

Nuclear Architecture from Chromosomes to Motifs

By

Ankit Agrawal

LIFE 10201304002

The Institute of Mathematical Sciences, Chennai

A thesis submitted to the

Board of Studies in Life Sciences

In partial fulfillment of requirements

For the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE



July, 2019

Homi Bhabha National Institute

Recommendations of the Viva Voce Board

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by Ankit Agrawal entitled “Nuclear Architecture from Chromosomes to Motifs” and recommend that it may be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date:

Chairman - Prof. Sitabhra Sinha

_____ Date:

Guide/Convener - Prof. Gautam I. Menon

_____ Date:

Co-guide - Prof. Rahul Siddharthan

_____ Date:

Examiner - Prof. Ranjith Padinhateeri

_____ Date:

Member 1 - Prof. Satyavani Vemparala

_____ Date:

Member 2 - Prof. Areejit Samal

_____ Date:

Member 3 - Prof. Pinaki Chaudhari

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to HBNI.

I hereby certify that I have read this thesis prepared under my direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date:

Place: IMSc, Chennai

Co-guide

Guide

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from, or reproduction of this manuscript in whole or in part, may be granted by the Competent Authority of HBNI when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Ankit Agrawal

DECLARATION

I hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree/diploma at this or any other Institution/University.

Ankit Agrawal

List of Publications arising from the thesis

Journal

1. **Chromatin as active matter**

Ankit Agrawal, Nirmalendu Ganai, Surajit Sengupta and Gautam I. Menon
Journal of Statistical Mechanics: Theory and Experiment, 2017, 014001

2. **ThiCweed: fast, sensitive detection of sequence features by clustering big datasets**

Ankit Agrawal, Snehal V. Sambare, Leelavati Narlikar and Rahul Siddharthan
Nucleic Acids Research, 2018, 46(5), e29

Preprint

1. **A first-principles approach to large-scale nuclear architecture**

Ankit Agrawal, Nirmalendu Ganai, Surajit Sengupta and Gautam I. Menon
bioRxiv:10.1101/315812

Seminars and posters presented

1. “A biophysical model of higher-order chromatin architecture”

presented as colloquium talk at “Department of Physics” Bar-Ilan University, Israel, 27th June, 2018

2. “A computational model of large-scale nuclear architecture” and “ThiCweed: fast, sensitive detection of sequence features by clustering big data sets”

presented as posters at The European Molecular Biology Laboratory, Heidelberg, Germany, 30th August, 2017

3. “Large-scale nuclear architecture model”

presented as invited talk at “3rd BSSE annual research symposium” Indian Institute of Science, Bangalore, 27th January, 2017

To my family and friends

Acknowledgements

The success and final outcome of this thesis required a lot of guidance and assistance from many people and I am extremely fortunate that I got these throughout my Ph.D. duration. Firstly, I would like to express my sincere gratitude to my advisor Prof. Gautam I. Menon and co-advisor Prof. Rahul Siddharthan for their continuous support and for establishing an excellent research environment for computational biology at IMSc. Both my advisors motivated me through their vast experience, immense knowledge, and patience in such a way that in future I can work as an independent researcher. Their consistent guidance helped me throughout my research and right up to the writing of this thesis.

Besides my advisors, I would like to thank my collaborators Prof. Leelavati Narlikar, Prof. Surajit Sengupta, Nirmalendu and Snehal for taking keen interest in our joint projects and for guiding us till their completion by providing all the necessary information to make them successful. I would like to thank the rest of my doctoral committee members: Prof. Sitabhra Sinha, Prof. Satyavani Vemparala, Prof. Areejit Samal, and Prof. Pinaki Chaudhuri, for their insightful comments and encouragement which enabled me to widen my research perspectives. I am fortunate to be the first student of the computational biology group at IMSc. Credit for this goes to my advisors, collaborators and doctoral committee members.

I owe my profound gratitude to many people and online platforms (for example, Stack Exchange, ResearchGate, MATLAB Answers, and LAMMPS mailing list)

whom I contacted via email on numerous occasions. Each time they were able to provide me with solutions. They helped me by providing data, debugging codes, and also by clearing my doubts on various topics.

I have been blessed with a friendly and cheerful group of postdocs and PhD fellows: Anil, Kamal, Vinod, Able, Rakesh, Vivek Vyas, Pritam, Dheeraj, Vinay, Tanmay Mitra, Devanand, Pulak, Rajesh, Rishu, Jaykumar, Kamalakshya, Sourav, Bijoy, Aradhana, Pallavi, Sebastien, Diptapriyo, Sanjoy, Abinash, Arnab, Anirban, Sagnik, Dipanjan, Arindam Mallick, Avijit, Priyamvad, Prosenjit, Trilochan, Ashraf, Dhruv, Prafulla, Ria, Vivek Ananth, Pavitra, Sreevidya, Chandrani, Farhina and Amir at IMSc. Their participation in stimulating discussions on all kinds of topics, encouragement, support, providing sweets and chocolates at various festivals, was our daily routine of IMSc life since 2013.

I also want to thank the administration, security guards, canteen pantry staff, and housekeeping staff for providing a 24 x 7 friendly working environment. Also, I thank my colleagues in other institution Amit, Himanshu, Ahmad, Sakshi, Kawstov, Manish, Girdhari, Raj, Broto, Bipin, Anubhooti, Krishna Kant, Pravin, Ram Vivek, Harish, Ashreya, Soma, Vasu, Sarath, Pankaj, Deepak, Sudeep, Camellia, and Sanjiv for fruitful discussions on various topics.

Last but not least, I would like to thank my family: my parents, bhaiya, bhabhi, Anusha, Atharva, mama, and mami for supporting me spiritually throughout my life in general. A special thanks to my mother for all her love, encouragement and support my decisions. Also thanks to my fiancée Pavi for pushing me to write the thesis on time so that we can make the plan for our future. And finally, this thesis is the result of all the prayers which kept me healthy, gave me wisdom and the ideas to complete the research.

Contents

Synopsis	v
List of Figures	xv
List of Tables	xix
1 Introduction	1
1.1 Chromatin Organization	3
1.2 Gene Regulation	5
1.3 Experimental Approaches to Chromatin Organization	10
1.3.1 Key Observations From FISH	11
1.3.2 Key Observations From 3C-based Methods	15
1.4 Nuclear Subcompartments	16
1.5 Contact Probability	18
1.6 Single Cell and Cell Type-specific Features	19
1.7 Theoretical Models of Chromatin Fiber	20
1.7.1 Polymer Models of Chromatin	22
1.8 Noise and Fluctuation In the Nuclear Environment	25
1.9 Active Matter	26
1.10 Effective Temperature Estimates	28
1.11 Conclusion	29
2 A First-principles Model for Large-scale Nuclear Architecture	31
2.1 Description of Model	32

2.1.1	Gene Density Model	35
2.1.2	Gene Expression Model	37
2.1.3	Combined Model	39
2.2	Incorporating Contact Information from Hi-C Data into Model Chromosome Loops	42
2.3	Simulation Methodology	43
2.3.1	Bond Interaction Parameters	44
2.3.2	Pair Interaction Parameters	46
2.3.3	Interaction of monomers with the nuclear envelope	48
2.3.4	Methodology, Units and Normalization	48
2.3.5	Summary of Analysis	50
2.4	Calculation of Distribution Functions	51
2.5	Geometric Properties of Chromosome Territories	53
2.5.1	Calculation of Three-dimensional Irregular Shape for a Given Chromosome	53
2.5.2	Comparison of Methods for Calculating Three-dimensional Irregular and Regular Shapes	57
2.5.3	Calculation of Volume and Surface Area of a Chromosome	58
2.5.4	Calculation of Asphericity and Prolateness Parameters	59
2.5.5	Calculation of Ellipticity and Regularity in 2d Projection	60
2.5.6	Calculation of Contact Probability	62
2.5.7	Calculation of Distance Maps and Contact Maps	62
2.5.8	Statistical error calculation for relative center of mass position data	63
2.6	Conclusion	63
3	Results from a First-principles Approach to Large-scale Nuclear Architecture	65
3.1	Overview of models and parametrization	68
3.2	Results from the Gene Density Model	69
3.3	Results from the Gene Expression Model I	73
3.4	Results from the Gene Expression Model II	79

3.5	Results from the Combined Model	81
3.5.1	S(R) and $S_{CM}(R)$	82
3.5.2	Relative Centre of Mass Positions	89
3.5.3	Distribution of Active and Inactive Monomers	94
3.5.4	Monomer Distribution across Cell types	95
3.5.5	Ellipticity and Regularity in Two-dimensional Projection	97
3.5.6	Three-dimensional Volume and Surface Area of Chromosome	98
3.5.7	The Differential Positioning of the Xa and Xi Chromosome in the Presence of Superloops in Xi	102
3.5.8	Contact Probability and Spatial Distribution	105
3.5.9	Asphericity and Prolatensess	108
3.5.10	Distance Maps and Contact Maps	110
3.6	Conclusions	112
4	Motif Identification Through Clustering of ChIP-Seq Data	115
4.1	Identification of Transcription Factor Binding Sites (TFBS)	116
4.1.1	Experimental Approaches	116
4.1.2	Computational Approaches	118
4.2	THiCweed: Introduction	122
4.3	THiCweed: Methods	123
4.3.1	Algorithm	126
4.3.2	Benchmarking: Synthetic Data	129
4.3.3	ENCODE Data	132
4.4	Results	133
4.4.1	Synthetic Data	133
4.4.2	Running Times: Synthetic Data	135
4.4.3	Running Times: ENCODE data	136
4.4.4	ChIP-Seq Data from the ENCODE Project	136
4.5	Discussion	142
5	Discussion and Conclusion	145

5.1 Future Directions	152
Bibliography	157

SYNOPSIS

The total length of DNA in a human cell is about 2 meters, divided across 46 chromosomes in somatic cells, and confined to a nucleus whose radius is typically a few microns. A representative sequence of DNA in humans, the human genome, was mapped about 2 decades ago and contains about three billion base pairs. This thesis studies the organization of DNA at multiple scales, ranging from binding regions for regulatory proteins that are between 6 – 30 base pairs, to entire chromosomes containing $\sim 10^7$ base pairs.

In eukaryotic cells, the term chromatin describes DNA as found *in vivo*, where it binds to a host of accessory proteins and specific short RNA sequences. Stretches of the genome that are associated to relatively open DNA conformations, referred to as euchromatin, usually contain actively transcribed genes. In contrast, relatively compact heterochromatin regions, where DNA is tightly bound, are typically associated to gene-poor or non-coding regions of the genome. Chromatin structure must locally be plastic enough to be able to accommodate transcription, DNA replication, and DNA repair machinery. The biophysical machinery responsible for these are largely non-equilibrium “active” processes, since they transduce energy derived from ATP hydrolysis into work. Since ATP concentration in the cell is held out of equilibrium, models for chromatin structuring must incorporate activity.

Genes are stretches of DNA that encode instructions for the synthesis of proteins through RNA. At any given time, the amount of a particular protein in a cell

is determined by the synthesis and degradation of RNA, mainly mRNA. RNA is synthesized from DNA by a protein complex called RNA polymerase in the first step of gene expression, called transcription. Cell-type specific functions are reflected by the amount and types of RNA molecules in a cell. The initiation of transcription is the key control point of gene expression. Transcription can be switched on or off when specific DNA-binding proteins bind to regulatory sequences. These sequence-specific DNA-binding proteins are also known as transcription factors (TFs).

Upstream regions on genes where transcription is initiated are called promoters. A variety of cell or region-specific TFs bind to promoter regions. Similarly, an enhancer is a region of DNA that is located far away (up to $> 1\text{Mb}$) upstream or downstream of genes, where TFs also bind. Once bound, these TF complexes interact with the transcriptional machinery at the promoter to enhance (or diminish) the transcription rate of the gene. Regulatory DNA sequences are capable of increasing or decreasing the expression of particular genes and both promoter and enhancer sequences have a regulatory activity. TFs recognize and bind specific regions of the regulatory sequence called transcription factor binding sites (TFBSs). The regulatory sequences are typically between a hundred and several thousand base pairs in length and can harbor many TFBS [Tuğrul et al., 2015]. Understanding how TFs are assembled and how they recognize binding sites and control transcription is key to understanding cell-type specific gene regulation.

Chromatin organization spans multiple scales. At the smallest scales, we are interested in what determines the binding of individual TFs. At intermediate scales, we are interested in the local organization of DNA, e.g. compartments, loops etc that bring combinations of regulatory proteins into proximity. At the largest scales, relevant to the study of nuclear architecture, we are interested in the shapes, locations and other structural properties of whole chromosomes. This thesis presents work that addresses both the small scale and large scale properties of chromatin, ranging

from TFBS recognition to large-scale nuclear architecture.

This thesis is divided into 5 chapters. In Chapter 1, the Introduction, we summarise the necessary background to the work described in this thesis, including a discussion of major features of nuclear architecture, the importance of non-equilibrium activity for biophysical models of such architecture, and the background to understanding gene regulation through DNA-binding proteins that associate to specific binding sites. Chapter 2 provides details of our model for large-scale nuclear architecture and a description of its computational implementation. In Chapter 3, we present *ab-initio* simulation predictions of a number of features of large-scale nuclear architecture. In Chapter 4, we describe an algorithm, called THiCweed, for clustering TFBS in ChIP-Seq data. This tool outperforms other existing tools in terms of its speed and its ability to capture biologically significant motifs. Finally, in Chapter 5, we end with a conclusion and describe how these studies can be further extended. A more detailed chapter-wise summary of our basic results follows below.

Chapter 1 surveys broad features of chromatin organization in metazoan nuclei as well as summarizes our current understanding of gene regulation via TF binding. We first discuss what is known about nuclear architecture in humans. We describe how information obtained from key experiments: fluorescent in-situ hybridization (FISH) experiments, chromosome conformation capture experiments, and chromatin immunoprecipitation coupled to high throughput sequencing (ChIP-Seq) experiments, inform our current view of chromatin organization. We also briefly describe experiments, as well as theoretical ideas, that suggest that non-equilibrium activity plays an important role in cellular organization.

Our work addresses the following observations of chromosome organization in metazoan nuclei, concentrating on human cell types: (i) chromosomes are territorial in nature, forming chromosome territories (CTs); (ii) euchromatin regions, which contain less condensed DNA and are mostly gene-rich, tend to be found towards the

nuclear interior; (iii) heterochromatin regions, which contain more condensed regions of DNA and are mostly gene-poor, tend to be found near the nuclear envelope; (iv) More active chromosomes, with larger transcriptional output, are rougher and more elliptical in shape than less active chromosomes; (v) active genes tend to locate toward the surface of CTs, whereas silenced genes tend to remain within CTs. Also, (vi) homologous chromosomes have similar properties, but; (vii) the two copies of X chromosomes in female cells tend to be positioned differently, with the inactive X chromosome occupying a more peripheral position than the active X chromosome. Finally, (viii) both size-dependent and activity-dependent radial positioning of chromosome have been described, although gene-density dependent positioning is most often seen in more spherical cells [Bickmore and van Steensel, 2013].

These are generic features, seen across cell types. We summarize previous models for these generic aspects of nuclear architecture. We then go on to emphasize the biophysical context: living cells are far from equilibrium, since they use chemical energy to drive active biological processes such as transport and metabolism. The biophysical consequences of these ATP-fueled active processes acting on chromatin have been ignored in all earlier models of large-scale nuclear architecture. Such active processes can be modeled via theories of “active matter” [Menon, 2010, Ganai et al., 2014]. Following standard approaches, such active processes are best described in terms of inhomogeneous, stochastic forces acting on chromatin, equivalent to a local “effective” temperature [Loi et al., 2011]. These ideas are at the core of our work in modeling large-scale nuclear architecture from first principles.

TFBS are generally characterized by short conserved patterns or motifs, commonly represented by position-weight-matrices (PWMs), a probabilistic representation where each position within a binding site is described by an independent categorical distribution over (A, C, G, T) nucleotides. At each base position of a TFBS, each nucleotide has a score that is proportional to the probability that it occurs. Multi-

plying these scores for each base of sequence yields a likelihood for observing that sequence under a given PWM model. PWMs are conveniently visualised using sequence logos. A PWM of a given TF is often used to scan regulatory sequences to identify potential TF binding sites. Most TFBS are small, usually 6-20 bases in length, and flexible.

ChIP-Seq is a method widely used for *in vivo* genome-wide identification of TFBS [Johnson et al., 2007]. In this method, *in vivo* DNA-protein complexes are crosslinked using formaldehyde, sonicated to break the DNA, and treated with a TF-specific antibody to precipitate the protein of interest. By then reversing the crosslinks, sequencing the DNA fragments and mapping them to a reference genome, a genome-wide map of TFBS with a resolution of 100-200 bp can be obtained.

Finding motifs in such large ChIP-Seq datasets is challenging. Most existing *ab initio* motif finding algorithms do not scale to large datasets. They also fail to report many motifs, such as those which are associated with cofactors or where the proteins being studied do not bind to DNA directly or are present only in a small fraction of sequences. We developed a program, THiCweed, which can address such questions [Agrawal et al., 2018b].

In **Chapter 2**, we describe our model for large-scale nuclear architecture in metazoans. This model provides a biophysical way of incorporating non-equilibrium activity, associated to the intensity of local transcriptional processes, into a polymer model for chromosomes. We base our study on 3 different approaches to incorporating activity, assigning activity based on gene density, gene expression as well as via a combined model that takes both gene density and gene expression into account.

In our model, chromosomes are described as polymers comprised of spherical monomers connected by non-linear springs. Our model chromosomes are dynamic and explore different configurations, based on the forces they experience. Such forces arise from the dense, non-equilibrium and fluctuating environment of the cell nucleoplasm, the

interactions of chromosomes and chromosome-nuclear envelope interactions [Ganai et al., 2014, Agrawal et al., 2017, Agrawal et al., 2018a]. Each monomer represents a coarse-graining to the 1Mb scale in the chromatin system. We apply overdamped Langevin dynamics to this system, but generalize this dynamics to account for active fluctuations. After associating activity to each monomer via an effective, monomer-dependent active temperature, our model adds a FENE bond potential between bonded monomers, a Gaussian repulsive potential between non-bonded monomers, specific long-range interactions coupling monomers inferred from Hi-C data, and a short-ranged Lennard-Jones potential between monomers and the simulated nuclear envelope. Our computational model then involves a system of 6086 monomers, divided across 46 polydisperse polymers confined to a hollow sphere. We simulate our model using the well-known LAMMPS package [Plimpton et al., 2007].

In our first approach, we associate inhomogenous activity within 1Mb segments to the number of genes contained in that segment. We term this the gene density model. Each such 1Mb segment maps to a monomer. Such 1 Mb segments, if they have a high gene density, are associated with a higher effective temperature. Monomers having a low gene density are associated with low effective temperatures. We take gene density data from the GENCODE database and examine a variety of ways of assigning active temperatures to monomers. The gene density model yields predictions in agreement with experiments on chromosome distribution functions. However, it cannot be generalized to examine cell-type specific variations in nuclear architecture. Our second approach assigns inhomogenous activity using gene expression data. We term this the gene expression model. Gene expression profiles vary across cell types and should provide a better reflection of cell-type-specific transcription levels. We used RNA-Seq data from the ENCODE project for 5 cell types, relating expression levels to FPKM values. We calculate the amount of gene expression associated with each 1 Mb interval of chromosome. We choose structured effective temperature assignments that reflect the overall shape of the gene expres-

sion curve. However, while incorporating gene expression led to some differences with the gene density model, we decided that an overall better description ought to combine features of both models. Accordingly, our most comprehensive simulations are for a third approach, for what we term the “combined model” . This model includes features of both the gene density and gene expression models.

