variation of chromosome structural ensemble

Building on the observation that regions with different A/B spatial compartment identity had different structural landscape SSI, we further examined the influence of local epigenetic state on the resulting structural landscape. Recent studies have shown the association of epigenetic states with the ensemble genome organization (Di Pierro et al., 2017; Di Pierro et al., 2018; Szabo et al., 2019). Figure 24a shows chromatin states along the 28-30 Mb genomic region in the three different cell types – IMR90, K562 and A549 (detailed steps in the Methods section). We observe that this region in IMR90 mostly has active marks while in K562 the region is mixed between repressive polycomb and heterochromatin marks as well as active marks. The A549 region shows a mixture of active, repressive, and poised marks. By qualitatively comparing this chromatin state data with the ensemble SSI distributions for these three cell types, we observe that the cell type with the more uniform epigenetic state (IMR90) has the highest ensemble SSI while cell types with mixed epigenetic states exhibit lower SSI.

In addition to the overall dynamic properties of the region, we sought to quantitatively compare local genomic bin epigenetic state with the degree of structure fluctuation experienced by that bin. We first calculated the bin variability for each cell type (Methods) which represents the structure variability of a particular genomic region across the population of cells (Fig. 24b). We do not see a direct bin-by-bin correlation between epigenetic state and bin variability. Instead, we observe that a higher proportion of active states within the region tends to correspond to higher bin variability across the region while a larger proportion of inactive states, even if focused in one part of the locus, corresponds to restricted bin variability across the whole region. In this comparison, it is important to remember that we are comparing a population averaged epigenetic state with single cell structural fluctuation. It is possible that additional relationships could become evident if single cell variation in epigenetic state were considered.

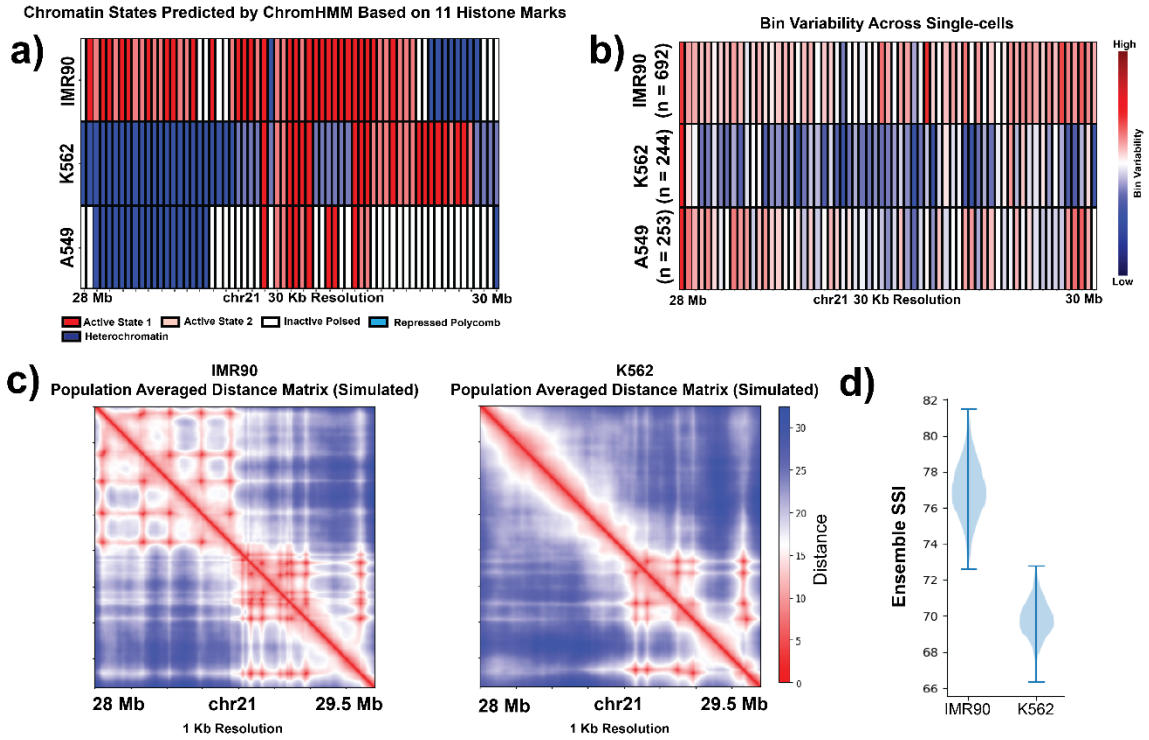


Figure 24 Epigenetic properties (histone modifications) contribute to the cell-type-specific variation in the chromosome structural ensemble.

a) Chromatin states of chr21:28-30 Mb genomic region from three different cell-types – IMR90, K562 and A549 at 30 kb resolution. The chromatin states are inferred based on 11 different histone marks for each cell-type by ChromHMM tool. b) Bin variability of the chr21:28-30 Mb region from IMR90, K562 and A549, where a higher value represents higher degree of structural variation of that genomic position across the population and vice versa. c) Population-averaged distance maps of IMR90 and K562 chr21:28-29.5 Mb genomic region from polymer simulations (see Methods) at 1 kb resolution. d) Chromosome structural ensemble SSI distributions for chr21:28-29.5 Mb from simulated IMR90 and K562. For resampling, 500 single-cell structures are selected randomly each time.

To test our hypothesis that epigenetic state differences contribute to cell-type-specific variation in the chromosome structural ensemble, we performed a set of polymer simulations of the desired genomic region of 28-29.5 Mb at a resolution of 1 kb, as described in the Methods section. We generated the models for IMR90 and K562 separately by assigning interactions between the polymer beads based on their specific chromatin states (Appendix: Fig. 35a). Once we obtained the 3D conformations from the simulations, we constructed pairwise Euclidean distance matrices for each conformation and further quantified the structural ensemble using our structural similarity approach. For the resampling purpose, we combined all the snapshots from 20 independent simulations and then selected 500 conformations randomly from this set 1000 times. When we compared the simulated IMR90 ensemble SSI distribution with simulated K562 data, we found that the K562 structural ensemble shows a lower similarity, which is the same trend observed in our analysis of the Bintu et al. imaging data (Fig. 23c and Fig. 24d). Since the only difference between cell types in these simulations was the encoded epigenetic state of the polymer beads, this result corroborates the idea that the epigenetic state is one factor that can contribute to the variation in chromosome structural ensemble.

Structure heterogeneity properties are associated with epigenetic state and boundary positions

We further evaluated the contribution of epigenetic factors to structural heterogeneity by analyzing single-cell structures of six genomic regions, each 2 Mb in length, obtained from IMR90 chr21 whole q-arm imaging data (Fig. 25a) (Su et al., 2020). For simplicity, the regions will be denoted as S1 to S6 based on their starting genomic coordinates. The regions S1 (chr21:18-20 Mb) and S2 (chr21:28-30 Mb) were selected in order to match the two IMR90 genomic regions data analyzed in Bintu et al. The three regions S3, S5 and S6 were selected based on their epigenetic state similarity with epigenetic state of the chr21:28-30 Mb region in K562 cells (Appendix: Fig. 35b). To evaluate epigenetic state similarity, we designed a sliding window-based approach where we move a window of size 40 (2 Mb genomic region with 50 kb resolution) over the whole q-arm of chr21 and for each window compare the IMR90 epigenetic state with the K562 28-30 Mb region epigenetic state using binary comparison (Appendix: Fig. 35c). However, as the resolution of the Bintu et al. data (65 bins per 2 Mb region) is different from the Su et al. data (40 bins per 2 Mb region), we take the least common multiple of the two bin numbers and expanded each data to 250 bins by repeating each bin element the required number of times. Here, a higher comparison value shows higher similarity with the K562 epigenetic state. Among the three regions, S3 has the highest score, S6 has the second highest and S5 has the third highest one (Appendix: Fig. 35c). The region S4 was selected randomly.

When we quantified the variation in the structural ensemble of these six regions, we first observed that the S1 and S2 regions have ensemble SSI distributions that are consistent with the results from the same regions in the Bintu et al. data (Fig. 25b). Next, we found that the S3 region has the lowest ensemble SSI distribution among all the six regions which

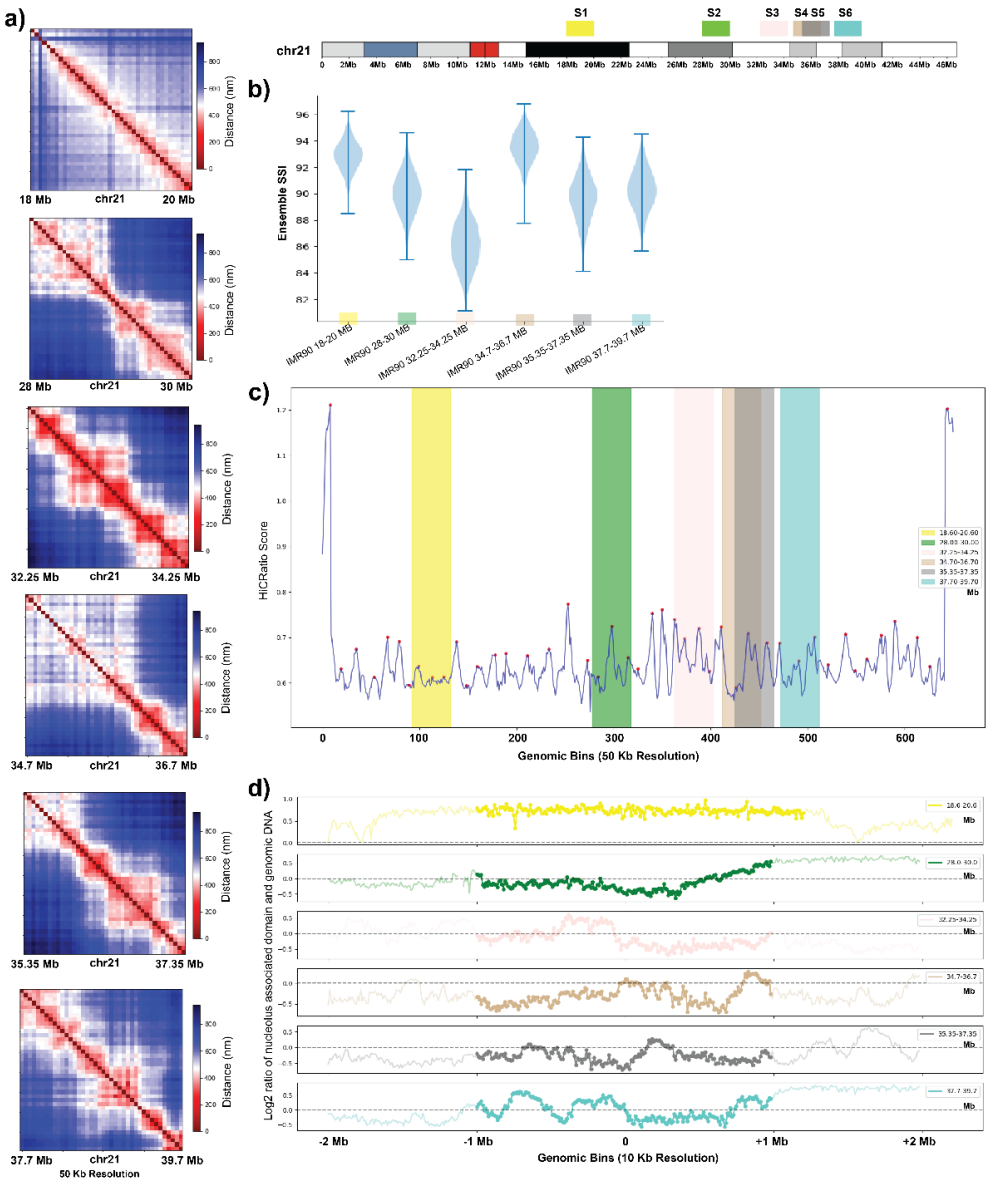


Figure 25 Evaluation of structural variation of regions with different epigenetic, genomic, and structural properties.

a) Population-averaged distance maps of 6 different genomic regions of IMR90 chr21 at 50 kb resolution obtained from Su et al. (Su et al., 2020). Locations of these 6 regions are indicated on the top right chromosome ideogram and labeled with colors used throughout the figure. b) Chromosome structural ensemble SSI distributions for the same 6 regions of IMR90 chr21. For resampling, 300 single-cell structures are selected randomly each time. c) HiCRatio score of each genomic bin of chr21 q-arm at 50 kb resolution (calculated with a 300 kb window size). The red points represent the boundary positions inferred by HiCRatio. d) Nucleolus association level of the same 6 regions of IMR90 chr21 (62). For each genomic region, the 1 Mb upstream and downstream regions are also shown. A value greater than 0 represents higher levels of nucleolus association and vice versa. Colors in each panel correspond to chromosome locations indicated above panel b.

is also consistent with the low ensemble SSI of the Bintu et al. K562 region. In addition, the ordering of S6 and S5 are also consistent with their similarity with K562 epigenetic state. Finally, the S4 region, which has a very low similarity score with K562 state, exhibits a very high ensemble SSI distribution. Again, this analysis supports the idea that epigenetic patterns contribute to the degree of variation present in the chromosome structural ensemble. In addition to the epigenetic states, we also identified domain boundaries along the whole q-arm of chr21 using the HiCRatio approach (with a 300 kb window size) (Lazaris et al., 2017). Since this boundary calling software is designed for contact data, we first converted the single-cell chromosome structural distance data into contact data using a cutoff of 500 nm, as reported here. We found that regions with a higher number of strong domain boundaries exhibit lower ensemble similarity and vice versa (Fig. 25c). For example, the S3 region, which has the lowest ensemble SSI, has three strong domain boundaries and one weak. On the other hand, regions like S1 and S4 which exhibit higher ensemble SSIs have fewer or weaker domain boundaries. This effect is visible across the entire chromosome, in that the average boundary strength in a given region is negatively correlated with the ensemble SSI of that region (Appendix: Fig. 35d, e).

Finally, we asked whether chromosome association with the nucleolus is associated with changes in the chromosome structural ensemble SSI (Fig. 25d). The NAD data source and processing steps are described in the Methods section. We observe that S1 and S4, the regions with the highest structural ensemble similarity, are the most uniform in their NAD state: S1 is completely associated with the nucleolus, while S4 has almost no association with the nucleolus. As seen with epigenetic state, here we find that regions with uniform nucleolus association exhibit a higher continuity level in the structural ensemble. In contrast, S3 has the lowest ensemble SSI and also alternating levels of nucleolus association in the first half of the genomic region, followed by an unconstrained region. Thus, variation in constraints across the region is again associated with specific distinct structures and thus a lower ensemble similarity.

Loop extrusion alone does not reproduce experimentally detected chromosome structure variation patterns

Above, we observed that chromosome regions with a higher number or strength of domain boundaries showed less continuous structural landscapes and lower ensemble SSI values. Boundaries detected by approaches like HiCRatio can result from cohesin-mediated loop extrusion that is periodically blocked by factors such as CTCF (Davidson & Peters, 2021), as well as from the borders between A and B compartment regions of chromatin. We next therefore asked whether the loop extrusion process, blocked by different numbers of boundaries, would be expected to generate the patterns of structure variation we observe. We simulated loop extrusion using the chr21:28-30 Mb segment using experimental cohesin binding data from K562 and IMR90 (see Methods). This region has fewer boundaries in K562 (likely due to the portion of the region that is heterochromatic in this cell type). The structure landscapes resulting from a single simulation run show that the

cell type with more loop extrusion boundaries (IMR90) would be predicted to have a lower ensemble SSI than the cell type with fewer boundaries if only loop extrusion were at work (Appendix: Fig. 36a and 36b). Once we start the simulation from several different initial conditions, this difference is diminished (Appendix: Fig. 36b) due to variation in structure between different runs. While this result matches our expectations that more boundaries associate with lower ensemble SSI, it does not agree with the experimental observations from these cell types, in which IMR90 actually has the higher ensemble SSI. This suggests that epigenetic state-specific interactions must be considered along with loop extrusion to explain the chromosome variability observed.

The effect of cohesin depletion on structural heterogeneity depends on the underlying epigenetic state of the chromosome region

To further examine the effect of cohesin on the chromosome structural landscape, we analyzed chr21 single-cell structural data from HCT116 cells in wildtype (WT) and cohesin depleted (+Auxin) conditions (Bintu et al., 2018). This analysis included two different genomic regions (chr21:28-30 Mb and chr21:34-37 Mb) with different underlying structural properties. The jointly constructed chromosome structural landscape for both genomic regions shows that along PC1 the structures are again ordered based on their R_g (Appendix: Fig. 36c and 36d). Thus, the chromosome structural landscape approach orders single-cell structures based on their R_g irrespective of cell-type, genomic regions, and perturbation condition. The WT and cohesin-depleted structural landscapes from the first genomic region mostly overlap with a slight bias for the +Auxin structures in the higher R_g regime (Fig. 26c). This suggests that removal of cohesin makes the structures less compact to some extent, as previously observed (Hao et al., 2021).

On the other hand, for the second region, the jointly constructed landscape shows a shift upwards along PC2 along with the shift towards the large R_g regime for cohesin-depleted structures (Fig. 26d). In addition, we found that after cohesin removal, both genomic regions exhibit more structural variation at the genomic bin level according to BSI standard deviation (Appendix: Fig. 36e and Fig. 36f). When we examine the structural ensemble SSI of these two regions, we see apparently conflicting effects of cohesin removal: the ensemble SSI is decreased in the 28-30 Mb region but mostly unchanged in the 34-37 Mb region (Fig 26e). Note that the initial ensemble SSI of these regions in the WT follows the expected pattern for regions of different lengths. We observed that this divergent ensemble SSI effect corresponds to a difference in the underlying compartment state of the regions. The 28-30 Mb region is divided into A and B compartment segments (which in earlier examples tends to correspond to lower SSI) (Fig. 26a) while the 34-37 Mb region is all in the A compartment with no compartment boundaries (which in earlier examples tends to correspond to higher SSI) (Fig. 26b).

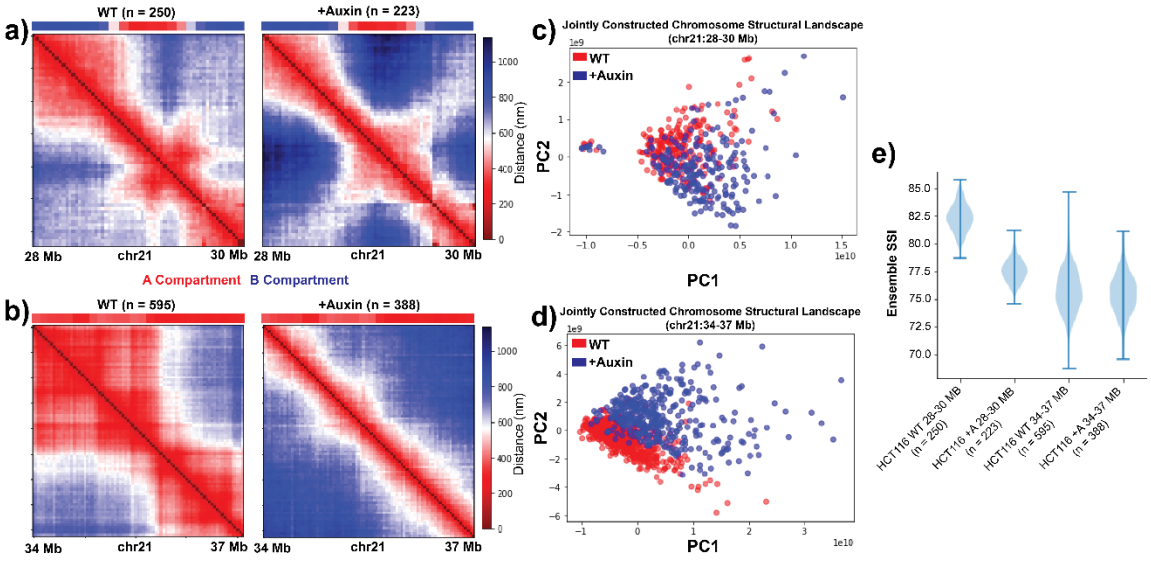


Figure 26 Cohesin depletion alters chromosome structural heterogeneity depending on underlying epigenetic state constraints.

a,b) Population-averaged distance maps of HCT116 chr21:28-30 Mb (a) or chr21:34-37 Mb (b) at 30 kb resolution for wild-type (WT) and cohesin-depleted (+Auxin) conditions based on data obtained from Bintu et al. (Bintu et al., 2018). Compartment status from Hi-C data at these locations shown above each matrix. c,d) Jointly constructed chromosome structural landscapes of chr21:28-30 Mb (c) or chr21:34-37 Mb (d) of HCT116 for WT and +Auxin conditions. e) Chromosome structural ensemble SSI distributions for different local chromosome regions of HCT116 for WT and +Auxin conditions. 200 single-cell structures are selected randomly each time for resampling.

Discussion

As the number and quality of single cell chromosome structural datasets continues to increase, a diverse toolbox of methods to capture properties of chromosome structure variation is needed. Our statistical analysis of single cell chromosome structures allows us to characterize both the major modes of chromosome structure fluctuation and the properties of the structure fluctuation landscape. While some approaches focus on classifying single cell structures into discrete clusters, other recent work has shown that chromosome structures can exist along a continuum (Cheng et al., 2020; Conte et al., 2020; Sawh et al., 2020; Takei et al., 2021; Xie et al., 2020). Our analysis does not require a decision to be made about how many different clusters of structures exist, but instead can capture the overall types of structure variations and can measure the degree of continuity between the structures.

For all observed length scales of chromosome structures, from local TAD scale to whole chromosome arms, we find that the primary type of single cell structure variation is the overall compaction of the region. For each structure assayed, there exist some cells in which the region being examined is very compact and others in which it is much more extended. This overall size fluctuation is true of both active and inactive regions, though inactive regions and the inactive X chromosome show a shift toward smaller sizes overall. However, we find that this overall region compaction is largely independent of a secondary fluctuation mode: the reorganization of domains. No matter how compact or extended the overall region is, we observe a similar progressive reorganization of the domain boundaries as a secondary type of structure variation.

By quantifying the network-based similarity properties of the structural ensemble, we can measure how discrete vs. continuous the structures adopted by a certain chromosome region are. Combining all of our observations of ensemble SSI in different regions and conditions, we propose that the ensemble SSI of a structural ensemble is influenced by the epigenetic state constraints put on the region by its epigenetic state. These constraints may take the form of domain boundaries, where different parts of the chromosome regions are constrained in self-interacting patterns, or by associations with nuclear structures such as the nucleolus. Overall, constraints on the chromosome structure and boundaries between different types of regions decrease structural ensemble SSI and result in more discrete structures while regions with more uniform linear chromatin states exhibit a more continuous structural landscape with higher SSI.

It might seem likely that the active or inactive chromatin status of a region would be the primary contribution of epigenetics to the structural fluctuation of a region. That is, one might expect an inactive region to have high structural ensemble SSI and an active region low ensemble SSI or vice versa. However, we find that this is not the case. Instead, the uniformity of the epigenetic state appears to be a key factor, such that a uniformly active region shares structural landscape properties with a uniformly inactive region (both have high SSI) while a region divided between alternating states behaves differently (low SSI). This leads us to the idea that alternating epigenetic states place constraints on the possible

conformations and thus create a more disjointed, rather than continuous, structural ensemble.

The idea that epigenetic state-mediated constraints are key to shaping dynamic structural landscapes is echoed again by the results we observe in loop extrusion simulations and cohesin depletion data. Previous work has shown that compartmentalization and loop extrusion can act in opposition to each other (Nuebler et al., 2018). When cohesin is removed, we see that a region consisting of multiple epigenetic states can form subdomains and is thus more likely to adopt more discrete structures, resulting in a decrease in ensemble SSI compared to the state where cohesin is active. When cohesin is active, there is a chance of the entire region being pulled into the same loop, creating a more continuous landscape of possible structures (higher SSI). In contrast, when a region exists in a uniform epigenetic state, we see that the removal of cohesin does not result in a reduction of SSI. As before, cohesin removal is expected to reduce the active movement of chromatin, but in this case, the region is entirely in the same compartment, so there are no additional constraints that would lead to the formation of discrete states in the cohesin-depleted structural landscape. Our observations are consistent with the previously published idea that cohesin loop extrusion acts to increase chromosome conformation dynamics, promoting distal contacts and countering the tendency of domains to settle into more discrete sub-structures (Nuebler et al., 2018; Schwarzer et al., 2017). Our loop extrusion simulations suggest that in a uniform epigenetic state, the region with more boundaries would adopt more discrete structures and thus have less landscape continuity. But, loop extrusion alone, without consideration of interactions between similar chromatin states, cannot reproduce the differences in structure variability we see between different cell types.

The analysis of cohesin depletion results also reiterate the value of considering both the structural landscape and the ensemble SSI of that landscape. The structural landscape of a region can change (adopting a different range of radius of gyration or domain structures) while the continuity (SSI) stays the same. Alternately, the ensemble SSI can change while the structural landscape is overall conserved, indicating cases where the range and types of structures a region can adopt are the same, but the region is more often in certain discrete structures within that range rather than continuously sampling the range.

Availability

Example analysis code is provided on the ChromosomeStructuralEnsembleCharacterization GitHub repository (https://github.com/rpmccordlab/ChromosomeStructuralEnsembleCharacterization).

Supplementary data

Supplementary Data are available online.

Acknowledgements

The authors would like to thank Dr. Rebeca San Martin for valuable suggestions and insightful discussions in the development of the research project. We thank Dr. Tian Hong for sharing of computational GPU resources.

Funding

This work was supported by the National Institutes of Health [NIGMS grant R35GM133557 to R.P.M]. Funding for open access charge: National Institutes of Health/ R35GM133557.

Conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

Arrastia, M. V., Jachowicz, J. W., Ollikainen, N., Curtis, M. S., Lai, C., Quinodoz, S. A., Selck, D. A., Ismagilov, R. F., & Guttman, M. (2021). Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nature Biotechnology*, 1-10.

Banerji, C. R., Miranda-Saavedra, D., Severini, S., Widschwendter, M., Enver, T., Zhou, J. X., & Teschendorff, A. E. (2013). Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Scientific reports*, *3*(1), 1-7.

Bersaglieri, C., Kresoja-Rakic, J., Gupta, S., Bär, D., Kuzyakiv, R., & Santoro, R. (2020). Genome-wide maps of nucleolus interactions reveal distinct layers of repressive chromatin domains. *bioRxiv*, 2020.2011.2017.386797. https://doi.org/10.1101/2020.11.17.386797

Bintu, B., Mateo, L. J., Su, J.-H., Sinnott-Armstrong, N. A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A. N., & Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, *362*(6413), eaau1783. https://doi.org/10.1126/science.aau1783

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., & Speicher, M. R. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, *3*(5), e157.

Braccioli, L., & de Wit, E. (2019). CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays in biochemistry*, *63*(1), 157-165.

Chanda, P., Costa, E., Hu, J., Sukumar, S., Van Hemert, J., & Walia, R. (2020). Information theory in computational biology: where we stand today. *Entropy*, *22*(6), 627.

Chen, H., Tian, Y., Shu, W., Bo, X., & Wang, S. (2012). Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS One*, *7*(7), e41374.

Cheng, F., Liu, C., Shen, B., & Zhao, Z. (2016). Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. *BMC systems biology*, *10*(3), 301-311.

Cheng, R. R., Contessoto, V. G., Aiden, E. L., Wolynes, P. G., Di Pierro, M., & Onuchic, J. N. (2020). Exploring chromosomal structural heterogeneity across multiple cell lines. *ELife*, *9*, e60312.

Cheng, Y., Liu, M., Hu, M., & Wang, S. (2021). TAD-like single-cell domain structures exist on both active and inactive X chromosomes and persist under epigenetic perturbations. *bioRxiv*, 2021.2005.2012.443887. https://doi.org/10.1101/2021.05.12.443887

Chiang, M., Michieletto, D., Brackley, C. A., Rattanavirotkul, N., Mohammed, H., Marenduzzo, D., & Chandra, T. (2019). Polymer modeling predicts chromosome reorganization in senescence. *Cell reports*, *28*(12), 3212-3223. e3216.

Clapier, C. R., Iwasa, J., Cairns, B. R., & Peterson, C. L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology*, *18*(7), 407-422.

Conte, M., Fiorillo, L., Bianco, S., Chiariello, A. M., Esposito, A., & Nicodemi, M. (2020). Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. *Nature communications*, *11*(1), 1-13.

Cremer, M., Hase, J., Volm, T., Brero, A., Kreth, G., Walter, J., Fischer, C., Solovei, I., Cremer, C., & Cremer, T. (2001). Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Res*, *9*. https://doi.org/10.1023/A:1012495201697

Davidson, I. F., & Peters, J.-M. (2021). Genome folding through loop extrusion by SMC complexes. *Nature Reviews Molecular Cell Biology*, *22*(7), 445-464.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., & Narayanan, A. K. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research*, *46*(D1), D794-D801.

Di Pierro, M., Cheng, R. R., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, *114*(46), 12126-12131.

Di Pierro, M., Potoyan, D. A., Wolynes, P. G., & Onuchic, J. N. (2018). Anomalous diffusion, spatial coherence, and viscoelasticity from the energy landscape of human chromosomes. *Proceedings of the National Academy of Sciences*, *115*(30), 7753-7758.

Dillinger, S., Straub, T., & Nemeth, A. (2017). Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PLoS One*, *12*(6), e0178821.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376-380.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, *9*(3), 215-216.

Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B. R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., & Solovei, I. (2019). Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, *570*(7761), 395-399.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell reports*, *15*(9), 2038-2049.

Hao, X., Parmar, J. J., Lelandais, B., Aristov, A., Ouyang, W., Weber, C., & Zimmer, C. (2021). Super-resolution visualization and modeling of human chromosomal regions reveals cohesin-dependent loop structures. *Genome biology*, *22*(1), 1-20.

Jerkovic, I., & Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, 1-18.

Kadauke, S., & Blobel, G. A. (2009). Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1789*(1), 17-25.

Kempfer, R., & Pombo, A. (2020). Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, *21*(4), 207-226.

Kim, J. M., Visanpattanasin, P., Jou, V., Liu, S., Tang, X., Zheng, Q., Li, K. Y., Snedeker, J., Lavis, L. D., & Lionnet, T. (2021). Single-molecule imaging of chromatin remodelers reveals role of ATPase in promoting fast kinetics of target search and dissociation from chromatin. *bioRxiv*.

Kim, Y., & Yu, H. (2020). Shaping of the 3D genome by the ATPase machine cohesin. *Experimental & Molecular Medicine*, *52*(12), 1891-1897.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., Vries, S. S., Janssen, H., Amendola, M., Nolen, L. D., Bickmore, W. A., & Steensel, B. (2013). Single-cell dynamics of genome–nuclear lamina interactions. *Cell*, *153*. https://doi.org/10.1016/j.cell.2013.02.028

Lazaris, C., Kelly, S., Ntziachristos, P., Aifantis, I., & Tsirigos, A. (2017). HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. *BMC genomics*, *18*(1), 1-16.

Lieberman-Aiden, E., Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., & Dorschner, M. O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*. https://doi.org/10.1126/science.1181369

Mateo, L. J., Murphy, S. E., Hafner, A., Cinquini, I. S., Walker, C. A., & Boettiger, A. N. (2019). Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*, *568*(7750), 49-54.

McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Molecular cell*, *77*(4), 688-708.

Mitter, M., Gasser, C., Takacs, Z., Langer, C. C., Tang, W., Jessberger, G., Beales, C. T., Neuner, E., Ameres, S. L., & Peters, J.-M. (2020). Conformation of sister chromatids in the replicated human genome. *Nature*, *586*(7827), 139-144.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., & Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, *502*(7469), 59-64.

Nichols, M. H., & Corces, V. G. (2015). A CTCF code for 3D genome architecture. *Cell*, *162*(4), 703-705.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., & Sedat, J. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381-385.

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., & Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, *115*(29), E6697-E6706.

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, *117*(1), 1-19.

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., Noble, W. S., Duan, Z., & Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nature methods*, *14*(3), 263-266.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, *42*(W1), W187-W191.

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. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665-1680.

Saha, P., & Sarkar, B. K. (2020). Entropy based analysis of genetic information. Journal of Physics: Conference Series,

Sanborn, A. L., Rao, S. S., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., & Li, J. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, *112*(47), E6456-E6465.

Sanulli, S., Justin, N., Teissandier, A., Ancelin, K., Portoso, M., Caron, M., Michaud, A., Lombard, B., da Rocha, S. T., & Offer, J. (2015). Jarid2 methylation via the PRC2 complex regulates H3K27me3 deposition during cell differentiation. *Molecular cell*, *57*(5), 769-783.

Sawh, A. N., Shafer, M. E., Su, J.-H., Zhuang, X., Wang, S., & Mango, S. E. (2020). Lamina-dependent stretching and unconventional chromosome compartments in early C. elegans embryos. *Molecular cell*, *78*(1), 96-111. e116.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., & Mirny, L. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, *551*(7678), 51-56.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, *148*(3), 458-472.

Shi, G., Liu, L., Hyeon, C., & Thirumalai, D. (2018). Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nature communications*, *9*(1), 1-13.

Stack, S. M., Brown, D. B., & Dewey, W. (1977). Visualization of interphase chromosomes. *Journal of cell science*, *26*(1), 281-299.

Stefko, M., & Douglass, K. M. (2017). *PolymerCpp: 2D and 3D wormlike chain generator for Python and written in C++*. In https://polymercpp.readthedocs.io/en/latest/

Strickfaden, H., Zunhammer, A., Koningsbruggen, S., Köhler, D., & Cremer, T. (2010). 4d chromatin dynamics in cycling cells: Theodor Boveri's hypotheses revisited. *Nucleus*, *1*.

Strom, A. R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., & Karpen, G. H. (2017). Phase separation drives heterochromatin domain formation. *Nature*, *547*(7662), 241-245.

Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B., & Zhuang, X. (2020). Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, *182*(6), 1641-1659. e1626.

Szabo, Q., Bantignies, F., & Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Science advances*, *5*(4), eaaw1668.

Takei, Y., Zheng, S., Yun, J., Shah, S., Pierson, N., White, J., Schindler, S., Tischbirek, C. H., Yuan, G.-C., & Cai, L. (2021). Single-cell nuclear architecture across cell types in the mouse brain. *Science*, eabj1966.

Tan, L., Ma, W., Wu, H., Zheng, Y., Xing, D., Chen, R., Li, X., Daley, N., Deisseroth, K., & Xie, X. S. (2021). Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell*, *184*(3), 741-758. e717.

Teschendorff, A. E., & Severini, S. (2010). Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC systems biology*, *4*(1), 1-15.

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., & Gros, M. A. (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci*, *113*. https://doi.org/10.1073/pnas.1512577113

Wang, L., Gao, Y., Zheng, X., Liu, C., Dong, S., Li, R., Zhang, G., Wei, Y., Qu, H., & Li, Y. (2019). Histone modifications regulate chromatin compartmentalization by contributing to a phase separation mechanism. *Molecular cell*, *76*(4), 646-659. e646.

Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C.-t., & Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, *353*(6299), 598-602.

West, J., Bianconi, G., Severini, S., & Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. *Scientific reports*, *2*(1), 1-8.

Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y., & Plewczynski, D. (2020). 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic acids research*, *48*(W1), W170-W176.

Xie, W. J., Qi, Y., & Zhang, B. (2020). Characterizing chromatin folding coordinate and landscape with deep learning. *PLoS computational biology*, *16*(9), e1008262.

Zhu, C., Zhang, Y., Li, Y. E., Lucero, J., Behrens, M. M., & Ren, B. (2021). Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature methods*, *18*(3), 283-292.

Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., Meister, P., Smallwood, S., & Giorgetti, L. (2021). Nonlinear control of transcription through enhancer-promoter interactions. *bioRxiv*, 2021.2004.2022.440891. https://doi.org/10.1101/2021.04.22.440891

APPENDIX

Table 3 ENCODE accession IDs for the eleven epigenetic marks from three different cell-types.

	IMR90	**K562**	**A549**
H3K4Me1	ENCFF812CWF	ENCFF352HXD	ENCFF269UHZ
H3K4Me2	ENCFF842IIU	ENCFF446FUS	ENCFF143DSJ
H3K4Me3	ENCFF021CLP	ENCFF855ZMQ	ENCFF756VHC
H3K9Ac	ENCFF733ZHB	ENCFF149MXA	ENCFF497JCB
H3K27Ac	ENCFF064XTS	ENCFF121RHF	ENCFF921QFJ
H4K20Me1	ENCFF951EYC	ENCFF923YUN	ENCFF841QBW
H3K36Me3	ENCFF994SJM	ENCFF880HKV	ENCFF347QGE
H3K27Me3	ENCFF547BFL	ENCFF392ZKG	ENCFF979OSH
H3K9Me3	ENCFF084CZB	ENCFF155UQU	ENCFF089MBN
H2A.Z	ENCFF640IYV	ENCFF874SMO	ENCFF159DMB
DNase	ENCFF530KLY	ENCFF257HEE	ENCFF539TQP
RAD21	ENCFF895JAW	ENCFF057JFH	-

Table 4 Parameters values for Lennard-Jones Potential $\boldsymbol{U_P(r)}$ used in the polymer model. Each entry represents the interaction strength and distance cutoff (separated by |) between beads of that two specific types. Energy values and distance cutoff values are represented in terms of $\boldsymbol{k_BT}$ and $\boldsymbol{\sigma}$ respectively.

	Active State 1	**Active State 2**	**Inactive Poised**	**Repressive Polycomb**	**Heterochromatin**
Active State 1	0.8 | 1.3	0.75 | 1.3	0.85 | 1.3	0.9 | 1.3	0.95 | 1.3
Active State 2	0.75 | 1.3	0.7 | 1.3	0.75 | 1.3	0.8 | 1.3	0.85 | 1.3
Inactive Poised	0.85 | 1.3	0.75 | 1.3	1.0 | 1.35	1.0 | 1.35	1.0 | 1.35
Repressive Polycomb	0.9 | 1.3	0.8 | 1.3	1.0 | 1.35	1.3 | 1.4	1.1 | 1.35
Heterochromatin	0.95 | 1.3	0.85 | 1.3	1.0 | 1.35	1.1 | 1.35	1.4 | 1.4

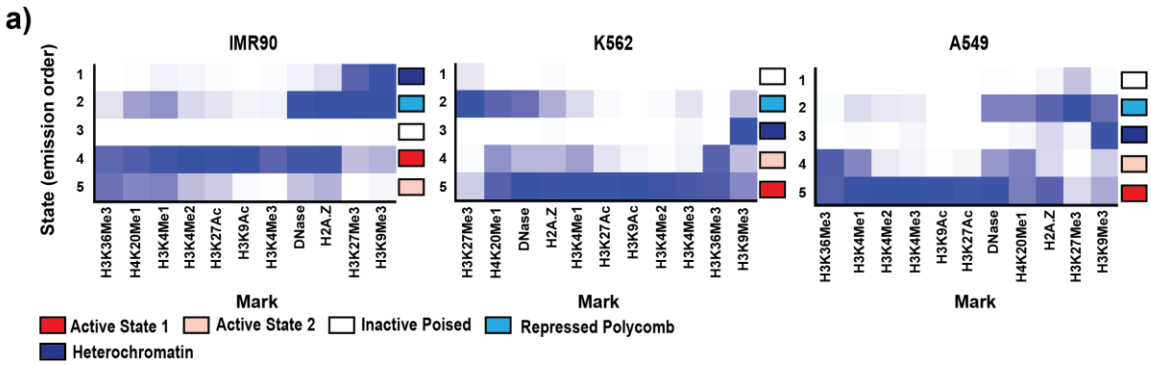


Figure 27 Assignment of different chromatin states based on relative enrichment of different histone marks using ChromHMM tool for chr21:28-30 Mb genomic region from three different cell-types - IMR90, K562 and A549 at 30 Kb resolution.

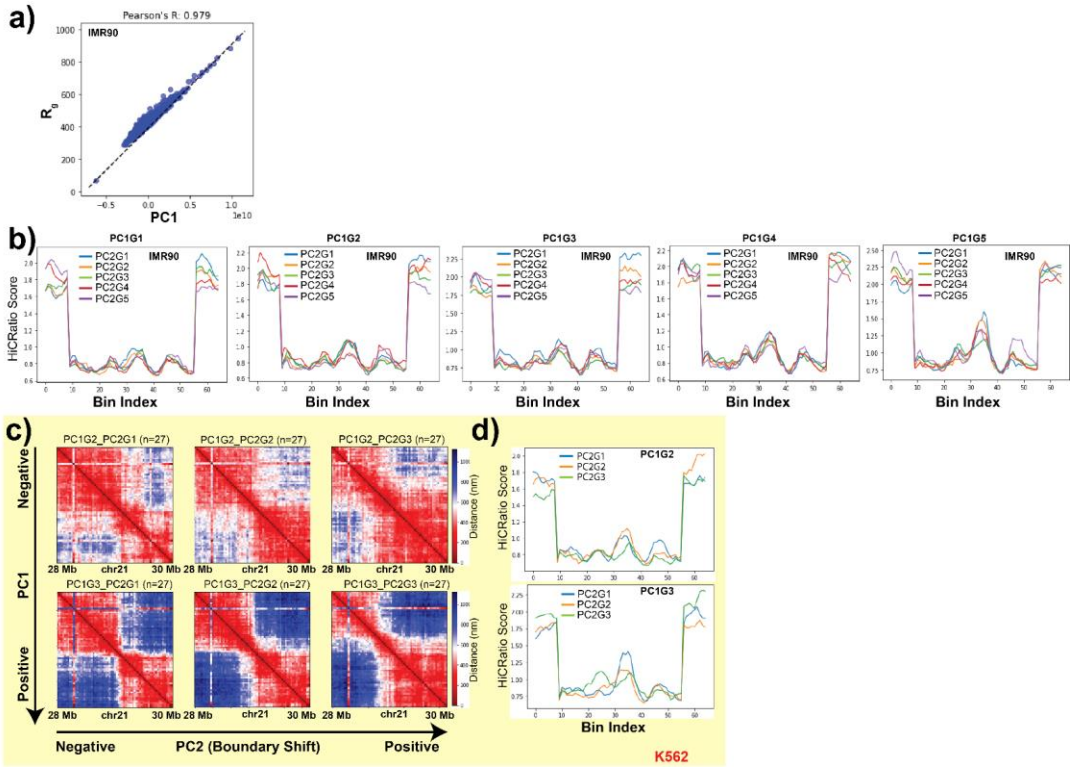


Figure 28 a) Pearson's correlation between chromosome structural landscape PC1 values and radius of gyration ($\boldsymbol{R_g}$) of single-cell structures (calculated from 3D position traces) for IMR90 chr21:28-30 Mb genomic region. b) HiCRatio score of PC2 ordered structures belonging to different PC1 groups for IMR90 chr21:28-30 Mb genomic region, calculated by HiCRatio approach (with a 300 kb window size). Here, PC1G1 refers to the group1 (G1; smallest $\boldsymbol{R_g}$ structures) based on PC1. Within PC1G1, then PC2G1 means the group1 (G1) based on PC2. Successive higher numbered PC1 groups have higher $\boldsymbol{R_g}$ structures. c) For K562 chr21:28-30 Mb genomic region, structures are first divided into three groups based on the PC1 ordering. Within each group, structures are further divided into three subgroups based on their PC2 ordering. To display the domain reorganization along PC2 in K562, averaged distance maps from only groups 2 and 3 from PC1 are shown here. d) For the same groups as in c, HiCRatio (with a 300 kb window size) scores of structures are shown. Colored lines represent HiCRatio scores from PC2 subgroups within two different PC1 subgroups (top = smaller $\boldsymbol{R_g}$ and bottom = larger $\boldsymbol{R_g}$).

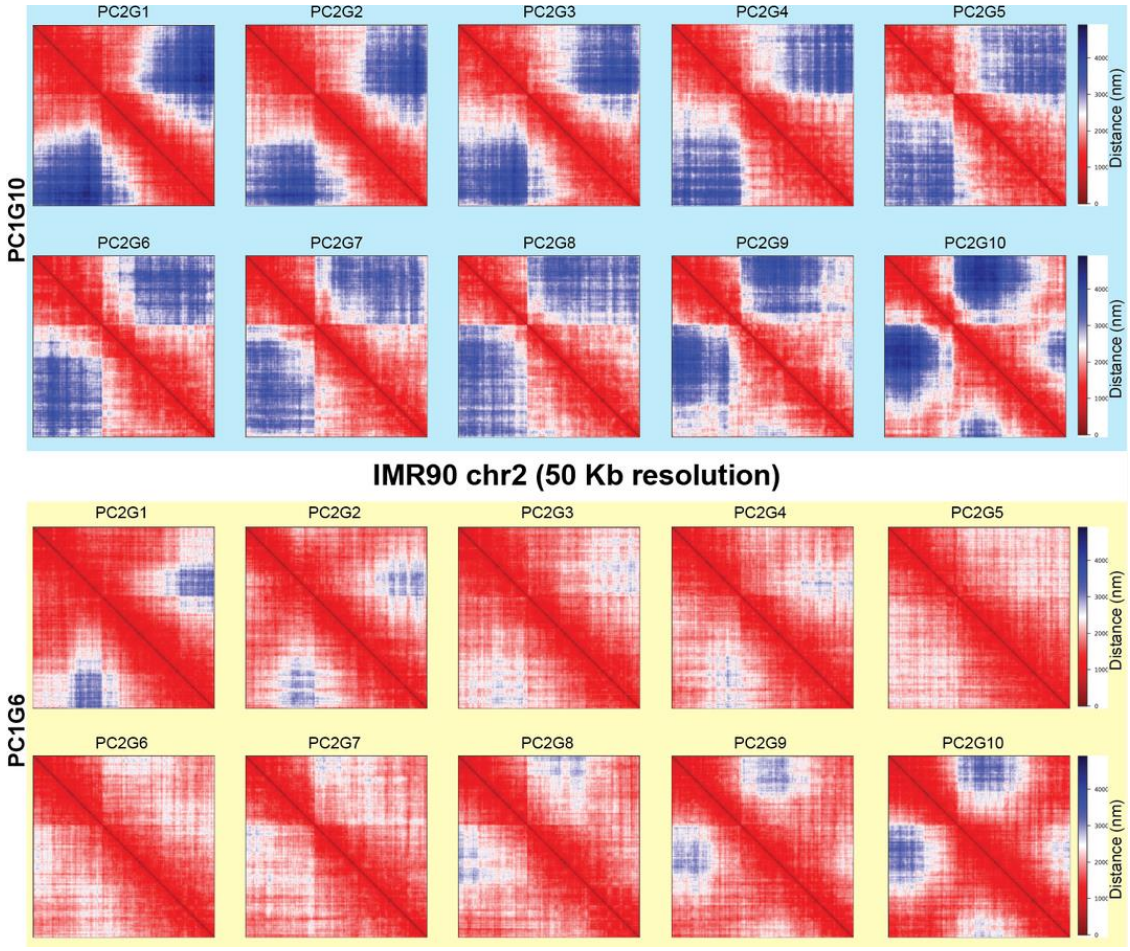


Figure 29 For IMR90 chr2 structure data, structures are first divided into ten groups based on the PC1 ordering. Within each group, structures are further divided into ten subgroups based on their PC2 ordering. To display the domain reorganization phenomenon along PC2 in IMR90 chr2, only PC1 subgroups 6 (lower $\boldsymbol{R_g}$) and 10 (higher $\boldsymbol{R_g}$) are shown here. Heatmaps display the average distances for structures within each PC2 subgroup in each PC1 subgroup.

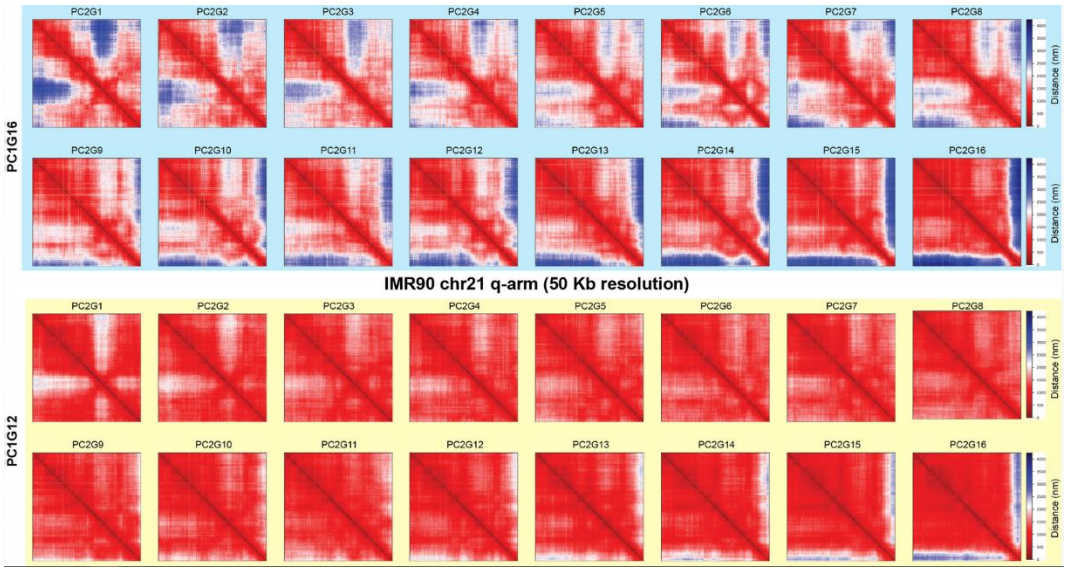


Figure 30 For IMR90 chr21 entire q-arm structure data, structures are first divided into sixteen groups based on the PC1 ordering. Within each group, structures are further divided into sixteen subgroups based on their PC2 ordering. To display the domain reorganization phenomenon along PC2 in IMR90 chr21, only PC1 subgroups 12 (lower $\boldsymbol{R_g}$) and 16 (higher $\boldsymbol{R_g}$) are shown here. Heatmaps display the average distances for structures within each PC2 subgroup in each PC1 subgroup.

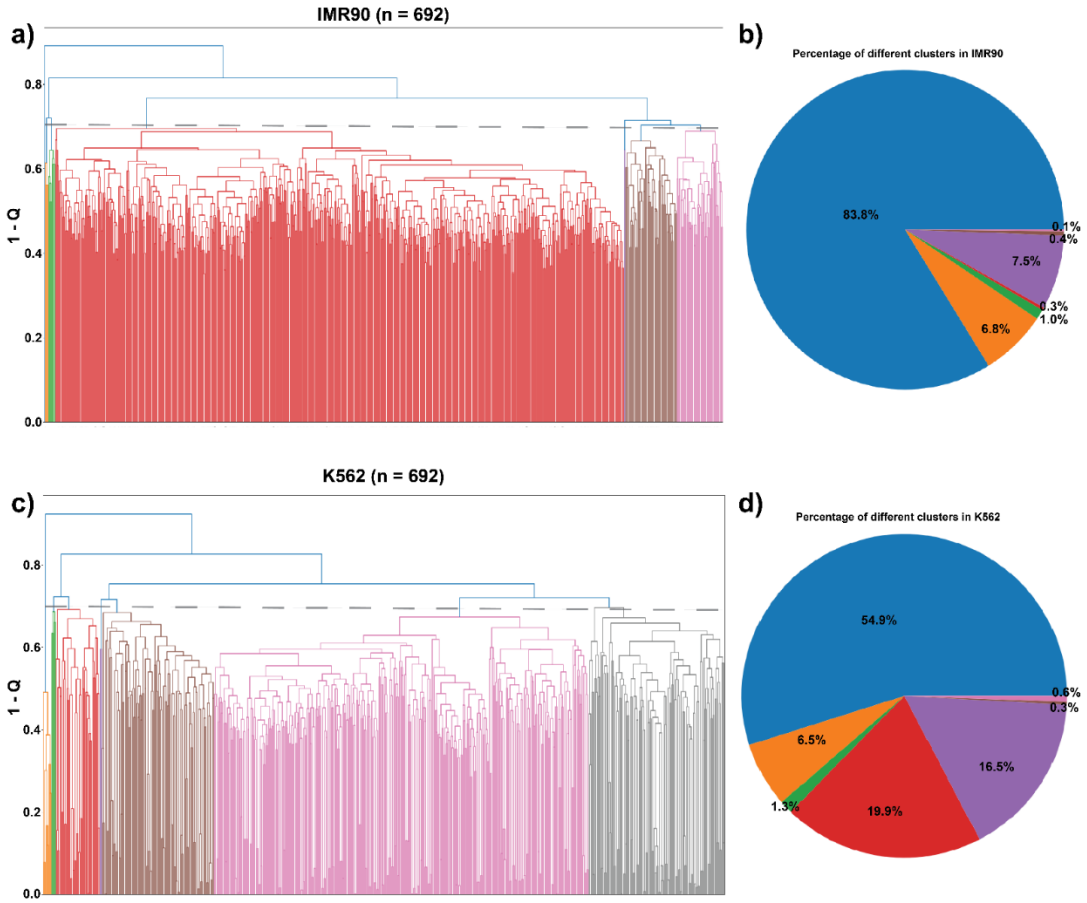


Figure 31 a,c) Hierarchical clustering of IMR90 (a) and K562 (c) structures for chr21:28-29.5 Mb genomic segment, based on ***Q*** similarity metric as described in Cheng et al. (1). To assign structures to the different clusters, the trees are cut at 0.7. b,d) Proportion of structures in different clusters obtained from IMR90 (b) and K562 (d) structural ensembles.

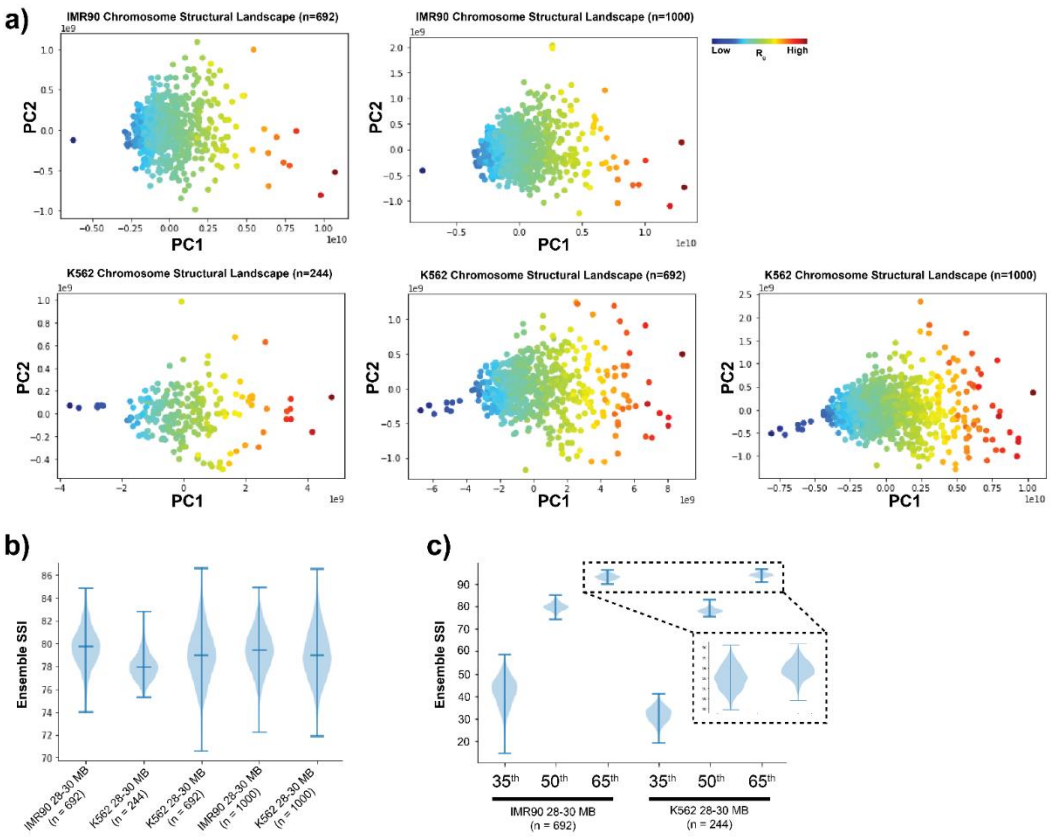


Figure 32 a) Chromosome structural landscape of IMR90 and K562 chr21:28-30 Mb chromatin regions constructed for different numbers of cells. Points are colored according to the radius of gyration ($\boldsymbol{R_g}$) of the corresponding structures (dark blue – smaller $\boldsymbol{R_g}$ and dark red – larger $\boldsymbol{R_g}$). b) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb segment from IMR90 and K562 for different number of cells (shown as n). Here, for resampling 200 single-cell structures are selected randomly each time. c) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb segment from IMR90 and K562 for different correlation matrix cutoff. Here, for resampling 200 single-cell structures are selected randomly each time.

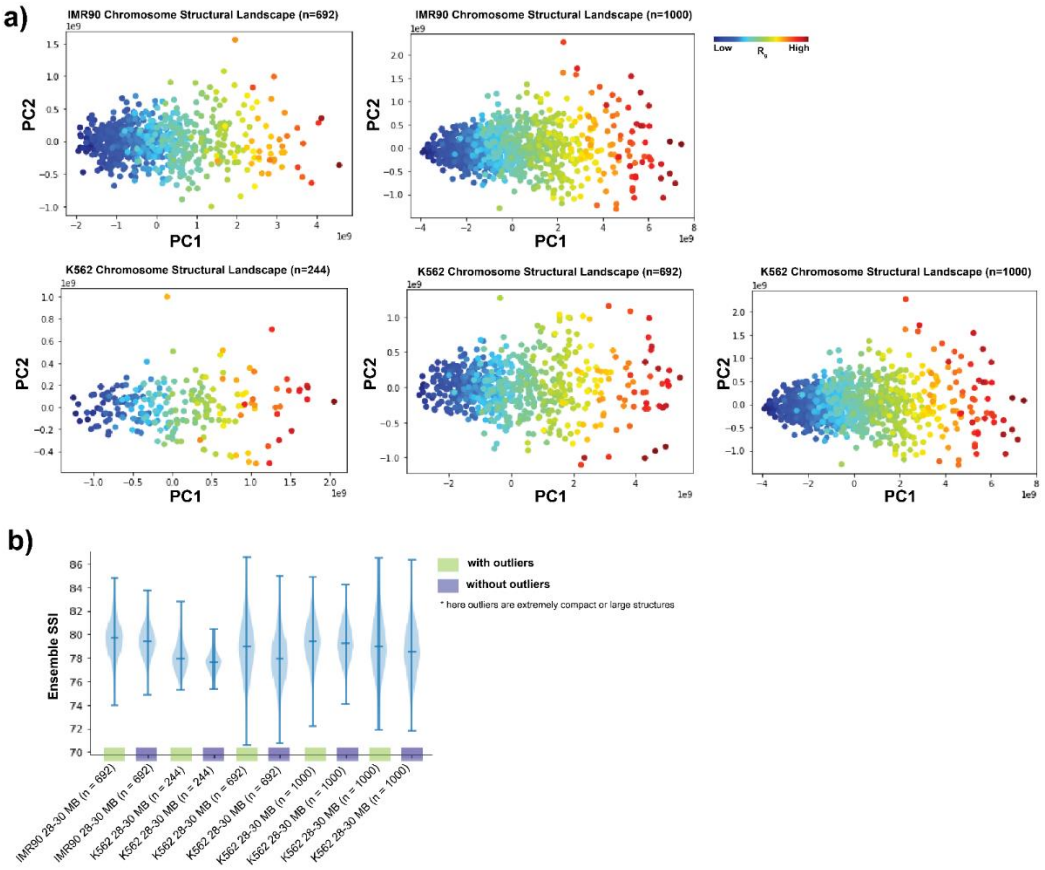


Figure 33 a) Chromosome structural landscape of IMR90 and K562 chr21:28-30 Mb chromatin regions constructed for different numbers of cells after removing extremely compact or large conformations. Points are colored according to the radius of gyration ($\boldsymbol{R_g}$) of the corresponding structures (dark blue – smaller $\boldsymbol{R_g}$ and dark red – larger $\boldsymbol{R_g}$). b) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb segment from IMR90 and K562 for different numbers of cells, before (green) or after (blue) removing extremely compact or large conformations ("outliers"). 200 single-cell structures are selected randomly each time for resampling.

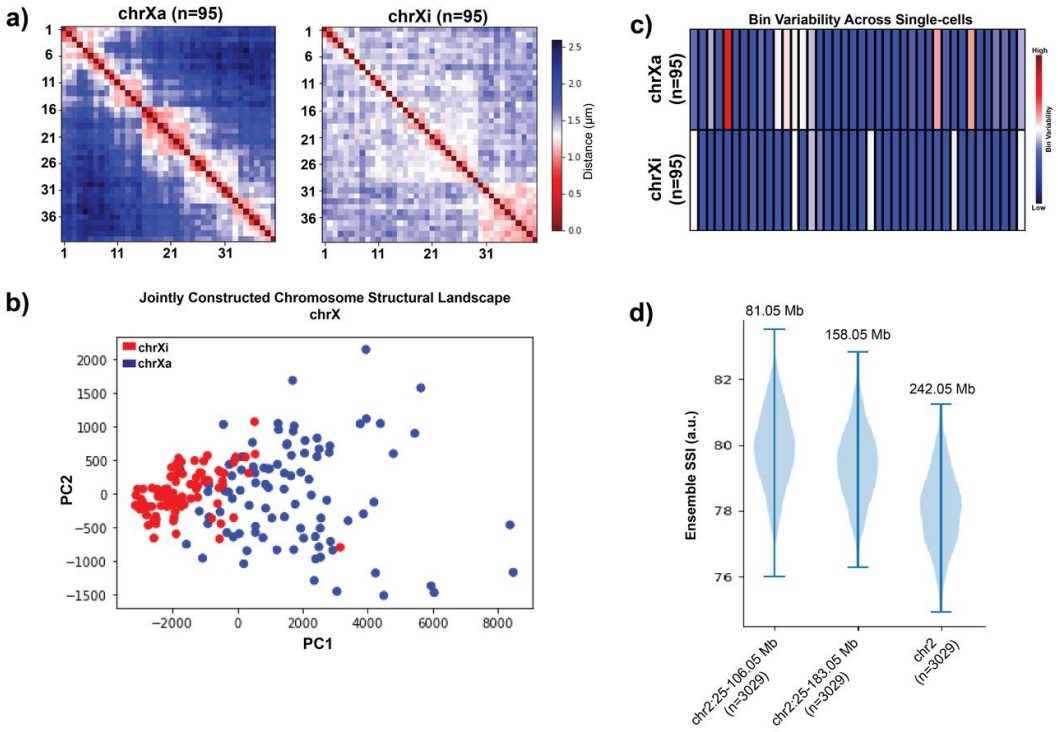


Figure 34 a) Population-averaged distance maps of active and inactive chrX from IMR90 cell at TAD resolution. b) Jointly constructed chromosome structural landscape of active and inactive chrX from IMR90 cell. c) Bin Variability (BSI standard deviation) of active and inactive chrX, where a higher value represents a higher level of structure variation for that bin across the population and vice versa. d) Chromosome structural ensemble SSI distributions for chr2 segments of varying length. Here, for resampling 300 single-cell structures are selected randomly each time.

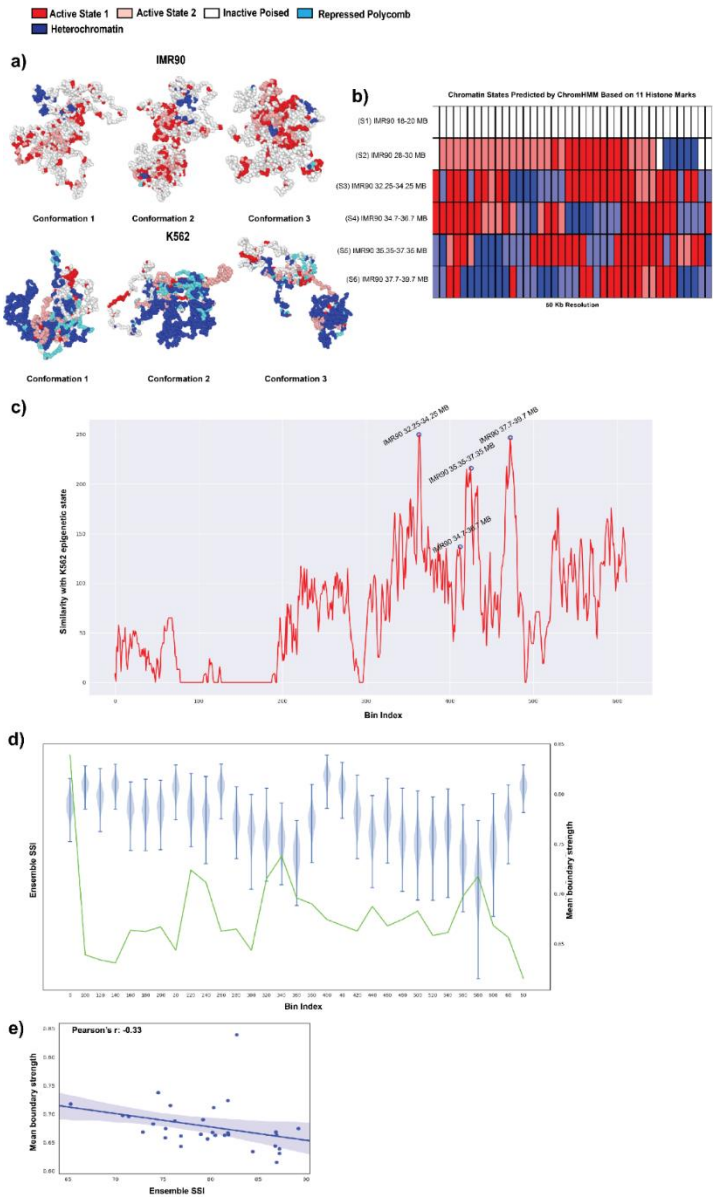


Figure 35 a) Representative polymer models of IMR90 and K562 for chr21:28-29.5 Mb genomic segment, obtained from different time points of the simulation. The color of the beads shows corresponding epigenetic states. b) Chromatin states of six different 2 Mb spanning genomic regions of IMR90 chr21 at 50 kb resolution. The chromatin states are inferred based on 11 different histone marks for each cell-type by ChromHMM tool. c) Epigenetic state similarity of different 2 Mb spanning genomic regions of IMR90 chr21 q-arm with the K562 28-30 Mb region using binary comparison. Here, a higher comparison value shows higher similarity with the K562 epigenetic state. d) Chromosome structural ensemble SSI distributions for 31 different 2 Mb spanning genomic regions of IMR90 chr21. For resampling, 300 single-cell structures are selected randomly each time. Mean boundary strength of those regions inferred by HiCRatio have been shown in green color. e) Pearson's correlation between the mean boundary strength and mean ensemble SSI of the 31 different genomic regions obtained from IMR90 chr21.

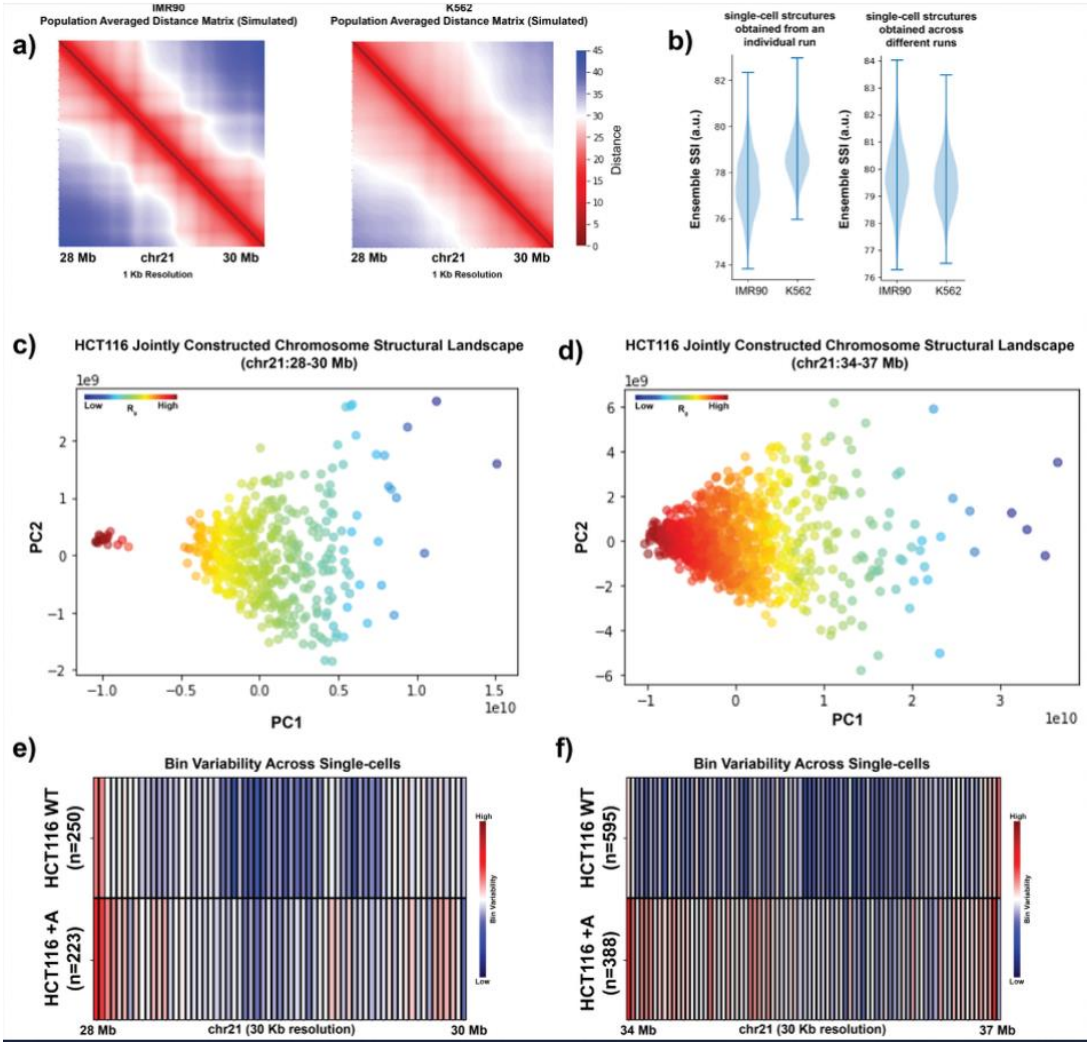


Figure 36 a) Population-averaged distance maps of IMR90 and K562 chr21:28-30 Mb genomic region from loop extrusion polymer simulations (see Methods) at 1 kb resolution. b) Chromosome structural ensemble SSI distributions for chr21:28-30 Mb from loop extrusion simulated IMR90 and K562. For resampling, 500 single-cell structures are selected randomly each time from a single run (left) or across multiple runs (right). c,d) Jointly constructed chromosome structural landscapes of chr21:28-30 Mb (c) or chr21:34-37 Mb (d) of HCT116 for WT and +Auxin conditions. Points are colored according to the radius of gyration ($\boldsymbol{R_g}$) of the corresponding structures (dark blue – smaller $\boldsymbol{R_g}$ and dark red – larger $\boldsymbol{R_g}$). e) Bin Variability of chr21:28-30 Mb genomic region for wild-type (WT) and cohesin-depleted (+Auxin) conditions. f) Bin Variability of chr21:34-37 Mb genomic region for wild-type (WT) and cohesin-depleted (+Auxin) conditions.