In **Chapter 3**, we provide results for all three versions of our model. We calculate two central structural quantities for each chromosome. The first is the distribution function $S(R)$ of chromosome specific monomer densities, as a function of the radial distance from the centre of the nucleus. The second is the distribution function of the center-of-mass $S_{CM}(R)$ of a chromosome, plotted as a function of the distance from the nuclear centre. We obtain such distribution functions for all chromosomes across all 5 cell types we study. Some general features of our results are the following: (i) $S(R)$ of the gene-rich chromosome 19 peaks at a more interior location in comparison to the gene-poor chromosome 18, although both have similar sizes but different gene densities; (ii) $S(R)$ of chromosome 12 is similar to that for chromosome 20. Both these chromosomes have different sizes but similar gene densities; (iii) $S_{CM}(R)$ of chromosome 19 peaks at a smaller R in comparison to chromosome 18; (iv) For female cells, $S(R)$ for active and inactive X chromosomes peak at different locations, with the inactive X chromosome found at a more peripheral location than the active X chromosome; (v) When $S_{CM}(R)$ is plotted with respect to increasing order of gene-density per chromosome or increasing order of chromosome sizes, we can fit a straight line in each case for a majority of the chromosomes. These observations suggest that observations of both size-and gene-density dependent chromosome positioning can be reconciled, depending on which chromosomes are fit to this behaviour. Both gene density and gene expression models fit specific aspects of the experimental data. However, a model description which appears to provide the most comprehensive fits to the data combines features of both these models. None of these features are obtained in models that do not account for non-equilibrium activity.

For the combined model, we calculate the 3D shape of each chromosome, as well as the shapes projected onto a 2d plane. We calculate 5 quantities: (i) ellipticity, from 2d chromosome images, where our results reveal that gene-poor chromosomes tend to be more elliptical in shape than gene-rich chromosome; (ii) regularity from 2d chromosome images, where our results indicate that gene-poor chromosomes are more regular, that is less rough, than gene-rich chromosomes; (iii) volume overlap of a given chromosome with other chromosomes in 3D space, where our results show that gene-rich chromosomes have a high overlap of volume with other chromosomes; (iv) contact probability $P(s)$ of chromosome as a function of genomic distance s , which we find follows a power-law $1/s^\alpha$ with α varying between 0.9 to 1.5 for different chromosomes; (v) prolateness (Σ) and asphericity (Δ) of individual chromosomes that reveal specific features across different cell types. These should be measurable in experiments. Our broad results provide a general biophysical way of understanding the origins of a number of patterns in large-scale nuclear architecture that have been identified in experiments.

In **Chapter 4**, we introduce an algorithm for ChIP-Seq data analysis, THiCweed, that takes an approach of clustering by sequence similarity rather than motif-finding. THiCweed uses a divisive hierarchical clustering approach based on sequence similarity. THiCweed's approach is purely based on clustering rather than traditional motif finding, and the clustering is based on stringent statistical criteria. It is specially geared toward data containing a mixture of motifs, which present a challenge to traditional motif-finders.

We report the following key observations when we applied THiCweed to ENCODE ChIP-Seq peaks for various TFs and cell types: (i) known canonical motifs are recovered in most datasets; (ii) motif variants, secondary motifs, widely-spaced dimer motifs, and additional sequence features over length scale much larger than a typical TFBS are observed; (iii) certain motifs such as CTCF-like, JUN-like, ETS-like,

and THAP11-like motifs occur frequently in different ChIP-Seq datasets, some of which were previously observed in Ref [Hunt and Wasserman, 2014]; (iv) we also see biological significance of these clusters using other genomic features, such as phylogenetic conservation, nucleosome occupancy and DNase-seq data.

THiCweed has many advantages over other existing tools: (i) It provides both speed and accuracy in finding multiple motifs in large datasets; (ii) It does not require any prior information about the length of the motifs or number of motifs; (iii) It is very fast in speed; (iv) For the ENCODE data, it successfully recovers literature motifs and also uncovers complex sequence characteristics in flanking DNA. It also able to find variant motifs and secondary motifs which found in less than 5% of the total given data.

In **Chapter 5**, we conclude by discussing further aspects of large-scale nuclear architecture that more detailed calculations can address. We discuss how we can make our model more realistic, through the inclusion of chromatin-nuclear lamina interactions, by incorporating the presence of the nucleolus, as well as through the study of dynamical aspects, such as nuclear envelope fluctuations in stem cells as a consequence of chromatin fluctuations. We also discuss how the TFBS problem can be modified if we incorporate chromatin interacting information and how TF co-regulation occurs between different cell types.

List of Figures

1.1	Cell cycle stages	2
1.2	Different levels of interphase chromatin organization.	3
1.3	ATP dependent chromatin remodeling.	4
1.4	Activation of genes through interaction between promoter and enhancer.	6
1.5	The eukaryotic transcriptional machinery.	8
1.6	Role of cohesin in looping out the DNA sequence.	9
1.7	Inferring chromatin contacts through 3C-related techniques.	14
1.8	Genomic interactions, between promoter, enhancer and boundary elements in A and B compartments of genome	17
1.9	Root-mean squared end-to-end distance as a function of genomic distance s between the ends of a subchain.	23
2.1	A schematic of female (XX) diploid genome in a typical cell nuclei.	34
2.2	Plot of log gene density in increasing order of monomers.	36
2.3	Plot of log gene expression in increasing order of monomers.	38
2.4	Extraction and fitting of transcriptomics data.	40
2.5	Assignment of effective temperature to each monomer for the combined model.	41
2.6	Comparison of FENE and harmonic potential.	45
2.7	The Gaussian core potential.	47
2.8	Lennard-Jones potential.	49
2.9	Model predictions for large-scale features of nuclear architecture.	50
2.10	Distribution function of the DNA density distribution $S(R)$ or centre-of-mass distribution $S_{CM}(R)$, if they are randomly distributed inside the nucleus.	52

2.11	Schematic of the grid method for computing chromosome territory shape.	54
2.12	Shapes of individual CT and all CTs in a nucleus.	56
2.13	Comparison of chromosome territory shape from grid method and the 3d ellipsoid fit method.	57
2.14	Schematic illustrating a 2D projection of a three-dimensional CT.	60
3.1	$S(R)$, the radial distribution of the monomer density associated to each chromosome from the gene density model.	70
3.2	$S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome from the gene density model.	71
3.3	Relative centre of mass position of all chromosomes in the gene density model.	72
3.4	$S(R)$, the radial distribution of the monomer associated to each chromosome from the gene expression model for different fractions of active monomers in HeLa cell line.	73
3.5	$S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome from the gene expression model for different fractions of active monomers in HeLa cell line.	74
3.6	Relative centre of mass position of all chromosomes from the gene expression model for different fractions of active monomers in HeLa cell line.	75
3.7	$S(R)$, the radial distribution of the monomer density associated to each chromosome from the gene expression model for different active temperatures of active monomers in HeLa cell line.	76
3.8	$S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome from the gene expression model for different active temperatures of active monomers in HeLa cell line.	77
3.9	Relative centre of mass position of all chromosomes from the gene expression model for different active temperatures of active monomers in HeLa cell line.	78
3.10	$S(R)$ for all chromosomes, within the gene expression and the combined model for the GM12878 cell type.	80
3.11	$S_{CM}(R)$ for all chromosomes, within the gene expression and combined model for the GM12878 cell type.	81
3.12	Schematic of effective temperature assignment to each monomer for chromosome 12 for the combined model.	82

3.13	S(R) for all simulated chromosomes across GM12878, HMEC, HUVEC, IMR90 and NHEK cell types for the combined model.	83
3.14	$S_{CM}(R)$ for all simulated chromosome across GM12878, HMEC, HUVEC, IMR90 and NHEK cell types for the combined model.	84
3.15	S(R) for all chromosomes with different combinations of activity and loops for the GM12878 cell type.	85
3.16	$S_{CM}(R)$ for all chromosomes with different combinations of activity and loops for the GM12878 cell type.	86
3.17	S(R) and $S_{CM}(R)$ for all chromosomes with presence of activity and loops and absence of activity and loops for the GM12878 cell type. . .	87
3.18	Predictions for the mean centre of mass location for each chromosome, computed for the GM12878, HMEC, IMR90, NHEK, and HUVEC cell types plotted as a function of chromosome gene density.	88
3.19	Predictions for the mean centre of mass location for each chromosome, computed for the GM12878, HMEC, IMR90, NHEK, and HUVEC cell types plotted as a function of chromosome sizes.	90
3.20	The relative centre of mass position of each chromosome for different combinations of activity and loops in the GM12878 cell type.	92
3.21	S(R) of active and inactive monomers of all chromosomes for the GM12878 cell type.	96
3.22	$S_M(R)$ of specific tagged monomers for GM12878, HMEC, IMR90, NHEK, and HUVEC cell types.	97
3.23	Ellipticity and Regularity for each chromosome as predicted by the combined model.	98
3.24	Fractional volume and fractional surface area of each chromosome for GM12878 and IMR90 cell types as predicted by grid method.	99
3.25	Summed volume overlap of chromosomes in GM12878 and IMR90 cell types as predicted by grid method.	99
3.26	Histogram of fractional volume and fractional surface area obtained using 3d ellipsoid fit method.	100
3.27	Fractional volume and fractional surface area for each chromosome across GM12878, HMEC, HUVEC, IMR90 and NHEK cell types as predicted by ellipsoid fit method.	101
3.28	Fractional volume and fractional surface area for different combination of activity and loops using the ellipsoid fit method.	101
3.29	S(R), for the Xi and Xa chromosome as obtained from simulations across GM12878, HMEC, IMR90, NHEK, and HUVEC cell types. . .	103

3.30	$S_{CM}(R)$ of the Xa and Xi chromosome as obtained from simulation across GM12878, HMEC, IMR90, NHEK, and HUVEC cell types. . .	103
3.31	Contact probability $P(s)$ vs s , for the active and inactive X chromosomes.	105
3.32	Contact probability $P(s)$ as a function of genomic distance for chromosome 1, fit to a power law in the range of 1-15 MB.	106
3.33	Spatial distance between monomers of each chromosome for GM12878 cell type.	107
3.34	Calculated average values of the prolateness parameter versus the asphericity parameter for each chromosome across different cell types and models.	109
3.35	Heatmap of mean distances between monomers, the distance map, and contact maps for GM12878, HMEC, HUVEC, IMR90, and NHEK cell types.	111
4.1	Difference between ChIP-seq and ChIP-exo workflow.	117
4.2	Consensus model of TF-DNA binding.	120
4.3	Position weight matrix (or frequency matrix) and sequence logo representation is shown for SP2 TF.	121
4.4	Flowchart of the THiCweed algorithm.	124
4.5	Flowchart of the THiCweed algorithm.	125
4.6	Performance of THiCweed on synthetic datasets for various value of T and r	130
4.7	Synthetic motifs and comparison of performance with other tools. . .	133
4.8	Running time of various programs as the size of the dataset.	135
4.9	Motifs that occur across multiple ChIP-Seq datasets.	137
4.10	Sample THiCweed output on four ChIP-Seq IRF1, NFYA, REST and FOXA1 datasets.	139
4.11	Comparison of clustering of 2,019 peaks for SP2 by various programs.	140
4.12	Biological relevance of sequence clusters.	141

List of Tables

2.1	Length of chromosomes in 1 Mb coarse-grained unit	33
3.1	Overview of our models	69
3.2	χ^2 and p-value of relative center of mass position to least square fits .	89
3.3	Different Model comparison	95
4.1	Commandline options for various tools	132

Chapter 1

Introduction

In eukaryotic cells, the term chromatin describes DNA as found *in vivo*, where it binds to scaffolding and other DNA-binding proteins as well as specific short RNA sequences. DNA in such cells is present in chromosomes, each a single molecule of DNA which is tightly packaged as chromatin. The total length of DNA in a human cell is about 2 meters, divided across 46 chromosomes in somatic cells, and confined to a nucleus whose radius is typically a few microns.

Each cell must grow and reproduce. This happens in a cyclic manner called the cell division cycle, shown in Figure 1.1. Chromatin is found in two major states, interphase and mitotic, across the cell cycle. Interphase is the longest phase of the cell cycle. In this phase, cells grow through G1, S, and G2 stages. Chromatin in interphase is less condensed, compared to chromatin in the mitotic stage, and acquires a cell type-specific spatial organization. In interphase, different regions of the chromosome become more compact or expanded, depending on whether access to those regions is required for a particular cell type at a given time [Belmont, 2002, Naumova et al., 2013]. Chromatin condensation begins during prophase and chromosomes remain condensed throughout the various stages of mitosis (prophase to telophase). During cell division, chromatin undergoes extensive spatial reorganization.

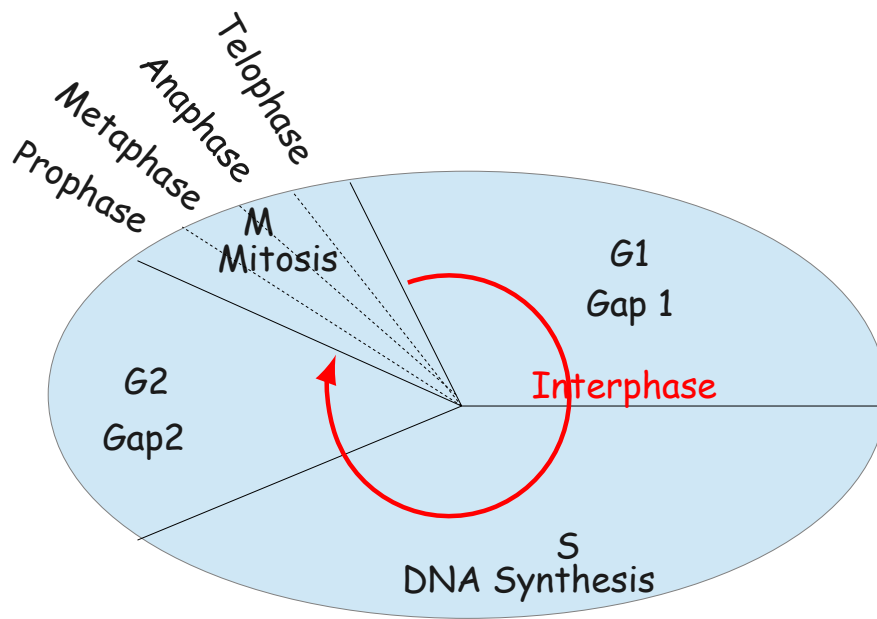


Figure 1.1: **Cell cycle stages.** Fraction of the cell cycle devoted to each of the stages. The interphase stage consists of G1, S, and G2 phase. In the G1 phase, the cell is metabolically active, duplicates organelles and cytosolic components and starts replicating centrosomes. In G2 phase, cell growth continues, enzymes and other proteins are synthesized and replication of centrosomes is completed. In S phase, DNA is replicated. The mitotic phase is comprised of mitosis and cytokinesis, where the cell nucleus divide into two separate nuclei through the stages of prophase, metaphase, anaphase and telophase.

Biologists and biophysicists have studied chromatin for at least the past hundred years, but significant improvements in our understanding have been made in the past three decades. This has mainly been due to advances in imaging and biochemical techniques, coupled to image and statistical analysis tools. We know now that the spatial organization of chromosomes is non-random. Each chromosome in the cell nucleus occupies a discrete region referred to a chromosome territory (CT) [Meaburn and Misteli, 2007]. Theodor Boveri introduced this term in 1909, while studying the interphase nuclei of *Ascaris megalocephala* worms [Cremer and Cremer, 2010]. Further evidence for complex nuclear organization came from electron microscopy, which reveals the presence of heterochromatin (dark-staining) and euchromatin (light-staining regions) in interphase chromatin [Straub, 2003, Sati and Cavalli, 2017].

Stretches of the genome containing relatively open DNA conformations usually contain actively transcribed genes. They are usually associated to euchromatin, as shown in Figure 1.2. In contrast, relatively compact heterochromatin regions, where DNA is tightly bound, are typically associated to gene-poor or non-coding regions of the genome.

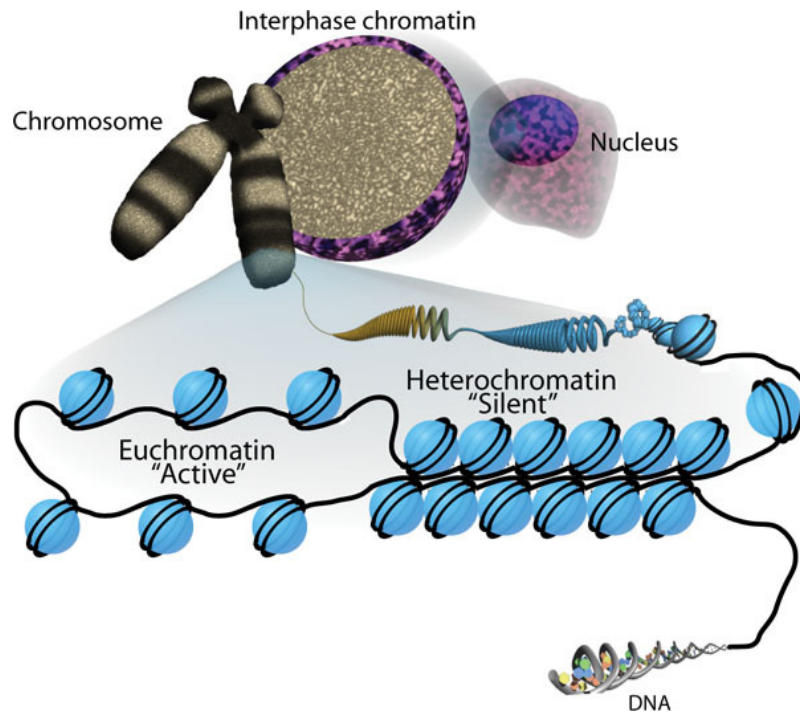


Figure 1.2: **Different levels of interphase chromatin organization.** The basic unit of organization is the nucleosome which is further organized into higher order structures. The level of packaging influences biological function. Euchromatin and heterochromatin are associated with open and compact chromatin regions respectively. This figure is reproduced from Figure 1 of NCBI bookshelf [Sha and Boyer, 2009] under Creative Commons Attribution License CC BY.

1.1 Chromatin Organization

A representative sequence of DNA in humans, the human genome, was mapped about 2 decades ago. It contains about three billion base pairs. The total content of DNA in a somatic human cell is divided into 46 chromosomes, consisting of 22 autosomal pairs, and 2 sex chromosomes that determine whether an individual is a

female (XX) or male (XY). DNA is made up of 4 types of nucleotides – A, C, G and T. Compacting DNA efficiently is an important requirement for chromatin, enabling DNA to be accommodated in a relatively small cell nucleus, while still being able to perform cellular functions.

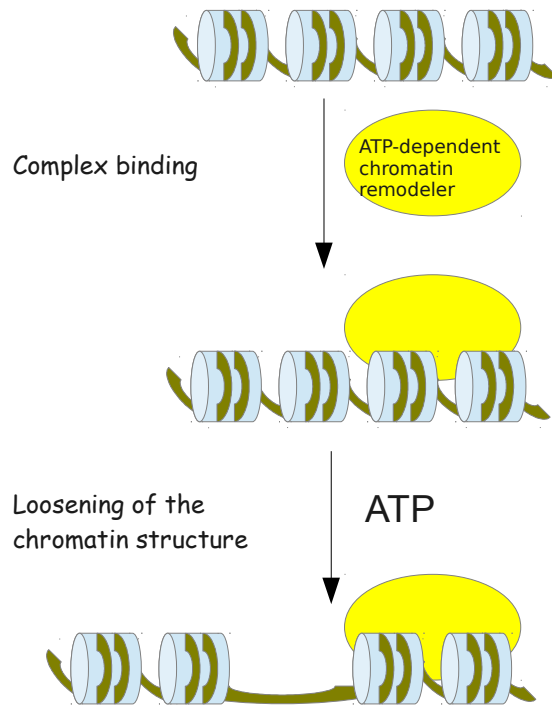


Figure 1.3: **ATP dependent chromatin remodeling.** Remodeling complexes, mainly SWI-SNF, ISWI and CHD, with specific binding domains for ATP, attach to chromatin. With the addition of ATP, the conformation of nucleosomes changes as interactions between histone and DNA are altered. Details of this mechanism are available at Ref. [Vignali et al., 2000].