CHAPTER III
USING CONTACT STATISTICS TO CHARACTERIZE STRUCTURE TRANSFORMATION OF BIOPOLYMER ENSEMBLES

A version of this chapter was originally published by Priyojit Das, Rosela Golloshi, Rachel Patton McCord and Tongye Shen:

Das, P., Golloshi, R., McCord, R. P., & Shen, T. (2020). Using contact statistics to characterize structure transformation of biopolymer ensembles. Physical Review E, 101(1), 012419.

My contributions to this work include: (1) conceived the project, (2) performed computational modeling and analysis, (3) wrote the manuscript, and (4) prepared the figures.

Abstract

As a unique subset of functional polymers, many biopolymers have a set of well-defined 3D structural characteristics that can be described by spatial contacts between monomers. Statistical analysis of the contacts has been extremely productive in characterizing the biopolymer structural ensemble, such as for 3D chromosome structures. Often, native contacts and compartment structures are the focus of the studies, while the generic polymer aspect, such as the overall decaying of contacts with increasing sequence distance, is analyzed separately or preemptively removed. Here, we explore significant insights that can be gained by performing "compartment analysis" that keeps the distance decay, which we believe is particularly useful for characterizing the structure transformation of biopolymers. We tested contact analysis on several such transformations under physical perturbation or biological processes, including (1) unfolding of proteins induced by thermal denaturation, (2) chromosome conformation transition during the cell cycle, and (3) chromosome unpacking by physicochemical perturbations. Useful score functions were developed to further quantitatively characterize the transformation judging from the contact analysis. We also find that the sinusoidal undertone of eigenvector patterns (the "unwanted", low frequency signal, in contrast to the detailed A/B compartment) that had previously been attributed to biological effects of centromere proximal and distal interactions may in fact reflect a universal feature of polymers that have relatively weaker long-range contacts.

Introduction

Many linear biopolymers, including well-understood examples of proteins and nucleic acids, are structured or semi-structured polymers, which means that they have defined structure(s) or at least well-defined structural components (Brown, 2018; Fersht, 1999). Often, their 3D folded conformations are essential to fulfill their biological functions and their structural features separate them from generic polymers. How these biopolymers are able to fold and sustain specific shapes has been an important question in biological physics research. The essential features are encoded by short-ranged favorable contact interactions between monomers that are not sequence neighbors. At the simplest level, even a model of "two-colored" beads-on-a-string reproduces basic features of structure formation for biopolymers (Doi, 1996; Michieletto et al., 2016). In protein folding, this type of model reflects a basic hydrophobic-polar (HP) dichotomy where H-H contact interactions are favorable compared to that of H-P (Dill, 1985). In chromosome folding, this two color model reflects a dichotomy of gene-rich/gene-poor regions (A/B compartments (Lieberman-Aiden et al., 2009)) where gene poor or gene rich regions tend to aggregate with like types (e.g., favorable B-B and A-A interaction compared to A-B interaction) (Tiana & Giorgetti, 2019). With a more refined biopolymer model containing bead types that accurately reflect the specific physical interactions between different parts of the polymers, other complicated and vital-for-function structures emerge.

How one should characterize the conformations of these structural ensembles is an essential question for the physical description of the biopolymers and the interpretation of their biological functions. For the structural ensemble of a highly structured biopolymer, the structural fluctuation can be characterized similar to a "vibrating" solid (Gō & Scheraga, 1969; Karplus & Kushick, 1981), where a set of Cartesian coordinates is conventionally used as the degrees of freedom (DOFs). At the other limit, a generic polymer that does not have specific non-neighboring interactions are best characterized by internal coordinates of bonding terms (bond, angle, and torsion angle) (Flory & Volkenstein, 1969). In between, when one wants to study a biopolymer's transformation, for example, from a structural to a disordered state, it is more suitable to use contact DOFs (Bryngelson & Wolynes, 1987; Go, 1983). The concept of contacts (spatial proximity formed between different parts of the polymer) is frequently used to characterize such (semi)structured conformations, and these contact interactions are often collectively displayed in the form of a matrix. In the case of chromosomes, such contact matrices often are derived from genome-wide chromosome conformation capture (Hi-C) data.

Statistical analysis methods, especially principal component analysis (PCA), have been used for studying the collective "modes" of biopolymer contacts. For example, a popular contact analysis method, known as the "compartment analysis" in the chromosome field, was designed to characterize and even amplify the "signature" of the well-folded structures (Lieberman-Aiden et al., 2009; Schmitt et al., 2016). On the other hand, studying the more flexible or even somewhat disordered structural ensembles can be important and have practical biophysical applications. Researchers often perturb the well-folded biopolymers

with various environmental perturbations to examine how their structural stability and dynamics alter, eventually moving these biopolymers out of their native conformations and into an unfolded state (Fersht, 1999; Poirier et al., 2001; Sun et al., 2018; Takahashi et al., 2018). Also, many biopolymers, from intrinsically disordered proteins to whole chromosomes at certain points in the cell cycle, may be viewed as weakly interacting polymers (with significantly less specific long-range contacts) that approach a more generic polymer state. Thus, it is desirable to use contact analyses not only for characterizing the ultimate structure aspect (the "enthalpic" aspect) of these biopolymers, but also for accessing the generic polymer aspect of the ensemble.

The schemes of using principal component analysis for studying contacts discussed in this work, E-PCA, I-PCA, and MI-PCA, are recapitulated in Fig. 37. As we have described previously, contact fluctuation analysis, E-PCA, characterizes the dominant fluctuation of the structural ensemble, while contact structural analysis, I-PCA (and its slight variation version MI-PCA) describes the consensus features of the ensemble such as domain and compartment classification (Lindsay, Pham, et al., 2018). Specifically, the matrix form of the contact information is used for these analyses. The column and row indices of these matrices are naturally derived from the linear indexing of the polymer, u_{ij}, which has a value of 1 if contact formed between monomers i and j, and 0 otherwise. A simple definition based on distance R_{ij} provides $u_{ij} = \Theta(R^C - R_{ij})$, where R^C is the cutoff. One needs further define $U_{ij} = \langle u_{ij} \rangle$ as the ensemble average of the contact interaction strength, also known as interaction frequency.

E-PCA treats each contact *explicitly* as an independent variable u_{ij} and for a matrix of $N \times N$, it largely contains an order of N^2 variables varying from one matrix to another. Here, N represents the dimension of the system, i.e., the number of basic building blocks of that biopolymer (e.g. for protein, these building blocks can be amino acids (Johnson et al., 2018) or coarse-grained segments that may even contain secondary structural elements (Lindsay, Siess, et al., 2018) and for chromosomes, genomic regions of different resolutions (bin size ranges from 10^3 to 10^7 base pairs (bp), with typically used values from 10^4 to 10^6 bp). In general, E-PCA has a covariance matrix ${C^E}_{ijkl} = \langle (u_{ij} - \langle u_{ij} \rangle)(u_{kl} - \langle u_{kl} \rangle) \rangle$. In contrast, for I-PCA and MI-PCA, which are methods frequently used for studying Hi-C data, one treats each row of contact matrix as a set of independent data and, one has total N *implicit* contact variables u_i that vary from row to row, ${C^I}_{ij} = \langle (u_i - \langle u_i \rangle)(u_j - \langle u_j \rangle) \rangle$. Specifically for MI-PCA, we first obtain contact average over rows $\langle U_j \rangle = \sum_i U_{ij} / N$ and a matrix δU defined by its elements $\delta U_{ij} = U_{ij} - \langle U_j \rangle$. Then, the covariance matrix for MI-PCA $C^I = \delta U^T \delta U$ is obtained. Once the covariance matrix C^I or C^E is obtained, principle component analysis (PCA) is applied to obtain the collective modes of fluctuation from the corresponding mean contacts $\langle u_i \rangle$ or $\langle u_{ij} \rangle$. These modes of variation for $\langle u_i \rangle$ reveal the domain information of the ensemble while variation modes of $\langle u_{ij} \rangle$ display dynamics.

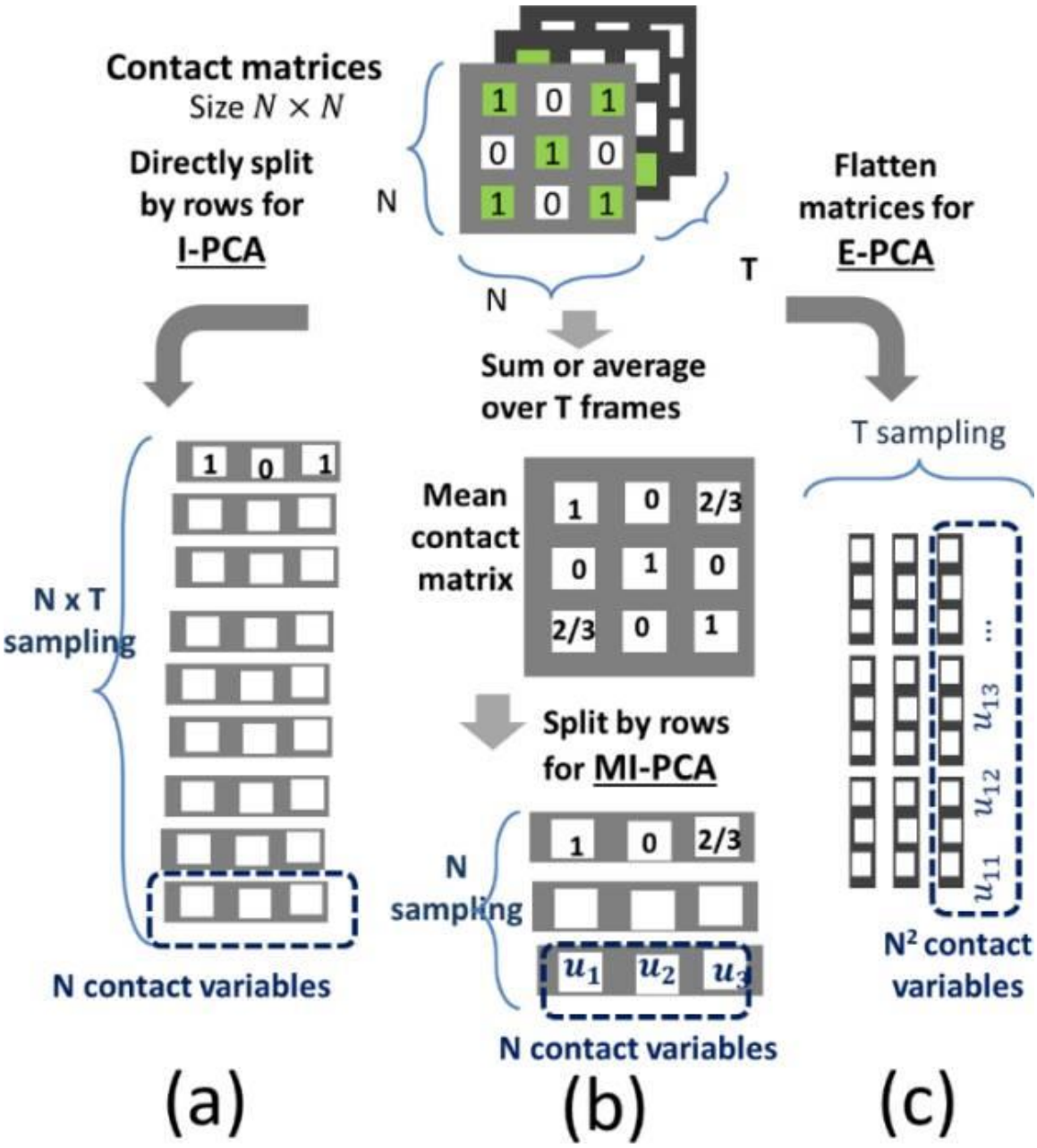


Figure 37 A cartoon illustration of three different analysis methods, (a) I-PCA, (b) MI-PCA, and (c) E-PCA, on contact matrices.

The generic polymer limit and the problem of removing distance decay in contact matrices

For chromosome contact data and particularly for the detection of A/B spatial compartments, instead of direct analysis of the "raw" contact interactions u_{ij}, often researchers manipulate the matrix to enhance the weight of the off-diagonal components (long-range contacts) and obtain an alternative definition of contact interaction, $v_{ij} = u_{ij}/\mathrm{b}(|i-j|)$. Here, scaling factor b is a function of genomic distance and scales the raw contact data u_{ij}. It was initially designed to preemptively "subtract" the random, generic polymer nature of the contact matrix and amplify the structured polymer aspect of these biopolymers(Belton et al., 2012; Lajoie et al., 2015; Mirny, 2011). Even though some early analysis recommended eigenvector decomposition of contact matrices without such modification (Imakaev et al., 2012), most applications of PCA to chromosome contact matrices (esp. intrachromosomal contacts) have continued to include such an operation (some of the recent examples can be found in (Miura et al., 2019; Ray et al., 2019; Zhou et al., 2019)) and the differences between the results obtained with and without this normalization have not been thoroughly explored in any previous work.

It is nontrivial to model factor $\mathrm{b}(|i-j|)$ and various function forms have been tested to best serve the signal enhancing purpose. There are largely two types of approaches. One is using a derived or empirical function form and previous studies on generic polymers have been quite extensive (de Gennes, 1979; Dewey, 1997; McKenzie, 1976). For example, factor b be can a power law form $|i-j|^{-D/2}$or an exponential form $\exp(-\alpha|i-j|)$. The other enhancing approach uses the statistics of raw data u_{ij} itself. The power law form was derived from the generic polymer limit, in which case the distribution of the end-to-end displacement $\vec{r}$, according to the random walk theory (de Gennes, 1979), is $\mathrm{p}(\vec{r};\mathrm{L})\mathrm{d}\vec{r} = (2\pi \mathrm{L a}^2)^{-3/2}\exp[-3\mathrm{r}^2/(2\,\mathrm{L}\,\mathrm{a}^2)]\,\mathrm{d}\vec{r}$, where N is the number of basic building blocks and a is the linear length of the building block unit. This equation can be generalized in a general D-dimensional polymer as $\mathrm{p}(\vec{r};\mathrm{L}) = (2\pi \mathrm{L a}^2)^{-\mathrm{D}/2}\exp[-\mathrm{D}\,\mathrm{r}^2/(2\mathrm{L}\,\mathrm{a}^2)]$. Thus, for a 3D random polymer (D=3), the chance of residues i and j forming a contact can be viewed as a loop or a polymer cyclization problem where the chance of a random walker revisiting the same location $\mathrm{r} = 0$ after distance $\mathrm{L} = |i-j|$ steps, which is $\mathrm{p}(0;\mathrm{i}-\mathrm{j})$, i.e., $p_{ij}\sim|i-j|^{-3/2}$. This random walker model can be used to describe the simplest case, an ideal generic polymer in a theta solvent. The practical situations can be more complicated. For example, a poor solvent condition packs the polymer, while on the other hand, polymers in good solvent are more extended. Researchers found that a fractal dimensional biopolymer (Dewey, 1997) might be better for the hierarchical structure of chromosome fold (Mirny, 2011). Using the generalized form of random walk models, one obtains a power law decay $\mathrm{p}(0;L)\sim L^{-D/2}$. Practical fitting for chromosome data even provides varying D depending on L, such as $\mathrm{D} \approx 2$ for longer L and as small as $\mathrm{D} \approx 1.5$ for shorter L (Sanborn et al., 2015). Other variety of empirical function forms with a gentler decay have been used as well, such as exponential form $\mathrm{p}(L)\sim\exp(-\alpha L)$, which has the advantage of

smoother short length scale. However, such exponential formula can also be limited by its applicable range (Sanborn et al., 2015).

As shown in Fig. 38(a), we calculated the top eigenvectors of MI-PCA using a directly generated mean contact matrix for exponential and power law decay and a computer simulated random polymer. One can observe that the results are always a symmetric odd function with respect to the middle point. The components of MI-PCA eigen vectors always have mixed signs, i.e., some are positive, while others are negative, if the diagonal elements of the matrix are larger than the off-diagonal elements. Practically, this is always true for a polymer's contact matrix since the contact frequencies of sequence neighbors are much higher. The location of sign change can be defined as a node or a boundary. Often these nodes for a structured polymer are important and can be used to separate biopolymers into distinct structural domains, often called A/B compartments for chromosomes (McCord et al., 2013). In contrast to MI-PCA, the elements of an E-PCA eigenvector can potentially have the same sign, i.e., a global "breathing" motion that have contacts forming and breaking in sync.

Besides mean contact matrices constructed using exact mathematical formulae of the exponential or power-law decay forms, we also obtain the mean contact matrix using computer simulation data of a 65-bead generic polymer that lacks specific contact interaction. For a fair comparison of polymers with different number of beads, we can project the result on the ideal mathematical formulation-based results of a 100-bead polymer using a scaling transformation: $x^* = \mathrm{x} \times (100/65)$ and $y^* = \mathrm{y} \times (100/65)^{-1/2}$. This random polymer simulation is a limiting case of a protein polymer simulation where the native contact interactions are removed. The rationale for scaling to account for the difference between different sizes of the mean contact matrix will also be given in the next section, where we discuss the conformations of a 65-residue protein polymer under thermal perturbation. Our simulation results matched quite well with the ones generated by mathematical formula, particularly, the exponential decay ones (Fig. 38a).

From Fig. 38b, we also can observe that high order eigenvectors (those with a smaller eigenvalue) have more nodes and in general, the n −th mode, MI-PCn, contains n nodes at the generic polymer limit. It is interesting to point out that the eigenvectors MI-PC1 to 5 resemble an expansion of cosine series, which is not accidental. In fact, one can locate a family of matrices having their eigenvectors being exactly a series of trigonometric expansion. The eigenvectors of Rouse polymer matrix A provide such an instance, where matrix A is essentially a second order differentiation operator(Rouse, 1953; Zimm, 1956). In general, matrix B = A + c x I (I being identity matrix and c a constant) shares the same eigenvectors with matrix A. Practically, all mean contact matrices with much stronger diagonal component and a very fast off-diagonal decay will have a covariance matrix C^I with similar characteristics.

We also examined the MI-PC1 signature of another type of contact matrix, band-diagonal matrices (Fig. 38c). One can see that with increasing band width, the signature of MI-PC1

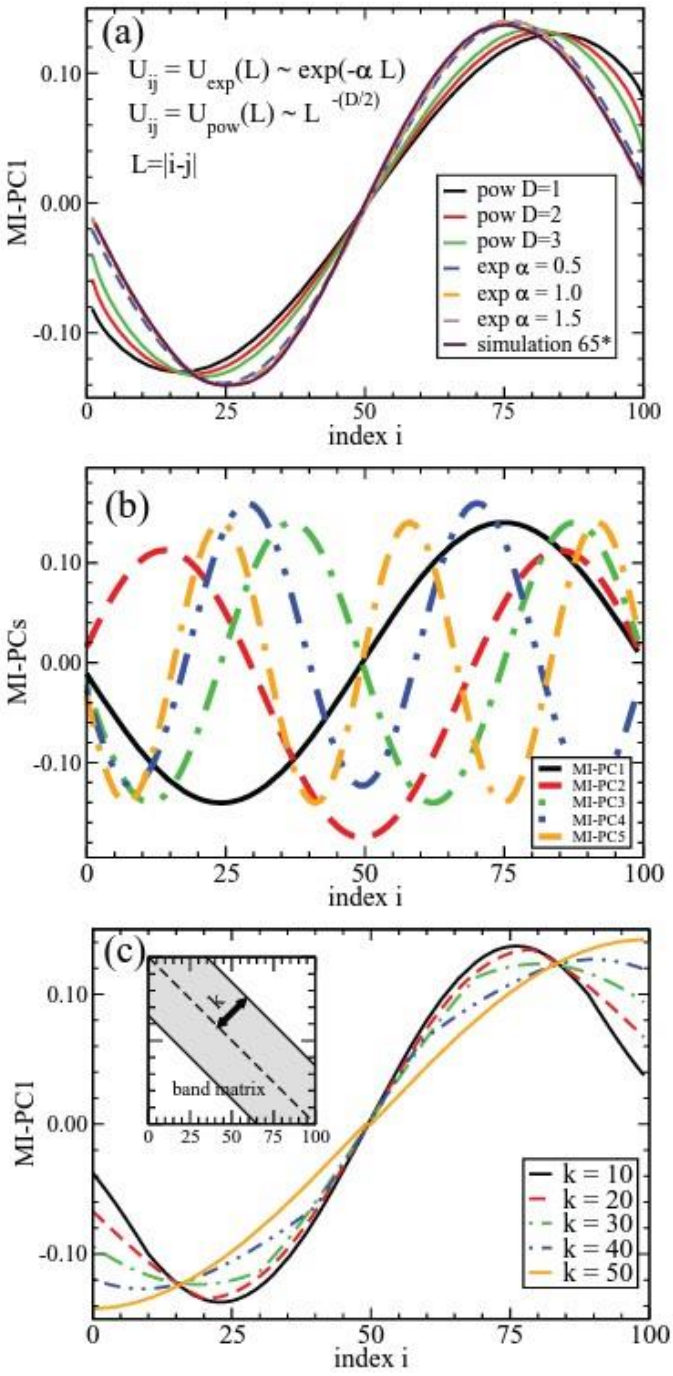


Figure 38 (a) The dominant mode of MI-PCA for a random polymer, MI-PC1, including power law form of mean matrix, exponential decay form of mean matrix, and a result of computer simulation of the polymer. (b) The top five eigenvectors of MI-PCA for the case of the exponential decay with α=1.5. Eigenvectors with smaller eigenvalues have increasing number of nodes.

deviates from the eigenvectors of the "ideal" non-interacting polymers displayed in Fig. 38a. As we will demonstrate below, some of the features of the deviation resemble certain observations of mitotic chromosome structures.

Because it is not clear which function form is best at removing the random polymer "background feature" in practical situations, the alternate common approach of treatment is normalization by an empirical average or median contact count at each length scale (Belton et al., 2012; Lajoie et al., 2015; Mirny, 2011; Pal et al., 2019). The common approach is directly normalizing by the average of the off-diagonal, more precisely, the average of k-diagonal elements, where k is a measurement of its relation to the main diagonal, i.e., $v_{ij} = u_{ij}/b_k$ with $b_k = \sum_{ij} u_{ij} \times \delta_{|i-j|-k} / \sum_{ij} \delta_{|i-j|-k} = \sum_i u_{i,i+k}/(N-k)$. Here Kronecker delta selects $k = |i - j|$. This way, the strength of matrix elements v_{ij} will not drop off as that of u_{ij} with increasing k. One drawback for this practical approach is that, when k is large, that is for those extreme super-diagonals and sub-diagonals, there are few matrix elements to contribute to the average.

While pre-treating a contact matrix can be useful to emphasize off-diagonal interactions, it has a few potential disadvantages. First, as we will demonstrate below, such treatment will remove or at least reduce the aspect of polymer (thermal) fluctuation, and thus make the examination of polymer conformations subjected to physical and chemical perturbations less direct. Secondly, the treatment will create false positive information on domain structure when the polymer is near the generic polymer limit. The reason is straightforward. At the generic polymer limit, a truly fully sampled u_{ij} will provide all elements $v_{ij} \equiv 1$ and thus, return a zero valued covariance matrix and thus a uniform domain. However, in any practical situation, there will be at least numerical noise that make v_{ij} less than perfect and thus create false-positive domains in the now nonzero covariance matrix. Thus, we find that treated matrices can be useful to reveal changes in folding patterns when the "signal" is strong enough, but they may not be best when one wants to compare a broad spectrum of conformation ensembles, as we will demonstrate below using specific examples.

It is also important to point out that, throughout this work, the pretreatment by scaling ($u_{ij} \rightarrow v_{ij}$) refers to off-diagonal distance normalization. This operation must not be confused with many correction schemes to remove biases from raw Hi-C data (Hu et al., 2012; Imakaev et al., 2012; Li et al., 2014; Meyer & Liu, 2014; Yaffe & Tanay, 2011). For example, the correction to remove experimental artifacts (which arise from Hi-C steps involving sequence-specific DNA cleavage and differential PCR amplification of sequences with different GC content) is achieved by ICE normalization. This ICE approach balances the matrix row and column sums by transformation $u_{ij} = \frac{u'_{ij}}{\beta_i \beta_j}$ where $\beta_i = \sum_i u'_{ik}$ and u'_{ik} is the raw number of contact hits (Imakaev et al., 2012; Yaffe & Tanay, 2011).

Contact analysis for partially structured proteins and chromosomes

Here we provide examples of biopolymer data that may exhibit a large range of dynamic (folding) motions along a spectrum of structure order, from relatively structured to the effective generic polymer limit, without specific contacts. The first example is a model protein system subject to increasing thermal noise which undergoes an unfolding transition. The second example is chromosome changes during the cell cycle, where chromosome structure dramatically reorganizes from interphase to metaphase. Last but not the least, we examined the less structured ensemble resulting from a physicochemical perturbation of chromosome structures, in which chemicals such as salt and histone deacetylase inhibitor drugs act to decondense chromosomes, which might destabilize chromosome structures and push them towards the direction of losing specific contacts.

Proteins approach the generic polymer limit with rising temperature

The first example is a structural contact analysis of a small protein (65 amino acid residues), chymotrypsin inhibitor CI-2 (PDB ID: 2CI2) (McPhalen & James, 1987) subjected to thermal perturbations. The advantage of using simulation data is that they provide information about individual structural fluctuation and we can perform E-PCA for a comparison with MI-PCA. They also provide direct 3D coordinate information that is not available for chromosome structures at high resolution. Here, we use MI-PCA for this pedagogical system to demonstrate how partially folded protein conformations can be characterized by contact statistics and how the results change with increasing temperature. CI-2 has a simple fold of several secondary structure elements (four β-strands and one α-helix), as shown in Fig. 39a. We adopted a simple coarse-grained Hamiltonian (Cα Go-model) of the protein that has frequently been used to study protein contact dynamics and folding (Dill et al., 2008; Portman et al., 1998). The configurations were sampled from the molecular dynamics simulation package Gromacs (Berendsen et al., 1995) and setup program SMOG(Noel et al., 2016). The simulation timestep dt is 0.0005 and total 2×10^8 steps were used for each simulation. Snapshots were taken at every 10,000 steps. Finally, 20,000 snapshots were converted to contact matrices. Here a simple distance scheme is used, i.e., a contact is being formed and $u_{ij} = 1$ when the distance between Cα atoms of a pair of amino acid residues r_{ij} is less than a distance cutoff 6.5 A, and $u_{ij} = 0$ otherwise. The contact matrix of the crystal structure using this definition is shown in Fig. 39a.

We display the results of I-PCA (particularly, MI-PCA) and E-PCA for this system at various temperatures ($T = 0.5, 1.0, 4.0,$ and $10.0\ T_f$) in Fig. 39b and 39c, respectively. Here, temperature T is expressed in the unit of the folding temperature T_f at which point the protein spends roughly 50% of time in the native, folded state and the other 50% unfolded. For a lower temperature $T = 0.5\ T_f$, we observed that MI-PC1 characterizes the mean contact map with clear domain features, while the corresponding E-PC1 describes the internal contact breaking and forming dynamics of this folded structure. At $T = T_f$, we see that the structural features revealed by MI-PCA are weakened. At even higher

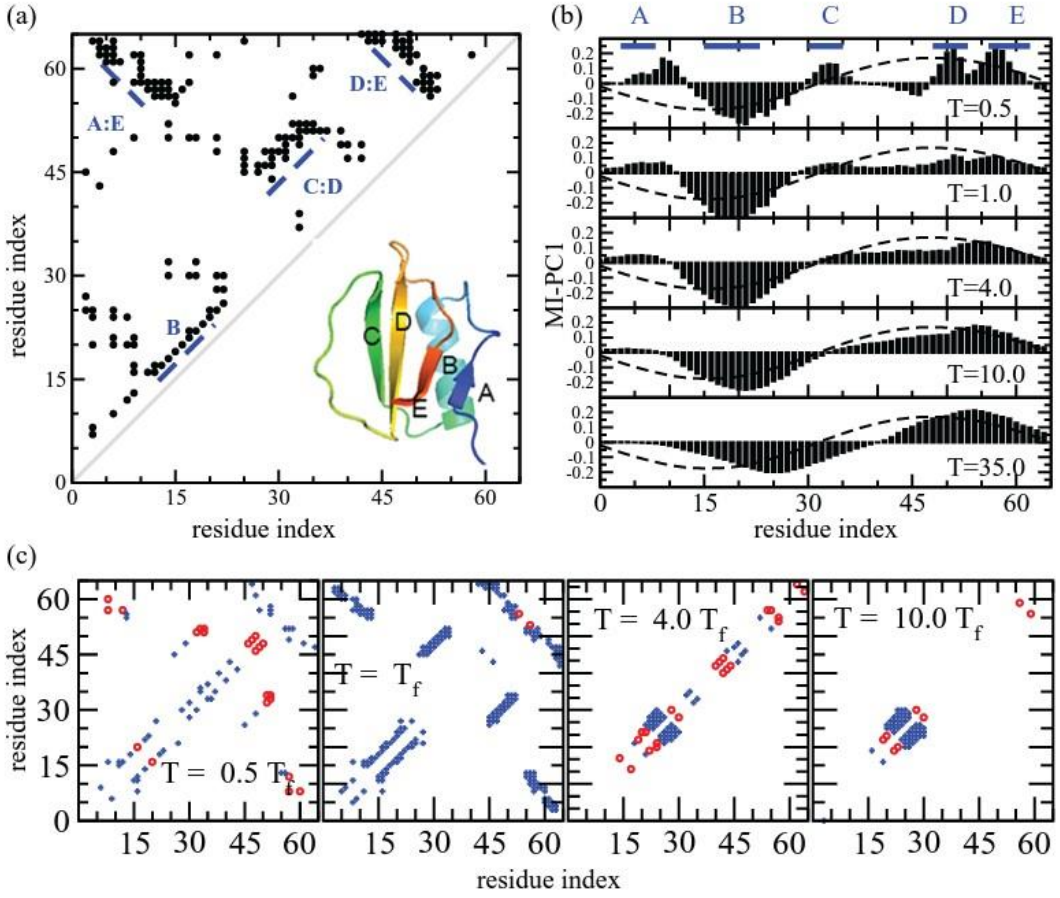


Figure 39 (a) Structure and contact map of protein CI2. (b) MI-PC1 of protein CI2 at different temperatures with the mean contact matrix at highest and lowest temperatures shown. The dashed line is the reference system of backbone interaction only. (c) The corresponding E-PC1 of protein CI2 at different temperatures are shown using its significant elements (a cutoff of $|d_{ij}| > 0.025$).

temperatures, the structural features disappear and the results approach that of the generic polymer limit shown in the previous section. Note that these high temperatures are not meant to be taken literally. Rather, they are used to demonstrate the effect of weakening native interactions. An interesting observation is that even at high T= 35 T_f, IPCA still reveals noticeable deviation from the ideal polymer behavior. Meanwhile, one could not easily distinguish such ensemble from a random unfolded ensemble by direct visualization, which indicates the sensitivity of the I-PCA method.

For a comparison and a demonstration of what E-PCA reveals, we show the top dynamic modes at the corresponding temperatures. Note the drastic differences of contact structural analysis I-PCA and contact dynamics analysis E-PCA's eigenvectors. One is expressed essentially by a linear plot, i.e., 1D function a_i while the other is often expressed in the form of the matrix a_{ij}that symbolizes a particular fluctuation mode of contact dynamics. For example, $T = 4.0\,T_f$ E-PC1, the dominant mode of fluctuation displays the forming/breaking of α-helix, which further nucleates with the rest of the secondary structures around T_f, which is consistent with experimental and theoretical studies (Jackson & Fersht, 1991; Ladurner et al., 1997).

Mitotic chromosomes approach the effective generic polymer limit during the cell cycle

Chromosomes are semi-structured biopolymers which change their conformations and compactness depending on the cell cycle stage (Naumova et al., 2013). The transition from interphase chromosome structure through prophase and into metaphase was recently captured in detailed Hi-C contact maps for synchronized DT-40 chicken cells (Gibcus et al., 2018). In G2 (t=0 min), before the cell enters mitosis, chromosomes show all the typical hallmarks of highly organized interphase structure, including topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012) and A/B compartments (Lieberman-Aiden et al., 2009)). As the cell progresses through mitosis, these conformations start to disappear. Upon entering prophase [t = 2 min after cyclin-dependent kinase 1 (CDK1) checkpoint release], compartments and TADs are lost, and the chromosomes start to become linearly organized structures. The chromatids become more shortened and thicker as the cell enters prometaphase (t = 10 min) and ultimately lead to the formation of fully condensed metaphase chromosomes. Here, we examine the contact matrices of the 39 chicken chromosomes at seven time points (t = 0, 2, 5, 7, 10, 30, and 60 min) starting from G2 phase to late prometaphase, binned at 40 kb and pre-processed with iterative correction (Gibcus et al., 2018).

We display the MI-PC results of chicken chromosome 21 (chr21) in Fig. 40. As shown, chr21 is quite structured at the G2 phase (t = 0 min). But as the prophase progresses, the MI-PC1 signals start to resemble that of a generic non-interacting polymer. From these observations, it can be concluded that chr21 transitions from well folded (highly ordered structure with specific contacts) to a linearly organized “effective polymer” (fiber) that lacks specific long-range contacts during the mitotic phase. It is important to discuss the structural nature of the mitotic chromosome here. As demonstrated by the imaging results

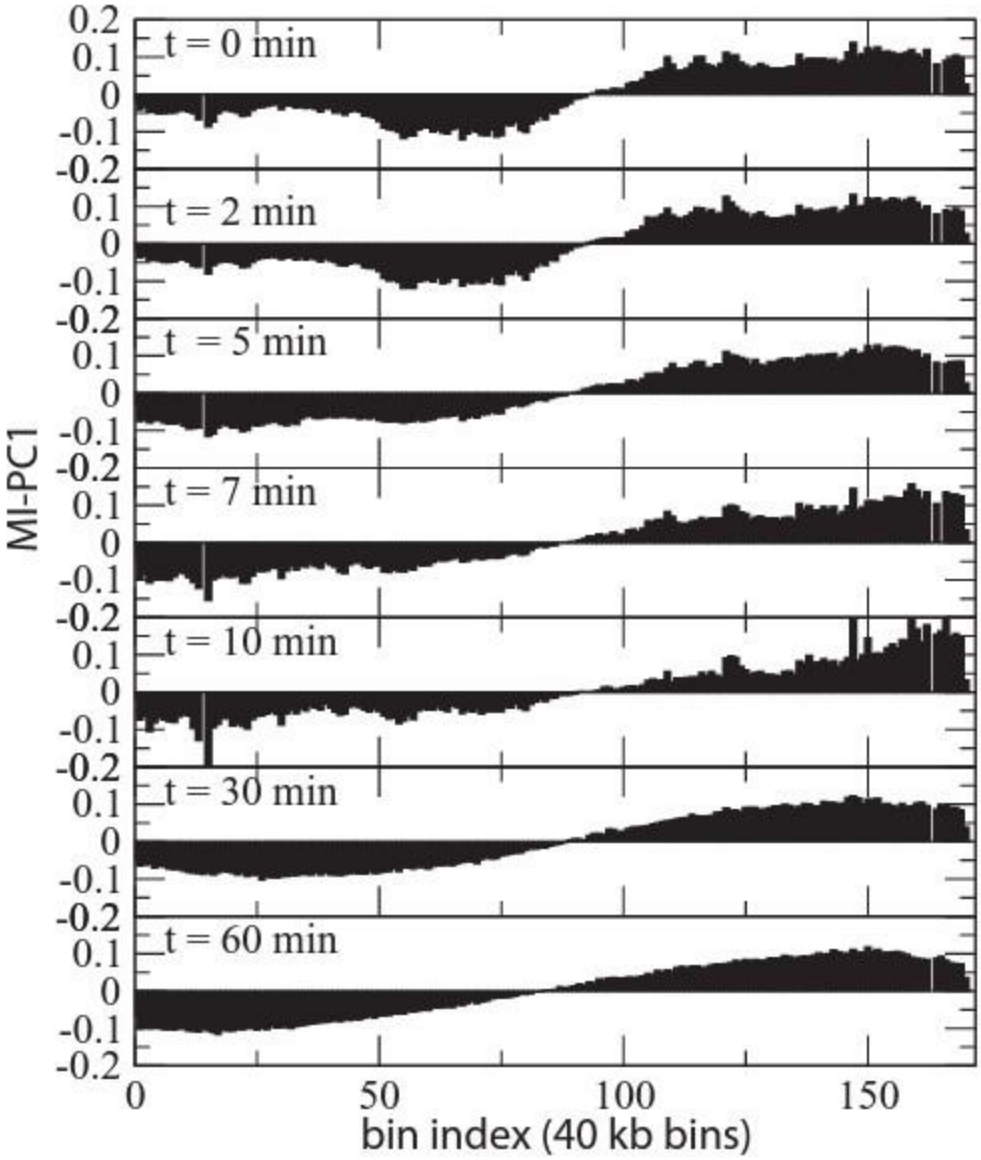


Figure 40 MI-PC1 of chicken chromosome 21 (chr21) at timepoints progressing from G2 to late prometaphase. The x-axis represents the 40 kb bin index along the chromosome.

at t =60 min in Ref. (Gibcus et al., 2018), chromosomes are shown as thick fibers with a thickness of 1 μm. It is important to point out that the close match to the MI-PCA curve of a generic noninteracting polymer during metaphase does not mean that the structure behaves as a linear DNA polymer at a much finer resolution. Instead, the compacted metaphase fiber as a whole acts as a coarse-grained polymer without specific distal contacts, while a sophisticated local structural model of regular loops being packed along a spiral "vine" to form such fiber was shown (Gibcus et al., 2018). Also, it is noteworthy that one can observe the underlying sinusoidal undertone throughout the time points of the cell cycle monitored here, even at t= 0 min where the chromosomes are still quite structured. Such undertone indicates the relatively weaker long-range contacts compared to the near diagonal ones.

Chromosomes maintain detailed structure under salt and drug treatments

We next asked whether chemical perturbations to chromosomes, analogous to increased temperature in protein simulations, would cause their structure to approach a more generic polymer state. We have examined two types of perturbations with Hi-C: changes in chromatin compaction via histone modifications and changes in cation concentration (Amat et al., 2019) (high and low salt). We find overall that these perturbations did not unpack the chromosome to the level of an effective generic polymer. However, we note that this approach can distinguish the degrees of partial chromosome folding that otherwise would look nearly identical in a distance decay normalized compartment analysis.

Histone deacetylases (HDACs) directly affect local chromatin structure by removing acetyl groups from histone tails and in turn making chromosomes more packed (Seto & Yoshida, 2014). Their inhibitors, HDACi, such as trichostatin A (TSA) allow the chromosome to exist in a more acetylated state, an open conformation which is more easily accessible to transcription factors (Wang et al., 2009). The decondensation of chromosomes by TSA has been shown to affect the overall physical stiffness of nuclei (Stephens et al., 2018), but it is not clear how it affects the polymer properties of chromosomes as measured by Hi-C contacts. We have analyzed the Hi-C chromosome contacts of cells treated with 0.5 μM TSA for 2 hours, a timepoint at which local histone modifications can be detected (Wang et al., 2009). As demonstrated in Fig. 41 using human chr21, MI-PC1 is heavily structured under these perturbations and does not resemble the signature of a generic polymer. However, one can clearly identify interesting regions (such as 21q22.2–21q22.3) that show significant differences in MI-PCA profiles that is not noticeable with a decay-normalized analysis. A full discussion of the important biological implications of these specific changes are beyond the scope of the current study, whereas here we emphasize that applying MI-PCA without matrix pre-treatment reveals variations between two structures that are not apparent in traditional "compartment analysis".

Perturbations such as altering concentration of salts may also affect the level of packing (Amat et al., 2019). For example, depleting salt perturbs the electrostatic interaction, leading to stronger electrostatic repulsion and swelling of the chromosomes. On the other

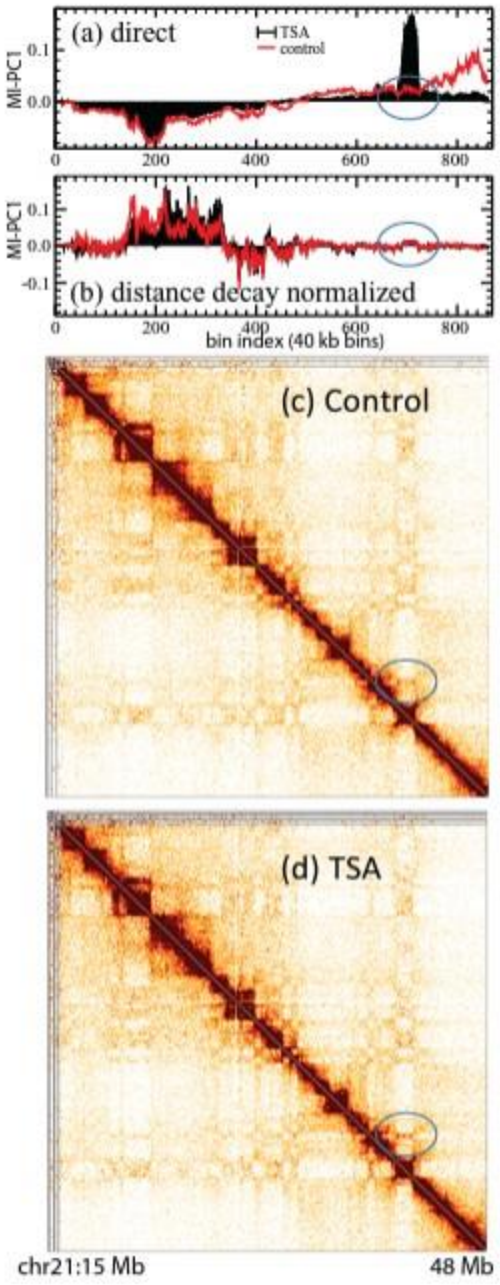


Figure 41 The comparison of drug (TSA) treated human chromosome 21 and the corresponding control using direct contact analysis, i.e., MI-PCA on u_{ij} (a) and the conventional approach with the distance-decay scaled (b), i.e., MI-PCA on v_{ij}.

Here the bin size used for the analysis is 40 kb. Region displayed is chr21:15–48 Mb. The corresponding contact interaction matrices u_{ij} (ICE normalized, but no distance normalization applied, bin size 100 kb) for the control (c) and TSA treated (d) chr21. The region where direct contact analysis detected a difference between TSA and control is highlighted (blue circles).

hand, stronger electrostatic shielding at high salt concentration will compact chromosomes (Widom, 1986). However, exploratory study of MI-PC1 analysis of Hi-C contacts under these perturbations shows that higher order topology is maintained, and these chromosomes do not behave more like a generic noninteracting polymer. (data not shown)

Cross comparison using score functions

We demonstrate here that score functions based on the MI-PC1 can further characterize the generic polymer nature (or lack thereof) of biopolymers such as proteins and chromosomes. Before we present the application with concrete data, we first introduce two empirical score functions (based on MI-PC1 as the input) to measure how close a particular conformational ensemble is to the two limiting cases: at one end, the native, folded structure and at the other end, the generic, noninteracting polymer.

For comparing whether a structure is close to a specific folded structure, we must use information from that specific structure as a reference state, as there is not a universal structural signature for all structured polymers. Specifically, we define a structure deviation score $R_{str} = \sum_i \left(a_i - a_i^{ref}\right)^2$. Here the reference state is chosen as the MI-PC1 of a highly structured ensemble. For example, for protein ensemble, we can use low temperature limit as a reference state. A high value of R_{str} may indicate a large deviation from that structure. On the other hand, we would like to define how an ensemble is deviated from to the presumed generic polymer limit with no specific structures. Here we define an ad hoc function, termed the interacting polymer score, $R_{ip} = \sum_{i=1}^{N-1}(a_{i+1} - a_i)^2 \times {N^2}/{4\pi^2}$, which works out well practically. Here, the linear size of the chromosome contact matrix, N is defined as the ratio of the total size of chromosome over the bin size. The rationale behind the resolution-free definitions of R_{ip} and R_{str} is as follows. Ideally, we want our score functions to be independent of how we choose the bin size. When we perform the MI-PCA, we can describe the normalized top eigenvector with its components a_i with $\mathrm{i} = 1, \ldots, \mathrm{N}$. However, this definition of a_i is not resolution free and depends on bin size. By the definition of a normalized eigen vector, we have $\sum_{i=1}^{N} a_i^2 \equiv 1$, so with increasing N, the value of element a_i decreases as $1/\sqrt{N}$. This point was well demonstrated in the transformation of $y^* = \mathrm{y} \times \sqrt{65/100}$ in Fig. 38a so we can project I-PCA results of a hypothetical 100-bead ideal polymer using a 65-bead polymer simulation. Thus, a resolution-independent definition of normalized eigenvector is necessary. Thus, we define a transformation $a_i \to \tilde{a}(\mathrm{x})$ which makes $\int_{-\pi}^{\pi} \tilde{a}(x)^2 \mathrm{dx} \equiv 1$
Where we define

$$\mathrm{x} \equiv 2\pi \times (\mathrm{i}/\mathrm{N}) - \pi \qquad (1)$$

Here, we map the discrete index range $\mathrm{i} \in [1,\mathrm{N}]$ to the normalized continuous range $\mathrm{x} \in (-\pi, +\pi]$. We choose to use π in our definition of a normalized range because, as demonstrated above, the IPCA eigenvectors of polymers approaching the non-interacting

limit have the nature of a cosine function series. Comparing the normalization condition in the integral form and in the discrete sum form and since $\mathrm{dx} \to \Delta x = 2\pi/N$, we can define a new resolution-free function

$$\tilde{a}(\mathrm{x}) \equiv a_i \times (2\pi/N)^{-1/2} \tag{2}$$

Finally,

$$\begin{aligned} R_{ip} \equiv \int \left(\frac{d\tilde{a}(\mathrm{x})}{dx}\right)^2 dx &= \int \left(\frac{\tilde{a}(x+\Delta\mathrm{x}) - \tilde{a}(x)}{\Delta x}\right)^2 dx \\ &= \sum_i \left(\frac{a_{i+1} - a_i}{\Delta x}\right)^2 \times (2\pi/N)^{-1} \times \Delta x \\ &= \sum_i (a_{i+1} - a_i)^2 \times (2\pi/N)^{-2} \end{aligned} \tag{3}$$

Thus, we have a factor $(N/2\pi)^2$ in the definition of R_{ip}. This factor essentially accounts for the slope being steep in the normalized function $\tilde{a}(\mathrm{x})$since the length is 2π instead of N. While on the other hand, for R_{str}, we do not need to explicitly compensate when comparing different resolutions N, since

$$R_{str} \equiv \int \left(\tilde{a}(\mathrm{x}) - \tilde{a}_{ref}(\mathrm{x})\right)^2 dx = \sum_i \left(a_i - a_{i,\,ref}\right)^2 \tag{4}$$

We first apply our scoring functions to the protein folding of CI-2 at different temperatures. As shown in Fig. 42a, with rising temperature, the polymer gradually loses long-range contacts (measured by a smaller interacting polymer score) and gradually loses its nativeness (measured by an increasing of deviation from the native reference MI-PC1). Here the native structure reference state is defined by $\mathrm{T} = 0.5\, T_f$. For the second example of chromosomes during the cell cycle, we can measure the corresponding properties for chr21, and we obtain an interesting curve with the evolution of time (Fig. 42b). It first decreases the structural order by losing native contacts and forming new contacts then increases nativeness again at t = 10-30 min before finally approaching the noninteracting limit at t = 60min. Other chicken chromosomes show a similar dynamic signature as well, suggesting that this multi-phase progression towards the mitotic chromosome is a fundamental biological feature. The detection of this multi-phase structure transition feature showcases the usefulness of the contact analysis as this feature is not immediately apparent in the original Hi-C contact matrices.

Also, the conclusion that the nearly noninteracting, generic effective polymer state of prometaphase chromosomes at t = 60 min, holds throughout the chicken genome not only for chr21. In Fig. 43, we display the MI-PC1 (Fig. 43a) and the corresponding interacting

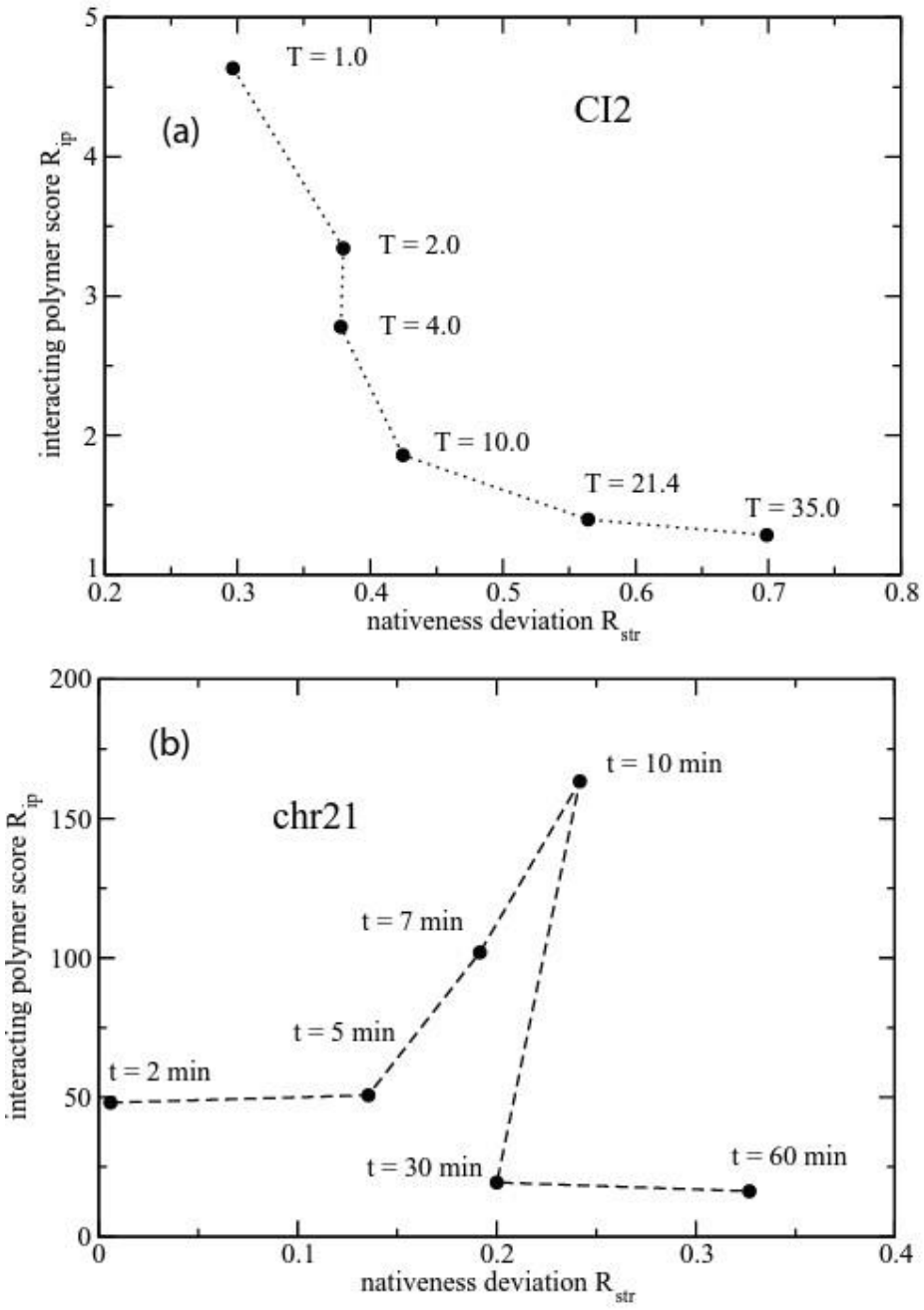


Figure 42 (a) Protein CI-2 at different temperatures characterized by two parameters R_{str} and R_{ip}, where the x-axis measures against a specific folded structure at the low temperature limit and the y-axis represents how well the structure resembles the features of the corresponding generic polymer. (b) Chicken chromosome 21 at different time points during the cell cycle characterized by two parameters R_{str} and R_{ip}.

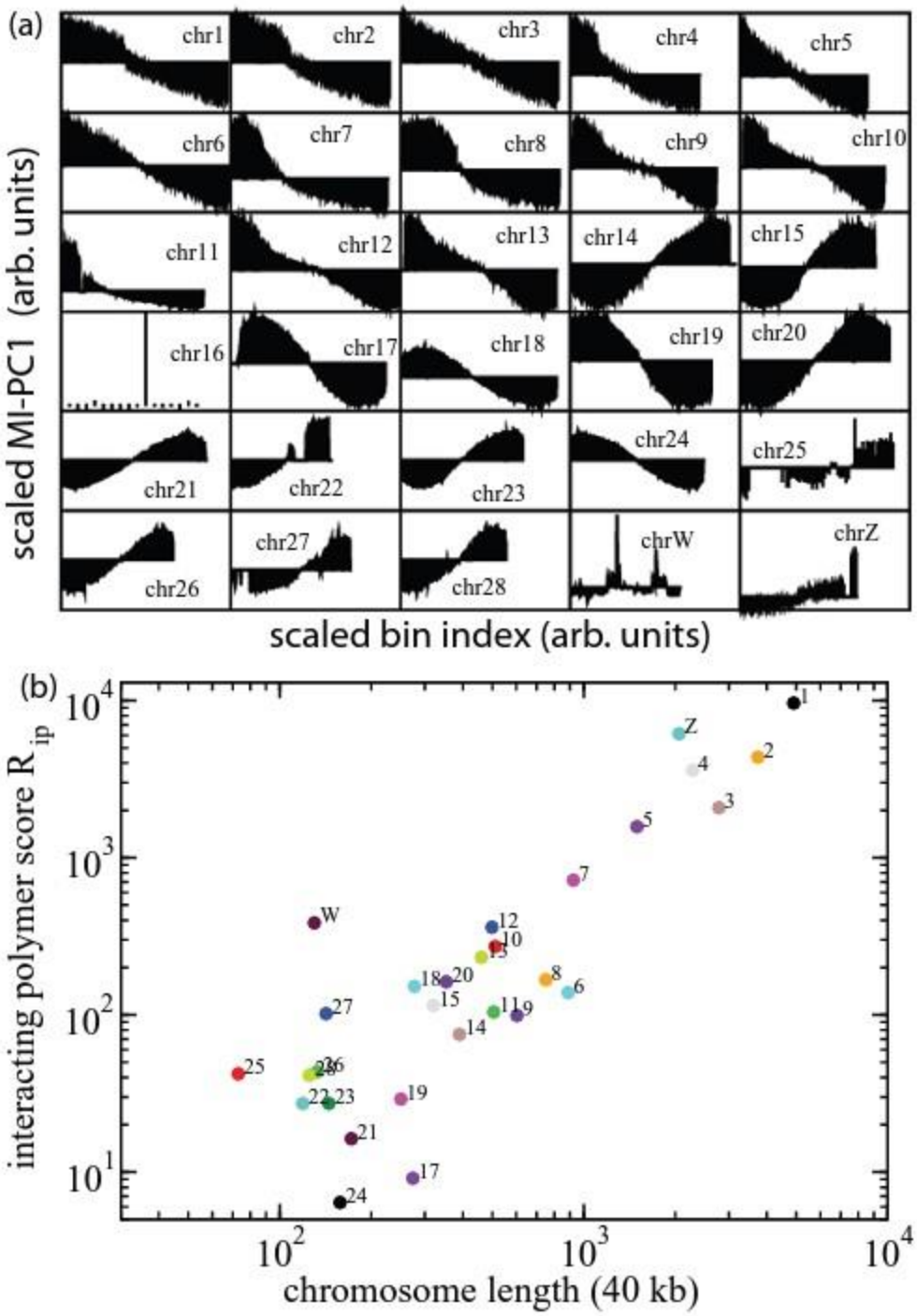


Figure 43 (a) The structural features of chicken chromosomes at t = 60 min expressed by the MI-PC1s of the unpretreated contact matrices. (b) The R_{ip} versus the corresponding chromosome length.

polymer score (Fig 43b) across chicken chromosomes at prometaphase. We observed an interesting feature: the longer chromosomes have a larger R_{ip} than the shorter ones. A few chicken chromosomes, chr16, 22, 25, and Z, display gaps in their contact maps which may affect their scores, but the differences observed between others cannot be simply explained by data artifacts. Further, this trend cannot be explained by the more trivial aspect of polymer chain length dependency, as the score function R_{ip} has been properly normalized in its definition. It is interesting to speculate why larger chromosomes have larger R_{ip} at $\mathrm{t} = 60$ min and are visually different from the sinusoidal curves. One possible explanation is that larger chromosomes lag behind shorter ones in term of cell cycle dynamics, taking longer to reach their final mitotic structure. Another scenario could be that with a fixed persistent length, larger chromosomes make more long-range contacts and thus effectively have a wide band-diagonal like matrix, as illustrated in Fig. 38c.

Another interesting observation is that allosome chrW appears to be the only chromosome highly structured at t = 60 min, both from a direct observation of MI-PC1 and from its position as a clear outlier on the score function plot. Upon closer inspection of the contact map, we speculated that these "structures" in fact represented errors in the sequencing assembly of the chicken genome chrW as represented in genome database galgal5 (Warren et al., 2017). This mis-assembly, confirmed by galgal6, would have been much less noticeable without this current method of examination.

Concluding remarks

Statistical analysis of contact matrices has been used to characterize the structure and dynamics of biopolymers. Particularly, I-PCA type of contact structural analysis is used to describe chromosome structures. Often, the matrix is pretreated to scale up the long-range contacts and enhance the signature of the structure. We explore the advantages of performing similar analyses with untreated contacts, especially for those structural ensembles of biopolymers with little specific structure and thus few long-range contacts. We also further suggest secondary tools to characterize the transformation from structured polymers to a generic polymer limit based on the first eigenvector of I-PCA. We found that certain operations on chromosomes will make them approach the noninteracting polymer limit (such as those that occur during certain phases of the cell cycle) while other perturbations are not dramatic enough. The methods and score functions developed can be useful to quantitatively study a wide range of largely unstructured biopolymer systems, from intrinsically disordered protein domains to mitotic chromosomes.

Acknowledgement

We specially thank R. J. Lindsay for helpful discussions on coarse-grained protein simulations using Gromacs. A portion of the research reported in this publication was supported by NIGMS of the National Institutes of Health under award number R35GM133557 to R.P.M. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Amat, R., Böttcher, R., Le Dily, F., Vidal, E., Quilez, J., Cuartero, Y., Beato, M., de Nadal, E., & Posas, F. (2019). Rapid reversible changes in compartments and local chromatin organization revealed by hyperosmotic shock. *Genome research*, *29*(1), 18-28.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268-276.

Berendsen, H. J. C., van der Spoel, D., & van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, *91*(1), 43-56. https://doi.org/https://doi.org/10.1016/0010-4655(95)00042-E

Brown, T. A. (2018). *Genomes 4*. Garland science.

Bryngelson, J. D., & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, *84*(21), 7524-7528.

de Gennes, P. G. (1979). *Scaling Concepts in Polymer Physics*. Cornell University Press. https://books.google.com/books?id=ApzfJ2LYwGUC

Dewey, T. G. (1997). *Fractals in Molecular Biophysics*. Oxford University Press. https://books.google.com/books?id=EAEPYhH6NwsC

Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, *24*(6), 1501-1509.

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The Protein Folding Problem. *Annual Review of Biophysics*, *37*(1), 289-316. https://doi.org/10.1146/annurev.biophys.37.092707.153558

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376-380.

Doi, M. (1996). *Introduction to polymer physics*. Oxford university press.

Fersht, A. (1999). *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. Macmillan.

Flory, P. J., & Volkenstein, M. (1969). Statistical mechanics of chain molecules. In: Wiley Online Library.

Gibcus, J. H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M. T., Xie, L., Paulson, J. R., & Earnshaw, W. C. (2018). A pathway for mitotic chromosome formation. *Science*, *359*. https://doi.org/10.1126/science.aao6135

Go, N. (1983). Theoretical studies of protein folding. *Annual review of biophysics and bioengineering*, *12*(1), 183-210.

Gō, N., & Scheraga, H. A. (1969). Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules. *The Journal of Chemical Physics*, *51*(11), 4751-4767.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., & Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, *28*(23), 3131-3133. https://doi.org/10.1093/bioinformatics/bts570

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization [10.1038/nmeth.2148]. *Nature Methods*, *9*(10), 999-1003. https://doi.org/http://www.nature.com/nmeth/journal/v9/n10/abs/nmeth.2148.html#supplementary-information

Jackson, S. E., & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry*, *30*(43), 10428-10435. https://doi.org/10.1021/bi00107a010

Johnson, Q. R., Lindsay, R. J., & Shen, T. (2018). CAMERRA: An analysis tool for the computation of conformational dynamics by evaluating residue–residue associations. *Journal of Computational Chemistry*, *39*(20), 1568-1578. https://doi.org/doi:10.1002/jcc.25192

Karplus, M., & Kushick, J. N. (1981). Method for estimating the configurational entropy of macromolecules. *Macromolecules*, *14*(2), 325-332.

Ladurner, A. G., Itzhaki, L. S., & Fersht, A. R. (1997). Strain in the folding nucleus of chymotrypsin inhibitor 2. *Folding and Design*, *2*(6), 363-368. https://doi.org/https://doi.org/10.1016/S1359-0278(97)00050-3

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods*, *72*, 65-75. https://doi.org/https://doi.org/10.1016/j.ymeth.2014.10.031

Li, W., Gong, K., Li, Q., Alber, F., & Zhou, X. J. (2014). Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics*, *31*(6), 960-962. https://doi.org/10.1093/bioinformatics/btu747

Lieberman-Aiden, E., Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., & Dorschner, M. O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*. https://doi.org/10.1126/science.1181369

Lindsay, R. J., Pham, B., Shen, T., & McCord, R. P. (2018). Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic Acids Research*, *46*(16), 8143-8152. https://doi.org/10.1093/nar/gky604

Lindsay, R. J., Siess, J., Lohry, D. P., McGee, T. S., Ritchie, J. S., Johnson, Q. R., & Shen, T. (2018). Characterizing protein conformations by correlation analysis of coarse-grained contact matrices. *The Journal of Chemical Physics*, *148*(2), 025101. https://doi.org/10.1063/1.5004141

McCord, R. P., Nazario-Toole, A., Zhang, H., Chines, P. S., Zhan, Y., Erdos, M. R., Collins, F. S., Dekker, J., & Cao, K. (2013). Correlated alterations in genome organization, histone methylation, and DNA–lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Research*, *23*(2), 260-269. https://doi.org/10.1101/gr.138032.112

McKenzie, D. S. (1976). Polymers and Scaling [Review article]. *Physics Reports*, *27*(2), 35-88.

McPhalen, C. A., & James, M. N. G. (1987). Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry*, *26*(1), 261-269. https://doi.org/10.1021/bi00375a036

Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology [Review Article]. *Nature Reviews Genetics*, *15*, 709. https://doi.org/10.1038/nrg3788

Michieletto, D., Orlandini, E., & Marenduzzo, D. (2016). Polymer model with epigenetic recoloring reveals a pathway for the de novo establishment and 3D organization of chromatin domains. *Physical Review X*, *6*(4), 041047.

Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell [journal article]. *Chromosome Research*, *19*(1), 37-51. https://doi.org/10.1007/s10577-010-9177-0

Miura, H., Takahashi, S., Poonperm, R., Tanigawa, A., Takebayashi, S.-i., & Hiratani, I. (2019). Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nature Genetics*, *51*(9), 1356-1368. https://doi.org/10.1038/s41588-019-0474-z

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., & Dekker, J. (2013). Organization of the mitotic chromosome. *Science*, *342*(6161), 948-953.

Noel, J. K., Levi, M., Raghunathan, M., Lammert, H., Hayes, R. L., Onuchic, J. N., & Whitford, P. C. (2016). SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLOS Computational Biology*, *12*(3), e1004794. https://doi.org/10.1371/journal.pcbi.1004794

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., & Sedat, J. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381-385.

Pal, K., Forcato, M., & Ferrari, F. (2019). Hi-C analysis: from data generation to integration. *Biophysical reviews*, *11*(1), 67-78.

Poirier, M. G., Nemani, A., Gupta, P., Eroglu, S., & Marko, J. F. (2001). Probing chromosome structure with dynamic force relaxation. *Physical review letters*, *86*(2), 360.

Portman, J. J., Takada, S., & Wolynes, P. G. (1998). Variational Theory for Site Resolved Protein Folding Free Energy Surfaces. *Physical Review Letters*, *81*(23), 5237-5240. https://doi.org/10.1103/PhysRevLett.81.5237

Ray, J., Munn, P. R., Vihervaara, A., Lewis, J. J., Ozer, A., Danko, C. G., & Lis, J. T. (2019). Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proceedings of the National Academy of Sciences*, *116*(39), 19431-19439. https://doi.org/10.1073/pnas.1901244116

Rouse, P. E. (1953). A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers. *The Journal of Chemical Physics*, *21*(7), 1272-1280. https://doi.org/10.1063/1.1699180

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., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., & Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences, USA*, *112*(47), E6456-E6465. https://doi.org/10.1073/pnas.1518552112

Schmitt, A. D., Hu, M., & Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, *17*(12), 743-755.

Seto, E., & Yoshida, M. (2014). Erasers of histone acetylation: the histone deacetylase enzymes. *Cold Spring Harbor perspectives in biology*, *6*(4), a018713.

Stephens, A. D., Banigan, E. J., & Marko, J. F. (2018). Separate roles for chromatin and lamins in nuclear mechanics. *Nucleus*, *9*(1), 119-124.

Sun, M., Biggs, R., Hornick, J., & Marko, J. F. (2018). Condensin controls mitotic chromosome stiffness and stability without forming a structurally contiguous scaffold. *Chromosome Research*, *26*(4), 277-295.

Takahashi, T., Okeyo, K. O., Ueda, J., Yamagata, K., Washizu, M., & Oana, H. (2018). A microfluidic device for isolating intact chromosomes from single mammalian cells and probing their folding stability by controlling solution conditions. *Scientific reports*, *8*(1), 1-10.

Tiana, G., & Giorgetti, L. (2019). *Modeling the 3D conformation of genomes*. CRC Press.

Wang, Z., Zang, C., Cui, K., Schones, D. E., Barski, A., Peng, W., & Zhao, K. (2009). Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, *138*(5), 1019-1031.

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., Markovic, C., Bouk, N., Pruitt, K. D., Thibaud-Nissen, F., Schneider, V., Mansour, T. A., Brown, C. T., Zimin, A., Hawken, R., Abrahamsen, M., Pyrkosz, A. B., Morisson, M., Fillon, V., Vignal, A., Chow, W., Howe, K., Fulton, J. E., Miller, M. M., Lovell, P., Mello, C. V., Wirthlin, M., Mason, A. S., Kuo, R., Burt, D. W., Dodgson, J. B., & Cheng, H. H. (2017). A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3: Genes Genomes Genetics*, *7*(1), 109-117. https://doi.org/10.1534/g3.116.035923

Widom, J. (1986). Physicochemical studies of the folding of the 100 Å nucleosome filament into the 300 Å filament: cation dependence. *Journal of molecular biology*, *190*(3), 411-424.

Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, *43*, 1059. https://doi.org/10.1038/ng.947

https://www.nature.com/articles/ng.947#supplementary-information

Zhou, Y., Gerrard, D. L., Wang, J., Li, T., Yang, Y., Fritz, A. J., Rajendran, M., Fu, X., Schiff, R., Lin, S., Frietze, S., & Jin, V. X. (2019). Temporal dynamic reorganization of 3D chromatin architecture in hormone-induced breast cancer and endocrine resistance. *Nature Communications*, *10*(1), 1522. https://doi.org/10.1038/s41467-019-09320-9

Zimm, B. H. (1956). Dynamics of Polymer Molecules in Dilute Solution: Viscoelasticity, Flow Birefringence and Dielectric Loss. *The Journal of Chemical Physics*, *24*(2), 269-278. https://doi.org/10.1063/1.1742462

CHAPTER IV
DIFFERENTIAL CONTRIBUTIONS OF NUCLEAR LAMINA ASSOCIATION AND GENOME COMPARTMENTALIZATION TO GENE REGULATION

A version of this chapter was originally published by Priyojit Das, Rebeca San Martin and Rachel Patton McCord:

Das, P., San Martin, R., & McCord, R. P. (2022). Differential contributions of nuclear lamina association and genome compartmentalization to gene regulation. bioRxiv.

My contributions to this work include: (1) conceived the project, (2) performed bioinformatics data analysis, (3) wrote the manuscript, and (4) prepared the figures.

Abstract

Interactions of chromatin with the nuclear lamina play a significant role in properly organizing the genome in 3D space and in regulating gene expression. Genome wide studies have inferred the global association between the lamina, heterochromatin, gene repression and the B genomic compartment, and repositioning genes to the lamina can result in their repression. However, there are scenarios in which these features are discordant and, in those cases, the relative contribution to gene regulation of genomic compartment, chromatin, and lamin association status can be examined. Here we compared datasets from cell lines representing different states of differentiation across different cell type lineages to examine the relationships between changes in genomic compartmentalization, lamin association, and gene expression. With these data, we could examine, for example, what gene expression changes occur when a B compartment region is moved from the nuclear interior to the nuclear lamina and what differences exist between lamin associated and internal A compartment regions. In general, we observed an additive rather than redundant effect in which lamin association and compartment status both contribute to gene expression state. However, we found that cell type lineages differed in whether compartment status or lamin association had a dominant influence on gene expression. Finally, we identified conserved trends of how compartment and lamin association status influence the likelihood that gene expression will be induced or repressed in response to a physiochemical treatment.