DNA is condensed using histone proteins. These proteins exist together with DNA in each individual cell nucleus. Figure 1.2 shows different levels of packaging in DNA for a interphase chromosome. When positively charged histone proteins interact with negatively charged DNA, a structure called the nucleosome is formed. The nucleosome is the first level of packaging, where ~ 147 bps of DNA wrap around histone protein octamers. Consecutive nucleosomes are connected by linker DNA and form an ~ 11 nm chromatin fiber, described as a "beads-on-a-string" structure as shown in Figure 1.2. The length of linker DNA fluctuates between ~ 20 -90 bp and varies among cell types, tissues and species. The second level of organization can be

described by a ~ 30 nm chromatin fiber, although there is no complete consensus on what form it takes. Various proposals exist for its structure. These include three different models, the solenoid model, the helical ribbon model and cross-linker model [Wu et al., 2007]. The nature of still higher levels of organization remains somewhat unclear.

Chromatin structure must be dynamic, since transcription, DNA replication, and DNA repair all require rearrangements of chromatin structure. For these activities, biological machinery needs to access specific DNA sequences. The packaging of DNA in nucleosomes reduces sequence accessibility. To address this problem, the cell uses chromatin remodelling to expose key regions of the DNA required by transcriptional apparatus and genome maintenance machinery [Armstrong, 2013].

Chromatin remodelling is facilitated by two principal processes, histone modification and ATP dependent chromatin remodelling. Histone modification complexes post-translationally modify the N-terminal histone tails to alter the structure of chromatin, thus provide binding sites for regulatory proteins. ATP-dependent chromatin remodelling complexes use the energy of ATP hydrolysis to disrupt nucleosome DNA contacts, move nucleosomes along DNA, and remove or exchange nucleosomes (an example of loosening of chromatin structure can be seen in Figure 1.3)[Hargreaves and Crabtree, 2011]. In general, these processes are reversible, so modified or remodelled chromatin can return to its compact state, once transcription and/or replication are complete.

1.2 Gene Regulation

Genes are stretches of DNA that encode instructions for synthesis of proteins through RNA. Messenger RNA (mRNA) is produced from the DNA template, by a mechanism called transcription. At any given time, the amount of a particular protein

in a cell is determined by synthesis and degradation of mRNA. The function of the cell is reflected by the amount and types of functional RNA transcript it contains.

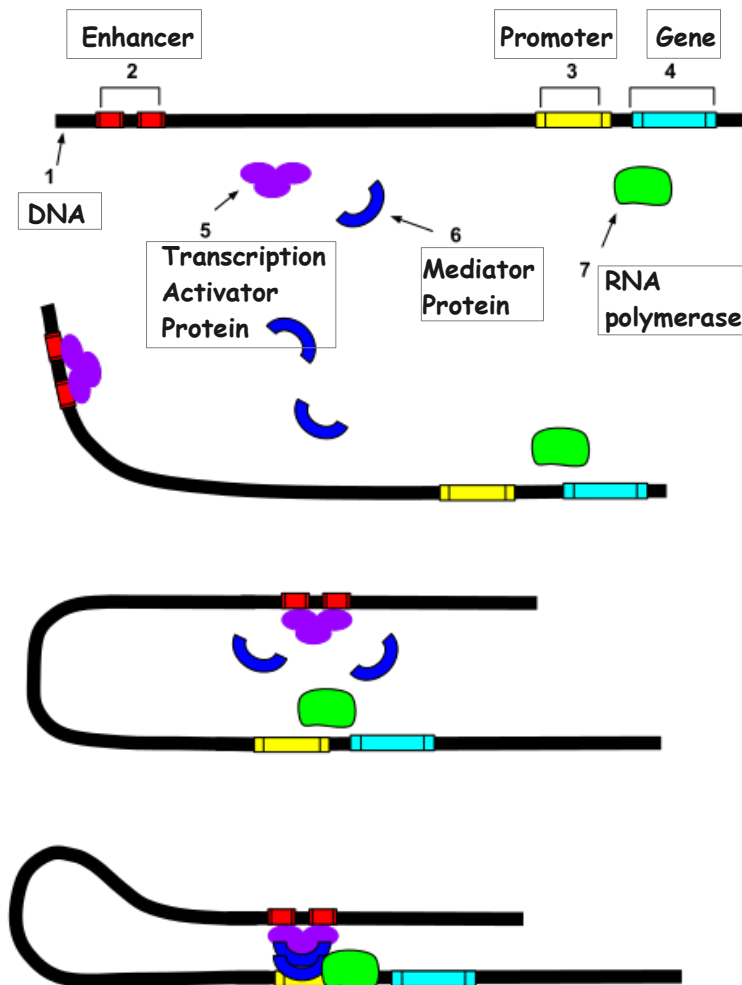


Figure 1.4: **Activation of genes through interaction between promoter and enhancer.** An activator protein (TFs) bind to DNA at an upstream enhancer sequence can attract proteins to the promoter region that activate RNA polymerase and initiate transcription. The DNA can loop around on itself to cause this interaction between an activator protein and other proteins (mediator) that mediate the activity of RNA polymerase. (Figure from Wikipedia, reproduced under Creative Commons CC BY-SA 4.0).

The expression of eukaryotic genes can be regulated at several steps. These steps include transcription initiation and elongation, mRNA processing, transport, translation and through the control of RNA stability. A good fraction of regulation occurs at the transcriptional initiation level. Transcription is performed by a pro-

tein complex called RNA polymerase. Genes transcribed by RNA polymerase can be controlled by two types of cis-acting regulatory sequences: (i) a promoter (composed of a core promoter and nearby proximal regions) which are near upstream regions on the genes where transcription is initiated, and (ii) distal regulatory sequences, for example enhancers, silencers, insulators etc. A schematic representation of such regulatory interaction is shown in Figure 1.4.

Transcription can be switched on or off when specific DNA-binding proteins bind to these regulatory sequences. It is regulated by such sequence-specific DNA-binding proteins, known as transcription factors (TFs). The term TF has been used to describe any protein involved in transcription and/or capable of altering gene-expression levels in both prokaryotic and eukaryotic cells. TFs regulate cellular processes either by binding to DNA sequence, either directly or via cofactor interactions.

TFs can be classified into three groups: general transcription factors (GTF), activators and coactivators. RNA polymerase requires the presence of GTF to recognize the transcription start sites (TSS) of a protein-coding gene. GTF includes a variety of protein complexes, for example: RNA polymerase itself, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. In addition to GTFs, *in vivo* transcription also requires a highly conserved, large protein complex called Mediator [Maston et al., 2006].

GTFs assemble on the core promoter, which includes the TSS as well as on other binding sites recognized by different subunits of the GTFs. RNA polymerase binds to a complex of GTFs and assemble on the core promoter in an ordered manner to further form a complex, called the preinitiation complex (PIC). A symbolic representation of the eukaryotic transcriptional machinery is shown in Figure 1.5. The assembly of a PIC on the core promoter initiates transcription at only a low level. Activators are sequence-specific DNA binding proteins, whose binding sites are present in sequences upstream of core promoters. These are responsible for enhanced

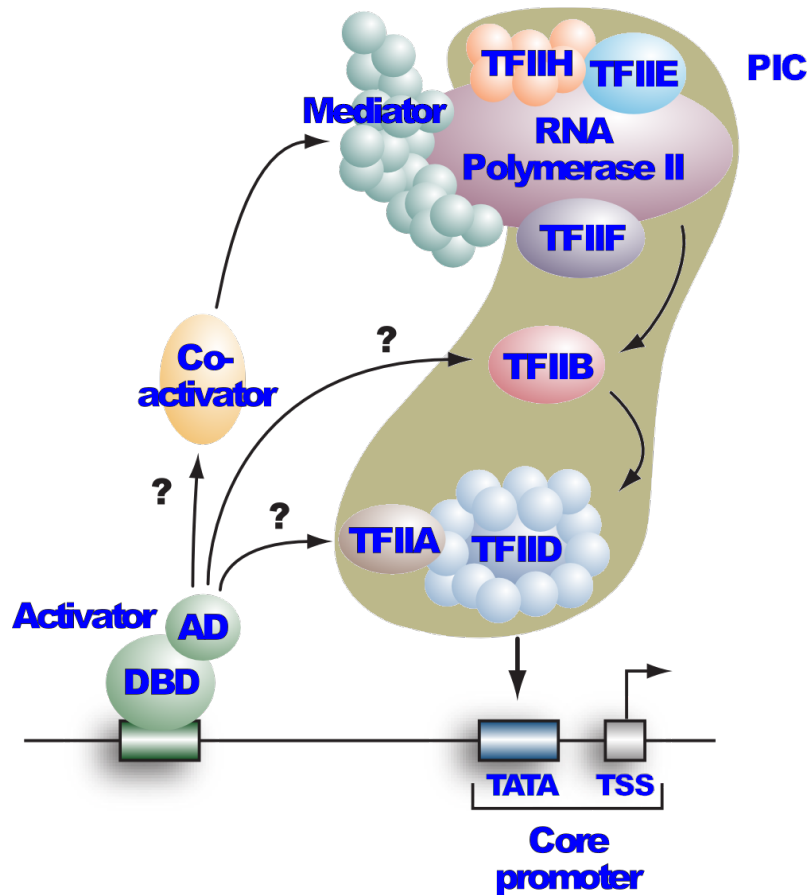


Figure 1.5: **The eukaryotic transcriptional machinery.** Transcriptional activity is greatly stimulated by activators, which bind to upstream regulatory sequence. Activators consist of a DNA-binding domain (DBD) and a separable activation domain (AD) that is required for activator to stimulate transcription. The DNA-binding sites for activators also called transcription factor-binding sites (TFBSs) are generally small, in the range of 6-12 bp. Figure by Sara Deibler from University of Massachusetts Medical School, used and modified with permission.

transcriptional activity. Activators increase PIC formation either through direct or indirect (co-activators) interactions with one or more component of transcriptional machinery. The DNA-binding sites for activators, also called transcription factor-binding sites (TFBSs), are generally short, in the range of 6-12 bp. The TFBS of a specific activator are generally described by a consensus sequence. Certain positions in TFBS are relatively constrained while others are more variable [Maston et al., 2006].

The main players of regulating PIC are classified as cis-acting and trans-acting

elements. One kind of cis-acting element are enhancers, which interact with the promoter and can be located up to more than 1 Mbp away from the gene without affecting genes that are closer in distance along the DNA sequence. Enhancers do not necessarily act on the closest promoter but can bypass neighbouring genes to regulate genes located more distantly along a chromosome. These interactions involve formation of chromatin loops, where enhancers physically contact the promoter regions by looping out the intervening DNA sequence, a process mediated by CTCF, cohesin and mediator [Spielmann et al., 2018]. This is illustrated in Figure 1.6.

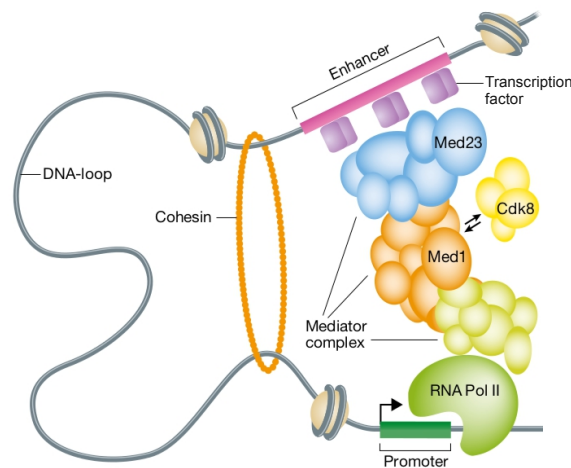


Figure 1.6: **Role of cohesin in looping out the DNA sequence.** Cohesin looping out the DNA-sequence to bring enhancer and promoter regions in close spatial proximity. Figure reproduced with permission from [Streubel and Bracken, 2015] EMBO.

TFs are involved in many functions. They control, among other processes, the processes that specify cell types, developmental patterning and controlling pathways like immune response [Lambert et al., 2018]. Mutations in TFs and TFBS are often associated with human diseases. The protein sequence of TFs, regulatory regions they bind to, and physiological roles are often conserved among metazoans, suggesting that global gene regulatory networks may also be conserved. TFs evolve over longer timescales and so may get duplicated and diverge. It is seen that the same TF can regulate different genes in different cell types, indicating that regulatory

networks are dynamic [Lambert et al., 2018]. How TFs are assembled and how they recognize binding sites and control transcription are key to understanding cell-type specific gene regulation.

1.3 Experimental Approaches to Chromatin Organization

Chromatin structure undergoes a cycle of condensation and decondensation as cells divide [Dekker, 2014]. In mitotic chromosomes, each chromosome is organized into a recognizable X-shaped structure and each chromosome can be identified separately through optical means. In interphase, genomes are largely decondensed but their organization into distinct chromosome territories can be inferred [Cremer and Cremer, 2001].

Chromatin organization in eukaryotes has been investigated by two extensively used experimental approaches. These are DNA fluorescence in situ hybridization (FISH) and chromosome conformation capture (3C) based methods. The development of chromosome paints that consist of fluorescent-dye labeled chromosome-specific probes enabled direct visualization of individual chromosomes in interphase nuclei [Fritz et al., 2016]. FISH usually involves fixation and permeabilization of cells, followed by hybridization of fluorescently labeled DNA probes to specific loci, but can also be applied to live cells. It is able to track the location of a few tagged loci (or even entire chromatin regions) of chromosomes.

3C (or similar Hi-C) based methods measure the frequency of ligation between DNA fragments that are in proximity and can be cross-linked. Such methods enable the detection of physical proximity between multiple genomic loci (and eventually across the whole genome) simultaneously [Williamson et al., 2014]. In 3C-

based approaches, crosslinking probabilities of chromosome contacts are generally cell population-averaged. On the other hand, FISH is measured at the single cell level, and can be used to obtain a 3D distance between a small number of genomic loci [Giorgetti and Heard, 2016].

3C based technologies have revolutionized the current view of the genome. DNA FISH, which was once a state-of-the-art technique, is now considered an accessory tool to validate 3C based predictions [Giorgetti and Heard, 2016]. The key observations from these two methods are described below.

1.3.1 Key Observations From FISH

All metazoan chromosomes are arranged in a non-random manner. Their territorial organization in interphase constitutes a basic feature of nuclear architecture [Cremer and Cremer, 2001, Meaburn and Misteli, 2007]. Genes are arranged in non-random positions within CTs [Meshorer and Misteli, 2006]. The preferential positions of genes are functionally important because they vary across tissues, during development and across cell types [Meshorer and Misteli, 2006].

In many cell types, gene-poor and late-replicating chromatin is localized close to the nuclear periphery, whereas gene-rich and early-replicating chromatin are located towards the nuclear centre [Croft et al., 1999, Boyle et al., 2001]. In human cells, the gene-rich chromosome 19, containing a large number of housekeeping genes, is distributed more centrally across several cell types than the similarly sized but gene-poor chromosome 18 [Croft et al., 1999, Boyle et al., 2001]. This observation has been suggested to generalize to a gene-density dependent radial organization for all chromosomes. It observed that active alleles are found more internally located compared to the inactive alleles within the same nucleus [Takizawa et al., 2008]. Heterochromatin regions in chromosomes are generally found at the nuclear

periphery or around nucleoli [Bickmore, 2013].

In non-human primates, gene-density dependent radial positioning of chromosomes has been found to be conserved irrespective of karyotype [Tanabe et al., 2002]. In one exceptional case, in the nuclei of rod photoreceptors of nocturnal animals, the conventional radial arrangements of euchromatin and heterochromatin have been found to be inverted, such that heterochromatin occupies the nuclear centre and euchromatin is found towards the nuclear periphery [Solovei et al., 2009]. For flatter nuclei, a chromosome size based radial positioning scheme has been suggested [Bridger et al., 2000, Bolzer et al., 2005]. Gene-rich chromosomes are preferentially located at the interior of the nucleus, though the preference disappeared after inhibition of transcription, suggesting that chromosome positioning may depend on transcriptional activity [Kalmárová et al., 2007].

In human cell nuclei within interphase, chromosome territories intermingle with one another, mostly at the boundaries. Intra-chromosomal interactions favour chromosome discreteness whereas inter-chromosome interactions favour intermingling. Such interactions are cell type specific and likely depend on the transcriptional activity of the loci [Branco and Pombo, 2007]. Any specific organization of chromosomal neighbourhoods is not apparent in the early G1 phase of cell division, but daughter cells eventually reestablish the general chromosomal organization pattern found in their mother's nuclei. This suggests that an active mechanism could play a role in establishing chromosomal neighbourhoods [Essers et al., 2005].

High-resolution 3D FISH with probes against transcripts can detect long-range genomic interactions, referred to as large multi-Mb chromatin loops. These loops represent a direct physical interaction between promoters and enhancers [Osborne et al., 2004]. Lamina-associated domains (LADs), identified by the DamID technique, cover approximately 40% of gene-poor regions of the genome. LADs are associated with low levels of gene expression and mostly found towards the nuclear

periphery [Guelen et al., 2008, Bickmore, 2013]. Analysis of single cells by FISH indicates that not all LADs mapped in cell populations are found at the nuclear periphery in all cells. This suggests that LADs constitute domains that dynamically anchor to, or detach from, the nuclear lamina [Briand et al., 2018].

Human ribosomal genes found in the acrocentric chromosomes 13, 14, 15, 21 and 22 are organized at particular chromosomal sites in clusters termed nucleolus organizer regions (NORs). Nucleoli in the cell nucleus are surrounded by a heterochromatin layer. This layer is roughly similar in appearance to heterochromatin found adjacent to the nuclear lamina [van Steensel and Belmont, 2017]. Late-replicating chromatin is distributed at both the nuclear periphery and around nucleoli [Bickmore, 2013]. Nucleolus-associated domains (NADs) are obtained from a genome-wide identification of DNA sequences associated with nucleoli [Németh et al., 2010]. NADs partially overlap with LADs. NADs are indeed found to be located near nuclear lamina in a subset of cells [van Steensel and Belmont, 2017]. Some regions of chromosomes associated to the nucleolus in mother cell can be repositioned to the nuclear periphery in the daughter cells. Thus, at least a subset of LADs is variably positioned at either the nuclear lamina or in close association with nucleoli [van Steensel and Belmont, 2017].

The spatial organization of chromatin is related to cellular processes such as replication, transcription, splicing and DNA repair. Several studies show that the spatial organization of chromosomes, centrosomes, centromeres, heterochromatin structure, nuclear lamina and nuclear speckles, change during the cell differentiation process and across development [Bártová et al., 2008, Meshorer and Misteli, 2006].