Introduction

Inside the nucleus, chromosomes interact among themselves, with other different nuclear substructures and the nuclear lamina, leading to the formation of 3D spatial genome organization (McCord et al., 2020). Several aspects of this organization have been implicated in gene regulation (Bersaglieri et al., 2020; Gupta & Santoro, 2020; Lochs et al., 2019; Morimoto & Boerkoel, 2013; Van Steensel & Belmont, 2017), including a repressive role of both chromatin tethering to the nuclear lamina and spatial segregation into the B compartment (Lochs et al., 2019; Van Steensel & Belmont, 2017). The nuclear lamina is a protein meshwork underneath the nuclear membrane which contributes to the structural stability of the nucleus (Dechat et al., 2008; Vahabikashi et al., 2022). The chromatin regions tethered to the lamina (lamina-associated domains; LADs) are typically between 0.1 and 10 Mb in size and are generally gene-poor, transcriptionally repressed, heterochromatin-rich regions (Van Steensel & Belmont, 2017). Several studies have shown that tethering an active region to the lamina tends to lead to decreased expression of the genes located in that region (Finlan et al., 2008; Harr et al., 2015; Robson et al., 2016; Rønningen et al., 2015; Zullo et al., 2012). Targeted activation or inactivation of genes can promote chromosome region detachment from or attachment to the nuclear lamina, respectively (Brueckner et al., 2020). However, it is not clear what purpose would be served by relocating an already inactive region to the lamina or an already active region toward the interior. Genome-wide chromosome conformation capture experiments (Hi-C) have demonstrated that chromosomal regions spatially segregate into different compartments in the nucleus (Lieberman-Aiden et al., 2009; Rao et al., 2014). The regions classified as belonging to the B compartment are enriched in heterochromatic marks, gene-poor regions, and inactive genes and are often associated with the nuclear lamina. Regions that switch into the B compartment often experience gene repression and vice versa (San Martin et al., 2021). But, despite the frequently reported correlations between compartment status, lamin association, and gene expression, the combinatorial effects of different repressive or activating environments on gene expression are not fully understood. How are genes affected when a typically active A compartment region is located at the lamina? Does an already repressed B compartment region become further stably repressed when located at the nuclear periphery as compared to the interior? Detailed analyses of the relationships between lamina association, gene expression, and compartmentalization are often performed in a single cell type, but does the relative contribution of these factors vary by cell type or cell lineage?

One approach to shed light on these questions is to consider cells across the spectrum of differentiation status in different cell type lineages. As a stem-like cell progresses through differentiation, finally committing to a particular fate, gene expression, genomic compartmentalization, and LAD organization also change (Bonev et al., 2017; Cuomo et al., 2020; Dixon et al., 2015; Grancharova et al., 2021; Keough et al., 2021; Miura et al., 2019; Peric-Hupkes et al., 2010). Thus, in this work, we integrated publicly available genomic data to compare cell lines in different lineages and stages of differentiation to investigate the relationship between compartment identity, LAD reorganization, and gene expression differences between cell lines.

Genome structure and lamina association can influence not only the stable expression or repression of genes in a particular cell type but also how likely a gene is to be induced by a stimulus (Leemans et al., 2019; Stik et al., 2020). Given the same perturbation, different cell types will activate or repress different genes (Honda et al., 2022), and we hypothesize that the pre-existing compartment and lamin organization state of the chromatin contributes to this differential gene expression. Thus, we also examined the impact of compartment status and lamina association on gene expression changes after stimulus or perturbation.

In this study, we present evidence of the existence of both conserved and lineage-specific trends of gene expression regulation. We found that there is most frequently an additive effect of LAD and compartment status on gene expression. However, for specific regions, depending on their underlying compartment identity and lamina association in different cell type lineages and states, we observed differences in whether compartment or LAD status dominate the gene expression effect. Further, across different cell types and perturbations, genes that were either induced or repressed were tended to be enriched in the interior A compartment state, suggesting that this positioning is most favorable for altering gene expression in either direction.

Results

Chromosome compartmentalization and lamina association segregate cells by lineage

To investigate how the LAD reorganization contributes to the regulation of gene expression located in the different compartments of the genome, we required data representing: a) genome-wide lamina association (to assess whether a region is associated with the lamina), b) compartmental organization (to assess whether a region belongs to A or B compartment) and c) transcriptomics (for gene expression) data. Therefore, we selected cell types for which publicly available lamina association (mainly DamID data, unless otherwise mentioned), genome organization (Hi-C) and gene expression (RNA-seq) data were available. In total, eleven different cell types consisting of both primary cells and cell lines were collected for this study and classified based on developmental lineage and malignancy status based on known characteristics. Considering embryonic stem cells (H1ESC) as the node of origin, cells were separated into three lineages: a) mesoderm, derived from mesenchymal stem cells (HMSC), b) endoderm, primarily composed of epithelial cells and c) hematopoietic cells (Fig 44a) ("Chapter 13 - Stem Cells and Regenerative Medicine," 2021; Keller, 2005). The mesoderm-derived cell lines were sub-divided into two families: fibroblasts (WI38, IMR90, Tig3, and Hffc6) and osteoblast-derived osteosarcoma (U2OS) (Caplan, 2005; Hayflick, 1965; Matsuo et al., 1982; Nichols et al., 1977). Among the fibroblast cells, WI38, IMR90, and Tig3 are all female fetal lung fibroblast cell lines and, for the purpose of this analysis, considered interchangeable (Hayflick, 1965; Matsuo et al., 1982; Nichols et al., 1977). The epithelial cell lines were classified as either potentially secretory (colon carcinoma – HCT116) or non-secretory (RPE) (Brattain et al., 1981; Peterson & Artis, 2014; Yang et al., 2021). A special case of epithelial fibrosarcoma (HT1080) was considered a potential midpoint between a bona fide sarcoma (U2OS) and

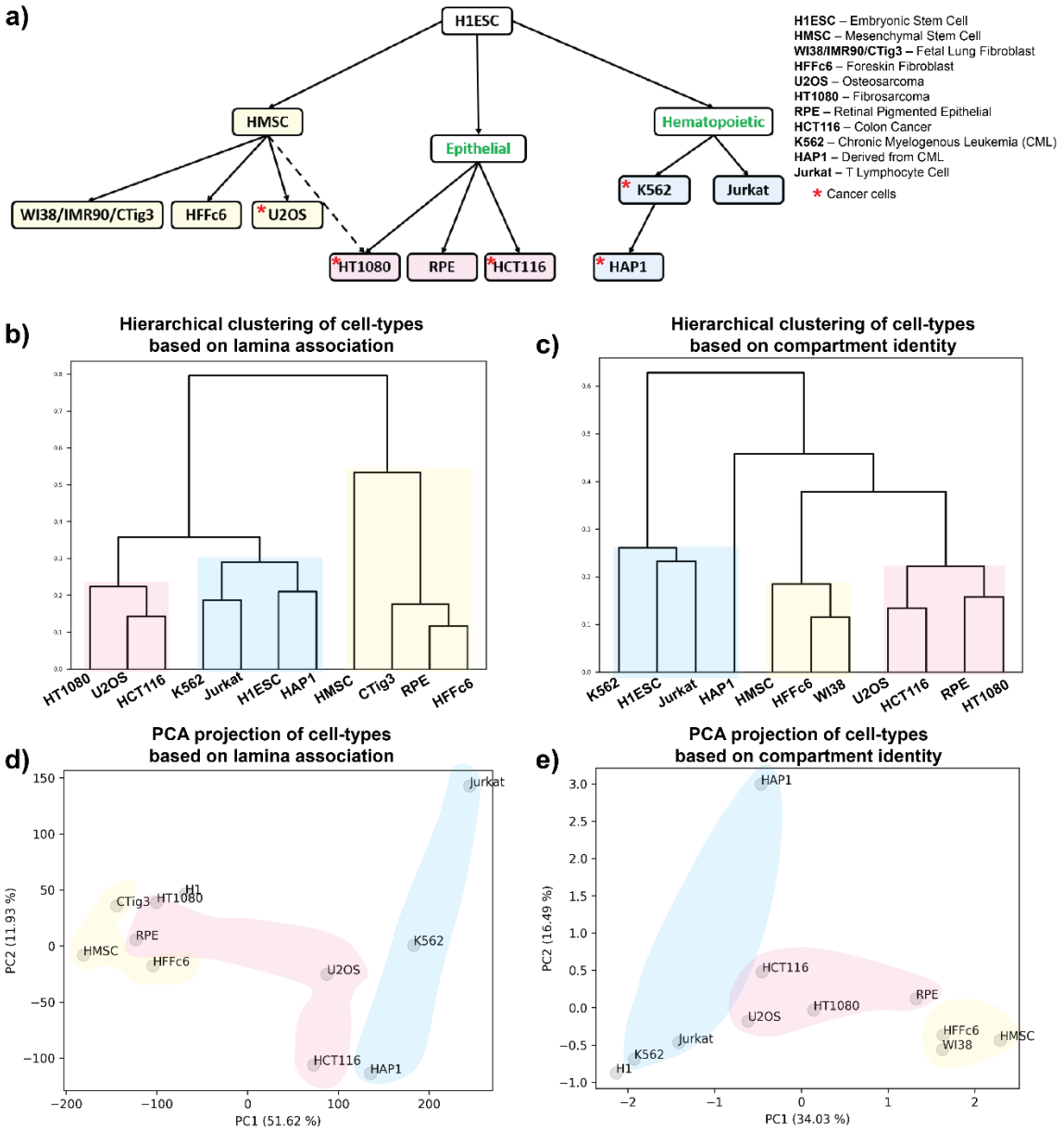


Figure 44 Cell lineage relationships are detected in both compartmentalization and lamin association patterns.

a) Human primary cells and cell lines are classified based on their developmental lineage relationships and malignancy status based on known characteristics. Yellow: mesenchymal, pink: epithelial, blue: hematopoietic, green labels indicate intermediate states in lineages which do not have specific samples representing them. b, c) Hierarchical clustering of genome-wide lamina association (b) and compartment identity (c) data for different cell types at 40 kb and 250 kb resolutions respectively. Sub-trees are colored based on their lineages as shown in (a). d, e) Principal component analysis of genome-wide lamina association (d) and compartment identity (e) data for different cell types at 40 kb and 250 kb resolutions respectively. Cells are colored based on their lineages as shown in (a).

an epithelial line (Ponten & Saksela, 1967). Finally, the hematopoietic cell lines were classified as either derived from a myeloid progenitor (K562 and HAP1) or a lymphoid progenitor (Jurkat) (Kotecki et al., 1999; Lozzio & Lozzio, 1975; Schwenk & Schneider, 1975).

Once we had ordered the cell types based on prior domain knowledge, we compared this ordering to the unsupervised classification of the cell types based on either the lamina association or compartment identity data as described in the Methods. Briefly, for lamina association data, each 40 kb genomic region was assigned a lamina association value, where a value greater than zero signifies lamina association and vice versa. For compartmental organization, positive and negative values assigned to each 250 kb region indicate A and B compartment status respectively. In the case of the lamina association, hierarchical clustering produced a cell type grouping similar to what we anticipated based on published cell characteristics, with a few interesting exceptions (Fig. 44b). We found that U2OS, which is considered epithelioid-like osteosarcoma, presents a similar type of lamina association to bona fide epithelial cells. Surprisingly, the RPE (retinal pigment epithelial) cells grouped with the mesenchymal lineage cells rather than with other epithelial cells. This may be explained by the relatively small nuclear size and the incidence of polynucleated cells in RPE (Pollreisz et al., 2020; Starnes et al., 2016). Compartment organization clustering produced similar cell type ordering trends (Fig. 44c). In both datatypes, we observed H1ESC clustering with the hematopoietic lineage cells (Fig. 44b and 44c). Interestingly, principal component analysis (PCA) on both types of genomic data, showed that the principal component 1 (PC1) axis ordered the cell types based on their lineages, where specifically the mesenchymal and hematopoietic lineages can be found at the ends of the spectrum, with cells of epithelial lineage in the middle (Fig. 44d and e). This recapitulates the relationships observed by unsupervised clustering and suggests that mesenchymal and hematopoietic lineages differ significantly in both lamina and compartmental organization. Furthermore, we observed that cell types segregated according to the overall proportion of their genome associated with the lamina or B compartment (Appendix: Fig. 50). Notably, H1ESC cells show a similar proportion of genomic regions with peripheral localization as the hematopoietic lineage cells. Similarly, U2OS and RPE cells show a similar global level of lamina association to epithelial and mesenchymal cells respectively (Appendix: Fig. 50). Overall, we find that lamina association and genomic compartmentalization vary in a lineage-specific manner.

Lamina association adds another layer of gene regulation to compartment identity

We next evaluated how gene expression changes as active or inactive regions alter their spatial position inside the nucleus. We compared pairs of cell types that share a directional connection in our cell lineage tree (Fig. 44a). For the analyses, we started by comparing regions that keep their compartment identity unaltered but change their lamina association among the compared cell types. We then checked the gene expression fold change between cell types within those regions (Fig. 45 and Appendix: Fig. 51). We found that genomic

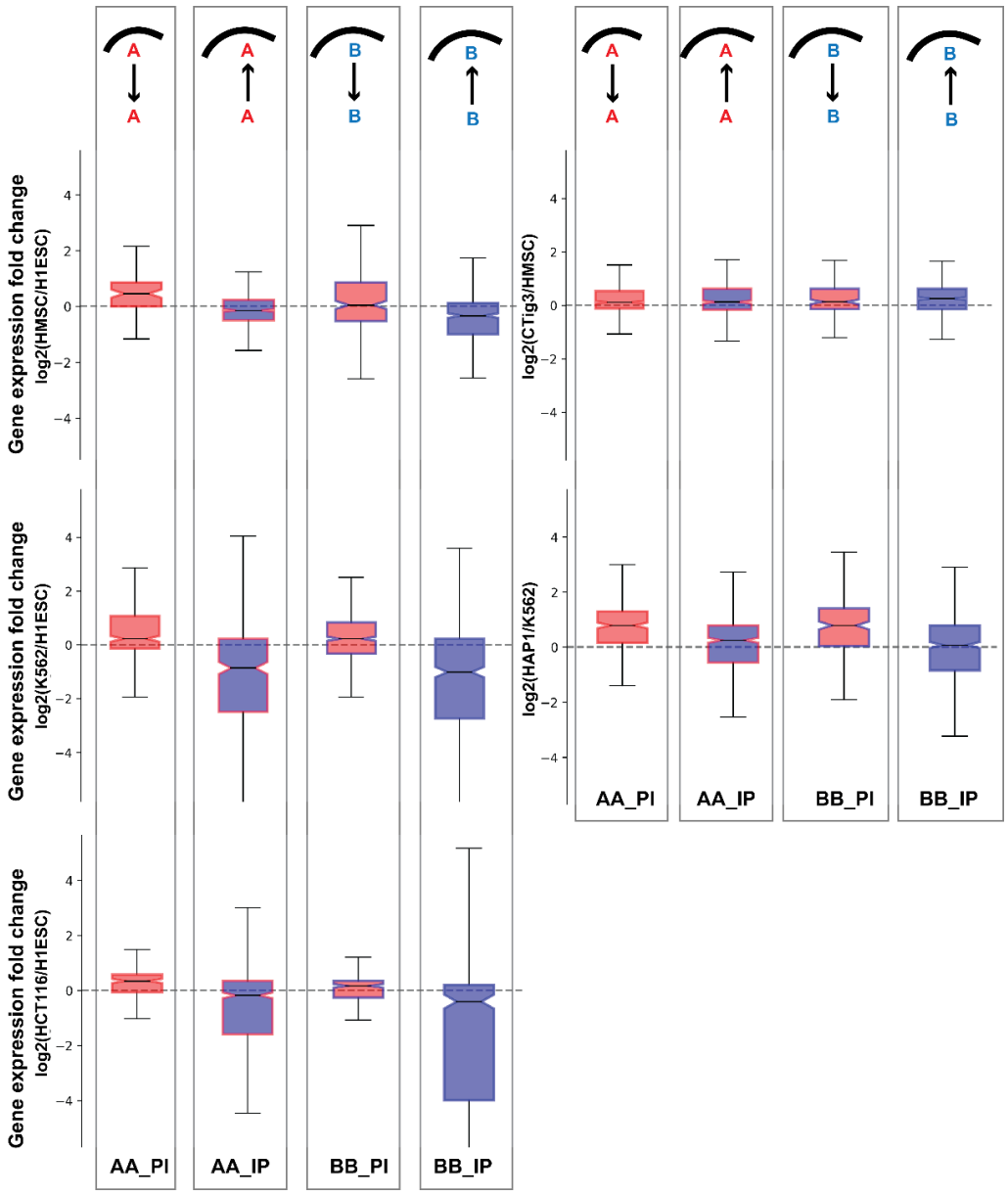


Figure 45 Gene expression change between cell types follows alterations in lamina association for genomic regions with unchanged compartment identity.

Cell types being compared are indicated in y-axis labels. Comparisons are represented using acronyms. For example, 'AA_IP' indicates regions which belong to A compartment in both cell types but switch from no lamina association ("Internal") to lamin associated ("Peripheral") from the first to the second cell type. The compartment identity of each set of regions is also shown by box plot border color (red = A, blue = B) while the internal shading indicates the direction of lamin association switch (red = shift to non-LAD, blue = shift to LAD). The transition from one cell type to another for the set of regions is also shown in a cartoon over the boxplot where the curved line indicates the nuclear lamina. Boxplots show minimum, 1st quartile, median, 3rd quartile, and maximum.

regions that have a similar compartment identity between cell types exhibit decreased or increased expression depending on their lamina association. Further, the association between lamin state and gene expression follows more strongly in the direction that would typically be associated with the underlying compartment state: gene upregulation is more likely to accompany a loss of lamina association for regions already in the A compartment and, conversely, gene repression is more likely to accompany relocation to the nuclear lamina for regions already in the B compartment. In contrast, a region that stays in the B compartment but shifts away from the lamina does not necessarily also show gene activation. Overall, these results suggest that the lamin association status of a genomic region further modulates the gene expression profile of that region, though the compartment identity remains unaltered.

Coupling and decoupling of lamina association and compartmentalization effects on gene expression

We next extended our gene expression analysis to all possible combinations of lamina association and compartment identity alterations for each of the cell type comparisons (Fig. 46 and Appendix: Fig. 52). For convenience, we refer to comparisons by acronyms such as BA_PI, indicating the compartment identity in each cell type (BA = B in the first cell type and A in the second cell type) as well as the LAD status in each cell type (PI = peripheral, LAD associated in the first cell type and internal, non-LAD associated in the second cell type). These comparisons further confirm that when compartment identity remains unaltered, changes in gene expression follow the lamina association. Since B compartment regions are often lamin associated and vice versa, we refer to B/P and A/I associations as "congruent". In all cases where the second cell type in the comparison showed congruent compartment and LAD status, (e.g., AA_PI, BA_II, AB_PP, and BB_IP), gene expression changes follow LAD status changes. On the contrary, in regions for which lamina association is incongruent with compartment identity (e.g., AB_II, BA_PP, AA_IP, BB_PI, AA_PP, AB_PI, BA_IP, and BB_II) some cell types appear to be more governed by LAD type while others favor compartment identity control. For example, in the case of mesenchymal lineage cells and RPE cells, the gene expression of these incongruent regions matches the compartment identity more than the lamina association (Appendix: Fig. 52). In contrast, the incongruent regions in comparisons among all other lineages show gene expression changes matching the lamin association status.

The strength of compartment identity correlates with lamina association and vice versa

Through the combined analysis of lamina association and compartmental organization data for different cell types, we identified several megabases of genomic regions which change their lamina association while their compartment identity remains unaltered. Since it is not known how the compartment identity strength (CIS), represented by the magnitude of the compartment analysis eigenvector, is affected by nuclear positioning, we compared the

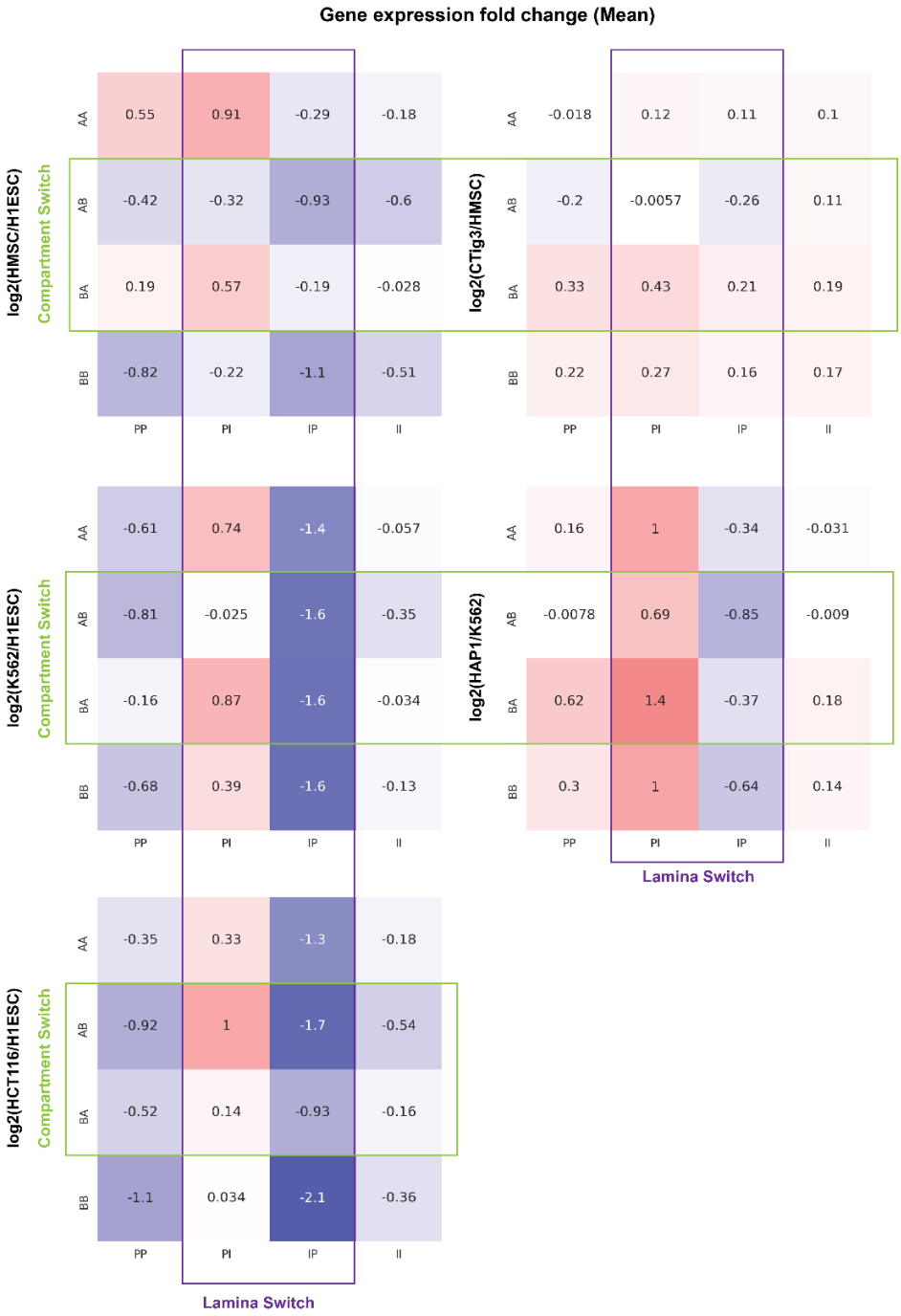


Figure 46 Mean gene expression fold change for different categories of genomic regions based on compartment identity and lamina association.

Color shading corresponds to gene expression fold change as indicated by the numbers in each box. Acronyms AB and PI are as defined in Fig. 45. Cell type comparisons are indicated to the left of each panel. Green box highlights comparisons where regions switch compartments between cell types while purple box highlights regions that change lamina status.

distribution of the CIS values for the regions which show peripheral localization in one cell type and internal in another while keeping their compartment identity unaltered (Fig. 47 and Appendix: Fig. 53). We found that irrespective of the compartment identity, the CIS values tend to decrease as the regions become more peripheral and increase when regions get internalized (Fig. 47).

In the converse analysis, we considered the strength of the lamin association signal (DamID-seq signal) for regions that change compartment status while their LAD status remains constant between two cell types (Appendix: Fig. 54). In the majority of cases, the LAD signal becomes stronger when the compartment shifts from A to B. Especially if the region was already classified as lamina associated in the first cell type, a shift to the B compartment often resulted in shifts to stronger lamina association (see H1->HT1080 comparison in Appendix: Fig 54a). In contrast, in some cell type comparisons, non-LAD regions shifted to even weaker lamina signal despite a change to the B compartment (see H1->Jurkat or K562 comparison in Appendix: Fig. 54c), suggesting that these B compartments are not mediated by lamin association. For B to A compartment shifts, a similar phenomenon was observed. For regions that were already internal, the compartment shift to A tended to reinforce the lamin status with a shift to even lower lamin association values (see for example comparisons in Appendix: Fig. 54c). However, LAD-associated regions were less likely to show a decrease in lamin association signal with a shift to the A compartment (see comparisons in Appendix: Fig. 54a).

Lamina-associated regions with similar compartment identity exhibit differential histone modifications between cell types

Epigenetic marks are known to be associated with both compartment identity and lamina association (Falk et al., 2019; Leemans et al., 2019; Poleshko et al., 2019; Smith et al., 2021). Therefore, we looked into the enrichment of different histone marks in those regions which exhibit similar compartment identity but differential peripheral organization between different cell types. To this end we focused on three different types of broad histone marks: H3K4me1 (active; enhancer), H3K27me3 (facultative heterochromatin), and H3K9me3 (constitutive heterochromatin). Details regarding the calculation of the enrichment of these different histone marks can be found in the Methods. We found that genomic regions with similar compartment identity but different lamina association status between cell-types exhibit differential histone modifications in a lineage-specific manner (Fig. 48 and Appendix: Fig. 55). For example, in the case of mesenchymal lineage comparisons, we noticed an enrichment of H3K4me1 active mark as the regions move towards the periphery, which holds for both A and B compartment regions. In contrast, for the other cell lineage comparisons, we observed a decrease in the H3K4me1 level as the regions move towards the periphery. Likewise, we noticed a lineage-specific trend for H3K9me3, where a decreased level of this constitutive heterochromatin mark was associated with the regions that move towards the periphery in most of the cell type comparisons except mesenchymal lineage cells, which show the opposite trend. And

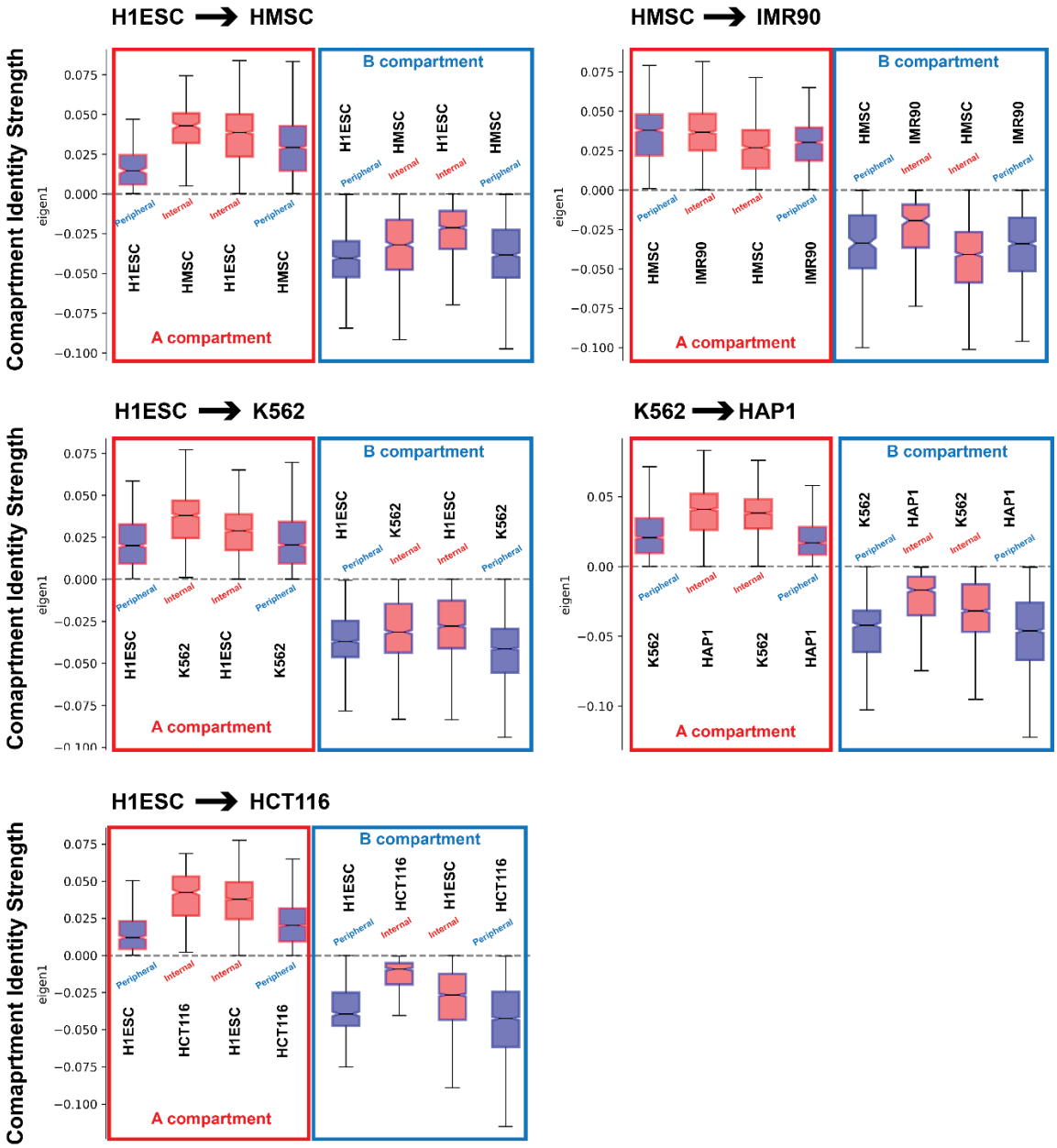


Figure 47 Compartment identity strength shifts with changes in LAD status even when compartment identity remains unchanged.

The state of each set of regions is indicated by box plot border color (compartment identity: A = red, B = blue) and shading (LAD = blue, non-LAD = red). Boxplots show minimum, 1st quartile, median, 3rd quartile, and maximum.

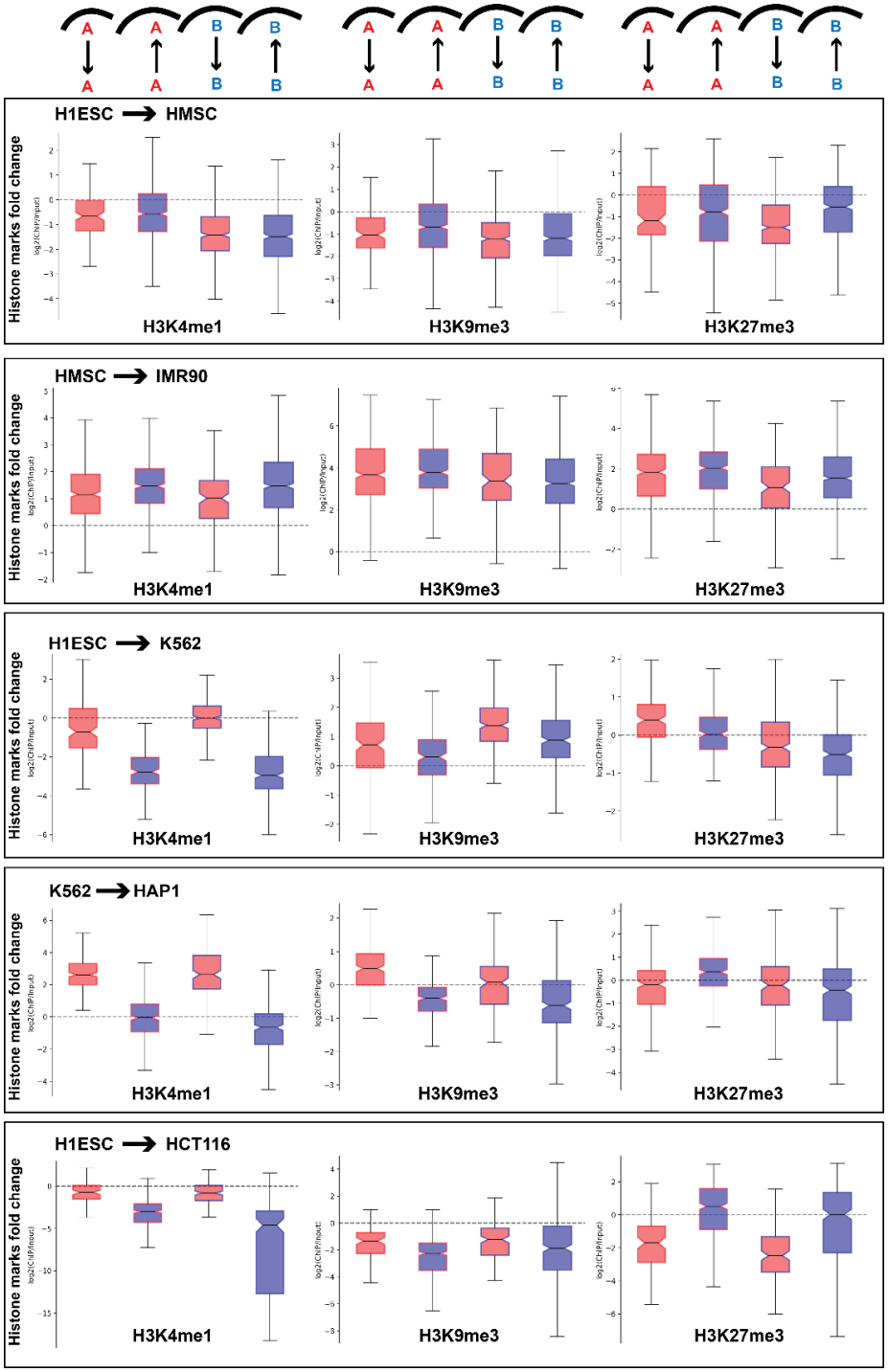


Figure 48 Fold change of different histone marks for the similar compartment identity regions which exhibit differential lamina association between cell types.

Shading and labels are as defined in Fig. 45. Boxplots show minimum, 1st quartile, median, 3rd quartile, and maximum.

finally, the H3K27me3 facultative heterochromatin mark most often increases in regions that show a relocation to the periphery while maintaining the same compartment status.

Lamina association reduces gene expression inducibility after physicochemical treatment

To examine the combined effect of compartmental and spatial localization on gene expression regulation in response to a stimulus, we analyzed transcriptomics data before and after different physicochemical treatments for different cell types with corresponding control lamina association (either DamID or laminB1 ChIP-seq), and compartment identity (Hi-C) data. Using data derived from an acute heat shock exposure of epithelial cells (HCT116), we asked how gene expression regulation was affected by compartment and lamin association in this well-characterized model for rapid and widespread transcriptional response (Kourtis et al., 2015). We found that internal positioning and A compartment status combinatorically increase the likelihood that genes will be upregulated after heat shock (Fig 49a). In contrast, downregulation occurred most frequently in lamina-associated A compartment regions.

To evaluate the effect of genome-wide histone deacetylation inhibition, in the context of LAD association and compartmentalization, we analyzed gene expression regulation in K562 cells treated with suberoylanilide hydroxamic acid (SAHA) (Fig. 49b) (Frank et al., 2016). This system models the increase in acetylation level of the chromatin at genes regulated by histone deacetylases (HDACs), which can in turn lead to changes in gene expression. In this condition, internal positioning, and A compartment identity combinatorically contributed to differential regulation, for both up- and down-regulation.

Finally, to assess how gene expression regulation can be affected by changes in transcription factor occupancy, we analyzed the effect of exogenous expression of androgen receptor splice variant 7 (AR-V7) in prostate adenocarcinoma LNCaP and bone metastatic VCaP cells (Basil et al., 2022). As with heat shock and histone deacetylase inhibitor treatment, both internal positioning and A compartment status increased the amount of gene activation or repression observed in LNCaP cells after AR-V7 induction (Fig. 49c). Gene expression changes were more modest in VCaP cells, with the percentage of upregulated genes correlated with the compartment status. In contrast, the fraction of genes downregulated was more associated with positioning relative to the lamina (Fig. 49d).

Discussion

The nuclear lamina has a well-recognized function of providing mechanical support to the nucleus and acting as a backbone to genome organization (Buchwalter et al., 2019; Cho et al., 2019; Fischer et al., 2020; Lammerding et al., 2006; Solovei et al., 2013; Srivastava et

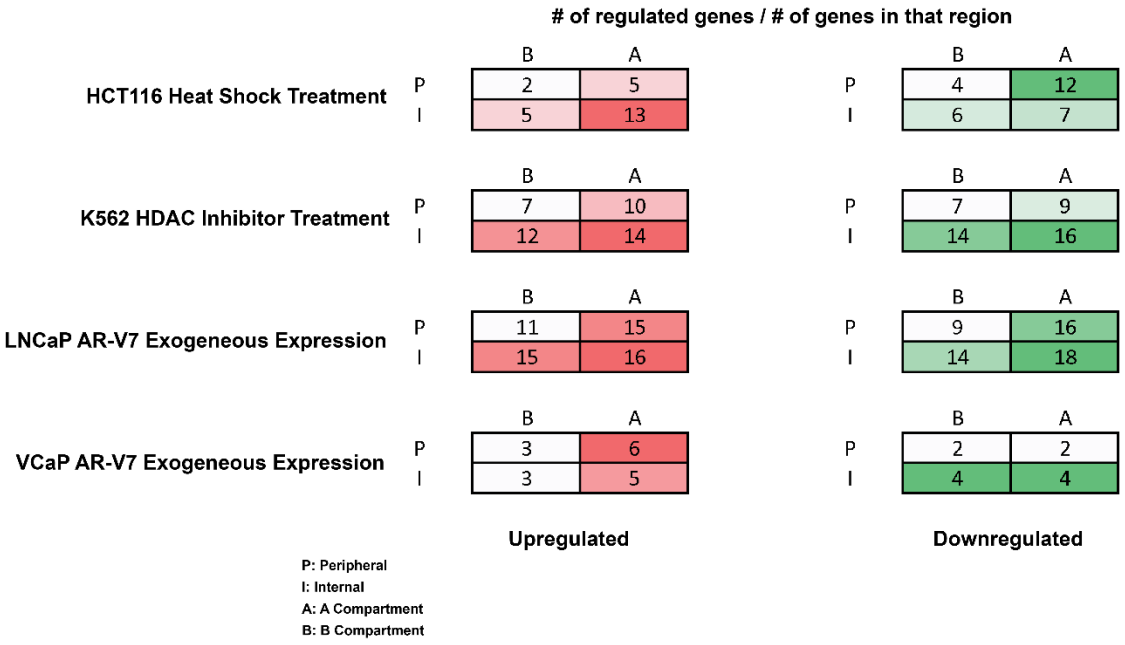


Figure 49 Effect of compartment identity and lamina association on differential gene regulation after different treatments for various cell types.

Labels indicate compartment and LAD of each set of regions before treatment and numbers and colors inside the box indicate the percentage of genes in that set of regions that are differentially upregulated (red) or downregulated (green). The differentially regulated genes are selected based on an adjusted p-value cutoff of 0.01.

al., 2021; Stephens et al., 2018; Swift et al., 2013). Functionally, chromatin contacts with the nuclear lamina have primarily been regarded as a mechanism for gene expression suppression. By creating a heterochromatic microenvironment underneath the nuclear membrane, the nuclear lamina contributes to silencing of the chromatin regions (Amendola & van Steensel, 2015; Snyder et al., 2016; Van Steensel & Belmont, 2017). Accordingly, genomic regions with heterochromatic properties are often localized to the nuclear periphery and euchromatin regions are found to be primarily located in the nuclear interior. Moreover, following that trend, when a region switches from a euchromatic to a heterochromatic state, it is more likely to assume a more peripheral localization. However, if we set aside the global trends and look systematically along the genome, at significant numbers of genomic regions, switches in either category do not follow the changes in the other. By looking into those different types of incongruent regions, here, we uncovered additional observations about the relationship between genome compartmentalization, lamina association and gene expression. For example, the genomic regions that do not alter their compartment identity between cell types exhibit gene expression changes concordant with their lamina association. Further, this effect is strongest when the lamina association changes in a direction concordant with the pre-existing compartment identity, suggesting that lamina association can reinforce pre-existing chromatin state. In further agreement with this idea, we found a relationship between lamin association strength and compartment identity strength such that, for example, increased lamina association associates with increased compartment identity strength for B compartment regions.

When compartment identity and final lamin association are discordant, some cell types show a strong association between lamin association and gene expression while others do not. This echoes previous observations that radial positioning can be uncoupled from gene expression in some cases (Brunet et al., 2019; Forsberg et al., 2019; Robson et al., 2017; Zuleger et al., 2013). Specifically, in the case of the mesenchymal lineage cells, compartment identity is found to be more strongly associated with gene expression regardless of lamin state. For other cell types, the changes in lamina association of those incongruent regions show a stronger association with gene expression. It is possible that in the case of the mesenchymal lineage cells, the genome compartmentalization is relatively decoupled from the lamina association compared to other cell types, as a result of the nuclear geometry in this cell lineage. It can be speculated that due to the flat ellipsoidal geometry of the mesenchymal nuclei, a higher proportion of the genome must be positioned near the periphery regardless of chromatin or compartment state. A nuclear geometry effect may also influence the relationships we observe for the RPE cell type, which behaves more like a mesenchymal cell type in our analyses despite its classification as an epithelial cell. This cell type is often observed to have a small and segmented nucleus, which may again lead to more and different chromosome regions being brought close to the nuclear periphery by geometric constraints (Pollreisz et al., 2020; Starnes et al., 2016).

When considering these results, it is important to keep in mind that the relationships we observe do not prove the directionality of the relationship between these changes. For example, while it could be that a shift of a B compartment region away from the nuclear

lamina enables the gene upregulation that we observe in a given cell state transition (Chen et al., 2014; Kohwi et al., 2013), it could instead be that the genes in that region were activated first and this led to relocation away from the nuclear lamina, as has been documented in some gene activation experiments (Brueckner et al., 2020). Or, both factors could work in tandem to produce the changes between cell types: the activation of some genes relocates chromatin away from the lamina, which in turn makes the activation of other genes more likely.

Our analyses of gene expression response to physicochemical treatments provide evidence that pre-existing lamina association and compartmentalization state influence the ability to alter gene expression in a region. But, while nuclear lamina contacts are often correlated with gene repression, we found that both induction and repression of genes was generally more likely in genomic regions that were located in the A compartment and away from the nuclear lamina. However, specifically in case of heat shock treatment, peripheral A compartment regions exhibit a higher level of downregulation compared to other regions. This may represent regions that are actively transcribed and accessible under normal conditions (thus in the A compartment), but which are poised to be shut down by lamina associated factors in the case of a stimulus.

In our results, while relationships between compartmentalization, lamin association, and gene expression tended to follow principles that often applied across different cell type comparisons, shifts in histone modifications were less consistent between cell types. Some cell lineages showed active and inactive histone modifications changing in concert with lamin association and dissociation while others did not. This may reflect the different relationships that can exist between chromatin modifications and lamin contacts in different regions and contexts. The nuclear lamina can promote alterations in chromatin modifications but also alterations in chromatin state can influence the likelihood of lamina tethering6. Since several of our cell lines are cancers derived from the lineages we analyze, it is also important to note that the differences in histone modification relationships we observe may relate to the dysregulation of epigenetic factors common to many cancers (Clapier et al., 2017; Dawson & Kouzarides, 2012; Mitchener & Muir, 2022). It appears that regardless of divergence of particular combinations of histone marks that are involved in compartment and lamin association status in a given cell type, the resulting spatial organization state produces consistent gene regulatory principles.

Methods

Cell culture

Prostate adenocarcinoma (LNCaP), and bone metastatic (VCaP) cell lines were acquired from the Physical Sciences Oncology Network Bioresource Core Facility, through ATCC (Manassas, VA). These cell lines were cultured according to standard protocols, with subculturing at 80% confluence. LNCaP media was made of RPMI (Gibco;11835030), supplemented to match the ATCC formula (4.5 g/L glucose, 2.383 g/L HEPES, 0.11 g/L

sodium pyruvate, 10% FBS). VCaP were cultured in DMEM F12: Ham 1:1 (Gibco; 11-320-033) containing 10% FBS. Media for all cell lines was supplemented with 100 mg/ml penicillin streptomycin (Gibco; 15-140-122).

Prostate cancer cell line Lamin B1 ChIP-seq

LNCaP and VCaP cells growing under continuous culture conditions in T75 flasks were allowed to reach 90% confluency. LNCaP cells were grown in RPMI (Gibco; 11835030) supplemented with 4.5 g/liter glucose, 2.383 g/liter HEPES, 0.11 g/liter sodium pyruvate and 10% FBS (Corning; 35–010-CV). VCaP cells were cultured in DMEM F12: Ham 1:1 (Gibco; 11–320-033) with 10% FBS. After media removal, the cell monolayer was washed once with HBSS, which was promptly removed by aspiration. Cells were fixed in-situ by incubation in 1% formaldehyde (37% formaldehyde diluted in HBSS) for 10 minutes, under gentle shaking. Crosslinking was quenched by adding glycine to a final concentration of 0.14 M (MP biomedicals ICN19482591), followed by a 5 min incubation at room temperature, with shaking. Flasks were then cooled down on ice for 15 min, after which the formaldehyde solution was aspirated from the plate and substituted with 10 ml of ice-cold HBSS, supplemented with 1X Halt Protease Inhibitor cocktail (Thermo PI78438). Cells were then collected by centrifugation and snap frozen in liquid nitrogen. A replicate flask, cultured under identical conditions, was trypsinized and cells counted, using trypan blue exclusion, to estimate the number of cells per pellet.

Cell pellets were lysed and chromatin was fragmented using a Covaris M220 Sonicator, and the Covaris truChIP Chromatin shearing kit (Covaris; 520154), according to the manufacturers' instructions, using 1 ml AFA tubes (Covaris). Briefly, cell pellets were thawed on ice and resuspended in 1 ml of 1X lysis buffer, supplemented with protease inhibitors. Samples were incubated on ice for 10 minutes, with gentle vortexing every 2 min. Nuclei were collected by centrifugation at 1700 g for 5 min (at 4 o C). The nuclei pellet was then resuspended in 1 ml of 1X wash buffer (supplemented with protease inhibitors) and collected by centrifugation as before. Samples were then washed twice, using centrifugation as before, with 1 ml of complete shearing buffer. Pellets were resuspended in 1 ml of fresh shearing buffer and transferred into a cold 1 ml AFA tube (Covaris). Sonication conditions were as follows: PIP 75, duty factor 10%, CBP 200, 4 o C, 12 minutes.

To verify sonication, aliquots of time-zero and sonicated samples were treated with RNase A (30 min at 37 o C) and reverse crosslinked overnight with proteinase K via incubation at 65 o C. After reverse crosslinking, DNA was purified using QIAquick PCR Purification Kit (Qiagen; 28104), according to the manufacturer's instructions. Shearing efficiency was verified via agarose gel electrophoresis.

Chromatin immunoprecipitation was done using the MAGnify Chromatin immunoprecipitation system (Life Technologies, 49-2024), according to the

manufacturer's protocol, diluting the sheared chromatin at a 1:3 ratio in the kit's dilution buffer and using three reactions per sample. Three micrograms of rabbit polyclonal antibody against Lamin B1 were used, per reaction (Abcam; ab16048). After pulldown and purification, DNA fragments were sonicated for a second time, using the following parameters in a Covaris M220 sonicator: 130 microliter AFA tube, 10 o C, PIP 50%, duty factor 20%, CPB 200, for 45 seconds.

Library preparation and sequencing services were provided by Azenta/Genewiz (South Plainfield, NJ).

DamID-seq data processing

DamID-seq data were processed as reported in Leemans et al. 2019, except the trimming step (Leemans et al., 2019). Briefly, at first, bwa mem tool (https://github.com/lh3/bwa) was used to map gDNA reads starting with GATC to a combination of hg19 reference genome with a ribosomal model. For further processing, only the mapped reads having a mapping quality of at least 10 are considered as GATC fragments. Next, the reads were combined into the bins of 40 kb resolution depending on the middle of the GATC fragments and then scaled to 1M reads. For normalization, log2-ratio of the scaled target over the scaled Dam-only bins was calculated with a pseudo count of 1. Furthermore, a Hidden Markov Model was used to call LADs from the normalized values using HMMt R package (https://github.com/gui11aume/HMMt).

Hi-C data processing

In this study, the HiC-Pro pipeline (https://github.com/nservant/HiC-Pro) was used with default parameters to process all the Hi-C sequencing data. The reads were aligned to hg19 human reference genome (Servant et al., 2015). To improve the quality of the contact data, reads from all the replicates of same cell-type or condition were combined together and used for further downstream processing. For all the Hi-C datasets, for each individual chromosome, the contacts were binned at 250 kb resolution and represented in a symmetric square matrix format. Each of the matrices was then further analyzed to obtain the compartmental identity of the chromosome structures at 250 kb resolution using the 'matrix2compartment.pl' script from the cworld-dekker (https://github.com/dekkerlab/cworld-dekker) tool suite. Briefly, principal component analysis was performed on the distance-decay normalized contact matrix and A/B compartment identities were assigned based on the sign of the eigenvectors (positive – A and negative – B).

RNA-seq data processing

For all the RNA-seq data, the fastq reads were first processed for adapter trimming and then for quality trimming using the BBDuk tool (https://github.com/kbaseapps/BBTools) (Bushnell et al., 2017). The parameters used for those adapter and quality trimming steps were "ktrim=r k=23 mink=11 hdist=1" and "qtrim=r trimq=28" respectively. By performing that quality trimming step, reads with quality score lower than 28 were discarded. After having the adapter trimmed and low-quality reads removed, the remaining reads were aligned to human reference genome hg19 with the help of STAR aligner (https://github.com/alexdobin/STAR) and for that step both the '--outFilterScoreMinOverLread' and ' –outFilterMatchNminOverLread' parameters values set to 0.2 (Dobin et al., 2013). Finally, the mapped aligned reads were sorted based on their genomic coordinates and then gene level feature count step was performed on those reads using HTSeq-Counts (https://github.com/simon-anders/htseq) (Anders et al., 2015).

Batch effect removal

The RNA-seq files used in this study were generated at different laboratories using various types of transcriptome profiling techniques. As a result of that the raw gene counts calculated by HTSeq-Counts tool suffer from batch effect, where the samples segregate based on their origin lab and experiment types without showing underlying true biological signal, when analyzed further for downstream processing. To uncover the biological signal, the ComBat-seq tool (https://github.com/zhangyuqing/ComBat-seq) was applied on the raw counts for the batch effect adjustments (Zhang et al., 2020). Un this step, the files originated from the same laboratory were assigned same batch number.

Differential expression analysis

Differential gene expression analysis was performed between different condition samples of the same cell type (for example, before and after AR7 induction in LNCaP and VCaP cells) using DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) (Love et al., 2014). To identify significantly differentially regulated genes, we use an adjusted p-value cutoff of 0.01.

Histone ChIP-seq data processing

BBDuk tool (https://github.com/kbaseapps/BBTools) was used to process all the histone ChIP-seq data for adapter removal and quality trimming (Bushnell et al., 2017). After performing those steps, the reads were mapped to hg19 human reference genome using STAR aligner (https://github.com/alexdobin/STAR) (Dobin et al., 2013). Finally, the peaks were called from the aligned data using MACS2 (https://pypi.org/project/MACS2/) peak

calling software (Zhang et al., 2008). Since, in this study, we are mainly focusing on broad histone marks, here we used MACS2 broad peak calling parameters "-p 1e-2 --broad --nomodel --shift 0 --keep-dup all -B --SPMR". In addition to all those parameters, we also supplied the estimated fragment length obtained from the ChIP-seq data as "--extsize" parameter. To calculate the estimated fragment length, 'run_spp.R' script from the phandompeakqualtools (https://github.com/kundajelab/phantompeakqualtools) was used (Kharchenko et al., 2008; Landt et al., 2012). Once we had the broad peak locations for a particular histone mark, the enrichment of that mark for different genomic regions was measured by calculating the base pair overlaps using the 'intersect' function of bedtools (https://bedtools.readthedocs.io/en/latest/index.html) (Quinlan & Hall, 2010).

LaminB1 ChIP-seq data processing

LaminB1 ChIP-seq files were processed first using the BBDuk tool (https://github.com/kbaseapps/BBTools) for adapter removal and quality trimming (Bushnell et al., 2017). After that, STAR aligner (https://github.com/alexdobin/STAR) was used to map the reads to the hg19 genome (Dobin et al., 2013). Once, both the target and input mapped reads were mapped, they were binned at 40 kb resolution and the log2-ratio of the target over the input bins was calculated using 'bamCompare' function of deepTools (https://deeptools.readthedocs.io/) with parameters "--operation log2 -bs 40000 --ignoreDuplicates --minMappingQuality 30 --scaleFactorsMethod SES --effectiveGenomeSize 2864785220" (Ramírez et al., 2014). The signal is then further smoothed using a rolling window averaging (120 kb) followed by LAD prediction by hidden markov model as described in the DamID-seq data processing.

Hierarchical clustering of genome-wide lamina association and compartmental organization data

For both types of data, we performed hierarchical clustering to cluster the cell types based on similarity of the respective data type. To do that, we first calculated the pairwise similarity between all the 11 cell types using Pearson's correlation. We further converted the similarity data into distance data by subtracting the values from 1. Following that, finally we used the distance data for hierarchical clustering using 'ward' linkage.

Combine lamina association and compartmental organization data

We combined the LAD and compartment identity data together for each cell type. For that, we used the 'intersect' function of bedtools (https://bedtools.readthedocs.io/en/latest/index.html) to combine genome-wide LAD and compartment identity data together with a minimum 50% overlap as a fraction of LAD data (Quinlan & Hall, 2010). This allowed us to categorize the genomic regions based on

whether they belong to LAD or not and also simultaneously whether those regions are part of A or B compartment.

Acknowledgements

The authors would like to thank all the members of McCord lab for their fruitful suggestions in developing the research project.

Disclosure statement

The authors declare no competing interests exist.

Data availability

Datasets used for this project were obtained from NCBI GEO, EMBL ENA, 4DN Data Portal and ENCODE repositories and their accession IDs are given in Appendix: Table 5 and 6. Newly generated LaminB1 ChIP data for VCaP and LNCaP are available at GEO with accession number GSE213032.

Author contributions

P.D. conceived the study and performed data analysis and prepared figures. R.S.M. performed the Lamin B1 ChIP experiments and assigned cell types to lineages. R.P.M. supervised the project, provided funding, and prepared figures. P.D. and R.P.M. wrote the manuscript with input from R.S.M.

Supplementary material

The supplementary material associated with the paper can be found and accessed online.

Funding

This work was supported by the National Institutes of Health NIGMS grant R35GM133557 to R.P.M and American Cancer Society postdoctoral fellowship 134060-PF-19-183-01-CSM to R.S.M.

REFERENCES

Amendola, M., & van Steensel, B. (2015). Nuclear lamins are not required for lamina-associated domain organization in mouse embryonic stem cells. *EMBO reports*, *16*(5), 610-617.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169.

Basil, P., Robertson, M. J., Bingman, W. E., Dash, A. K., Krause, W. C., Shafi, A. A., Piyarathna, B., Coarfa, C., & Weigel, N. L. (2022). Cistrome and transcriptome analysis identifies unique androgen receptor (AR) and AR-V7 splice variant chromatin binding and transcriptional activities. *Scientific reports*, *12*(1), 1-18.

Bersaglieri, C., Kresoja-Rakic, J., Gupta, S., Bär, D., Kuzyakiv, R., & Santoro, R. (2020). Genome-wide maps of nucleolus interactions reveal distinct layers of repressive chromatin domains. *bioRxiv*, 2020.2011.2017.386797. https://doi.org/10.1101/2020.11.17.386797

Bonev, B., Cohen, N. M., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., & Tanay, A. (2017). Multiscale 3D genome rewiring during mouse neural development. *Cell*, *171*(3), 557-572. e524.

Brattain, M. G., Fine, W. D., Khaled, F. M., Thompson, J., & Brattain, D. E. (1981). Heterogeneity of malignant cells from a human colonic carcinoma. *Cancer research*, *41*(5), 1751-1756.

Brueckner, L., Zhao, P. A., van Schaik, T., Leemans, C., Sima, J., Peric-Hupkes, D., Gilbert, D. M., & van Steensel, B. (2020). Local rewiring of genome–nuclear lamina interactions by transcription. *The EMBO journal*, *39*(6), e103159.

Brunet, A., Forsberg, F., Fan, Q., Sæther, T., & Collas, P. (2019). Nuclear lamin B1 interactions with chromatin during the circadian cycle are uncoupled from periodic gene expression. *Frontiers in genetics*, *10*, 917.

Buchwalter, A., Kaneshiro, J. M., & Hetzer, M. W. (2019). Coaching from the sidelines: the nuclear periphery in genome regulation. *Nature Reviews Genetics*, *20*(1), 39-50.

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge–accurate paired shotgun read merging via overlap. *PLoS One*, *12*(10), e0185056.

Caplan, A. I. (2005). Mesenchymal stem cells: cell–based reconstructive therapy in orthopedics. *Tissue engineering*, *11*(7-8), 1198-1211.

Chapter 13 - Stem Cells and Regenerative Medicine. (2021). In S. R. Goodman (Ed.), *Goodman's Medical Cell Biology (Fourth Edition)* (pp. 361-380). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-817927-7.00013-2

Chen, H., Zheng, X., & Zheng, Y. (2014). Age-associated loss of lamin-B leads to systemic inflammation and gut hyperplasia. *Cell*, *159*(4), 829-843.

Cho, S., Vashisth, M., Abbas, A., Majkut, S., Vogel, K., Xia, Y., Ivanovska, I. L., Irianto, J., Tewari, M., & Zhu, K. (2019). Mechanosensing by the lamina protects against nuclear rupture, DNA damage, and cell-cycle arrest. *Developmental cell*, *49*(6), 920-935. e925.

Clapier, C. R., Iwasa, J., Cairns, B. R., & Peterson, C. L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology*, *18*(7), 407-422.

Cuomo, A. S., Seaton, D. D., McCarthy, D. J., Martinez, I., Bonder, M. J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., & Buettner, F. (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature communications*, *11*(1), 1-14.

Dawson, M. A., & Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. *Cell*, *150*(1), 12-27.

Dechat, T., Pfleghaar, K., Sengupta, K., Shimi, T., Shumaker, D. K., Solimando, L., & Goldman, R. D. (2008). Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. *Genes & development*, *22*(7), 832-853.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., & Xie, W. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331-336.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B. R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., & Solovei, I. (2019). Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, *570*(7761), 395-399.

Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J. R., & Bickmore, W. A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS genetics*, *4*(3), e1000039.

Fischer, T., Hayn, A., & Mierke, C. T. (2020). Effect of nuclear stiffness on cell mechanics and migration of human breast cancer cells. *Frontiers in cell and developmental biology*, *8*, 393.

Forsberg, F., Brunet, A., Ali, T. M. L., & Collas, P. (2019). Interplay of lamin A and lamin B LADs on the radial positioning of chromatin. *Nucleus*, *10*(1), 7-20.

Frank, C. L., Manandhar, D., Gordân, R., & Crawford, G. E. (2016). HDAC inhibitors cause site-specific chromatin remodeling at PU. 1-bound enhancers in K562 cells. *Epigenetics & chromatin*, *9*(1), 1-17.

Grancharova, T., Gerbin, K. A., Rosenberg, A. B., Roco, C. M., Arakaki, J. E., DeLizo, C. M., Dinh, S. Q., Donovan-Maiye, R. M., Hirano, M., & Nelson, A. M. (2021). A comprehensive analysis of gene expression changes in a high replicate and open-source dataset of differentiating hiPSC-derived cardiomyocytes. *Scientific reports*, *11*(1), 1-21.

Gupta, S., & Santoro, R. (2020). Regulation and roles of the nucleolus in embryonic stem cells: from ribosome biogenesis to genome organization. *Stem Cell Reports*, *15*(6), 1206-1219.

Harr, J. C., Luperchio, T. R., Wong, X., Cohen, E., Wheelan, S. J., & Reddy, K. L. (2015). Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins. *Journal of Cell Biology*, *208*(1), 33-52.

Hayflick, L. (1965). The limited in vitro lifetime of human diploid cell strains. *Experimental cell research*, *37*(3), 614-636.

Honda, A., Hoeksema, M. A., Sakai, M., Lund, S. J., Lakhdari, O., Butcher, L. D., Rambaldo, T. C., Sekiya, N. M., Nasamran, C. A., & Fisch, K. M. (2022). The Lung Microenvironment Instructs Gene Transcription in Neonatal and Adult Alveolar Macrophages. *The Journal of Immunology*, *208*(8), 1947-1959.

Keller, G. (2005). Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes & development*, *19*(10), 1129-1155.

Keough, K. C., Shah, P. P., Gjoni, K., Santini, G. T., Wickramasinghe, N. M., Dundes, C. E., Karnay, A., Chen, A., Salomon, R. E. A., Walsh, P. J., Nguyen, S. C., Whalen, S., Joyce, E. F., Loh, K. M., Dubois, N., Pollard, K. S., & Jain, R. (2021). An atlas of lamina-associated chromatin across twelve human cell types reveals an intermediate chromatin subtype. *bioRxiv*, 2020.2007.2023.218768. https://doi.org/10.1101/2020.07.23.218768

Kharchenko, P. V., Tolstorukov, M. Y., & Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, *26*(12), 1351-1359.

Kohwi, M., Lupton, J. R., Lai, S.-L., Miller, M. R., & Doe, C. Q. (2013). Developmentally regulated subnuclear genome reorganization restricts neural progenitor competence in Drosophila. *Cell*, *152*(1-2), 97-108.

Kotecki, M., Reddy, P. S., & Cochran, B. H. (1999). Isolation and characterization of a near-haploid human cell line. *Experimental cell research*, *252*(2), 273-280.

Kourtis, N., Moubarak, R. S., Aranda-Orgilles, B., Lui, K., Aydin, I. T., Trimarchi, T., Darvishian, F., Salvaggio, C., Zhong, J., & Bhatt, K. (2015). FBXW7 modulates cellular stress response and metastatic potential through HSF1 post-translational modification. *Nature cell biology*, *17*(3), 322-332.

Lammerding, J., Fong, L. G., Ji, J. Y., Reue, K., Stewart, C. L., Young, S. G., & Lee, R. T. (2006). Lamins A and C but not lamin B1 regulate nuclear mechanics. *Journal of Biological Chemistry*, *281*(35), 25768-25780.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., & Cayting, P. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, *22*(9), 1813-1831.

Leemans, C., van der Zwalm, M. C., Brueckner, L., Comoglio, F., van Schaik, T., Pagie, L., van Arensbergen, J., & van Steensel, B. (2019). Promoter-intrinsic and local chromatin features determine gene repression in LADs. *Cell*, *177*(4), 852-864. e814.

Lieberman-Aiden, E., Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., & Dorschner, M. O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*. https://doi.org/10.1126/science.1181369

Lochs, S. J., Kefalopoulou, S., & Kind, J. (2019). Lamina associated domains and gene regulation in development and cancer. *Cells*, *8*(3), 271.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21.
Lozzio, C. B., & Lozzio, B. B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome.
Matsuo, M., Kaji, K., Utakoji, T., & Hosoda, K. (1982). Ploidy of human embryonic fibroblasts during in vitro aging. *Journal of gerontology*, *37*(1), 33-37.
McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Molecular cell*, *77*(4), 688-708.
Mitchener, M. M., & Muir, T. W. (2022). Oncohistones: Exposing the nuances and vulnerabilities of epigenetic regulation. *Molecular cell*, *82*(16), 2925-2938.
Miura, H., Takahashi, S., Poonperm, R., Tanigawa, A., Takebayashi, S.-i., & Hiratani, I. (2019). Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nature genetics*, *51*(9), 1356-1368.
Morimoto, M., & Boerkoel, C. F. (2013). The role of nuclear bodies in gene expression and disease. *Biology*, *2*(3), 976-1033.
Nichols, W., Murphy, D., Cristofalo, V., Toji, L., Greene, A., & Dwight, S. (1977). Characterization of a new human diploid cell strain, IMR-90. *Science*, *196*(4285), 60-63.
Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R. M., & van Lohuizen, M. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, *38*(4), 603-613.
Peterson, L. W., & Artis, D. (2014). Intestinal epithelial cells: regulators of barrier function and immune homeostasis. *Nature Reviews Immunology*, *14*(3), 141-153.
Poleshko, A., Smith, C. L., Nguyen, S. C., Sivaramakrishnan, P., Wong, K. G., Murray, J. I., Lakadamyali, M., Joyce, E. F., Jain, R., & Epstein, J. A. (2019). H3K9me2 orchestrates inheritance of spatial positioning of peripheral heterochromatin through mitosis. *ELife*, *8*, e49278.
Pollreisz, A., Neschi, M., Sloan, K. R., Pircher, M., Mittermueller, T., Dacey, D. M., Schmidt-Erfurth, U., & Curcio, C. A. (2020). Atlas of human retinal pigment epithelium organelles significant for clinical imaging. *Investigative ophthalmology & visual science*, *61*(8), 13-13.
Ponten, J., & Saksela, E. (1967). Two established in vitro cell lines from human mesenchymal tumours. *International journal of cancer*, *2*(5), 434-447.
Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842.
Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, *42*(W1), W187-W191.
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. (2014). A 3D map of

the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665-1680.

Robson, M. I., Jose, I., Czapiewski, R., Lê Thành, P., Booth, D. G., Kelly, D. A., Webb, S., Kerr, A. R., & Schirmer, E. C. (2016). Tissue-specific gene repositioning by muscle nuclear membrane proteins enhances repression of critical developmental genes during myogenesis. *Molecular cell*, *62*(6), 834-847.

Robson, M. I., Jose, I., Czapiewski, R., Sivakumar, A., Kerr, A. R., & Schirmer, E. C. (2017). Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome research*, *27*(7), 1126-1138.

Rønningen, T., Shah, A., Oldenburg, A. R., Vekterud, K., Delbarre, E., Moskaug, J. Ø., & Collas, P. (2015). Prepatterning of differentiation-driven nuclear lamin A/C-associated chromatin domains by GlcNAcylated histone H2B. *Genome research*, *25*(12), 1825-1835.

San Martin, R., Das, P., Dos Reis Marques, R., Xu, Y., Roberts, J. M., Sanders, J. T., Golloshi, R., & McCord, R. P. (2021). Chromosome compartmentalization alterations in prostate cancer cell lines model disease progression. *Journal of Cell Biology*, *221*(2), e202104108.

Schwenk, H.-U., & Schneider, U. (1975). Cell cycle dependency of a T-cell marker on lymphoblasts. *Blut: Zeitschrift für die Gesamte Blutforschung*, *31*(5), 299-306.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology*, *16*(1), 1-11.

Smith, C. L., Lan, Y., Jain, R., Epstein, J. A., & Poleshko, A. (2021). Global chromatin relabeling accompanies spatial inversion of chromatin in rod photoreceptors. *Science advances*, *7*(39), eabj3035.

Snyder, M. J., Lau, A. C., Brouhard, E. A., Davis, M. B., Jiang, J., Sifuentes, M. H., & Csankovszki, G. (2016). Anchoring of heterochromatin to the nuclear lamina reinforces dosage compensation-mediated gene repression. *PLoS genetics*, *12*(9), e1006341.

Solovei, I., Wang, A. S., Thanisch, K., Schmidt, C. S., Krebs, S., Zwerger, M., Cohen, T. V., Devys, D., Foisner, R., & Peichl, L. (2013). LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell*, *152*(3), 584-598.

Srivastava, L. K., Ju, Z., Ghagre, A., & Ehrlicher, A. J. (2021). Spatial distribution of lamin A/C determines nuclear stiffness and stress-mediated deformation. *Journal of cell science*, *134*(10), jcs248559.

Starnes, A. C., Huisingh, C., MCGWIN, G., Sloan, K. R., Ablonczy, Z., Smith, R. T., Curcio, C. A., & Ach, T. (2016). Multi-nucleate retinal pigment epithelium cells of the human macula exhibit a characteristic and highly specific distribution. *Visual neuroscience*, *33*.

Stephens, A. D., Banigan, E. J., & Marko, J. F. (2018). Separate roles for chromatin and lamins in nuclear mechanics. *Nucleus*, *9*(1), 119-124.

Stik, G., Vidal, E., Barrero, M., Cuartero, S., Vila-Casadesús, M., Mendieta-Esteban, J., Tian, T. V., Choi, J., Berenguer, C., & Abad, A. (2020). CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. *Nature genetics*, *52*(7), 655-661.

Swift, J., Ivanovska, I. L., Buxboim, A., Harada, T., Dingal, P. D. P., Pinter, J., Pajerowski, J. D., Spinler, K. R., Shin, J.-W., & Tewari, M. (2013). Nuclear lamin-A scales with tissue stiffness and enhances matrix-directed differentiation. *Science*, *341*(6149), 1240104.

Vahabikashi, A., Adam, S. A., Medalia, O., & Goldman, R. D. (2022). Nuclear lamins: Structure and function in mechanobiology. *APL bioengineering*, *6*(1), 011503.

Van Steensel, B., & Belmont, A. S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, *169*(5), 780-791.

Yang, S., Zhou, J., & Li, D. (2021). Functions and diseases of the retinal pigment epithelium. *Frontiers in Pharmacology*, 1976.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., & Li, W. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, *9*(9), 1-9.

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR genomics and bioinformatics*, *2*(3), lqaa078.

Zuleger, N., Boyle, S., Kelly, D. A., de Las Heras, J. I., Lazou, V., Korfali, N., Batrakou, D. G., Randles, K. N., Morris, G. E., & Harrison, D. J. (2013). Specific nuclear envelope transmembrane proteins can promote the location of chromosomes to and from the nuclear periphery. *Genome biology*, *14*(2), 1-20.

Zullo, J. M., Demarco, I. A., Piqué-Regi, R., Gaffney, D. J., Epstein, C. B., Spooner, C. J., Luperchio, T. R., Bernstein, B. E., Pritchard, J. K., & Reddy, K. L. (2012). DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell*, *149*(7), 1474-1487.

APPENDIX

Table 5 NCBI GEO, EMBL ENA, 4DN Data Portal and ENCODE accession IDs for the Hi-C, DamID-seq, RNA-seq and ChIP-seq data from thirteen different cell-types. RC = replicates combined for analysis, R1 = replicate 1 used for analysis. Specific conditions chosen from each accession number indicated where needed in parentheses.

Cell type	Hi-C	DamID-seq	RNA-seq
H1ESC	4DNESRJ8KV4Q (RC)	4DNESOFQR5FS (LaminB1; RC)	ENCSR588EJX (RC)
HMSC	GSE140782 (RC)	GSE146387 (Emerin; RC)	GSE108638 (RC)
IMR90	X	X	X
WI38	PRJEB8073 (Cycling; RC)	X	X
Tig3	X	GSE76594 (LaminB1; Cycling; RC)	GSE75643 (RC)
HFFc6	GSE165894 (RC)	4DNESXZ4FW4T (LaminB1; RC)	4DNESFH3EHTU (RC)
U2OS	GSE137469 (RC)	4DNESB2HMY7D (LaminB1; RC)	GSE131571 (RC)
HCT116	GSE133023 (RC)	4DNES24XA7U8 (LaminB1; RC)	GSE162160 (RC)
HT1080	GSE117582 (RC)	GSE87148 (LaminB1; RC)	ENCSR535VTR (RC)
RPE	ENCSR846DUS (RC)	4DNESHGTQ73M (LaminB1; RC)	GSE163315 (RC)
K562	ENCSR545YBD (RC)	4DNES6SM8UGO (LaminB1; RC)	ENCSR615EEK (RC)
HAP1	ENCSR390HMC (RC)	4DNESUK5H9Y8 (LaminB1; RC)	GSE160383 (RC)
Jurkat	GSE122958 (RC)	GSE94971 (LaminB1; Resting; RC)	GSE130140 (RC)

Table 6 NCBI GEO, EMBL ENA, 4DN Data Portal and ENCODE accession IDs for the ChIP-seq data from thirteen different cell-types. RC = replicates combined for analysis, R1 = replicate 1 used for analysis. Specific conditions chosen from each accession number indicated where needed in parentheses.