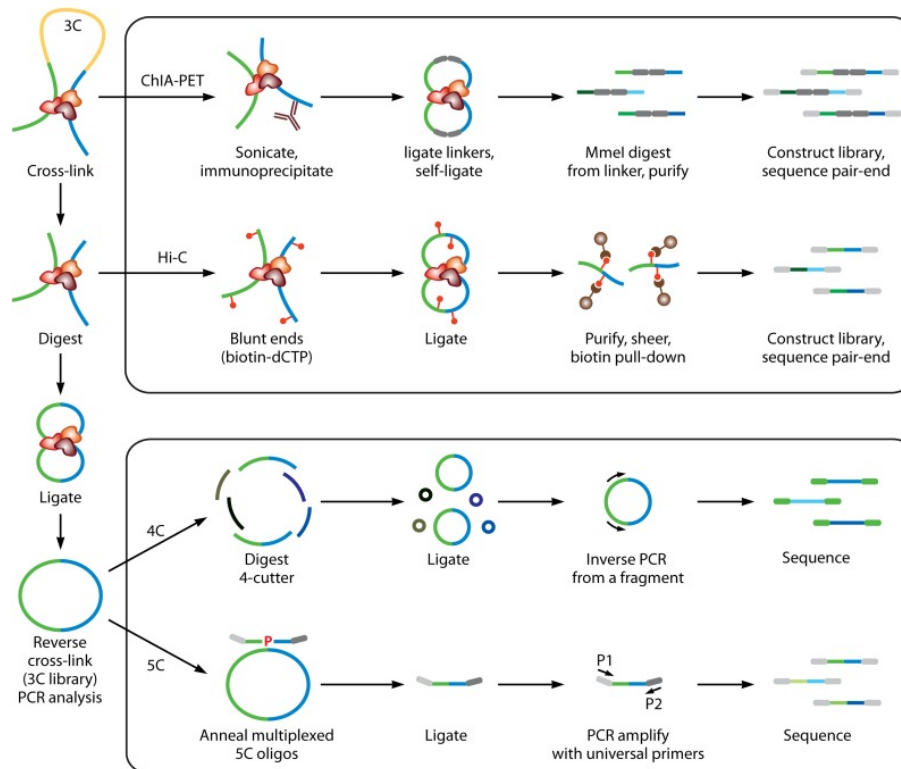


Figure 1.7: **Inferring chromatin contacts through 3C-related techniques.** The 3C method is shown on the left. Crosslinking with formaldehyde captures the interactions between chromatin (blue and green) segments contacted by protein complexes. The chromatin fragments are digested with a restriction enzyme. The free DNA ends are ligated under low DNA concentration such that ligation between non-crosslinked DNA is less likely. The genome-wide 4C, 5C, ChIA-PET and Hi-C techniques are shown on the right. ChIA-PET includes a chromatin immunoprecipitation (ChIP) step tagged by antibodies that enriches for those chromatin interactions which are mediated by specific protein. In Hi-C method, after digested with a restriction enzyme, DNA ends are marked with a biotinylated nucleotide (red dots), DNA in the crosslinked complexes are ligated to form chimeric DNA molecule, external biotin is removed from the ends of linear fragments, molecules are fragmented by shearing, internal biotin are pulled down with streptavidin (brown) magnetic beads and then quantification of chromatin interactions is achieved through massive parallel deep sequencing. The 4C method involves a second ligation step, to create self-circularized short DNA ligation products between a specific restriction fragment (the bait represent green arrows) and the rest of the genome. Inverse PCR is then used to amplify the sequence ligated to it. For 5C, computationally designed primers for the restriction site of each fragment used during the ligation-mediated amplification step are illustrated with green and blue lines, where the light and dark gray moieties represent universal primer sequences. Figure reprinted with permission from [Fraser et al., 2015] ©(2015) American Society for Microbiology.

1.3.2 Key Observations From 3C-based Methods

The quantification of long-range interactions between spatially proximal pairs of loci can be performed by different 3C-based methods. The common steps of 3C-based methods are the following: (i) Cells are cross-linked with formaldehyde; (ii) Chromatin is fragmented by restriction enzymes or sonication; (iii) Crosslinked fragments are ligated at low DNA concentrations; (iv) Such unique DNA junctions are quantified and analyzed [Fraser et al., 2015]. Different 3C-based methods differ in the last 2 steps.

While 3C, 4C and 5C approaches are unsuitable for finding genome-wide chromatin interactions, Hi-C is a genome-wide chromosome conformation capture method which allows unbiased identification of chromatin interactions [Dekker et al., 2013, Lieberman-Aiden et al., 2009]. Other genome-wide 3C methods are tethered conformation capture (TCC) [Kalhor et al., 2011], single-cell Hi-C [Nagano et al., 2013], ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) [Fullwood et al., 2009] and in situ Hi-C [Rao et al., 2014]. The key steps of these methods are shown in Figure 1.7. TCC observations indicate the noise from random inter-chromosomal ligations is considerably lower than Hi-C [Kalhor et al., 2011]. ChIA-PET is used for the *de novo* detection of global chromatin interactions arising from DNA-binding proteins.

Different 3C-based methods reproduce the observations that intra-chromosomal interactions are significantly higher than inter-chromosomal interaction. They confirm that chromosome positions are non-random, and that each individual chromosome is localized in spatially distinct volumes known as CTs [Dekker, 2014]. In lymphoblasts, larger chromosomes frequently interact with other larger chromosomes and are found in at the nuclear periphery while smaller gene-dense chromosomes interact preferentially with each other and are located more internally in the nucleus [Dekker, 2014].

1.4 Nuclear Subcompartments

Genome-wide Hi-C data suggest chromosomes are made up of large chromatin domains several Mb in sizes called compartments. Compartment A contains transcriptionally active, gene-rich, open chromatin domains while compartment B contains transcriptionally silent, gene-poor, closed chromatin domains [Lieberman-Aiden et al., 2009]. Among CTs, inter-chromosomal interactions are often between A compartment or, less frequently, between B compartments but rarely between A and B compartment [Lieberman-Aiden et al., 2009, Kalhor et al., 2011, Dekker, 2014]. Compartmentalization varies between cell types and across development [Dixon et al., 2016]. The DamID technique indicates that inactive B-compartments often reflect the clustering of loci at the nuclear lamina and nucleoli, consistent with these measurements [Gibcus and Dekker, 2013].

In a given cell population, chromatin contacts are observed with a wide range of frequencies, suggesting they may be present only in fractions of cells. This means that contact data describe an average over contacts of various genome structures in different cells. In a given population of cell conformations, only about 20% of contacts are shared between any two conformations [Kalhor et al., 2011]. Inter-chromosomal contact probabilities between pairs of chromosomes which are small and gene-rich (chromosomes 16,17,19,20,21 and 22) show that they preferentially interact with each other [Lieberman-Aiden et al., 2009, Kalhor et al., 2011] and that these chromosomes frequently colocalize in the centre of the nucleus. TCC data reveal that chromosome 19 is located closer to the centre of the nucleus, and that chromosome 18 is found more often towards the periphery. These observations are consistent with FISH, leading us to infer that chromosomes may be positioned by both gene density and by size [Kalhor et al., 2011].

Both A and B compartments are further composed of smaller domains. From two-dimensional interaction matrices at bin sizes less than 100 kb, highly self-interacting

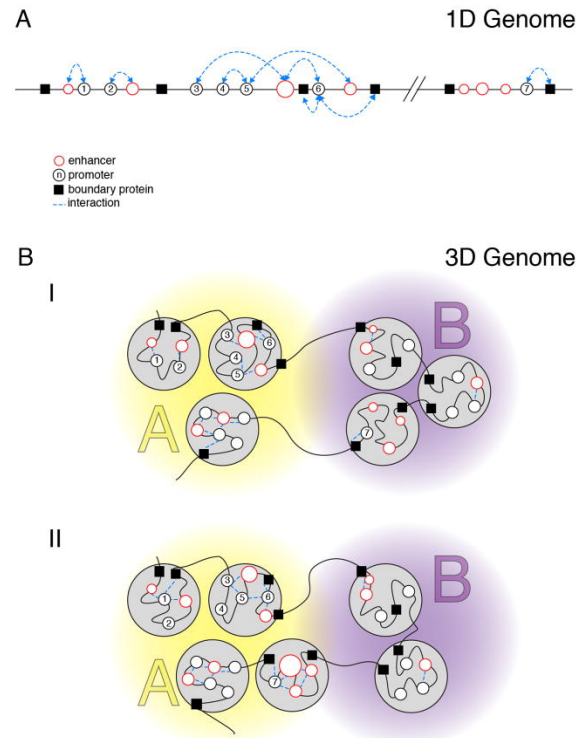


Figure 1.8: **Genomic interactions, between promoter, enhancer and boundary elements (for example CTCF) in A and B compartments of genome.** Genomic element interaction between promoters (black circle), enhancers (red circle) and architectural boundary proteins (black squares) are shown using a linear representation in Figure A and 3D representation in Figure B. The size of enhancers indicates the strength of their activity. Interactions relevant to gene expression are shown as dotted blue lines. In Figure B the interactions are largely confined to TADs (grey circles). TADs containing similar activities are arranged in same compartments (A or B). Altered gene expression due to change in promoter and enhancer interaction can lead to changes in the compartment but does not lead to change in TAD organization. Figure reprinted with permission from [Gibcus and Dekker, 2013] ©(license 4450740026637, October 16, 2018) Elsevier.

regions are found to emerge. These are difficult to identify by microscopy. Such contiguous, frequently interacting regions are called topologically associating domains (TADs). The typical organization of genomic elements in A/B compartments are shown in Figure 1.8. TADs range in size from several hundred kilobases to a few megabases. They appear to be the fundamental domain organization of chromatin and are found across cell types and across species [Dixon et al., 2012, Smith et al., 2016]. TADs are generally arranged hierarchically, with the hierarchy including several levels of smaller contact domains separated by weaker boundaries [Razin and

[Gavrilov, 2018](#)]. Loci located in adjacent TADs interact much less frequently than those within the same TAD, suggesting that TAD boundaries act as physical insulators [[Smith et al., 2016](#)]. These regions are bounded by narrow segments such that beyond them no chromatin interaction occurs. TAD boundaries are enriched for genomic features including promoters, insulator binding protein CTCF, housekeeping genes, tRNA and short interspersed element (SINE) retrotransposon elements [[Dixon et al., 2012](#)].

TADs and TADs inside the A or B compartment frequently interact with other TADs of the same compartment, but rarely with other TADs of the different compartment. The organization of TADs changes across different cell-types. The architectural proteins, cohesin, and CTCF play a crucial role in chromatin organization during interphase. They are believed to be colocalized at TAD boundaries [[Schwarzer et al., 2017](#)]. A recently proposed loop extrusion model explains a possible mechanism of TAD formation. In this model, CTCF-cohesin complex promotes extrusion of DNA through a cohesin ring until it reaches a pair of CTCF molecules in convergent orientation, where it can be retained until it dissociates [[Sanborn et al., 2015](#), [Fudenberg et al., 2016](#)].

1.5 Contact Probability

Population-based Hi-C data suggest that the averaged probability of contacts between a pair of loci on intra-chromosome decreases as the inverse of their genomic distance s raised to a power, i.e. $P(s) \sim 1/s^\alpha$ [[Lieberman-Aiden et al., 2009](#)]. The power-law scaling predicted in the fractal globule polymer model of chromatin contacts is consistent with Hi-C contact at length scales from several hundred kilobases to several megabases. The scaling of the contact probability in active compartments differs from that in passive ones [[Kalhor et al., 2011](#)]. The equilibrium globule model

predicts a value of $\alpha = 1.5$ whereas the fractal globule predicts a value $\alpha = 1$. This is to be compared to the experimental value of $\alpha \approx 1.08$ in the range between ~ 500 Kb and ~ 7 Mb [Lieberman-Aiden et al., 2009]. Further studies showed that chromosome X and 19 deviate from the fractal globule predictions significantly, with $\alpha \sim 0.93$ for the active chromosome X and $\alpha \sim 1.3$ for chromosome 19 [Barbieri et al., 2012]. The loop extrusion model predicts an exponent $\alpha \sim 1.27$ between ~ 300 Kb and ~ 3 Mb [Sanborn et al., 2015].

Studies of the scaling of $P(s)$ in other species provides an exponent $\alpha = 1.5$ for yeast [Duan et al., 2010], with $\alpha = 0.85$ for active domains and $\alpha = 0.7$ for repressive domains in Drosophila [Sexton et al., 2012]. Other studies found that contact frequency for chromatin is inversely proportional to the 4th power of the mean spatial distance [Wang et al., 2016]. Chromosome folding deviates from ideal fractal-globule model at larger length scales (several megabases) and the mean spatial distance follows the power law exponent ~ 0.17 for chromosome 20 and 22, with exponent ~ 0.074 for chromosome Xi and with exponent ~ 0.22 for chromosome Xa [Wang et al., 2016]. However, in another study scaling curves between genomic distance and contact probability were found to be organism specific. The calculated exponent also depended on parameters such as resolution and sequencing depth settings, even for the same organism [Ay et al., 2014].

1.6 Single Cell and Cell Type-specific Features

Despite some cell-to-cell variation in chromatin interacting regions, the scaling of the interaction probability with genomic distance is largely consistent between individual cells and also agrees with population data [Dekker and Mirny, 2013]. In oocyte cells, the value of $\alpha = 1.5$ for genomic separation $s > 1$ Mb, is consistent between individual cells. However, this exponent is significantly different from that

seen in interphase cells. The reason may lie in the large nuclei of oocytes [Flyamer et al., 2017].

There are thus three mechanisms, operative at different length scales, which largely define different cell-type specific features of chromatin structure. First, at the largest length scales, patterns of compartments are cell-type specific and correlate with chromatin state [Lieberman-Aiden et al., 2009]. Second, at intermediate length scale, studies reveal that large fraction of TADs (60-70%) are largely tissue invariant and highly conserved across species [Spielmann et al., 2018]. The regions within each TAD are dynamics and potentially take part in cell type-specific regulatory events [Dixon et al., 2012]. This led to a conclusion that TADs might be fundamental building blocks of chromosomes [Nora et al., 2012, Dixon et al., 2012]. Third, at still smaller length scale, chromatin loops significantly enrich for promoter-enhancer interactions. These are mostly cell-type dependent and such looping interactions directly related to structural differentiation within TADs [Dixon et al., 2016].

1.7 Theoretical Models of Chromatin Fiber

Theoretical models of chromatin can be approached in four ways: electronic, atomistic, coarse-grained(CG) and mesoscopic depending on the resolution of the study and what is intended to be modeled. Since different phenomena, or emergent properties, become apparent at different length-scales and time-scales of resolution, the basic physical models used to understand them must also depend on the chosen scale of study. These physical models range from quantum mechanical calculations, which deal with electron clouds, to polymer models of the ideal chromatin fiber, requiring polymer physics approaches [Dans et al., 2016].

At the electronic scale, quantum mechanical calculations can be used to study the electron distribution of nucleobase interactions, backbone rotamers in DNA and

RNA, the impact of ion polarization on the stabilization of certain quadruplexes, DNA hairpins, the catalytic reaction of a restriction enzyme and the prediction of photophysical and spectroscopic properties of DNA etc [Dans et al., 2016]. All-atom simulations at the atomistic scale use standard force fields, CHARMM or AMBER [Dans et al., 2016]. Coarse-grained models are effective at larger scales that cannot be dealt with by means of atomistic models.

The selection of an appropriate energy function for different atomistic and coarse-grained schemes is challenging. One such force field is the MARTINI-DNA force field. The main techniques to deal with CG systems are particle-based CG methods, molecular dynamics and Monte Carlo simulations. These can be applied to study DNA shapes, the salt-dependent persistence length of DNA, DNA hybridization, the formation of duplexes from short ssDNA oligomers and DNA curvature [Dans et al., 2016].

Coarse-graining at even large scales uses polymer models to address mainly chromatin folding and chromatin organization problem. At such scales, a polymer model of multiple nucleosomes with their linker DNA can address issues such as compaction, nucleosome-nucleosome interactions and the physical interactions of chains at high density.

3C-based chromosome conformation capture experiments provide an immense amount of data for the genome-wide interaction of contact loci in terms of interacting frequency or probabilities of interacting loci. These must be converted into models for the structure and shape of chromosomes. Some polymer models takes this 2D interacting matrix of size $(N \times N)$ as an input, where N is the resolution of the experiment. One can then apply either restraint based modelling or other statistical tools to produce single or ensemble structure ($3N$ coordinates) of chromatin. Other polymer models can be used at scale from a few Kb to 1 Mb, to see the distribution of long-range genome interactions, the effects of looping, the scaling behaviour of

contact probability vs. genomic distance, the effect of nucleus size or volume, the diffusion of chromosomal loci as well as epigenomic features [Dans et al., 2016, Junier et al., 2015].

1.7.1 Polymer Models of Chromatin

Polymer models come in two types. The first is a homopolymer model, in which all monomers are identical. The second are heteropolymer models in which monomers can differ from each other. A simple model for a homopolymer chain is that of a random walk. Such a random walk is ideal if there is no restriction on monomers occupying the same region of space, thus allowing the polymer to cross itself. The root-mean-square end-to-end distance ($\langle R^2 \rangle^{1/2} = R_m$) of a polymer chain should follow $R_m \sim aN^\nu$, where the length of polymer is N and the step size is a . For an ideal chain the value of $\nu = 0.5$.

Polymers behave differently in different solvents. A good solvent is one where the conformation of a polymer chain expands as it tries to increase the number of contacts with the solvent. It is energetically more favourable for monomers to increase contacts with solvent molecules. This situation effectively repels monomers from each other, since each monomer creates a region around it called the excluded volume where the chance of finding another monomer is very small. The excluded volume effect can be studied through self-avoiding walk (SAW) models where the monomer never visits the same site again. The value of the exponent ν for a good solvent is $\nu \sim 0.6$. For bad solvents, the chain forms compact globular conformations as it decreases the number of contacts with the solvent. The value of exponent for compact polymers is $\nu \sim 0.33$.

Two models for the organization of chromatin are defined by the “equilibrium globule” and the “fractal globule”. Both models follow the same exponent value

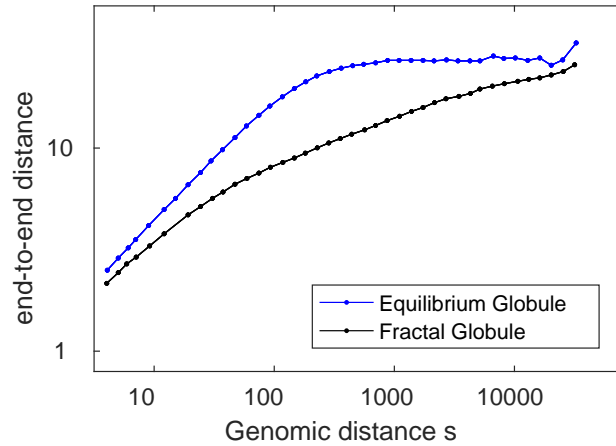


Figure 1.9: **Root-mean squared end-to-end distance R_m as a function of genomic distance s** between the ends of a subchain for equilibrium globule and fractal globule is shown with blue and red respectively. The data points of the figure are taken from Ref. [Mirny, 2011].

($\nu \sim 0.33$) for the end-to-end distance, shown in Figure 1.9, indicating large-scale compactness. The scaling at short distances, appropriate to the local globule structure, follows $\nu \sim 0.5$ for equilibrium globule and $\nu \sim 0.33$ for fractal globule [Mirny, 2011].

Models of chromatin architecture can be addressed using polymer models, via two approaches. In the first set of approaches, the aim is to develop simple polymer models by adding sequence-specific physical and biological interactions to a given polymer that explain experimental observations. The second approach aims at integrating experimental observations (specially Hi-C data) into a system of spatial restraints to be satisfied, thereby constraining possible structural models of chromatin organization [Marti-Renom and Mirny, 2011, Farré and Emberly, 2018].