Cell type	H3K4me1 ChIP-seq	H3K27me3 ChIP-seq	H3K9me3 ChIP-seq
H1ESC	ENCSR991BTY (R1)	ENCSR687FDK (R1)	ENCSR395USV (R1)
HMSC	ENCSR065HOR (R1)	ENCSR832JVP (R1)	ENCSR746CUY (R1)
IMR90	GSE106146 (Proliferating; R1)	GSE135088 (Growing; R1)	GSE53332 (OIS_Proliferating; R1)
WI38	X	X	X
Tig3	X	X	X
HFFc6	X	X	X
U2OS	GSE73742 (R1)	GSE35573 (R1)	GSE69809 (R1)
HCT116	ENCSR161MXP (R1)	ENCSR810BDB (R1)	ENCSR179BUC (R1)
HT1080	X	GSE153869 (R1)	GSE135580 (R1)
RPE	GSE163315 (R1)	X	GSE163315 (R1)
K562	ENCSR000AKS (R1)	ENCSR000AKQ (R1)	ENCSR000APE (R1)
HAP1	ENCSR450JTP (R1)	ENCSR891YFI (R1)	ENCSR460TQV (R1)
Jurkat	GSE119439 (R1)	GSE100693 (R1)	GSE162605 (R1)

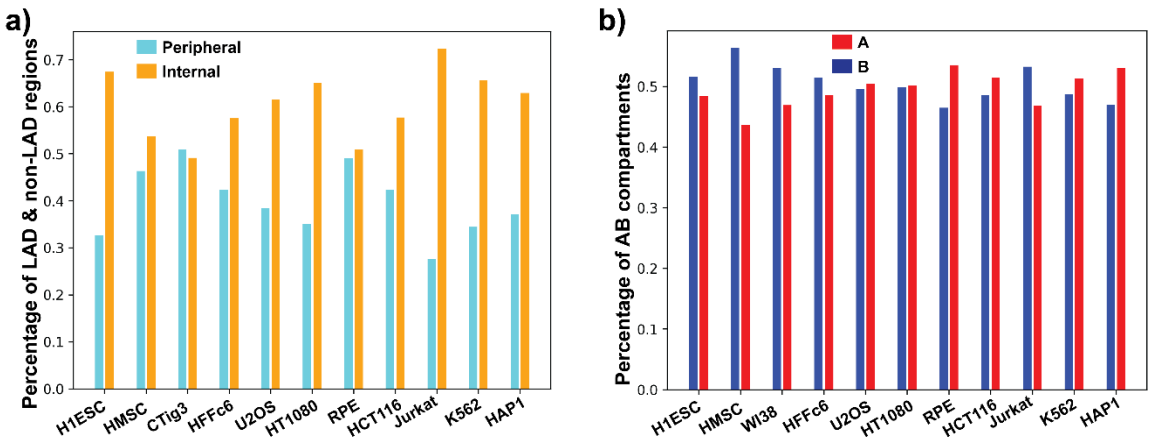


Figure 50 Proportion of LAD and compartment type for each cell type.

a) Percentages of peripheral and internal regions based on lamina association for different cell types. b) Percentages of A and B compartment regions based on genome organization data for different cell types.

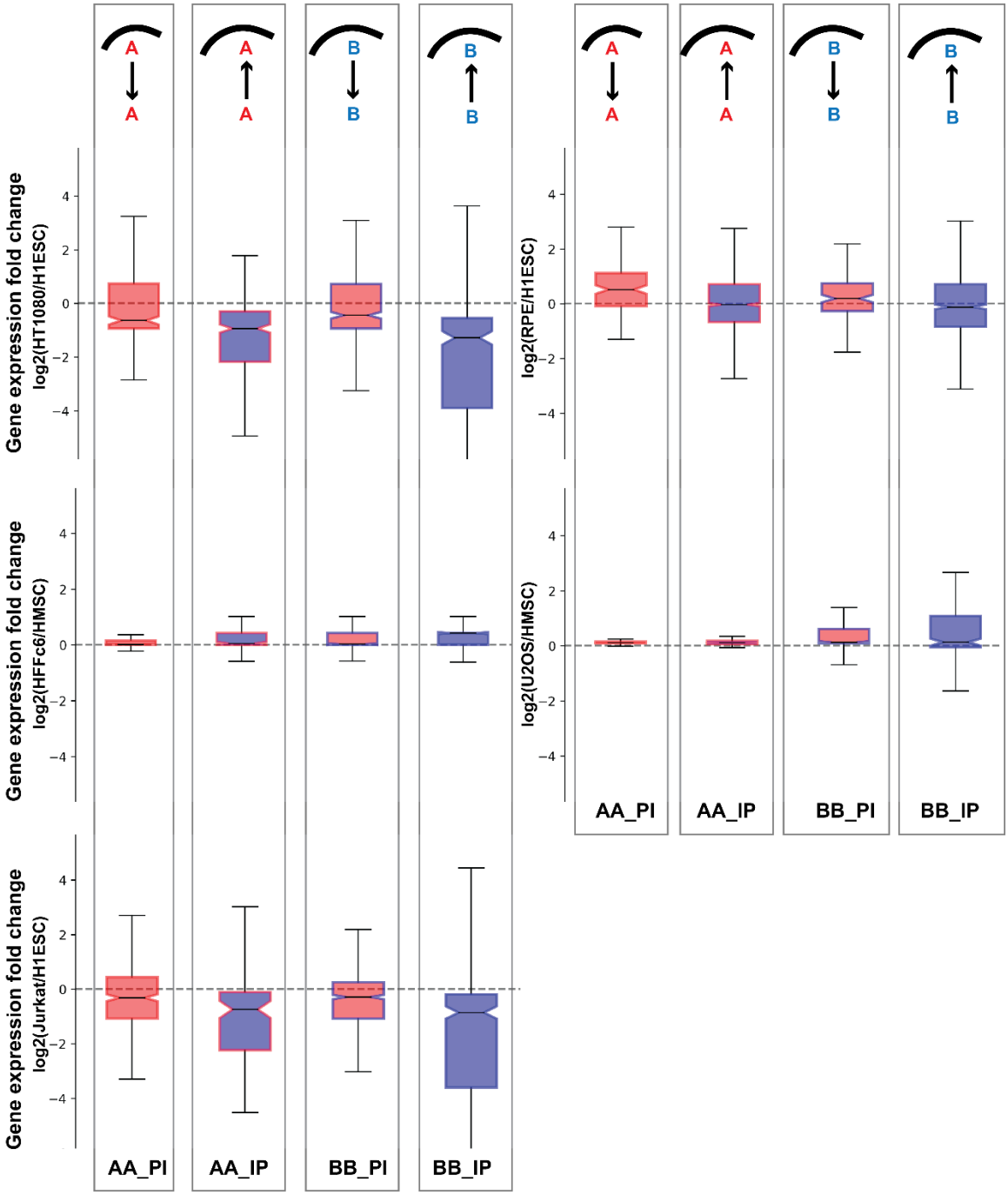


Figure 51 Gene expression changes with alteration in lamina association for similar compartment identity genomic regions.

Labels, coloring, and organization are as described in Fig. 45, for different cell type comparisons.

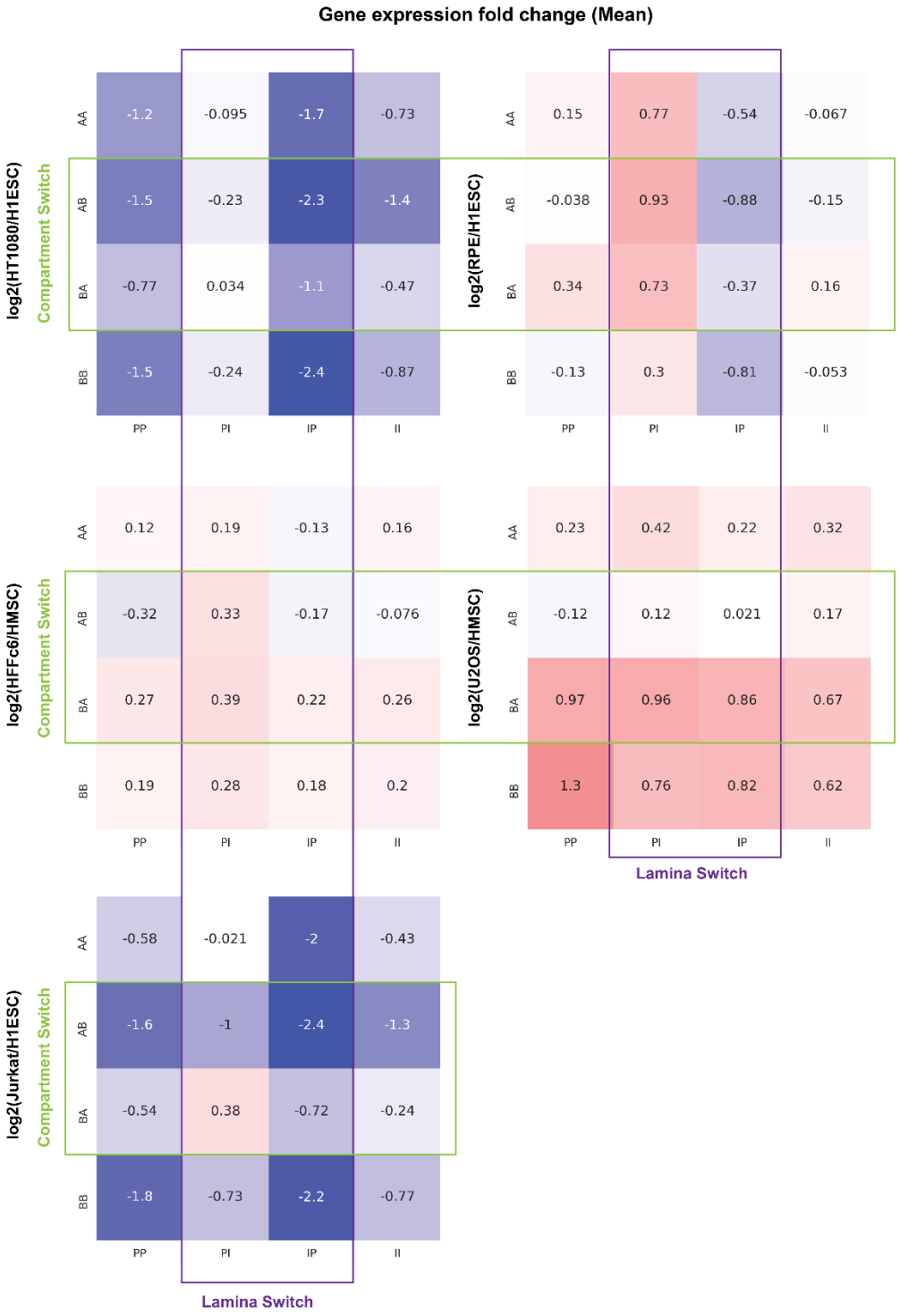


Figure 52 Mean gene expression fold change between cell types for different categories of genomic regions based on compartment identity and lamina association.

Labels, coloring, and organization are as described in Figure 46, for different cell type comparisons.

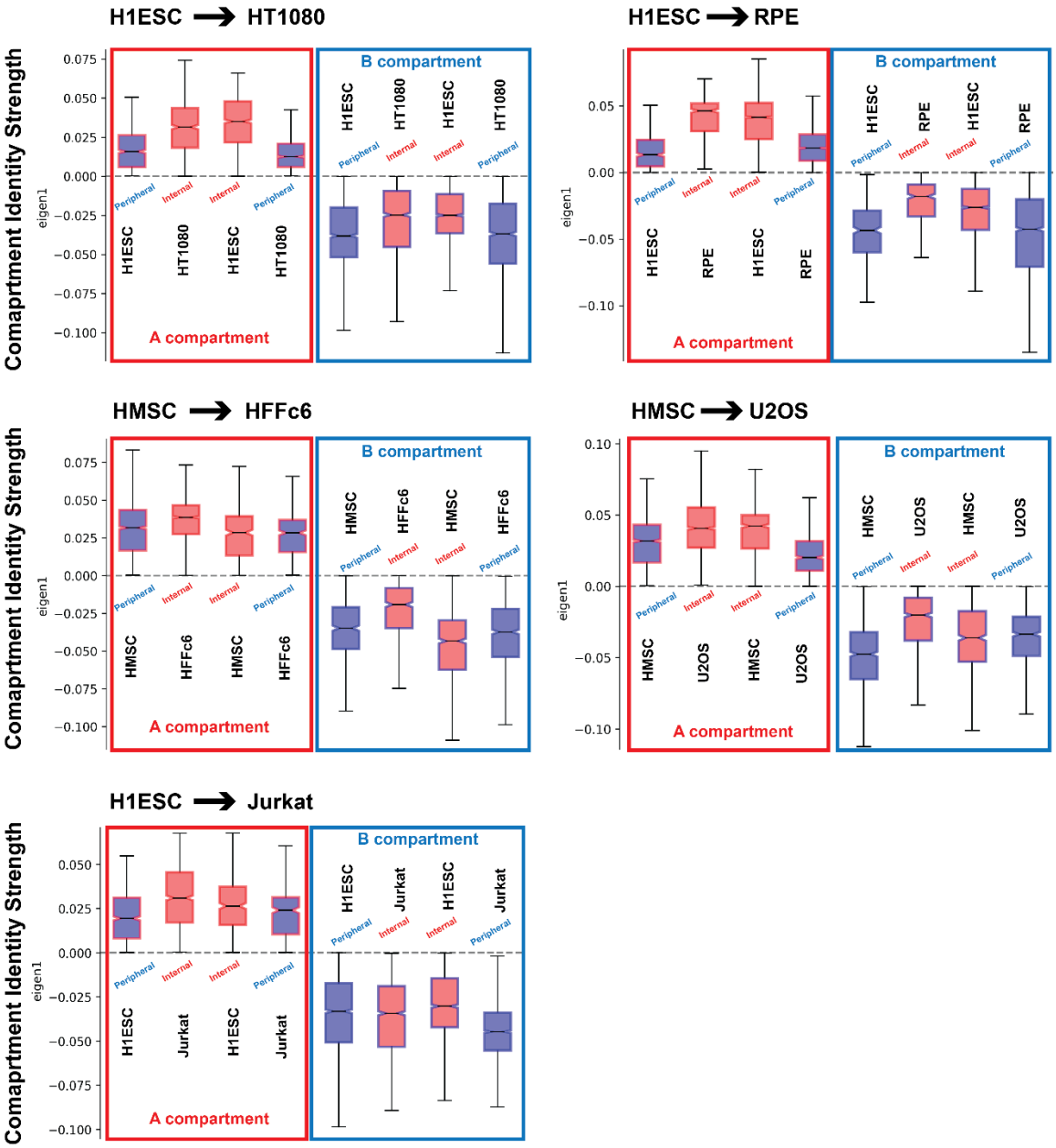


Figure 53 Compartment identity strength shifts with changes in LAD status even when compartment identity remains unchanged.

Labels, coloring, and organization are as described in Figure 47, for different cell type comparisons.

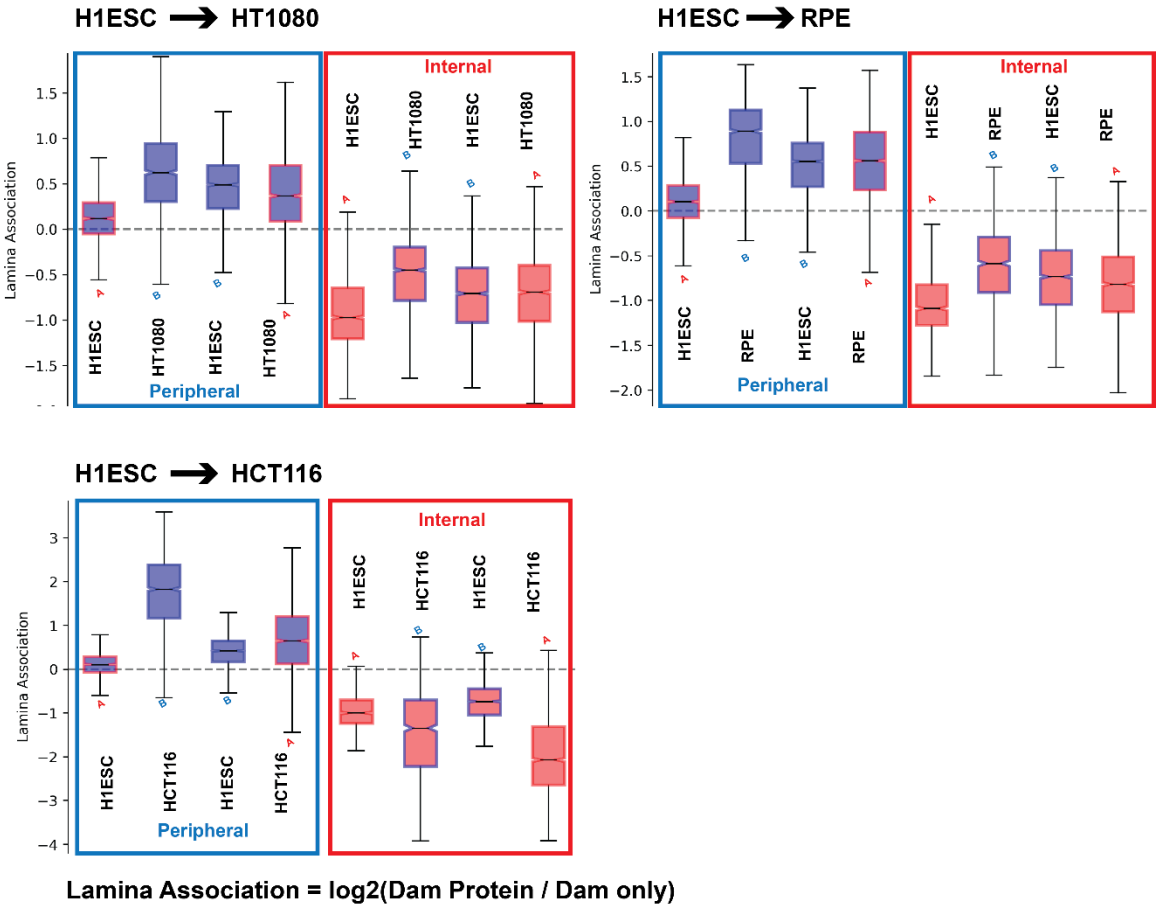


Figure 54 Lamina association strength shifts with changes in compartment status.

Cell type comparisons shown along the a) epithelial lineage, b) mesenchymal lineage, and c) hematopoietic lineage. Labels, coloring, and organization are as described in Fig. 47.

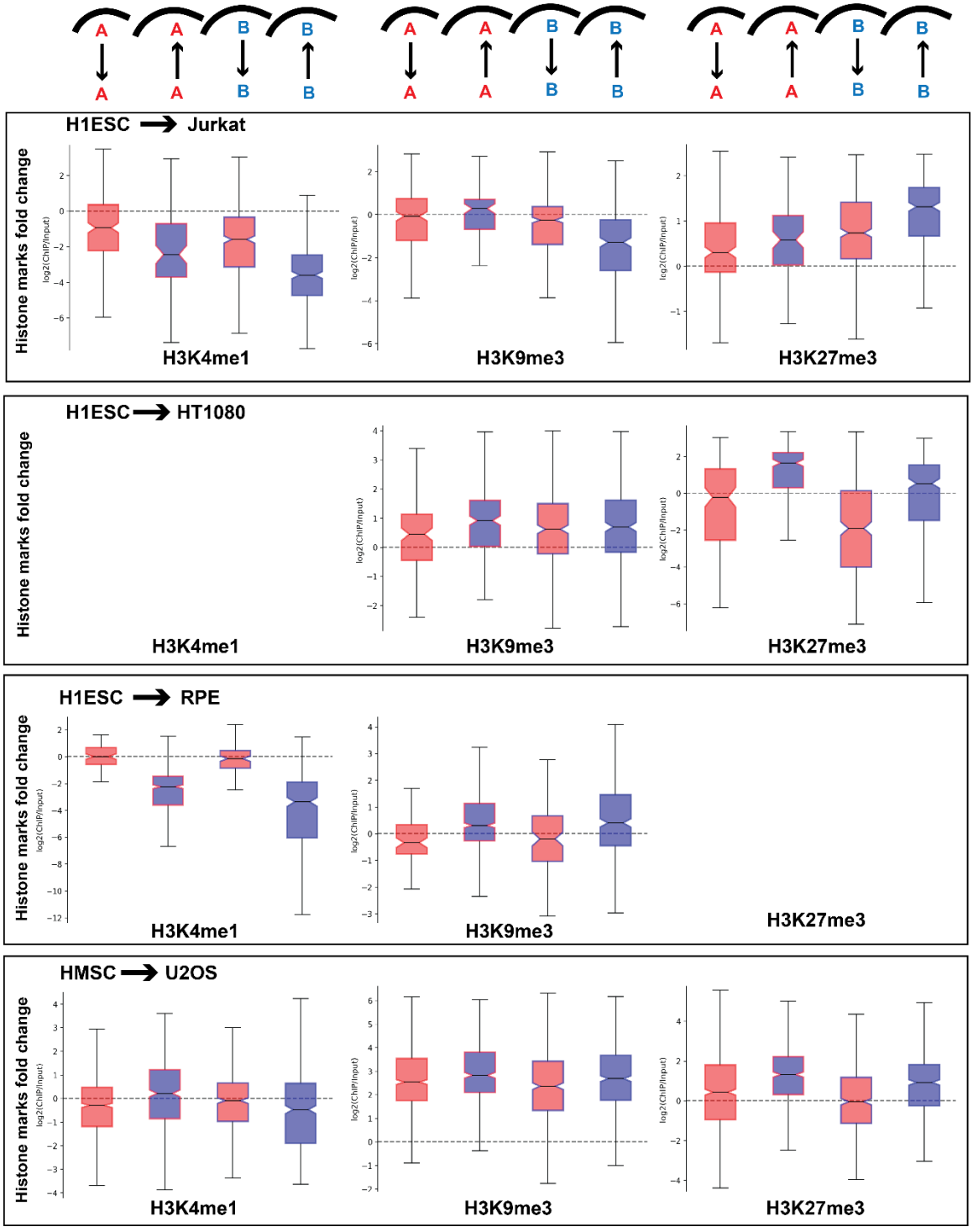


Figure 55 Fold change of different histone marks for the similar compartment identity regions which exhibit differential lamina association between cell types.

Labels, coloring, and organization are as described in Fig. 48, for different cell type comparisons.

CHAPTER V
REARRANGEMENT OF 3D GENOME ORGANIZATION IN BREAST CANCER METASTASIS ORGANOTROPISM

A version of this chapter is under preparation for publication.

My contributions to this work include: (1) conceived the project, (2) performed bioinformatics data analysis, (3) wrote the manuscript, and (4) prepared the figures.

Abstract

Breast cancer cells exhibit organotropism during metastasis in which they show a preference for homing to certain organs such as bone, lung, liver, or brain. One potential mechanism that helps explain this organotropism behavior is the 'seed and soil' theory, which hypothesizes that certain cancer cells (seed) have specific abilities to thrive in certain microenvironments (soil). Several studies have shown that the genesis of these specific metastatic traits are to some extent governed by the regulation of the gene signatures at the primary tumor site. On the other hand, recent advancements in the field of genome biology have found that the spatial genome organization plays a crucial role in oncogenic transformation and progression. However, it is still not known to what extent chromosome architecture contributes to organ specific metastatic traits. It is possible that similar to gene expression, the three-dimensional chromatin architecture also changes in metastasis specific manner due to the influence of the primary tumor and host organ microenvironment. Therefore, in this project, we decided to characterize the chromosome structural changes associated with organotropic metastasis traits. For that purpose, we compared genome organization Hi-C data from different subtypes of lung metastatic breast cancer cells with both lung normal and cancerous cells to find how the genome reorganizes to adapt to a new microenvironment. We observed that the lung metastatic cells exhibit a higher percentage of compartment switches which are consistent with a lung microenvironment compared to the cancer cells that remain at the primary site. Additionally, when we analyzed the genome organization data in an unsupervised fashion, we found that cells segregate based on epithelial-mesenchymal states.

Introduction

Breast cancer is the leading malignant disease in women worldwide. It is highly heterogeneous; based on molecular characteristics it can be divided into four broad subgroups: luminal A (estrogen receptor ER+ progesterone receptor PR+/- human epidermal growth factor receptor 2 HER2-), luminal B (ER+ PR+/- HER2+), HER2-enriched (ER- PR- HER2+), and triple-negative/basal (ER- PR- HER2-) (Fragomeni et al., 2018). About 20 to 30% of early-stage breast cancer patients have a chance to ultimately develop distant metastasis, which accounts for 90% of all breast cancer-related deaths (Chaffer & Weinberg, 2011). Overall, metastasis implies the dissemination of cancer cells from the primary tumor site into the circulation system and later extravasation to a new site which leads to the formation of a secondary tumor. However, the organs affected by breast cancer metastasis vary remarkably, depending on the disease subtype, molecular features, primary tumor location, and secondary site microenvironment. This phenomenon is known as 'metastasis organotropism' (Chen et al., 2018). For example, breast cancer cells are able to metastasize to bone, lung, liver, and brain. However, bone metastasis accounts for a major percentage of metastatic cases compared to other organs (Savci-Heijink et al., 2016). In addition, luminal subtypes show a higher percentage of bone metastasis whereas luminal B and basal subtypes preferentially metastasize to the lung (Smid et al., 2008). The mechanism behind this organotropic behavior can be explained in the light of the 'seed and soil' hypothesis, according to which a cancer cell (seed) having specific intrinsic abilities will only be able to survive in a host local microenvironment (soil) that is favorable to it (Langley & Fidler, 2011).

Secondary organ colonization by circulating tumor cells is a complex process, ruled by dynamic and evolving interactions between cancer cells and the host microenvironment. In most situations, the microenvironment is unfavorable, and the disseminated cells fail to survive in that soil. Despite the low survival rate of the cancer cells during the metastatic colonization process, cells with the intrinsic capabilities to interact with the host microenvironment can survive and proliferate in the secondary site (Obenauf & Massagué, 2015). Findings in breast cancer metastasis research show that if the breast primary tumor microenvironment is enriched in transforming growth factor beta (TGFB), it promotes lung metastasis through the induction of angiopoietin-like 4 (Padua et al., 2008). On the other hand, having a tumor stroma enriched with mesenchymal cytokines like C-X-C motif chemokine 12 (CXCL12) and insulin-like growth factor 1 (IGF1) predisposes primary cancer cell population to grow in the bone marrow (Zhang et al., 2013). But in most cases, the origin of these traits still poses an outstanding question.

The current view in the cancer biology field is that the gene signatures associated with the organ-specific metastatic traits are regulated in part by cues from the reactive microenvironment that accompanies primary tumorigenesis. However, whether the regulation of those genes involve changes in genome organization is still an open question. Recent advancements in imaging and molecular biology techniques have shown that the genome organizes in a non-random multi-layered fashion inside the nucleus and each layer

of the organization contributes both independently and synergistically to the gene expression regulation (McCord et al., 2020). For example, topologically associated domains (TADs), which are highly interacting local chromatin regions demarcated by the boundary elements such as CTCF, have a close linkage with the gene expression (Luppino et al., 2020; Nora et al., 2012; Szabo et al., 2020; Wutz et al., 2017). Disruption of these domains has been shown to contribute to the abnormal expression of oncogenes in breast and other cancers (Flavahan et al., 2019; Flavahan et al., 2016; Hnisz et al., 2016; Lebeau et al., 2022). In one such case for breast cancer, copy number loss of CTCF causes global TAD boundary weakening which enables the upregulation of different oncogenes and oncogenic pathways such as snail family transcriptional repressor 1 (SNAI1) and phosphatidylinositol 3-kinase (PI3K) pathway (Lebeau et al., 2022). Similar findings have also been observed in lymphoblastic leukemia, gliomas, and gastrointestinal cancers where disruption of a boundary element allows insulated neighborhoods containing proto-oncogenes and distal enhancers to interact (Flavahan et al., 2019; Flavahan et al., 2016; Hnisz et al., 2016). Apart from the TADs, at larger scale, the reorganization of the compartments, which represents the segregation between euchromatic and heterochromatic regions, are also found to be associated with the cancer progression (Achinger-Kawecka et al., 2020; Kim et al., 2022; San Martin et al., 2021). It has been shown that as breast cancer becomes endocrine resistant, the change in compartment organization is accompanied by gene expression changes and differential estrogen receptor binding (Achinger-Kawecka et al., 2020). In another study, the researchers have found that CTCF-dependent rearrangement of chromatin domains are associated with dysregulated chromatin accessibility and transcription in triple-negative breast cancer (Kim et al., 2022). Other studies have also found the rearrangement of chromatin architecture at different levels due to various drug treatments and therapies (Achinger-Kawecka et al., 2020; Barutcu et al., 2016; Yang et al., 2020). Though these types of studies are essential for elucidating the association between structural reorganization and oncogenic transformation, they do not necessarily focus on the host microenvironment specific structural changes associated with metastatic cancer cell genome organization. It can be hypothesized that similar to gene expression, metastatic cancer cells also exhibit distant host microenvironment specific genome rearrangements. Therefore, it is important to investigate to what extent the three-dimensional chromatin architecture of metastatic breast cancer cell mimics the host microenvironment. Here, we focused on the breast cancer lung metastasis system. We compared genome organization data from several breast normal, localized, and metastatic cancerous cells with the normal and diseased lung cells. We found that, indeed, metastatic breast cancer cells have a higher proportion of lung specific changes than localized breast cancer cells.

Results

Modeling of breast cancer lung metastasis system

In order to characterize distant organ microenvironment specific rearrangement of breast cancer spatial genome organization of the breast cancer genome, we required compartmental organization data from different types of normal and cancerous cells for

both breast and another secondary organ. Here, we used lung as our focus organ for metastasis as most of the breast cancer subtypes except luminal show a preference for lung metastasis (Bennett et al., 2022). Further, the breast cancer lung metastasis system has been widely studied, and a plethora of omics data is already available publicly. Therefore, different breast and lung normal and cancerous cell types for which publicly available genome organization (Hi-C) are available were selected for this work (Appendix: Table 7). A total of 15 cell types including both primary cells and cell-lines from both the organs were obtained. The cells were then arranged based on their malignancy status and other molecular signatures in an ordered tree fashion (Fig. 56). Briefly, for the breast cancer system we considered both the HMEC and MCF10A as normal breast epithelial cells and placed them at the starting node of the tree. Following that, the normal epithelial cell category was further divided into three subcategories based on breast cancer molecular signatures: luminal (ZR751, BT474, ZR7530, MCF7 and T47D), HER2-enriched (HCC1954 and SKBR3) and triple-negative (HCC70, BT549 and MDAMB231). Further within the luminal subtype, ZR751 and BT474 were designated as localized cancers and the remaining three cell-lines, ZR7530, MCF7 and T47D were used as ascites, lung, and lung metastatic breast cancers respectively. Similarly, for the HER2-enriched subcategory, HCC1954 and SKBR3 were considered as localized and lung metastatic cancers respectively. And finally, in case of the triple-negative subtype, HCC70 and BT549 were used to model localized cancer whereas MDA-MB-231 represents lung metastatic cancer. On the other hand, in the case of the lung system, HTBE primary cells were used to represent normal lung epithelium and both the A549 and H460 represented localized lung cancer.

Breast cancer compartmental organization exhibits subtype and epithelial-mesenchymal state based segregation

After having both the breast and lung system cells arranged based on their already known molecular and clinical properties, we decided to compare this with the in-silico predicted ordering based on genome-wide compartmental organization data. For compartmental organization, for each of the breast and lung cells, the whole genome was divided into bins of 250 kb resolution and compartment identity strengths, obtained from the eigenvector decomposition, were assigned to them. Here, a positive value signifies a euchromatic A compartment region, while a negative value represents a heterochromatic B compartment. When we applied hierarchical clustering on the genome-wide compartmental organization data, we observed a grouping of cells that was mostly consistent with the cell ordering that we constructed based on molecular and clinical features (Fig. 57a). For example, all the normal cell types, including both the breast and lung systems, were clustered together. Similarly, all the luminal subtype and the lung carcinoma cells were grouped in their respective clusters. However, we also observed a few interesting exceptions from the clustering result. One such is the clustering of MDA-MB-231 cell with the lung carcinoma cells. This implies that among all the lung metastatic breast cancer cells, MDA-MB-231 genome organization is closer to lung cells genome organization. We also found that

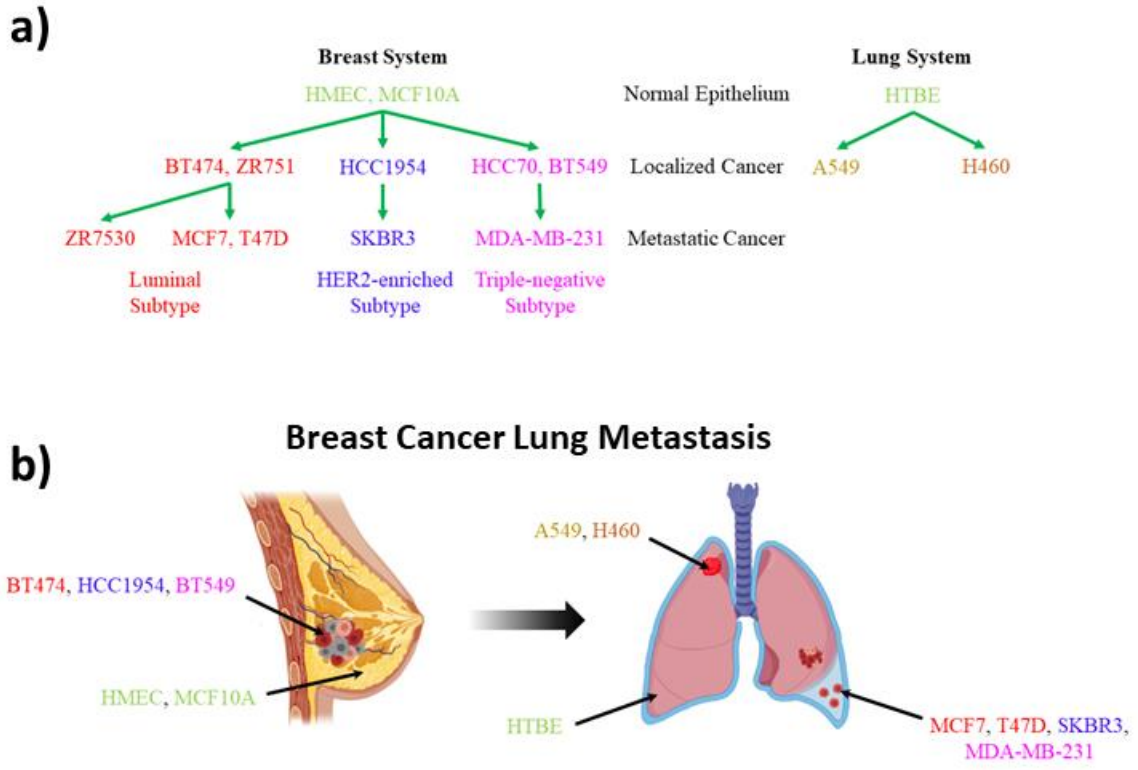


Figure 56 Modelling of breast cancer lung metastasis system.

a) Primary cells and cell lines obtained from human breast and lung organs are arranged depending on their molecular subtypes and malignancy status based on prior available characteristics. Green: normal breast and lung epithelium, red: luminal subtype, blue: HER2-enriched subtype, pink: triple-negative subtype and brown labels for localized lung cancer. b) Normal breast epithelium (HMEC and MCF10A), localized breast cancer (BT474, HCC1954, and BT549), lung metastatic breast cancer (MCF7, T47D, SKBR3, and MDA-MB-231) cells are used to model different stages of breast cancer progression. To model diseased lung, normal lung epithelium (HTBE) and localized lung cancer (A549 and H460) cells are used.