The most basic model of chromatin organization is a homopolymer model where sequence does not affect polymer conformation. This model explains the compact chromatin state through the inclusion of random loops [Mateos-Langerak et al., 2009]. These models can be made more refined based on experimental input, such as the specific scaling of internal distances argued for in e.g. the “fractal globule”

model [Lieberman-Aiden et al., 2009, Mirny, 2011]. Further improvements to this class of models included the “strings and binders” approach to incorporating loops. These aim at understanding broad details of chromatin compaction and organization but without reference to specific cell types.

These models fail to account for compartmentalization into TADs and other hierarchical structures, which depend on the local genomic composition. More refined models are then heteropolymer models which consider the local coupling between chromatin structure and function. This can be achieved by assigning varying labels to different chromatin regions and in addition including topological constraints. The heteropolymeric nature can be incorporated based upon transcriptome [Jerabek and Heermann, 2012], upon the known location of binding molecules [Barbieri et al., 2012], upon gene density [Ganai et al., 2014] and based upon epigenetic marks [Shi et al., 2018]. Topological constraints can be incorporated based upon the inclusion of specific chromatin loops between promoter, enhancer and insulators [Mukhopadhyay et al., 2011, Doyle et al., 2014, Tark-Dame et al., 2014], based upon supercoiling [Benedetti et al., 2013] and based upon experimental Hi-C contacts [Giorgetti et al., 2014, Zhu et al., 2018, Tiana et al., 2016, Zhang and Wolynes, 2015, Sanborn et al., 2015, Fudenberg et al., 2016].

On one side, heteropolymer polymer models based on Hi-C data and the knowledge of local constraints are complex, require detailed experimental input and are highly dependent across contact resolution. On the other, homopolymer models are simpler, do not require much experimental input and do not vary much across different resolutions. A detailed review of different models of chromatin polymers can be followed in Refs. [Halverson et al., 2014, Amitai and Holcman, 2017, Sazer and Schiessel, 2018].

1.8 Noise and Fluctuation In the Nuclear Environment

The cell nucleus contains both chromatin fibers as well as a fluid component which plays the role of solvent. The fluid component is best thought of as a highly viscous liquid, which contains free proteins, nucleotides, RNAs, small molecules and salts, nucleoli, Cajal bodies, paraspeckles and PML bodies. The solvent molecules are in constant motion due to thermal agitation, and their collisions with the chromatin fiber provides a viscous damping [Bruinsma et al., 2014]. Because we are not interested in the motion of solvent molecules (these are fast degrees of freedom), but only in the large particles or fibers (the slow degrees of freedom), we may assume that the solvent molecules exert random forces on the latter. We use Langevin dynamics to represent this system. The Langevin equation describes a system coupled to fast degrees of freedom, where the solvent is not explicitly taken into account, but enters implicitly by means of a random force. The equation for a single particle is given below [Schlick, 2010].

$$m\vec{a}(t) = \vec{F}(r) - \gamma\vec{v}(t) + \vec{\xi}(t) \quad (1.1)$$

Where m , \vec{a} , \vec{v} and \vec{F} are mass, acceleration, velocity, and the force of a single particle due to the interactions with its surroundings. γ is the friction coefficient, given by Stokes law, and related to diffusion constant D . This equation becomes the overdamped Langevin equation (represented by Brownian dynamics) if inertial effects due to the acceleration terms are disregarded. This is achieved in the large friction limit. Hydrodynamical calculations yield values for γ in the Stokes limit. For a sphere,

$$\gamma = 6\pi\eta a \quad \text{and} \quad D = \frac{k_B T}{6\pi\eta a} \quad (1.2)$$

Here, T is the temperature, η is the viscosity of the liquid, k_B is the Boltzmann constant and a is the radius of the particle. In general, the size of the solute particle

is much bigger than the size of the solvent particles, so the agitated motion of the solute particle is much slower than that of the solvent. Its random motion is the result of random and rapid collisions due to fluctuations in the surrounding liquid.

The term $\vec{\xi}$ represents stochastic forces. These stochastic forces can be assumed, given central limit theorem arguments, to be summarized by their first order moment for any spatial component i , $\langle \xi_i(t) \rangle$, which does not depend on t , and their second order moments $\langle \xi_i(t) \xi_j(t') \rangle$ which depends only on the time difference $t - t'$. The random noise is usually assumed to be distributed according to a Gaussian probability distribution, with cross correlation vanishing at all times irrespective of particle labels. The diagonal correlations at equal times and for same particle are non-zero, following from:

$$\langle \xi(t) \rangle = 0, \quad \langle \xi_i(t) \xi_j(t') \rangle = 2k_b T \gamma \delta_{ij} \delta(t - t'), \quad i, j = x, y, z \quad (1.3)$$

The Langevin dynamics effectively provide a thermostat with a given temperature that is controlled through the magnitude of the random forces. This equation governs the dynamics and contains both frictional and random forces. In a specific limit, it simulates the dynamics of a system in thermal equilibrium, but it can be used in more general contexts.

1.9 Active Matter

Equilibrium statistical mechanics is based on the idea of a statistical ensemble, following from the Gibbs-Boltzmann distribution. Ensemble averages are understood to be equivalent to solving the equations of motion of the system i.e. a time average equals an ensemble average. There is no obvious way to construct an ensemble for a non-equilibrium system. The idea is then to prescribe the equation of motion of the non-equilibrium system, to hope that these do in fact lead to a steady state,

and to evaluate all quantities of interest using those, including properties at steady state [Zwanzig, 2001].

Chromatin serves as a substrate for enzymatic activity which is fueled by the consumption of free energy obtained from ATP hydrolysis and other sources [Ganai et al., 2014, Bruinsma et al., 2014]. ATPases are enzymes which hydrolyze ATP to ADP and inorganic phosphate (Pi), using the energy released for various cell processes. The incorporation of NTP (or dNTP) into RNA (or DNA) by a RNA (or DNA) polymerase, releases RNA (or DNA) monophosphate complex and pyrophosphate (PPi). The monophosphate is used to establish the phosphodiester bond between the two NTP (or dNTP). The energy for the process is taken from the hydrolysis of the pyrophosphate which is independent of the nucleobases, so it gets energy from each added nucleotide. The RNA polymerase is thus not an ATPase. It metabolises ATP as well as the other nucleoside-triphosphates and uses the energy from the release of PPi, but it does not produce ADP nor Pi.

Cells are dynamic and far from equilibrium. They use chemical energy in the form of ATP (or GTP) to drive active biological processes like transport and metabolism. ATP-dependent fluctuations are known to be responsible for the motion of chromosomal loci [Weber et al., 2012]. Chromatin is thus driven both by Brownian motion (thermal fluctuation), and through ATP-dependent processes [Maeshima et al., 2016]. Activity which consumes ATP generates non-thermal fluctuations of greater magnitude than thermal fluctuations at physiological temperature.

Most generally, the random motion of molecules is driven by a combination of thermal fluctuations and athermal fluctuations. The maintenance of a non-equilibrium steady state requires energy from an external source. This source in biological systems is the use of chemical energy of ATP hydrolysis (or GTP) to drive active biological processes, such as transport and metabolism. Any molecular motion which uses the energy from ATP to perform work yields ATP-dependent fluctua-

tions. These are fluctuations because ATP consumption occurs at random intervals, results in energy transduction and the exertion of local forces on the surrounding environment. ATP-dependent fluctuations are known experimentally contribute to macromolecular motion in vivo. These fluctuations behave like thermal fluctuations but with greater magnitude and steeper temperature-dependence.

Such “active” processes can be modelled via biophysical theories of “active matter” [Menon, 2010, Ganai et al., 2014]. Following standard approaches, such active processes are best described in terms of inhomogeneous, stochastic forces acting on chromatin, equivalent to a local “effective” temperature [Loi et al., 2011].

The hydrolysis of ATP through the breaking of the tri-phosphate bond yields an energy which is roughly 20 times the energy available from thermal fluctuations. If we idealise the ATP consumption as the source of an effective temperature, it is easy to see that these contribute to fluctuations at a potentially much larger scale than can be obtained from thermal fluctuations alone. Regarding the source of dissipation, ATP-consuming enzymes are not perfectly efficient so not all of the chemical energy in ATP is converted into useful work. The excess energy is dissipated as heat into the cellular environment [Weber et al., 2015, Weber et al., 2012]. Also, forces acting on chromatin must be balanced by forces exerted on the surrounding nucleoplasm, and the net effect of this is a largely random, non-thermal contribution to a Langevin-like noise.

1.10 Effective Temperature Estimates

We use an effective temperature to describe the dynamical activity in active regions of chromosomes. In general, an effective temperature can be used to describe the net effects of active mechanical fluctuations. We assume in our case, these fluctuations are uncorrelated from monomer to monomer. Some idea of the effective temperature

scale for our monomers can be obtained using the following argument. We know that ATP-dependent chromatin remodelling complexes are present in large number in the cell nucleus, and that energy released in ATP hydrolysis can surmount barriers an order of magnitude larger than energy scales associated to physiological temperatures while positioning nucleosomes [Hargreaves and Crabtree, 2011].

Measures of fluctuations of individual chromosome loci suggest a variation in diffusion constants consistent with around a ten-fold variation in an effective noise temperature seen by these loci [Weber et al., 2012]. Since non-equilibrium energy input comes through the hydrolysis of γ -phosphate bond on ATP to generate a molecule of ADP and inorganic phosphate ion is approximately 40-60 kJ/mole \sim 50 kJ/mole = 50000 J/mole (T/300K) = 166.7 (J/K) (T/mole) = 166.7 ($\frac{K_b}{1.38 \cdot 10^{-23}}$) ($\frac{T}{6.022 \cdot 10^{23}}$) $\approx 20 k_B T$.

In general we expect the active temperature to be smaller than this value, although still different from, and larger than, physiological temperatures.

1.11 Conclusion

What structural principles underlie diverse cell type-specific chromosome organization? The structure of chromosome organization might also provide insights into essentials question at the core of epigenetics, such as how cellular function is governed by cell type-specific gene regulation programs and how chromatin structure affects diseases, including cancer. Linking genome structure and function is a fundamental open question concerning higher order chromatin organization in metazoan animals. But what driving mechanisms are important to the large-scale architecture of chromatin are still unknown [Dekker, 2014, Cremer et al., 2018].

As summarized, large-scale nuclear architecture exhibits generic features that are largely common across cell types. These should severely constrain potential mod-

els [Bickmore, 2013]. However, set against this stringent requirement, virtually all prior models for such architecture are incomplete: (i) these models fail to predict gene-density based or size-based positioning schemes; (ii) no simulations reproduce the chromosome-specific distribution functions for gene density or chromosome centre-of-mass that FISH-based experiments provide; (iii) the differential positioning of the active and inactive X chromosomes cannot be obtained using any model proposed so far and (iv), the spatial separation of heterochromatin and euchromatin, seen in interphase cell nuclei across multiple cell types, has not been reproduced in model calculations in which this information is not incorporated *a priori*. Understanding these discrepancies is an outstanding problem. It is this problem that we make an attempt to address in the two chapters that follow.

In the fourth chapter, we introduce a program, THiCweed, that performs clustering of subsequences of the ChIP-Seq peaks. ENCODE contains thousands of ChIP-Seq datasets for hundreds of transcription factors across many species and cell types. Each dataset contains thousands to hundreds of thousands of peaks. Traditional *ab initio* motif finders cannot scale to datasets of such sizes. Additionally, because TF-DNA binding may be indirect and cofactors may be involved, ChIP-Seq peaks may be enriched for multiple motifs, rather than one dominant motif. We present THiCweed (**T**op-down **H**ierarchical **C**lustering to **w**eed out the signals in ChIP-Seq peaks) to approach this problem.

Chapter 2

A First-principles Model for Large-scale Nuclear Architecture

A central challenge in scientific computing applied to multi-scale systems is to develop a model that bridges different spatial and temporal scales [Schlick, 2009]. Describing 23 pairs of chromosomes in human cell nuclei, contained within the densely crowded, fluid and confined environment of the nucleoplasm, at the atomistic scale is impossible. The model approach described in this chapter stresses a specific biophysical effect relevant to the modelling of chromosomes in living cells. This, the presence of non-equilibrium fluctuations arising out of activity, is an effect that all previous studies have ignored.

Our central assumption connects levels of inhomogeneous activity across different regions of chromosomes to their large-scale properties. This inhomogeneous activity is associated with non-equilibrium ATP consuming processes which act locally on chromatin. The formation of chromosome territories or their positioning is determined by such inhomogeneous activity across segments of chromatin. Such activity acts as a ‘fingerprint’ for each chromosome. Earlier work has shown that ignoring the effects of inhomogeneous activity leads to unstructured and essentially equiv-

alent distributions for the gene density associated with each chromosome [Ganai et al., 2014].

In the next section, we describe our models. We emphasize the importance of non-equilibrium inhomogeneous activity and incorporate it in 3 different but related models. These are a gene density-based model, a gene expression-based model and a combined model. We describe the detailed methodology underlying our simulations and describe how we calculate various statistical properties of chromosomes.

2.1 Description of Model

We model the human female diploid genome in interphase, represented by heteropolymer chains confined within the nuclear envelope. The length of each heteropolymer chain can be mapped to the length of each chromosome in our coarse grained units. The monomers in our simulation represent 1Mb sections of chromatin of diameter 500 nm.

We could have defined our model at the smaller scales of 0.1 or even 0.01 Mb. However, the averaging inherent in summing transcriptional output over a 1Mb scale renders the model relatively less sensitive to errors and noise in this input and the 1Mb scale is the fundamental level of chromatin territory organization [Malyavantham et al., 2008, Jackson and Pombo, 1998, Berezney et al., 2000]. Hi-C studies also suggest ~ 1 Mbp chromatin domains are stable and provide a reasonable description of chromosome territories [Dixon et al., 2012, Kölbl et al., 2012]. In our model 46 chromosomes are represented via 46 polymer chains.

The length of chromosomes in our 1 Mb coarse grained unit is shown in Table 2.1. The largest chromosome, chromosome 1, has 249 monomers while the smallest chromosome, chromosome 21, has 47 monomers. We have a total of 6086 monomers across 46 chains for a diploid genome.

Table 2.1: Length of chromosomes in 1 Mb coarse-grained unit

Chromosome Id	Length (in Mb)
1	249
2	243
3	199
4	191
5	182
6	171
7	160
8	146
9	139
10	134
11	136
12	134
13	115
14	107
15	102
16	91
17	84
18	81
19	59
20	65
21	47
22	51
X	157
Total	3043 (haploid)

The nucleus is an active environment where ATP-driven molecular machines act in conjunction with the ordinary thermal fluctuations of Brownian motion [Di Pierro et al., 2018]. All molecular machinery associated with chromatin remodelling, replication, transcription, recombination, segregation and DNA repair is energy-consuming, relying on the hydrolysis of ATP (or NTP) molecules [Flaus and Owen-Hughes, 2011]. This leads to the localised, irreversible consumption of energy at the molecular scale [Ganai et al., 2014]. This energy is transduced, through chemo-mechanical “active” processes, into mechanical work [Weber et al., 2012, Zidovska et al., 2013, Chu et al., 2017]. Active processes within the nucleus can be described in terms of inhomogeneous, stochastic forces acting on chromatin, equivalent to an

effective temperature reflecting local levels of activity, providing the right biophysical setting [Fodor et al., 2015, Hameed et al., 2012].

Describing each chromosome as a polymer composed of consecutive monomers where individual monomers are self-repelling, different monomers can then be expected to experience different effective temperatures correlating to local active processes [Ganai et al., 2014, Agrawal et al., 2017, Wang and Wolynes, 2011]. (Our own rationale for replacing self-avoidance by self-repulsion follows from considerations of the biological system. The cell employs a large number of enzymes that change DNA topology through active, energy-consuming processes. In this respect, the *in vivo* situation is far from the one encountered in conventional polymeric systems encountered in non-biological soft matter systems in equilibrium where the time-scales for topology changes far exceed any relevant experimental time-scale.)

Monomers experience forces from other monomers, arising from both bonded and non-bonded interactions. Additionally, each monomer experiences random forces arising from thermal as well as active fluctuations. We treat such active noise as analogous to thermal noise, drawing particular realisations of the noise from a Gaussian distribution with zero mean and a variance set by the effective temperature.

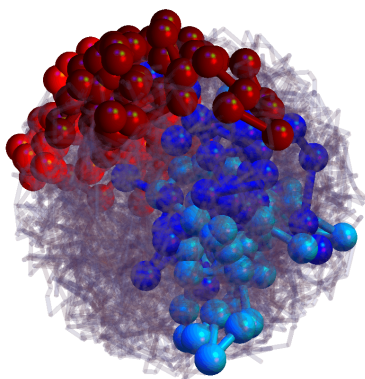


Figure 2.1: **A schematic of female (XX) diploid genome in a typical cell nuclei** is shown with pairs of chromosome 18 and 19 highlighted in the background of other chromosomes.

Overall, our model chromosomes are dynamic and explore different configurations, based on the forces they experience. Such forces arise from the dense, non-equilibrium and fluctuating environment of the cell nucleoplasm, the interactions of chromosomes

and chromosome-nuclear envelope interactions. We show a simulation snapshot of our model in Figure 2.1 where homologous pairs of chromosomes 18 and 19 are highlighted in the background of all other chromosomes represented in grey-scale. From such snapshots, we compute a variety of statistical properties of chromosomes accessed in experiments.

We propose an *ab initio* biophysical approach to predicting both cell-type-specific and cell-type independent features of large-scale nuclear architecture, where we incorporate inhomogeneous activity in the system as in the form of an effective temperature assignment to each monomer. To represent activity we assign an active effective temperature to monomers in three ways. In the simplest model, the “gene density” model, the temperature assigned to each monomer reflects the gene density associated with the specific region of the chromosome associated to that monomer. A second model, the “gene expression” model, assumes that the temperature assigned to each monomer is proportional to the amount of RNA transcript generated across that region of chromosome. A third model, providing the most comprehensive fits to the data, combines features of both gene density and gene expression models called as the “combined model”. We describe each of these models in detail.

2.1.1 Gene Density Model

To incorporate inhomogeneous activity correlated purely to gene density, the gene content of each such 1 Mb region is obtained from the GENCODE database [Harrow et al., 2012]. GENCODE version 24 contains a total of 60554 genes spread across chromosomes. We count the number of genes associated to each monomer whose mid gene positions lie in our 1 Mb interval range. Single monomers containing a number of genes which fall below a preset cutoff are termed as ‘inactive’ or ‘passive’ and are characterized by an effective temperature T equal to the physiological temperature $T_{ph} \approx 310\text{K}$. Monomers possessing a larger number of genes or number of genes

above the cutoff are termed as ‘active’ and assigned an effective temperature $T_a > T_{ph}$.

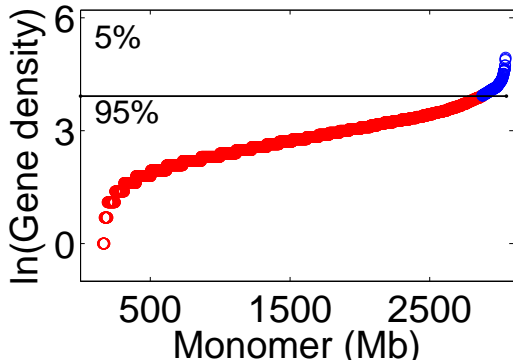


Figure 2.2: **Plot of log gene density in increasing order of monomers.** Monomers in the top 5% by gene density are active (blue) while the remaining are inactive (red). Monomers above the black line are assigned an effective temperature of $T = 12$. Those below are assigned $T = 1$.

We have experimented with several different choices of T_a as well as the cutoff, finding that a relatively small spread between physiological and active temperatures is sufficient to generate activity-dependent structuring. For concreteness, here we take the maximum value for the active temperature to be $T_a = 12$, where T_a is measured in units of T_{ph} . (Measurements of the diffusion constants of individual gene loci in bacteria and yeast provide evidence for a similar spread in local “effective” temperatures, as inferred from an Einstein relation. A related variation in local active forces has been suggested to explain observations from colloidal micro-rheology in the nucleus [Zidovska et al., 2013, Weber et al., 2012, Hameed et al., 2012]).

In Figure 2.2 we show the the logarithm of gene density associated with individual monomers along the y-axis, where we have initially sorted monomers in order of increasing gene density. Here, the top 5% monomers in blue colour are assigned to be active and experience associated temperatures in excess of the physiological temperature T_a . The remaining 95% of monomers, shown in red, are inactive, and are assigned as a physiological temperature T_{ph} .

2.1.2 Gene Expression Model

In the gene expression model, we emphasize the relevance of non-equilibrium effects arising from local transcriptional activity for descriptions of nuclear architecture. We propose that transcription levels provide a proxy for the intensity of active processes locally. We map a reasonable measure of local transcriptional activity, inferred from combining population-level measures of local RNA output from processed RNA-seq data from ENCODE [Consortium et al., 2012] on an ensemble of cells, into an effective temperature seen by each monomer.

The gene density largely correlates with gene expression across cell types. Thus, monomers that are labelled as active based on high levels of gene density tend to usually also be labelled as active according to gene expression levels, at about the 70% level. However, cell type-specificity comes from the fact that this is not true across all monomers.

All cell type have identical genes and genome. The identity of a given cell type derived from transcription regulation which switches on and off of the particular gene relevant to that cell type. Cell type GM12878 is a B-lymphocyte found in blood. Cell type NHEK is an epidermal keratinocyte found in skin. Cell type IMR90 is a fibroblast cell found in lung tissue. Cell type HUVEC is umbilical vein endothelial cells also found in blood vessel. Cell type HMEC is an epithelial cell found in breast tissue [Rouillard et al., 2016]. The differing levels of activity across these cell types is directly associated with differences in their biological function.

The sequence reads obtained from RNA-seq data are first mapped to a set of known genes obtain using GENCODE and, second, mapped to *de novo* genes. Transcripts generated across the human genome are quantified in terms of FPKM (Fragments Per Kilobase of transcript per Million mapped read, with “fragment” referring to a pair of reads for paired-end data) [Trapnell et al., 2010]. We consider all genes whose

FPKM value lies above a specified cutoff to reduce noise in the data for each cell type. We then summed the FPKM value for all these genes whose chromosome position (mid position of start and end coordinate of a gene) lies within our 1 Mb interval, to assign an activity value to that monomer. We assign effective temperatures proportional to such activity values using a derivative cutoff method as described below. Gene expression, as measured through FPKM, varies across a logarithmic scale. We can place appropriate cutoffs on the data, to map activity obtained from such gene expression data to a proxy for the active temperature.

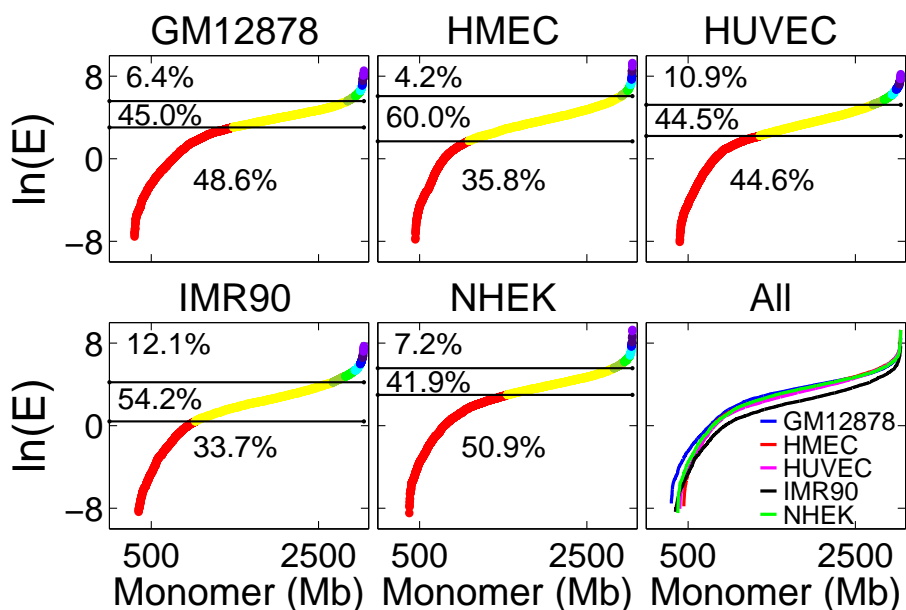


Figure 2.3: Plot of log gene expression in increasing order is shown for 5 cell types. Cell types are mentioned in the title of each subfigure. Each subplot is divided into 3 regimes demarcated by 2 lines. Monomers in the lowest region are assigned an effective temperature of $T = 1$. In the plateau region they are assigned $T = 6$. In the top region, they are assigned an effective temperature interpolating between $T = 7$ and $T = 12$. The percentage of monomers belongs to each region is also indicated. In the last sub-figure, all 5 cell types are plotted together to display the variability due to differing RNA-seq profile across different cell types.

We focused on transcriptomes across a number of model systems, exploring varied ways of associating transcript levels to effective temperatures. Figure 2.3 shows RNA-seq derived FPKM values summed over 1Mb intervals, indexing transcript levels, across GM12878, HMEC, HUVEC, IMR90 and NHEK cell types. The order of monomers across the x-axis in each subplot is different, since the assignment

of activity differs across cell type and the plot is ordered by activity. We chose structured effective temperature assignments such that they reflect the overall shape of this curve. To do this, we took a numerical derivative of the (sorted) logarithmic gene expression data, using the *diff* function of MATLAB. We set a cutoff of 0.02 for this slope but our results were largely insensitive to where this cutoff was chosen. Given the generic shape of the plot for the sorted activity, which has two regions in which activity increases sharply separated by a plateau region, this procedure automatically yields two cutoff lines and three demarcated regions. Monomers in the plateau region are assigned a constant value of active temperature, between T_{ph} and the maximum value of the active temperature $12T_{ph}$. For each cell type, given similar curves of sorted expression value, such cutoffs on the data represent the effects of activity on each monomer. We can translate this into an active temperature.

In Figure 2.3 for each cell type, we show these cutoff lines. We assign the lowest temperatures $T = 1$ to monomers whose activity falls below the value it takes in the plateau region. Monomers associated with the plateau are assigned a common temperature of $T = 6$. Finally, monomers with the highest expression values and thus the highest activity, are assigned a temperature which interpolates, in units of 1, between $T = 7$ and $T = 12$. The effects of variation of activity are strongest for these monomers, as is reasonable since activity increases steeply in this region.

We have plotted the distribution of (the logarithm of) the gene expression data in Figure 2.4. This data appear to fit a Gumbel form $f(x; \mu, \sigma) = \frac{1}{\sigma} \exp(\frac{x-\mu}{\sigma}) \cdot \exp(-\exp(\frac{x-\mu}{\sigma}))$, as shown in the figure.

2.1.3 Combined Model

Both the gene expression and gene density models fit specific aspects of the experimental data well. Surprisingly, the gene expression model did not yield appreciably

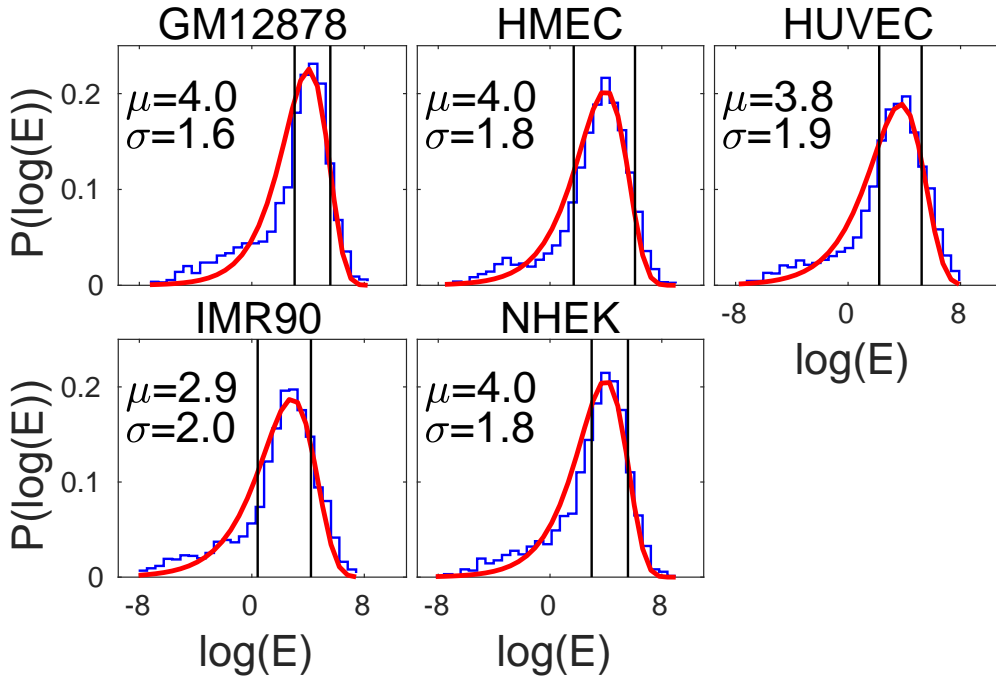


Figure 2.4: **Extraction and fitting of transcriptomics data from Figure 2.3.** The histogram of the log of gene expression values as obtained from transcriptome data for 5 cell types, shown in blue. Cell type names are provided for each subfigure. The sub-figure illustrates a fit of histogram values to an extreme value distribution, the Gumbel distribution, shown in red. The Gumbel distribution is of the form $f(x; \mu, \sigma) = \frac{1}{\sigma} \exp(\frac{x-\mu}{\sigma}) \cdot \exp(-\exp(\frac{x-\mu}{\sigma}))$. The best-fit parameters μ and σ are provided in each subfigure. Two black vertical lines, derived from the analysis that led to Figure 2.3 are shown. The left heavy tail distribution is the low expression region. The more sharply decaying right tail derives from the high expression regions. The middle region corresponds to the plateau.

better results overall than the much simpler gene density model. To address this, we noted that transcript levels need not directly correlate to activity, since FPKM values are controlled by the rate at which transcripts are both produced and degraded. This arises also because non-coding transcription is not fully captured in this version of our model, and because our description averages over the typical time-scales associated with transcriptional “bursts” [Fraser and Bickmore, 2007, Chubb et al., 2006].

We felt that a model which included features of both gene density and gene expression models might provide a more accurate representation of inhomogeneous cell-type-dependent activity [Murmann et al., 2005]. Accordingly, we decided to

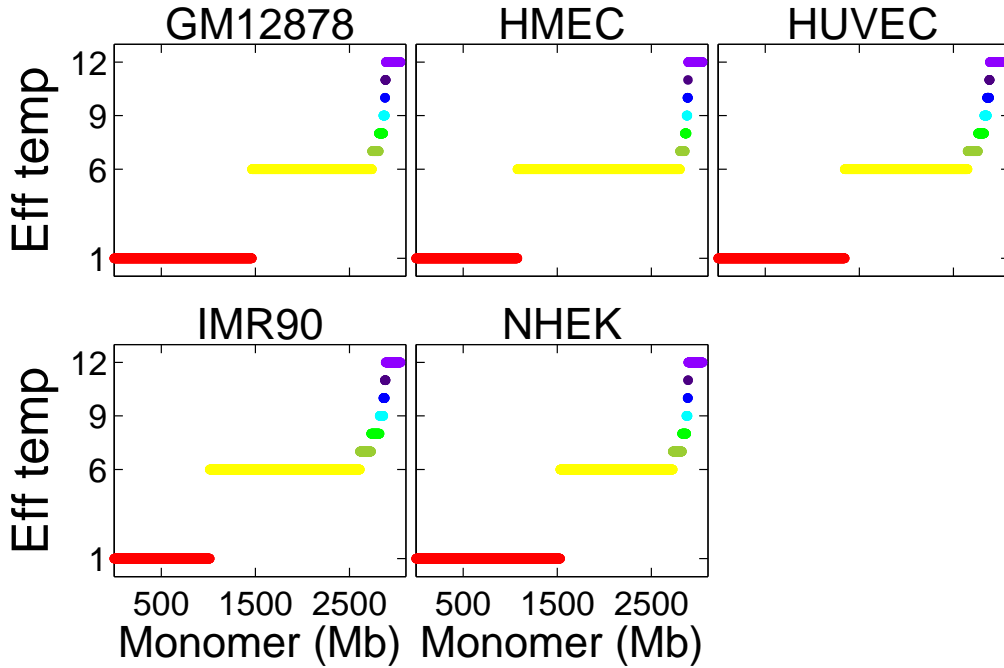


Figure 2.5: **Assignment of effective temperature to each monomer for the combined model, incorporating from both gene expression and gene density.** The red monomers are simulated at $T = 1$, yellow at $T = 6$, yellow-green at $T = 7$, green at $T = 8$, cyan at $T = 9$, blue at $T = 10$, indigo at $T = 11$ and violet at $T = 12$ times the physiological temperature.

assign monomers with a gene density above a present cutoff, the maximum active temperatures, as in the earlier gene density model. Such a model description appears to provide the most comprehensive fits to the data combines features of both these models. We will refer to this approach as the “combined model”, since it bases itself largely on the gene expression model but also assigns high activity to a fraction of monomers with the highest values of gene density.

Figure 2.5 shows temperature assignments, within the combined model, for 5 cell types. Such inhomogeneous effective temperature assignments, correlating both to gene density and transcription levels averaged over consecutive 1Mb sections of each chromosome, lie at the core of our work. For the combined model, we use the same temperature assignments as for the gene expression model but, in addition, also take the top 5% of monomers by gene density as inferred from GENCODE, promoting them to a temperature of $T = 12$.

We did not use Histone marks as a proxy for activity, although this would certainly be one approach. It seemed to us to be more direct to use RNA-seq data, since we felt that direct measures of local transcriptional output would provide an overall better proxy for activity. Although, although we use RNA-seq in this paper, we believe that more advanced methods like GRO-seq which are not just limited to steady state transcription levels might be more suitable to represent activity. We hope to explore these in future study.

2.2 Incorporating Contact Information from Hi-C Data into Model Chromosome Loops

An important part of our model is the incorporation of existing prior information regarding how chromosomes contact each other. We concentrate on contacts in *cis*, since these are far more prominent in the Hi-C data, representing these contacts in terms of permanent loops. In our model, loops correspond to the bonded interaction between non-consecutive monomers. Such loops generally form between the promoter and the enhancer region of chromatin.

There are two ways to incorporate permanent loops in our model, the first via a choice of random loops and the second by counting actual Hi-C loops. First, in the random loop model, we connect any two arbitrarily chosen monomers of the same chromosome that are not directly bonded to each other, with low probability, through a spring interaction [Mateos-Langerak et al., 2009]. The ‘random loop’ model used in [Ganai et al., 2014] actually leads to very accurate results for the DNA density distribution of chromosomes.

The probability cutoff on the order of $\sim 10^{-4}$ is used to connecting two monomers in a random loop model. We use such random loops to bring monomers into close

proximity, which are otherwise spatially far in distance from each other. For such reason, we use a higher strength of the bond in such loops, 10 times more than K of the normal FENE bond.

Second, chromosome conformation capture experiments directly measure the probability that different chromosome segments are in close proximity and such estimated loops which are larger than 1 Mb size are very few. We use Hi-C data on GM12878, NHEK, IMR90, HUVEC and HMEC cells, obtained from data made publicly available by the authors of Ref. [Rao et al., 2014], to represent the effects of long-range looping within a chromosome. We ignore loops smaller than the 1 Mb scale since these are folded into our description of a single monomer. Across these cell types, we have 236 (GM12878), 50 (NHEK), 116 (IMR90), 51 (HUVEC), and 13 (HMEC) loops that are larger than the 1 Mb size and which our model accounts for. These loops are represented by permanent FENE bonds, with an effective interaction strength that is the same as those of the springs for connected monomers. The number and nature of loops that we assume are important in determining the positioning of chromosomes. We experimented with several ways to turn off the activity or loops to check how the chromosome positioning was affected.

We used processed Hi-C data at kb resolution uploaded at GEO with accession identifier GSE63525 from Ref [Rao et al., 2014]. From these data, we simply removed those loops which are lesser than 2 Mbp in size. The remaining loops are incorporated in our model using FENE bond. The bond strength K of Hi-C loops are the same as the bond strength of normal FENE bond.

2.3 Simulation Methodology

Our numerical evolution of the system of monomers adapts the widely-used LAMMPS code implementing Brownian dynamics [Plimpton et al., 2007] for a polydisperse

polymer system with a FENE interaction between monomers. The effective temperature is incorporated as a local monomer-dependent effective temperature. The microscopic state of an monomer is characterized by its position and velocity, but also by an additional microscopic state, activity [Bellomo et al., 2007].

For each monomer, LAMPPS applies a Langevin thermostat, via the following overdamped equation of motion,

$$\zeta \frac{d\mathbf{r}_i}{dt} = \mathbf{F}_i + \eta_i \quad (2.1)$$

where \mathbf{r}_i represents the location of the i^{th} monomer, ζ is a drag coefficient, \mathbf{F}_i accounts for all non-stochastic forces acting on the monomer and η_i represents stochastic forces (gaussian, with vanishing cross-correlations) arising from both active and thermal fluctuations. The noise is assumed Gaussian distributed, with cross-correlations vanishing at all times irrespective of monomer labels. The diagonal correlators, at equal times and for the same monomer, are non-zero and obtained from $\langle \eta_i^x(t) \eta_j^x(t') \rangle = \langle \eta_i^y(t) \eta_j^y(t') \rangle = \langle \eta_i^z(t) \eta_j^z(t') \rangle = 2k_B T_i \zeta \delta_{ij} \delta(t - t')$. Here T_i is an “effective” temperature associated to each monomer, reflecting its local level of activity. We represent each of the components of $\eta_i / \sqrt{\zeta}$ as the product of a Gaussian random number with zero mean and unit variance with the quantity $\sqrt{2k_B T_i / \zeta}$. In thermal equilibrium, we have $T_i = T_{eq}$ for all monomers.