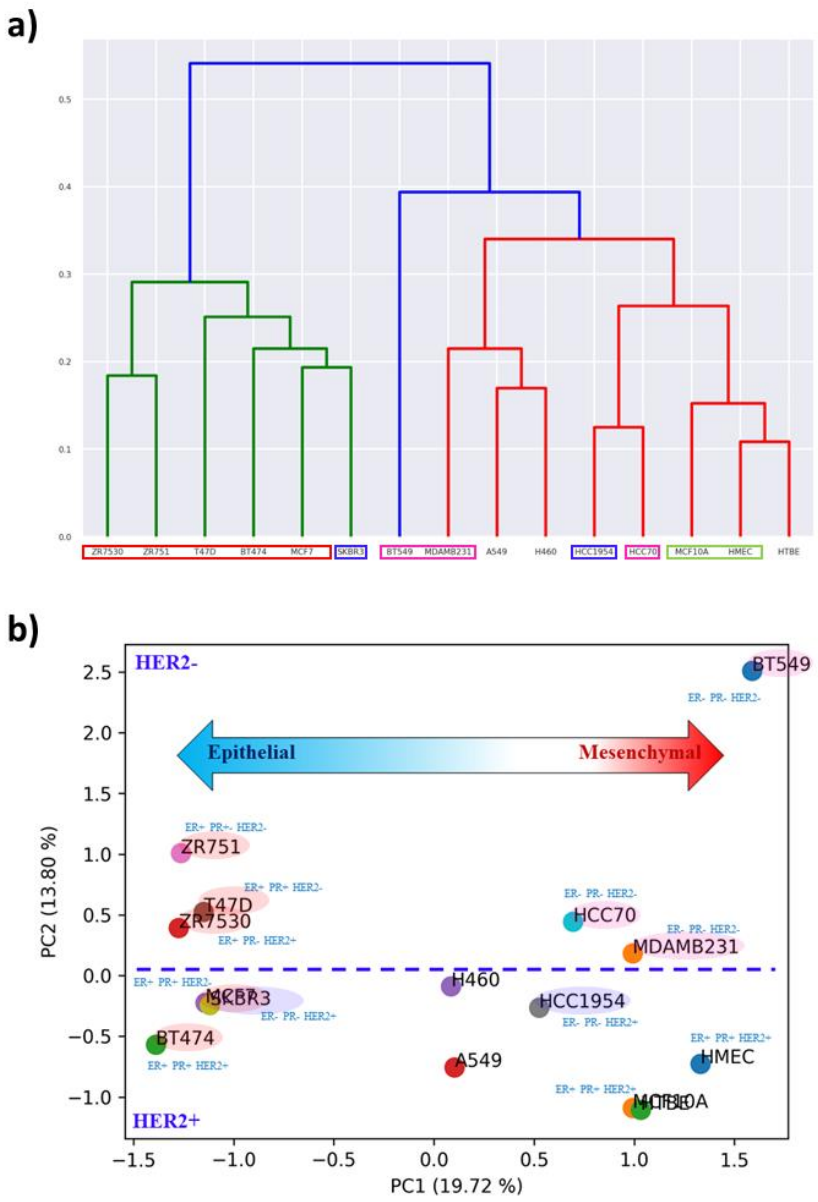


Figure 57 Unsupervised analysis separates breast cancer cells depending on their molecular subtypes and epithelial-mesenchymal states.

a) Hierarchical clustering of genome-wide compartment identity data for different breast and lung cell types at 250 kb resolutions. Breast cancer cells are colored based on their molecular subtypes as represented in 5.1a. b) Principal component analysis of genome-wide compartment identity data for different breast and lung cell types at 250 kb resolution. Breast cancer cells are colored based on their molecular subtypes as represented in 5.1a. For each breast cell, the status of estrogen receptor (ER), human epidermal growth factor receptor (HER2) and progesterone receptor (PR) are given. Also, the PC2 axis is divided into two subspaces based on HER2 status (blue thick dotted line).

HCC70 and HCC1954, which are both triple-negative localized cancer cells, clustered relatively close to the normal epithelial cells. Collectively, this unsupervised analysis shows that the genome organization is quite different between different types of cancer and even within the subtypes also. Moreover, performing principal component analysis (PCA) on the genome-wide compartmental organization data also revealed similar trends exhibited by the hierarchical clustering (Fig. 57b). Interestingly, when we looked at the ordering of the cell types along the principal component 1 (PC1) axis, we noticed a transition between epithelial and mesenchymal cells. Cells characterized as highly epithelial were found on the left extreme of the PC1 axis and as we move towards the right along PC1, the cells start to become more mesenchymal. For example, MDA-MB-231 cells are of epithelial origin but after going through an epithelial-mesenchymal transition, those cells show a mesenchymal-like phenotype and in our analysis these cells are characterized as highly mesenchymal. The transition along PC1 matches quite well with the previously characterized epithelial-mesenchymal transition based on the transcriptome data (Appendix: Fig. 59) (Le et al., 2018; Tan et al., 2014). The PC2 axis, on the other hand, was found to segregate the cell types depending on the presence or absence of the HER2 receptor. As the HER2 receptor is an epithelial growth factor receptor responsible for regulating cell growth and apoptosis, it is possible that presence or absence of this receptor affects the genome organization at the compartment scale. Overall, our analyses show that the genome organization is not only able to distinguish different subtypes of cancers but also exhibits changes along the epithelial-mesenchymal transition.

Proportions of lung specific changes are higher in lung metastatic breast cancers compared to the localized cancers

We next investigated how specific the metastatic genome organization changes are to the secondary organ microenvironment by examining the association of 3D genome organization between the metastatic breast cancer cells and lung cells. We first compared each of the metastatic breast cancer compartment identity data with the normal breast epithelial cells. For each comparison, this resulted in four possible compartment identity combination: AA (A compartment in both normal breast and metastatic breast cancer), AB (A compartment in normal breast and B compartment in metastatic breast cancer), BA (B compartment in normal breast and A compartment in metastatic breast cancer), and BB (B compartment in both normal breast and metastatic breast cancer). Similarly, the compartment identity of the localized lung cancer cells were also compared with the normal lung cell. Once we had those comparisons, we further performed a cross-comparison between the metastatic breast and localized lung cancer comparisons and that led to the emergence of a total of 16 compartment identity combinations. Out of those 16 combinations, we were only interested two specific combinations: AB_BB (A compartment in normal breast and B compartment in metastatic breast cancer; B compartment in both normal lung and localized lung cancer) and BA_AA (B compartment in normal breast and A compartment in metastatic breast cancer; A compartment in both normal lung and localized lung cancer). These particular combinations were termed as lung

permissive changes, as the compartment switches in metastatic breast cancer are congruent to stable lung compartments. This particular combinations allow us to distinguish lung-specific changes from the cancer-specific traits such as AB_AB and BA_BA. We then calculated the percentage of these lung permissive changes out of all possible compartment identity combinations for each of the cross-comparison, which represent the extent of that specific metastatic breast cancer's similarity to the lung microenvironment. Instead of performing all the comparisons in an individual fashion, we focused on the breast cancer subtype level comparisons. As we have more than one cell for some of the cancer subtypes and normal epithelium, for those cases we only considered compartment identity combinations which were consistent among their respective groups. In addition to metastatic breast cancer, we also repeated the analysis for the localized breast cancer. We then compared the percentages of lung permissive changes between metastatic and localized breast cancers for different subtypes. We found that for luminal and triple-negative subtypes, metastatic breast cancer exhibits a higher level of lung permissive changes compared to localized cancer (Fig. 58).

Discussion

During metastasis, breast cancer cells spread to distant sites in an organ-specific manner, which is widely known as metastasis organotropism. This phenomenon, also shown by other types of cancer, may be explained by the 'seed and soil' hypothesis to some extent. According to that, cancer cells will only be able to colonize at a distant site if the cells have intrinsic ability to survive at that particular host microenvironment. Though the process of metastasis is complex and involves other factors, researchers have shown that at the primary tumor microenvironment, cancer cells acquire the ability to survive at a particular distant organ through the regulation of gene signatures. However, there are many open questions related to the genesis of organ specific metastasis traits. Do those organ specific gene signatures occur as a result of the genome organization changes, as a vast number of recent studies have shown that rearrangement of the genome can lead to gene expression dysregulation? To what extent is alteration in gene expression associated with the rearrangement of the chromatin architecture? What is the role of the rearrangements in metastasis organotropism for which there are no significant gene expression alteration? And what is the similarity level of genome organization between metastatic cancers and host microenvironments? In this work, we have made a step toward answering the last question by analyzing compartmental organization data of different cells involved with breast cancer lung metastasis compared to and their host lung tissue.

When we analyzed the data using an unsupervised clustering technique, we found that the cell types are grouped together based on nested features ranging from molecular subtypes to tissue origin. For example, luminal subtypes are separated from the other breast cancer subtypes, and the triple-negative metastatic breast cancer cell with pleural effusion origin, MDA-MB-231, clusters close to the lung carcinoma cells (A549 and H460). The clustering result supports the conclusion that the breast cancer subtypes are not only quite different at the gene expression level, but also at the level of genome organization. Furthermore, it also

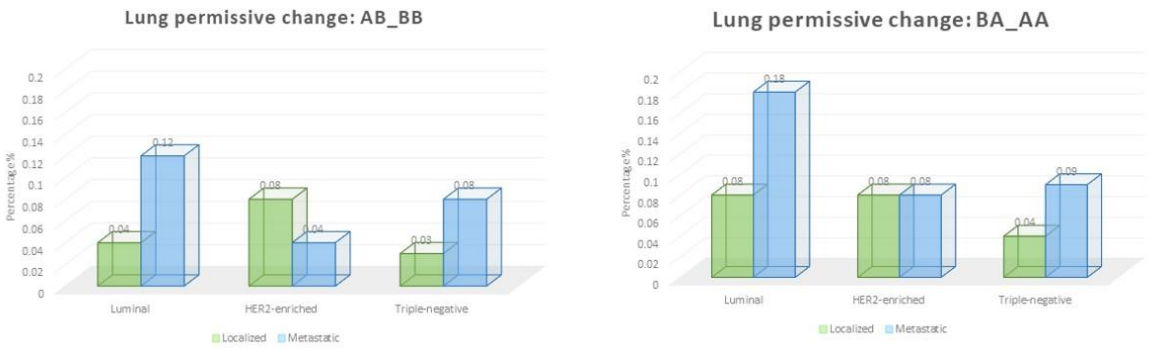


Figure 58 Percentages of different lung permissive changes shown by localized and metastatic cancers from different breast cancer subtypes.

shows that the genome organization is highly consistent among luminal subtype cancers, irrespective of the tissue from which the cells were collected, while the other subtypes such as triple negative, were more heterogeneous in their structure. This makes sense as studies involving gene expression have already shown that other breast cancer subtypes such as triple-negative is highly heterogeneous and has six further subtypes within themselves (Wang et al., 2019). Careful inspection of the result further highlights that the cells are also grouped based on their epithelial and mesenchymal states: luminal – epithelial-like, HER2-enriched – basal-like and triple-negative – mesenchymal-like (D'Amato et al., 2012). This observation becomes much more visible from the PCA transformed data as the cells are arranged based on epithelial-mesenchymal spectrum. This is super interesting because epithelial-mesenchymal states have primarily been associated with gene expression, but it also becoming clear that these states also exhibit characteristics genome architecture as well. Similarly, in a recent study, the researcher have shown that chromatin organization and states change during epithelial-mesenchymal transition (Pang et al., 2022). Another interesting observation is that the normal breast and lung cells are placed on the mesenchymal side of the axis. Although these cells are of epithelial origin, previous transcriptomics studies have characterized them having basal-like gene expression which is close to mesenchymal-like expression (D'Amato et al., 2012; Ince et al., 2007; Kao et al., 2009). This shows the possibility that similar to gene expression, basal-like genome architecture is also close to a mesenchymal state.

In addition to analyzing the data in an unsupervised fashion, we also quantified the level of compartmental organization changes showed by the different localized and lung metastatic cancers specific to the lung microenvironment. For both the luminal and triple-negative subtypes, we found the metastatic cells exhibit higher percentages of lung permissive changes compared to the localized cancers. It can be speculated that these metastatic cells already acquired lung specific compartmental organization to some extent at the primary tumor site in order to survive at the lung once they reach there. Otherwise, it can also happen that these changes were acquired at the metastatic site for their survival as all these cells were isolated from the pleural effusion. In order to answer this, we need temporal genome organization data of metastatic progression which is not available at this moment. We also notice that for the lung permissive changes the percentage of change is higher towards the A compartment compared to the B compartment. It may be possible that shifts toward the A compartment may be more important for matching the host microenvironment state because both activation and gene repression is more possible once a region is in the A compartment while a shift to the B compartment is not necessary to accomplish gene repression (Das et al., 2022; Hakim et al., 2011; Vieux-Rochas et al., 2015). Alternatively, this can also be related to the general loss of heterochromatin previously observed in cancer (Carone & Lawrence, 2013). We further notice that that the luminal subtype exhibits more compartment changes when metastasizing to lung compared to the other types. As the luminal subtypes are less likely to metastasize to lung, they may need to rearrange their genome more in order to acquire lung permissive changes. Overall, our results are consistent with the idea that genome compartmentalization can play a role in facilitating organ-specific metastasis.

Methods

Hi-C data processing

To process Hi-C sequencing reads, here in this work we used HiC-Pro (https://github.com/nservant/HiC-Pro) pipeline with default parameters (Servant et al., 2015). The reads were mapped to hg19 reference genome for the alignment purpose. Also, for each cell type, all the available replicated reads were merged together to improve the quality of contact data for further downstream processing. After obtaining the valid pair interactions, we binned the interactions at 250 kb resolution for each individual chromosome and represented them as square symmetric matrices. Once we had the raw contact matrices, the contacts were normalized using ICE technique (Imakaev et al., 2012). Following that, the number of normalized contacts of each chromosome was scaled to 1M using 'scaleMatrix.pl' script from the cworld-dekker (https://github.com/dekkerlab/cworld-dekker) tool suite. And finally, in order to get the compartmental organization for each individual chromosome, the principal component analysis was performed on the previously obtained normalized and scaled matrices with the help of 'matrix2compartment.pl' script from the cworld-dekker tool suite. The compartment analysis assigns A/B compartment identity to genomic regions based on the sign of the elements of the eigenvector (A-positive and B-negative).

RNA-seq data processing

The RNA-seq files were processed as per the steps described in the previous publication. Briefly, BBDuk (https://github.com/kbaseapps/BBTools) tool was used first for adapter and quality trimming (Bushnell et al., 2017). After that, the reads were mapped to hg19 human reference genome using STAR aligner (Dobin et al., 2013). Once aligned reads were generated, HTSeq-Counts (https://github.com/simon-anders/htseq) tool was used to perform gene level feature count (Anders et al., 2015). As the RNA-seq files were obtained from various sources and generated using different experimental techniques, they suffer from the batch effect which masks the biological signal. To remove the batch effect, ComBat-seq (https://github.com/zhangyuqing/ComBat-seq) tool was applied on the raw gene counts data (Zhang et al., 2020).

Hierarchical clustering of genome-wide compartmental organization and transcriptome data

To cluster different cell types based on their compartmental organization and transcriptome profile similarities, hierarchical clustering was performed on the genome-wide compartment identity and gene expression data respectively. For both the data types, the pairwise similarity score between all the cells were first calculated using Pearson's correlation. The similarity data was then further converted to distance data by subtracting from 1. And finally, hierarchical clustering was performed on the distance data using 'ward' linkage.

REFERENCES

Achinger-Kawecka, J., Valdes-Mora, F., Luu, P.-L., Giles, K. A., Caldon, C. E., Qu, W., Nair, S., Soto, S., Locke, W. J., & Yeo-Teh, N. S. (2020). Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. *Nature communications*, *11*(1), 1-17.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166-169.

Barutcu, A. R., Hong, D., Lajoie, B. R., McCord, R. P., van Wijnen, A. J., Lian, J. B., Stein, J. L., Dekker, J., Imbalzano, A. N., & Stein, G. S. (2016). RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1859*(11), 1389-1397.

Bennett, C., Carroll, C., Wright, C., Awad, B., Park, J. M., Farmer, M., Brown, E., Heatherly, A., & Woodard, S. (2022). Breast Cancer Genomics: Primary and Most Common Metastases. *Cancers*, *14*(13), 3046.

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge–accurate paired shotgun read merging via overlap. *PLoS One*, *12*(10), e0185056.

Carone, D. M., & Lawrence, J. B. (2013). Heterochromatin instability in cancer: from the Barr body to satellites and the nuclear periphery. Seminars in cancer biology,

Chaffer, C. L., & Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science*, *331*(6024), 1559-1564.

Chen, W., Hoffmann, A. D., Liu, H., & Liu, X. (2018). Organotropism: new insights into molecular mechanisms of breast cancer metastasis. *NPJ precision oncology*, *2*(1), 1-12.

D'Amato, N. C., Ostrander, J. H., Bowie, M. L., Sistrunk, C., Borowsky, A., Cardiff, R. D., Bell, K., Young, L. J., Simin, K., & Bachelder, R. E. (2012). Evidence for phenotypic plasticity in aggressive triple-negative breast cancer: human biology is recapitulated by a novel model system.

Das, P., San Martin, R., & McCord, R. P. (2022). Differential contributions of nuclear lamina association and genome compartmentalization to gene regulation. *bioRxiv*, 2022.2009.2012.507606. https://doi.org/10.1101/2022.09.12.507606

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

Flavahan, W. A., Drier, Y., Johnstone, S. E., Hemming, M. L., Tarjan, D. R., Hegazi, E., Shareef, S. J., Javed, N. M., Raut, C. P., & Eschle, B. K. (2019). Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature*, *575*(7781), 229-233.

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., & Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, *529*(7584), 110-114.

Fragomeni, S. M., Sciallis, A., & Jeruss, J. S. (2018). Molecular subtypes and local-regional control of breast cancer. *Surgical Oncology Clinics*, *27*(1), 95-120.

Hakim, O., Sung, M.-H., Voss, T. C., Splinter, E., John, S., Sabo, P. J., Thurman, R. E., Stamatoyannopoulos, J. A., de Laat, W., & Hager, G. L. (2011). Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. *Genome research*, *21*(5), 697-706.

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. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, *351*(6280), 1454-1458.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, *9*(10), 999-1003.

Ince, T. A., Richardson, A. L., Bell, G. W., Saitoh, M., Godar, S., Karnoub, A. E., Iglehart, J. D., & Weinberg, R. A. (2007). Transformation of different human breast epithelial cell types leads to distinct tumor phenotypes. *Cancer cell*, *12*(2), 160-170.

Kao, J., Salari, K., Bocanegra, M., Choi, Y.-L., Girard, L., Gandhi, J., Kwei, K. A., Hernandez-Boussard, T., Wang, P., & Gazdar, A. F. (2009). Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One*, *4*(7), e6146.

Kim, T., Han, S., Chun, Y., Yang, H., Min, H., Jeon, S. Y., Kim, J.-i., Moon, H.-G., & Lee, D. (2022). Comparative characterization of 3D chromatin organization in triple-negative breast cancers. *Experimental & Molecular Medicine*, 1-16.

Langley, R. R., & Fidler, I. J. (2011). The seed and soil hypothesis revisited—The role of tumor-stroma interactions in metastasis to different organs. *International journal of cancer*, *128*(11), 2527-2535.

Le, A. V.-P., Szaumkessel, M., Tan, T. Z., Thiery, J.-P., Thompson, E. W., & Dobrovic, A. (2018). DNA methylation profiling of breast cancer cell lines along the epithelial mesenchymal spectrum—Implications for the choice of circulating tumour DNA methylation markers. *International journal of molecular sciences*, *19*(9), 2553.

Lebeau, B., Jangal, M., Zhao, T., Wong, C. K., Wong, N., Cañedo, E. C., Hébert, S., Aguilar-Mahecha, A., Chabot, C., & Buchanan, M. (2022). 3D chromatin remodeling potentiates transcriptional programs driving cell invasion. *Proceedings of the National Academy of Sciences*, *119*(36), e2203452119.

Luppino, J. M., Park, D. S., Nguyen, S. C., Lan, Y., Xu, Z., Yunker, R., & Joyce, E. F. (2020). Cohesin promotes stochastic domain intermingling to ensure proper regulation of boundary-proximal genes. *Nature genetics*, *52*(8), 840-848.

McCord, R. P., Kaplan, N., & Giorgetti, L. (2020). Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Mol Cell*, *77*. https://doi.org/10.1016/j.molcel.2019.12.021

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., & Sedat, J. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381-385.

Obenauf, A. C., & Massagué, J. (2015). Surviving at a distance: organ-specific metastasis. *Trends in cancer*, *1*(1), 76-91.

Padua, D., Zhang, X. H.-F., Wang, Q., Nadal, C., Gerald, W. L., Gomis, R. R., & Massagué, J. (2008). TGFβ primes breast tumors for lung metastasis seeding through angiopoietin-like 4. *Cell*, *133*(1), 66-77.

Pang, Q. Y., Tan, T. Z., Sundararajan, V., Chiu, Y.-C., Chee, E. Y. W., Chung, V. Y., Choolani, M. A., & Huang, R. Y.-J. (2022). 3D genome organization in the epithelial-mesenchymal transition spectrum. *Genome biology*, *23*(1), 1-31.

San Martin, R., Das, P., Dos Reis Marques, R., Xu, Y., Roberts, J. M., Sanders, J. T., Golloshi, R., & McCord, R. P. (2021). Chromosome compartmentalization alterations in prostate cancer cell lines model disease progression. *Journal of Cell Biology*, *221*(2), e202104108.

Savci-Heijink, C. D., Halfwerk, H., Koster, J., & van de Vijver, M. J. (2016). A novel gene expression signature for bone metastasis in breast carcinomas. *Breast cancer research and treatment*, *156*(2), 249-259.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology*, *16*(1), 1-11.

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G., Foekens, J. A., & Martens, J. W. (2008). Subtypes of breast cancer show preferential site of relapse. *Cancer research*, *68*(9), 3108-3114.

Szabo, Q., Donjon, A., Jerković, I., Papadopoulos, G. L., Cheutin, T., Bonev, B., Nora, E. P., Bruneau, B. G., Bantignies, F., & Cavalli, G. (2020). Regulation of single-cell genome organization into TADs and chromatin nanodomains. *Nature genetics*, *52*(11), 1151-1157.

Tan, T. Z., Miow, Q. H., Miki, Y., Noda, T., Mori, S., Huang, R. Y. J., & Thiery, J. P. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO molecular medicine*, *6*(10), 1279-1293.

Vieux-Rochas, M., Fabre, P. J., Leleu, M., Duboule, D., & Noordermeer, D. (2015). Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proceedings of the National Academy of Sciences*, *112*(15), 4672-4677.

Wang, D.-Y., Jiang, Z., Ben-David, Y., Woodgett, J. R., & Zacksenhaus, E. (2019). Molecular stratification within triple-negative breast cancer subtypes. *Scientific reports*, *9*(1), 1-10.

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., & Hossain, M. J. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO journal*, *36*(24), 3573-3599.

Yang, Y., Choppavarapu, L., Fang, K., Naeini, A. S., Nosirov, B., Li, J., Yang, K., He, Z., Zhou, Y., & Schiff, R. (2020). The 3D genomic landscape of differential response

to EGFR/HER2 inhibition in endocrine-resistant breast cancer cells. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1863*(11), 194631.

Zhang, X. H.-F., Jin, X., Malladi, S., Zou, Y., Wen, Y. H., Brogi, E., Smid, M., Foekens, J. A., & Massagué, J. (2013). Selection of bone metastasis seeds by mesenchymal signals in the primary tumor stroma. *Cell*, *154*(5), 1060-1073.

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR genomics and bioinformatics*, *2*(3), lqaa078.

APPENDIX

Table 7 NCBI GEO or Encode accession codes for different breast and lung cell types.

Cell type	GSE or Encode Accession	Replicates
HMEC	GSE167150	Replicates Combined
MCF10A	GSE109229	Replicates Combined
BT474	GSE109229	Replicates Combined
ZR751	GSE128676	Replicate 1
ZR7530	GSE167150	Replicates Combined
MCF7	GSE109229	Replicates Combined
T47D	GSE167150	Replicates Combined
HCC1954	GSE167150	Replicates Combined
SKBR3	GSE109229	Replicates Combined
HCC70	GSE167150	Replicates Combined
BT549	GSE167150	Replicates Combined
MDA-MB-231	GSE143678	Replicate 1
HTBE	GSE113703	Replicates Combined
A549	ENCSR444WCZ	Replicates Combined
H460	ENCSR489OCU	Replicates Combined

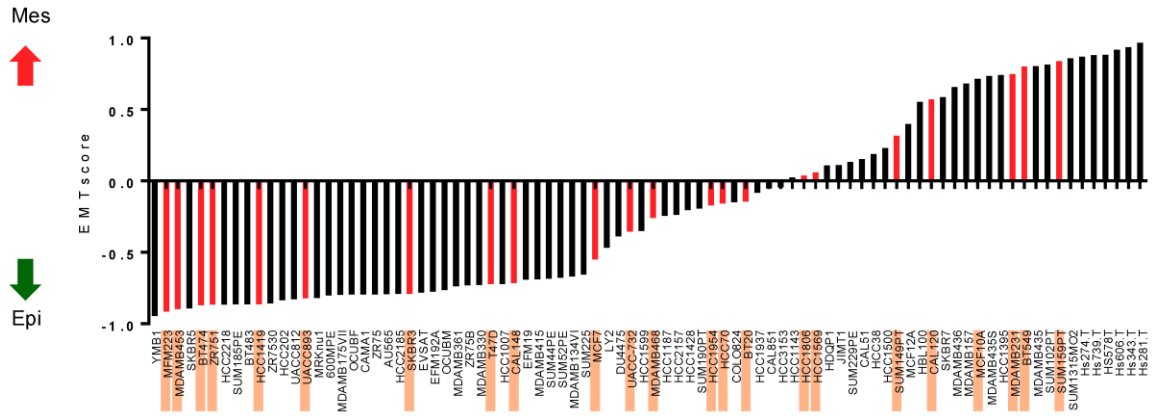


Figure 59 Breast cancer cell lines are arranged depending on their epithelial-mesenchymal status.

The states are calculated based on their gene expression profile using the method described in Tan et al. (Tan et al., 2014). This figure has been adopted from Le et al. (Le et al., 2018).

CONCLUSION

The primary objective of the works presented in this dissertation is to understand how the chromosome structural organization and dynamics are associated with the proper functioning of the cellular system. And also, to characterize the extent of alteration shown by these structural and dynamical properties during different physiological and pathological conditions. To answer these questions, statistical, network and polymer modelling based approaches have been developed through which different types of chromosome structural organization data are analyzed at different resolutions.

At the global scale of the genome organization, we find that the non-random arrangement of the chromosome territories is influenced by several factors such as gene density and chromosome length (Chapter 1). Although, these observations are not entirely new, the interesting thing is here, chromosome conformation capture Hi-C data is used instead of microscopy imaging. This allows us to use Hi-C data to investigate the rearrangement of the radial organization of the chromosomes during different physiological cell states (e.g., proliferating, and senescent) and pathological conditions such as progeria. This also has several other potential applications and can be combined readily with other genome analysis methods, as Hi-C and related techniques have become a household name for the genome organization study. For example, this can be used to study the effect of ionizing radiation on chromosome organization. Recently, researchers have found that lymphoblastoid and fibroblast cells exhibit different genome reorganizations upon exposure to the same ionizing radiation (Sanders et al., 2020). Since these cell types have quite distinct territorial organization, it is possible that depending on the radial localizations of the chromosomes, different DNA damage related reorganization is observed both at the whole genome and individual chromosome level.

Another potential application of Hi-C inferred chromosome territorial positioning can be studying the contribution of other factors (e.g., nuclear substructures and genomic properties) in maintaining the radial organization of the chromosome territories. The nuclear lamina is one such substructure, disruption of which leads to the misregulation of the peripheral heterochromatin. However, the effect of the nuclear lamina in maintaining the global territorial organization has not been investigated thoroughly. Only a handful of microscopy-based studies have been performed to study the mislocalization of the chromosome territories due to the disruption of different lamina components, which further focus on the selective set of chromosomes (Grigoryan et al., 2021; Mewborn et al., 2010; Pradhan et al., 2018). On the other hand, several Hi-C datasets from the lamina perturbed system are available, analysis of which can reveal the rearrangement of the territories associated with different components of the nuclear lamina. Furthermore, using the Hi-C data using my network layout approach also reproduces the radial distributions of the individual chromosomes revealing single-cell variability associated with territorial organization (Chapter 1). This can be really important for situations where the average radial positioning of a chromosome remains same at the population level but exhibits higher variability at the single-cell level.

Another observation which has implications across my several research projects is that the nuclear shape is associated with the non-random organization of the chromosome territories. It is known that the gene poor heterochromatin regions are preferentially located to the nuclear periphery whereas the gene rich euchromatin can be found internally. However, this primarily holds for the spherical nuclei which follow a gene density based radial organization. In case of the flat ellipsoidal fibroblast nuclei, we observe a lesser agreement between genomic compartmentalization and lamina association, as euchromatin regions are more likely to be peripherally located compared to the spherical nuclei. This may possibly be explained by the fact that as the nucleus becomes flatter, the internal euchromatin gene rich regions also get close to the periphery due to space constraint, though they do not tether directly to the nuclear lamina. This also indicates that for lamina tethering other modifications are needed. Similarly, in a breast cancer system, we observe a clear difference between the genome organization of different subtypes (Chapter 5). Prior studies have also shown that for those subtypes, the nuclear morphology is also highly distinctive. Therefore, it is possible that the distinct morphologies associated with the different subtypes modulate the breast cancer genome organization separately. This can also be extended to the epithelial-mesenchymal transition where the genome organization may be regulated by the distinct nuclear morphologies of epithelial and mesenchymal cells.

Similar to characterizing variability of chromosome territories at the single-cell level, we also analyzed variabilities associated with the individual chromosome structures among a population of cells. We find that the underlying epigenetic properties of chromatin regions are a major contributing factor to the structural variation (Chapter 2). We observe that epigenetically homogenous regions are more likely to exhibit a smooth continuum of structures at the single-cell level in comparison to epigenetically heterogeneous regions. Separately, we also notice that chromatin architectural proteins and nuclear substructures also contribute to the structural heterogeneity present at the single-cell level. We, however, still do not know whether these different factors act antagonistically or synergistically to regulate the structural variation. In contrast, it may also be possible that one factor regulates the other: either epigenetic states regulate the interaction with nuclear substructures or the other way around. As the answer to this question is still debated, first analyses often consider the contribution of each factor separately, though it is more likely that these factors act synergistically. For example, a peripherally located B compartment region is more constrained compared to an internally localized one, due to both heterochromatin marks and chromatin tethering to the lamina. The peripheral localization also increases the structural similarity by favoring the repressed compact structures, and as a result the transcription of the genes from the region gets further repressed. We observe a similar result across different cell types where heterochromatic regions become further transcriptionally silenced once they localize peripherally (Chapter 4). Following the same trend, we also find that the peripherally located inactive regions are less likely to be perturbed upon different physicochemical treatments compared to other regions (Chapter 4). Again, it can be speculated that the tethering of chromatin to nuclear lamina in those regions contributes to that observation. Apart from the single-cell level, analyzing the

individual chromosome structures at the population level without distance-decay normalization reveals the generic polymer aspect of the structures. We find that chromosomes approach a generic noninteracting polymer limit as cell enters mitosis phase (Chapter 3). It is known that as the cell enters mitosis phase, the chromosomes gradually lose interaction with many nuclear substructures such as nuclear lamina as well as many of the architectural proteins (e.g., cohesin). Therefore, this again supports the idea that chromatin architectural proteins and interaction with the nuclear substructures are important for the proper structural organization of the chromosomes.

In summary, the works presented in this dissertation show us that not only the multi-layered structural organization but also dynamics of each of the layer contribute to the proper function of the biological processes.

REFERENCES

Grigoryan, A., Pospiech, J., Krämer, S., Lipka, D., Liehr, T., Geiger, H., Kimura, H., Mulaw, M. A., & Florian, M. C. (2021). Attrition of X chromosome inactivation in aged hematopoietic stem cells. *Stem Cell Reports*, *16*(4), 708-716.

Mewborn, S. K., Puckelwartz, M. J., Abuisneineh, F., Fahrenbach, J. P., Zhang, Y., MacLeod, H., Dellefave, L., Pytel, P., Selig, S., & Labno, C. M. (2010). Altered chromosomal positioning, compaction, and gene expression with a lamin A/C gene mutation. *PLoS One*, *5*(12), e14342.

Pradhan, R., Ranade, D., & Sengupta, K. (2018). Emerin modulates spatial organization of chromosome territories in cells on softer matrices. *Nucleic acids research*, *46*(11), 5561-5586.

Sanders, J. T., Freeman, T. F., Xu, Y., Golloshi, R., Stallard, M. A., Hill, A. M., San Martin, R., Balajee, A. S., & McCord, R. P. (2020). Radiation-induced DNA damage and repair effects on 3D genome organization. *Nature communications*, *11*(1), 1-14.

VITA

Priyojit Das was born in Kolkata, India. He completed his undergraduate study from Heritage Institute of Technology Kolkata, where he received Bachelor of Technology degree in Computer Science and Engineering. After finishing his undergraduate program, he decided to pursue higher studies in computational sciences with a focus on quantitative biology research. He later obtained a Master of Technology degree in Computer Science and Engineering from National Institute of Technology Calicut. During his master's, he performed research under the guidance of Dr. KA Abdul Nazeer where he worked on designing clustering algorithms to identify different cell types within the tumor cell population based on gene expression data. Upon graduation, he decided to move to the United States and joined the Genome Science and Technology graduate program at University of Tennessee, Knoxville where he pursued his doctoral research under the supervision of Dr. Rachel Patton McCord. At McCord lab, he studied the structural and dynamical properties of chromosome structural ensemble and how those properties are altered due to different perturbations and diseased conditions (salt stress, cancer, and progeria). He presented his work at multiple conferences including Biophysical Society, American Chemical Society, Cold Spring Harbor Meeting, Keystone Symposia and European Molecular Biology Laboratory Meeting. During his doctoral career, he published several first- and co-authored research papers in reputed peer-reviewed journal including Physical Review E, BMC Bioinformatics, PLoS Computational Biology, Journal of Cell Biology, and EMBO Reports. For his postdoctoral research, he will be joining Dr. Jeannie T. Lee's lab, which is part of both Massachusetts General Hospital and Harvard Medical School.