2.3.1 Bond Interaction Parameters

Our model chromosomes occupy the interior of a spherical shell. The radius of this shell is R_0 , which we take to be 17.2 in the reduced units we derive below. The interaction between neighbouring monomers (labeled as $i, i + 1$, with position coordinates $\mathbf{r}_i, \mathbf{r}_{i+1}$) is of the FENE form

$$V_{\text{neighbour monomers}}(\mathbf{r}_i, \mathbf{r}_{i+1}) = -\frac{1}{2} K r_0^2 \ln \left[1 - \left(\frac{r}{r_0} \right)^2 \right], \quad (2.2)$$

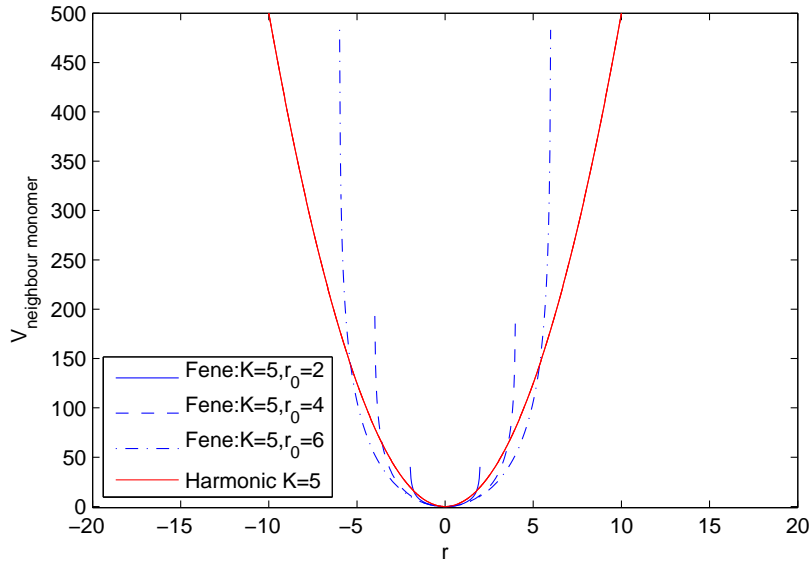


Figure 2.6: **Comparison of FENE and harmonic potential.** For larger r_0 FENE becomes harmonic. This form is then Kr^2 . The definition of the FENE potential is provided in equation 2.2.

where K is a spring constant and r_0 is the maximum stretchable length of the bond. We take $r_0 = 10$ and $K = 4.17$, for both bonded neighbours and non-bonded monomers connected by long range loops. The FENE potential is shown in Figure 2.6 for different value of r_0 and $K = 5$, while the harmonic potential is shown only for $K = 5$.

The shape of the chromatin fiber in the absence of any external constraints is determined by the persistence length and the contour length. The value of K and r_0 is estimated from the model of persistence length L_p . The L_p of naked DNA is ~ 150 bp, equivalent to around 50 nm while L_p of 30 nm chromatin fiber varies between 170 – 210 nm, depending on the compaction level of the chromatin [Junier et al., 2010, Bystricky et al., 2004]. The following equation is used to determine packing density and L_p simultaneously using mean-squared distance $\langle R^2 \rangle$ [Kreth et al., 2004].

$$\langle R^2 \rangle = 2L_P L_C \left(1 - \frac{L_P}{L_C} (1 - \exp(-L_C/L_P)) \right), \quad (2.3)$$

where L_c is contour length between two monomers. In the limiting case when $L_C \gg L_P$, the relation becomes

$$\langle R^2 \rangle = 2L_P L_C \quad (2.4)$$

The $\langle R^2 \rangle$ of a random walk for a chain of N segments with the Kuhn segment length L_K according to the freely jointed chain model is defined as

$$\langle R^2 \rangle = N L_K \quad (2.5)$$

The SCD model takes $L_K = 300$ nm or $L_P = 150$ nm [Kreth et al., 2004]. The 120 kbp chromatin linker has a contour length of $L_C = 1200$ nm. The average bond length between adjacent monomers in the SCD model is $l_0 = (2L_P L_C)^{1/2} = (N L_K)^{1/2}$ and strength of FENE potential energy (entropic spring energy) $K = 6/l_0^2$. After putting the value of L_P and L_C , we get $l_0 = 600$ nm, Because the simulation units of all the parameters are scaled against our length unit of 500 nm, in scaled units $l_0 = 1.2$, and $K = 4.17$.

The FENE bond behaves like a spring at a low stretch but becomes very stiff at high stretches. The FENE parameter K is the stiffness, while r_0 is the maximum distance, at which the force will diverge. We plotted the histogram of the bond length for various r_0 value to check which r_0 gives histogram peak at $l_0 = 1.2$. Any value $r_0 > 5$ follow such criteria and will give similar results. So, we chose $r_0 = 10$ to avoid the ‘FENE bond too long’ warning and reduce the computation time in LAMMPS.

2.3.2 Pair Interaction Parameters

Monomers further interact with non-neighbouring monomers via a repulsive Gaussian interaction, the Gaussian core potential used earlier to model polymer brushes [Stillinger, 1976]. The Gaussian potential as the function of r is shown in Figure 2.7. A

higher or lower value of B makes the width of the Gaussian wider or narrower.

$$V_{monomer-monomer}(\mathbf{r}_i, \mathbf{r}_j) = V_0 \exp(-B|\mathbf{r}_i - \mathbf{r}_j|^2), \quad |\mathbf{r}_i - \mathbf{r}_j| < r_{cut} \quad (2.6)$$

The effective pair potential at zero separation, V_0 , is chosen to be of order $k_B T_{ph}$, with k_B the Boltzmann constant and T_{ph} the physiological temperature: $B = 1.0$, $V_0 = 1.5$ and $r_{cut} = 3.5$.

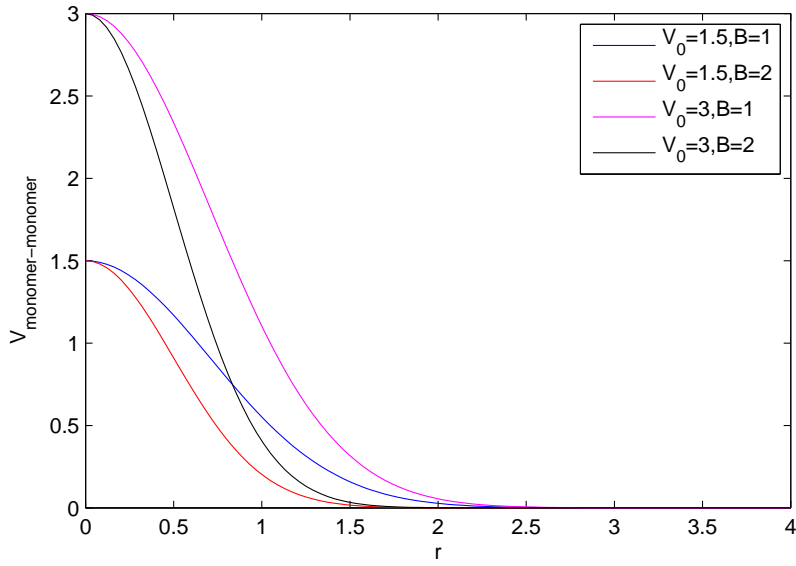


Figure 2.7: **The Gaussian core potential** for different values of V_0 and B .

The interactions between atoms or molecules contain a short-range repulsive component such that local molecular structure is dominated by excluded volume effects. Excluded volume effects dominate the interaction between compact colloidal particles, the effective forces between soft particles such as polymer coils or membranes cannot be modeled by hard cores. The effective interactions between soft particles are often of entropic origin and can be modeled by Gaussian repulsive potential [Louis et al., 2000].

The gaussian repulsive potential is used to represent the entropic repulsion between (the centers of mass of) self-avoiding polymer coils dispersed in a good solvent. The idea for the modeling of polymer coils is to map N polymers each made up of L

segments, replaced by N particles using Gaussian pair potential. This defined as the COM of two polymer coils duly averaged over internal conformations as [Prestipino et al., 2005].

$$v(r) = \epsilon \exp\left(\frac{-r^2}{\sigma^2}\right) \quad (2.7)$$

Here, the length σ is of the order of the gyration radius of the coils while the (positive) energy ϵ is of the order of $k_B T$. This potential is finite even at the full overlap between the particles and decays rapidly beyond the radius of gyration of the coils. In our case, $\epsilon = 1.5k_B T$ and $\sigma = 1$.

2.3.3 Interaction of monomers with the nuclear envelope

The interaction between each monomer and the confining sphere vanishes if the monomer centre falls within the sphere. Outside the sphere and within a cutoff r_c , the monomer experiences a Lennard-Jones potential that diverges as the distance to the cutoff is reduced. The parameters are $\epsilon = 250$, $\sigma = 1$ and $r_c = 1$ for this Lennard-Jones potential. The pure repulsive Lennard Jones potential is shown in Figure 2.8 for $\epsilon = 1$ and $\epsilon = 250$.

$$V_{wall}(\mathbf{r}_i) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 \right], \quad r < r_c \quad (2.8)$$

2.3.4 Methodology, Units and Normalization

We consider a chromatin volume fraction between $0.1 \leq \phi \leq 0.2$; see Refs. [Kreth et al., 2004, Ganai et al., 2014]. The monomer is assumed to have a diameter $d \simeq 500nm$; the equilibrium domain separation is $\ell_0 \simeq 600nm$. Both these quantities accord with computed Kuhn lengths of $\approx 300nm$ [Kreth et al., 2004, Rosa and Everaers, 2008]. Assuming that the radius of the nucleus is $R_0 \simeq 8.6\mu m$ yields a

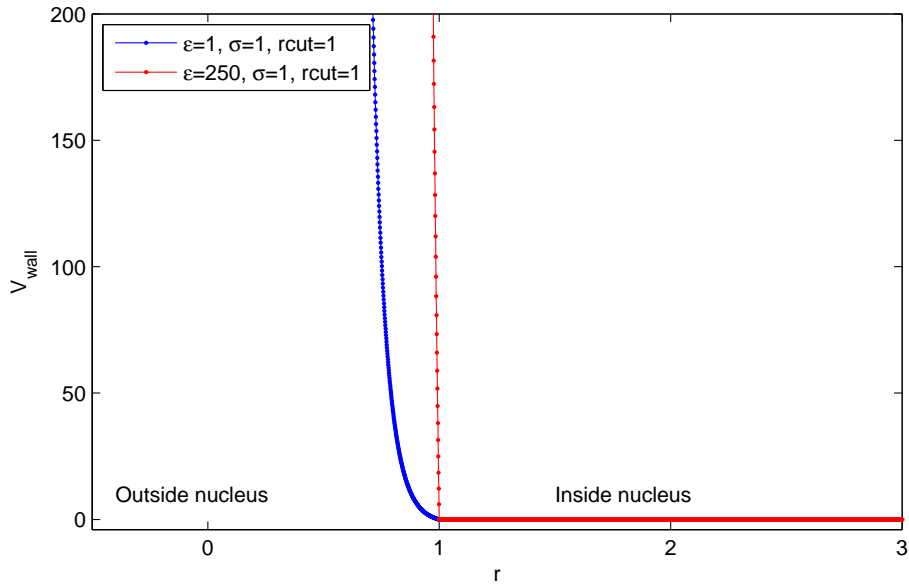


Figure 2.8: **Lennard-Jones potential for different values of ϵ .** This interaction between monomers and the nuclear wall is purely repulsive .

packing fraction of $\phi \simeq 0.15$. We ignore the marginal differences in nuclear volume across cell types; such volumes differ by at most a factor of 1.5 for the cell types we consider. We scale all lengths in units of d and measure energies in units of $k_B T_{eq}$. All chromosomes are fairly tightly confined to R_0 . We can choose units of time (τ) such that $\zeta = 1$.

We can approximate the value of ζ appropriate for this calculation from the Stokes relation: $\zeta \simeq 6\pi\eta_s R$ where R is the hydrodynamic radius appropriate to the monomer size. Assuming that the appropriate value of the viscosity at such scales is $\eta_s \sim 10\eta_w$, with η_w the viscosity of water ($8.9 \times 10^{-4} Pa \cdot s$), its numerical value is then $\zeta = 8.38 \times 10^{-8} N \cdot s/m$. With this choice, τ is then $(500.0)^2 \zeta / 6k_B T \simeq 8.16 \times 10^{-1} s$. Since $\tau \approx 10^{-1} s$; our choice of time-step of 0.001 thus corresponds to real-time evolution by $10^{-4} s$.

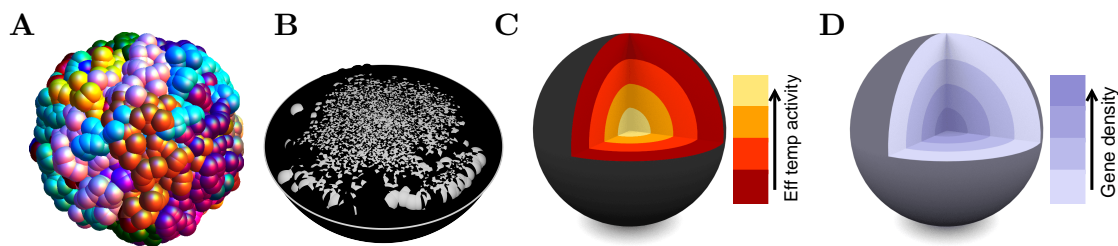


Figure 2.9: **Model predictions for large-scale features of nuclear architecture.** **A.** Typical image of chromosome territories computed in our simulations, with each chromosome represented as different colour. Note the tendency of each chromosome to overlap relatively little, visually representing emergence of territoriality. **B.** A cut-away sphere representation of the average spatial distribution of euchromatin (or active white) and heterochromatin (or inactive black) monomers as computed for the GM12878 cell type. Here, the active monomers are defined as those having an effective temperature in excess of the physiological one. Heterochromatin is found more peripherally compared to euchromatin which is located towards the nuclear interior. **C.** A cutaway sphere representation of average effective temperatures within the simulated nucleus, as computed for the GM12878 cell type. This illustrates the larger effective temperatures, indicating enhanced activity, obtained towards the centre of the nucleus, in comparison to a lower effective temperature in the nuclear periphery. **D.** A cutaway sphere representation of the average gene density within the simulated nucleus, computed for the GM12878 cell type. This illustrates the excess in gene density seen towards the centre of the nucleus in comparison to the gene density in the nuclear periphery. This separation of gene-dense and gene-poor 1Mb segments of chromatin correlates to the distinction in the spatial positioning of euchromatin and heterochromatin.

2.3.5 Summary of Analysis

After verifying that the system has achieved a non-equilibrium steady state, we compute all properties of interest, including the distribution of DNA density and of chromosome centre-of-mass, territorial organization, shape statistics and spatial distance maps from which we can infer potential contacts. All data are averaged over the two autosomal homologs, as their positioning was found to be equivalent.

Our simulations are run for 10^7 time steps, with around 4×10^6 steps discarded to ensure adequate equilibration. All data are averaged over at least 25 independently initialized configurations, with each initial configuration contributing 6000

independent measurements as the simulation proceeds. We verified that the same steady-state properties were achieved irrespective of initial (random) configuration. Since the probability of finding a chromosome at a radial separation \mathbf{r} from the origin depends only on the modulus of \mathbf{r} , *i.e.* $|\mathbf{r}| \equiv r$, we calculate the probability of finding a monomer belonging to a specific chromosome at a radial distance from the origin, for each chromosome.

We show a summary of predictions of large-scale features of nuclear architecture from our combined model in Figure 2.9. A snapshot of chromosome territories is shown in Figure 2.9A with the different coloured chromosome. Figure 2.9B represents the average distribution of euchromatin (white) and heterochromatin (black) monomers. Figure 2.9C shows that the effective temperature increases from the periphery to the centre of the nucleus. Similarly, the average gene density increases from the periphery to the centre of the nucleus as Figure 2.9D shows.

2.4 Calculation of Distribution Functions

Chromosome-specific distribution functions $S(R)$ are obtained experimentally using confocal slices of FISH images from an ensemble of fixed nuclei. We calculate $S_i(R) = 4\pi R^2 P_i(\vec{R})$, where $P_i(\vec{R}) dR$ is proportional to the probability of finding a monomer of chromosome i at a radial vector \vec{R} from the origin. For a uniform distribution, $S_i(R) = 4\pi R^2$. We compute $S_i(R)$ for every model chromosome indexed by i . We measure activity in successive radial shells by performing a configurational average over the effective temperature of every monomer in that shell. From these, we extract a quantity similar to $S(R)$ but normalize by $4\pi R^2$, so that the quantity plotted in the cut-away sphere representation simply represents the activity at radial distance R .

The quantity $S(R)$ measures the DNA density associated with a specific chromosome,

across a radial shell at distance R from the nuclear centre, averaged over a large number of nuclei. The quantity $S_{CM}(R)$ measures a similar distribution, but of the chromosome centre-of-mass. We calculate the distribution of centres of mass of each individual chromosome similarly. If the monomers are randomly distributed inside the nucleus, the $S(R)$ or $S_{CM}(R)$ should follow a quadratic rise with R which is shown in Figure 2.10. To visually examine configurations we colour-coded monomers belonging to individual chromosomes shown in Figure 2.9A.

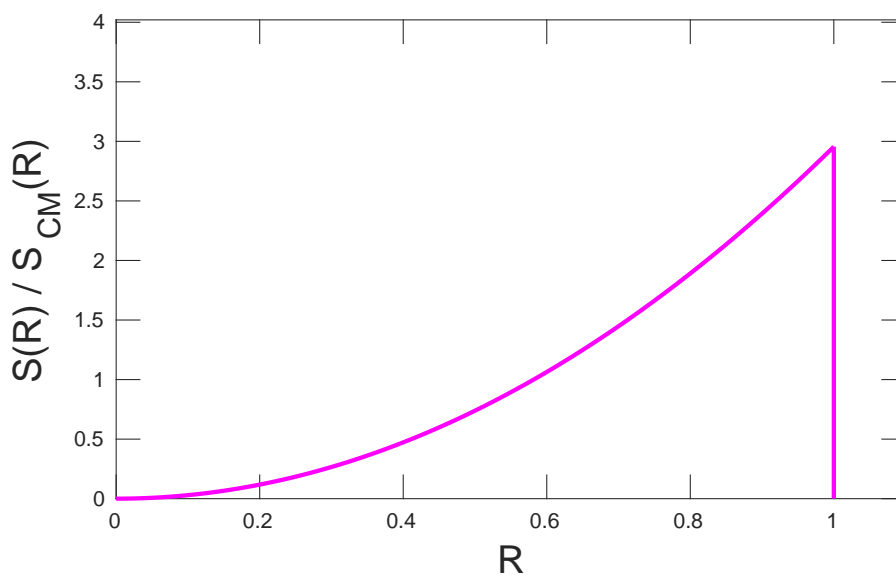


Figure 2.10: **Distribution function of the DNA density distribution $S(R)$ or centre-of-mass distribution $S_{CM}(R)$, if they are randomly distributed inside the nucleus.** The R^2 rise of $S(R)$ or $S_{CM}(R)$ towards the nuclear envelope, expected for uniformly distributed chromosomes is shown.

We simulate a diploid genome that means every chromosome has two identical copies (or homologs). These homologs are spatially independent of each other and they can be found anywhere in the nucleus. The common thing between the homologs is the activity of the monomers along the polymer chain except for X chromosome and length of the polymer chain. These two properties are enough to give an identical radial distribution of homologs in regard to the spatial constraint. So, wherever we computed any statistical quantity like $S(R)$ or $S_{CM}(R)$ or any we averaged over the homologs of autosomes(chromosomes 1 to 22).

2.5 Geometric Properties of Chromosome Territories

We have calculated a number of geometric properties of individual chromosomes, to compare to experimental results. Such geometric properties include the three dimensional shape characteristics of chromosomes, two dimensional shape parameters of chromosomes, a calculation of the smallest ellipsoid which can contains an individual chromosome. They also include properties such as the asphericity of chromosomes, prolateness of chromosomes, their volume and surface area, their contact maps of chromosomes and contact probability vs genome distance. The detailed description of each of the above mentioned properties are given below.

2.5.1 Calculation of Three-dimensional Irregular Shape for a Given Chromosome

The shape of a chromosome tells us about the surface that it exposes to solvent and about its physical proximity and contacts to other chromosomes. In our simulation, each chromosome is represented by a polymer chain. Finding the shape of a chromosome is equivalent to finding properties associated with the shapes of a polymer chain. Because polymer chains are irregular, finding the right statistical description of their shape statistics is complex and computationally challenging.

Finding a regular shape which contains a polymer chain inside is computationally feasible. One example of such a regular shape is an ellipsoid, which can be calculated by standard methods. If the length of a polymer is large and it is homogeneously distributed in space then we expect that the ellipsoid will enclose the polymer well. However, if the polymer is irregularly shaped, such a simple geometrical approach will fail. We use an alternative grid method to compute the properties of such

irregular shapes of chromosomes.

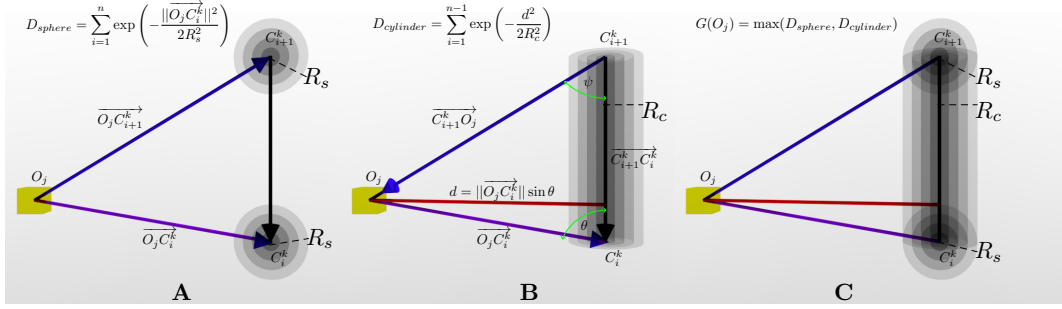


Figure 2.11: **Schematic of the grid method for computing chromosome territory shape.** (A) Monomer contribution to the density at the yellow grid point (D_{sphere}) is computed by adding the contribution from all monomers of a given chain. (B) Bond contribution to the density at the yellow grid point ($D_{cylinder}$) is computed by counting the contribution from cylinders representing all connected bonds in a given chain. (C) The actual density at yellow grid point $G(O_j)$ is the maximum of either D_{sphere} or $D_{cylinder}$

Here, we describe the grid method: For each chromosome, we begin by drawing a 3d grid across the nucleus with a coarse grid spacing in range of 0.2 – 0.6 in our scaled units. The larger grid spacing takes less time to compute but at the same time is less accurate. Computations are more time-consuming for smaller grid spacings, but the trade-off lies in improved accuracy. We calculate the DNA density associated with our polymer model for each chromosome individually on this grid as described below.

We represent chromosome configurations in each simulation snapshots in the following way. First, individual monomers are considered as spheres about which the density decays as a normalized gaussian in the radial direction. The characteristic scale of the Gaussian is set by the length scale R_s . For bonded monomers, we assume that the DNA configuration can be described as a cylindrical region with fixed radius, with the DNA density about the axis of each cylinder assumed to fall off also as a Gaussian with specified width R_c . We experimented with various choices of R_s and R_c to obtain the optimal fits of experimental data of 2d chromosome shape.

The density at any given grid point can then be computed by adding up the contri-

butions from all spherical and cylindrical regions associated to a single chromosome. If the chosen grid point spacing is larger, then it can be interpolated for smaller grid values, typically of spacing 0.10 – 0.15 using MATLAB’s INTERP3 function. Once such a density field is obtained, we can find the surfaces on which it attains a fixed value, the “implicit surface”.

We adjust the scales $R_s = 1\sigma$ governing the decay of the density distribution associated with the monomers and the cylindrical regions $R_c = 1\sigma$ connecting between bonded monomers as well as the isovalue constant specifying the implicit surface to optimise geometrical quantities associated with chromosome territories vis a vis experiments. Once fixed, these parameters remain the same for all chromosomes and all cell types.

Our calculation of the density at a given grid point proceeds as follows. Consider the location of the grid point O_j , where j indexes the specific grid point. Let n be the length of chromosome C^k where k is chromosome index. We denote the i th monomer of the k^{th} chromosome as C_i^k . The calculation is shown visually in Figure 2.11. The two consecutive monomers C_i^k and C_{i+1}^k of chromosome k are represented by small circles in Figure 2.11 along with lines representing the centreline of the associated cylindrical regions. The sphere and cylinder, for different value of R_s and R_c , are drawn in Figure 2.11A and 2.11B respectively. O_j is the grid point where we have to calculate the density due to sphere only D_{sphere} . The contribution comes from the two monomers C_i^k and C_{i+1}^k shown in Figure 2.11A with a similar contribution from the cylindrical density $D_{cylinder}$ along the centerline. This is shown in Figure 2.11B. The density at the grid point O_j arising from chromosome k is then computed by summing up all monomers and cylinders from the following formula. This is also shown in Figure 2.11C for two monomers.

$$G(O_j) = \max \left(\sum_{i=1}^{n-1} \exp \left[-\frac{d^2}{2R_c^2} \right], \sum_{i=1}^n \exp \left[-\frac{\|\overrightarrow{O_j C_i^k}\|^2}{2R_s^2} \right] \right) \begin{cases} \text{if } (\theta > \pi/2) | (\psi > \pi/2) \\ d = 1000 \\ \text{otherwise} \\ d = \|\overrightarrow{O_j C_i^k}\| \sin \theta \end{cases} \quad (2.9)$$

Here $\cos \theta = \frac{\overrightarrow{O_j C_i^k} \cdot \overrightarrow{C_{i+1}^k C_i^k}}{\|\overrightarrow{O_j C_i^k}\| \|\overrightarrow{C_{i+1}^k C_i^k}\|}$ and $\cos \psi = \frac{\overrightarrow{C_{i+1}^k O_j} \cdot \overrightarrow{C_{i+1}^k C_i^k}}{\|\overrightarrow{C_{i+1}^k O_j}\| \|\overrightarrow{C_{i+1}^k C_i^k}\|}$. These conditions ensure that (a) the contribution of the associated cylinder to the density at any grid point is cylindrically symmetric about the line joining neighbouring monomers and that (b) the contributions from the spherical regions is also accounted for.

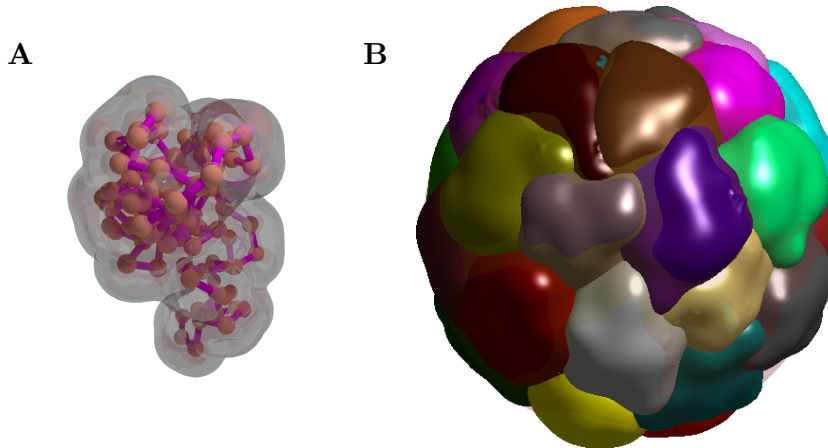


Figure 2.12: **Shapes of individual CT and all CTs in a nucleus.** (A) The shapes of individual chromosome territories extracted from simulation configurations. Such shapes are used to compute a number of geometrical properties of chromosome territories, e.g. their volume, surface area, asphericity and other shape parameters. (B) The shapes of all chromosome together in a nuclei is shown for one simulation configurations. Here each chromosome represented with different colour, illustrating the emergence of chromosome territoriality.

After the density values at all grid points are computed, the isosurface command from MATLAB is used to draw the implicit surface ($F(x, y, z) = c$) for the chromosome given a density isovalue c . A smaller value of c yields a loose cloud-like surface around chromosomes while larger values of c gives tighter, more well-defined

surfaces around them. We chose a value of c such that the chromosome territories it yields are visually equivalent to those obtained in experiments based on similar isosurface representations of experimental FISH data. The 3d surface representation of such individual chromosome is shown in Figure 2.12 A and for all chromosomes together is shown in Figure 2.12 B.

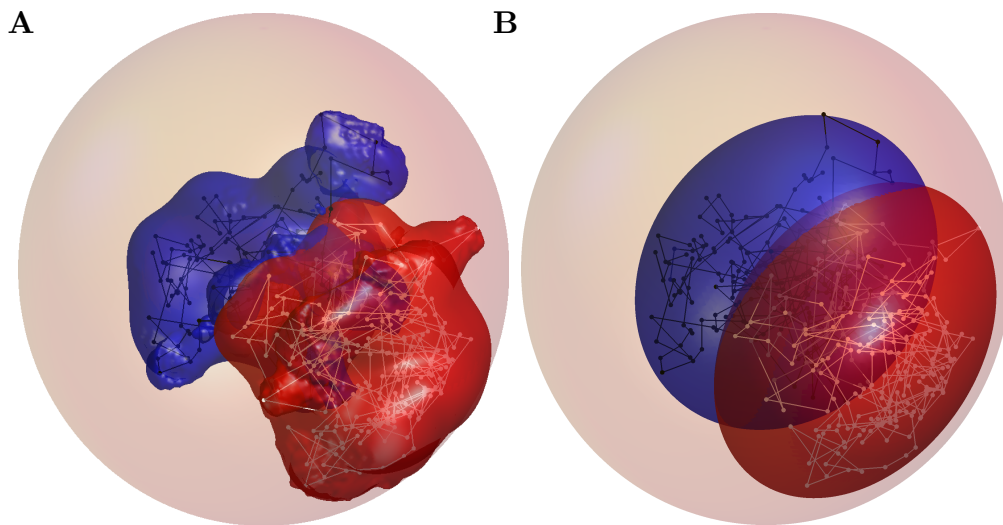


Figure 2.13: **Comparison of chromosome territory shape from grid method and the 3d ellipsoid fit method.** (A) Schematic representation of volume overlap of chromosome 1 and 3 by an implicit surface method. The polymer form of chromosome 1 and 3 are represented with white and black colours respectively. The 3d surfaces of chromosome 1 and 3, computed through the grid method, are shown with red and blue colour respectively. (B) Schematic representation of volume overlap of chromosome 1 and 3 by an ellipsoidal fit method. The polymer form of chromosome 1 and 3 is represented with white and black colour respectively similar to Fig. A. The 3d surfaces of chromosome 1 and 3, computed from the ellipsoid algorithm, are shown with red and blue colours respectively.

2.5.2 Comparison of Methods for Calculating Three-dimensional Irregular and Regular Shapes

An ellipsoid fit provides an easier and faster method for finding the approximate shapes of the polymer, in comparison to the grid method. In the ellipsoidal fit

method, the center, rotation, and principal radii of the smallest volume of ellipsoid which encloses all the data points of a polymer is calculated using the Khachiyan algorithm [Todd and Yildirim, 2007]. The distinction between these two methods and the relatively bad performance of the ellipsoidal method is due to the fact that the ellipsoidal method leads to larger overlaps between chromosomes. Further, for rougher chromosome territories, the ellipsoidal method of fitting a smooth three-dimensional shape is less useful. The difference of chromosome territories computation between the grid-based method and the ellipsoid fit method is shown in Figure 2.13. It is clearly visible from the figure that the ellipsoid fit method, although commonly used see e.g. Ref. [Uhler and Shivashankar, 2016, Wang et al., 2017] is more error prone and leads to some inconsistencies in the measurement of volume and surface area.

The pseudocode for the Khachiyan algorithm for fitting an ellipsoid of a minimum volume to a polymer chain X is given in Algorithm 1. The radii of an ellipsoid gives the a, b, c values of three principal axis. Once we know the center and rotation as well, then an ellipsoid can be drawn which encloses the polymer chain. The volume and surface area of ellipsoid are $\frac{4}{3}\pi abc$ and $4\pi\left(\frac{(ab)^p+(ac)^p+(bc)^p}{3}\right)^{\frac{1}{p}}$ respectively, where $p = 1.6075$.

2.5.3 Calculation of Volume and Surface Area of a Chromosome

In the three-dimensional case in above, we mentioned that once we associate an implicit surface to a chromosome, that surface can further be triangulated using standard methods, such as the ISOSURFACE command in MATLAB. The total surface area of the chromosomes is obtained by adding the area of these triangles. To calculate the volume of the chromosome we count the number of grid points whose grid density values are more than the given isovalue density c .

Algorithm 1: Khachiyan algorithm

- Input:** $X(d,N)$ N is number of monomers in a chain and d is the dimension
Output: center, radii, rotation
- 1 Make Q of size $(d + 1, N)$ augmented vectors of ones in X
 - 2 Initialize U with zeros matrix of size (N, N) with diagonal entry filled with value $1/N$
 - 3 Calculate $M = Q'[\{Q(UQ')\}^{-1}Q]$ which is size of (N, N)
 - 4 Find the largest entry $M(j, j)$ in diagonal position j of M
 - 5 Calculate step size $\delta = \frac{M(j,j)-d-1}{(d+1)(M(j,j)-1)}$
 - 6 Update U matrix with $U_{new} = (1 - \delta)U$
 - 7 Update j entry of $U_{new}(j, j) = U_{new}(j, j) + \delta$
 - 8 Calculate the norm of diagonal vector which is error
 $e = |diag(U) - diag(U_{new})|$
 - 9 Update $U = U_{new}$
 - 10 Repeat *Steps 3 to 9*, till $e < \epsilon$ where ϵ is some tolerance limit
 - 11 Once the algorithm stop then center c , radii and rotation of ellipsoid can be calculate by following formula
 - 12 center $c = X \cdot diag(U)$
 - 13 $A = (1/d)\{X(UX') - cc'\}^{-1}$
 - 14 $[U, s, rotation] = svd(A)$
 - 15 radii $r = 1./sqrt(diag(s))$
-

2.5.4 Calculation of Asphericity and Prolateness Parameters

Three measures of size and shape (radius of gyration R , asphericity Δ , prolateness Σ) can be derived from the moment of inertia tensor. The asphericity and the shape (or prolateness) parameter of individual chromosomes can be calculated from the 3 semi-principal radii (a,b,c) of the ellipsoid obtained for each ellipsoidal fit to a chromosome territory using the Khachiyan algorithm mentioned above. The asphericity Δ parameter given in Eq. 2.10b characterizes the average deviation of the chain conformation from spherical symmetry [Millett et al., 2009, Rawdon et al., 2008]. The shape Σ parameter measures the prolateness or oblateness of chromosomes and

is defined in Eq. 2.10c.

$$R(a, b, c) = \sqrt{\frac{a^2 + b^2 + c^2}{3}} \quad (2.10a)$$

$$\Delta(a, b, c) = \frac{(a - b)^2 + (b - c)^2 + (c - a)^2}{2(a + b + c)^2} \quad (2.10b)$$

$$\Sigma(a, b, c) = \frac{(2a - b - c)(2b - a - c)(2c - a - b)}{2(a^2 + b^2 + c^2 - ab - ac - bc)^{3/2}} \quad (2.10c)$$

The parameter Δ is bounded in the regime $0 \leq \Delta \leq 1$. The Δ value is 0 for the perfect sphere when $(a = b = c)$ and 1 for rod shape when $(b = c = 0)$. The shape or prolateness parameter Σ is bounded by $-1 \leq \Sigma \leq 1$. The Σ is -1 for perfect oblate shapes, e.g. when $(a = b > c)$ and 1 for perfectly prolate shapes, e.g. when $(a > b = c)$, providing a useful index for chromosome shapes.

2.5.5 Calculation of Ellipticity and Regularity in 2d Projection

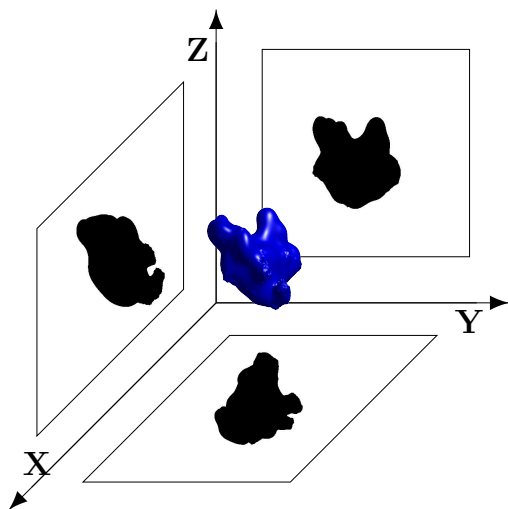


Figure 2.14: **Schematic illustrating a 2D projection of a three-dimensional CT**, projected along the XY , YZ , and XZ planes. The ellipticity and regularity parameters can be computed from such 2d projections, and compared to 2D FISH data.

To calculate the two-dimensional properties ellipticity and regularity of projected chromosome territories, we use a method described in Ref. [Sehgal et al., 2014]. To compare our simulation data with data from 2d FISH as obtained in those experiments and other similar ones, we project three-dimensional chromosome territories onto the xy plane for specificity. (Note that averaging over configurations which are rotationally symmetric implies that all projections should be equivalent.) A schematic illustrating how the projection of CTs has been taken is shown in Figure 2.14. We use the ellipticity algorithm of Ref. [Žunić and Žunić, 2013] which rely on the computation of the geometric moments M of the image I . For calculating the ellipticity and regularity of 2d CTs image, we first normalize, scale and rotate the complementary binary image I of our projected chromosome territories such that:

1. The area of shape is 1
2. The centroid of shape coincides with the origin and
3. The orientation of shape is 0 (implying that the long axis is parallel to the x-axis).

The scaled moments then are:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x, y) \quad (2.11a)$$

Where $I(x,y)$ is the pixel intensities of the grayscale image and $\{p, q\}$ can each be drawn from $0 \dots \infty$, and refer to the order of moments. We choose each of $\{p, q\}$ from 0,1,2 for specificity.

Given the moments, we compute the ellipticity ϵ from

$$\epsilon = \frac{1}{2} \times \frac{1}{a^2 \times M_{20} + b^2 \times M_{02}} \quad (2.11b)$$

where a and b are given by

$$a = \sqrt{2\pi^2 L - \pi \times \sqrt{4\pi^2 L^2 - 1}}, b = \sqrt{2\pi^2 L + \pi \times \sqrt{4\pi^2 L^2 - 1}} \quad (2.11c)$$

Here, $L = M_{20} + M_{02}$, as ϵ tends to 1, projected chromosome territories resemble true ellipses.

To calculate the regularity of these projected territories, the area ratio of each CT over its convex hull is determined, using the MATLAB functions BWAREA and

BWCONVHULL. If a chromosome is regular, this ratio should be close to 1. As the irregularity in projected CTs image increases, this ratio will decrease.

2.5.6 Calculation of Contact Probability

The contact probability is computed using numerical calculations of the contact frequency of monomers of a given chromosome, averaged over a large number of configurational snapshots. When two monomers i and j of the same chromosome are separated in genomic distance of $s = |i - j|$ and in 3d space at least less than of 2.5σ units in terms of our scaled unit distance, we assume that they are in contact. If two monomers are in contact, they are close in distance. However, measures that look at the frequency of contacts will assign a larger frequency to such monomers which are predisposed to be close by. The 3d distance between bonded monomers in our simulation is $\approx 1.2\sigma$ so choice of 2.5σ between the monomers which are in genomic distance of s is quite reasonable. We count the number of frequency when any two non-bonded monomers makes a contact within genomic distance of s for all possible configuration.

2.5.7 Calculation of Distance Maps and Contact Maps

The distance map is the heatmap plot of 2d matrix of the average distance over many configurations between every possible i and j monomers. Similarly, the contact map is the heatmap plot of the 2d matrix of the average frequency of contacts between every possible i and j monomers within some cutoff, averaged over many configurations. We consider that two monomers i and j are in contact when the Euclidean distance between them is less than 2.5σ units in our scaled unit distance.

2.5.8 Statistical error calculation for relative center of mass position data

We fit a straight line ($y_{fit} = mx + c$) using a least-squares method to the relative center-of-mass positions, where we minimize the sum of the squares of the difference between the observed value y and the fitting value y_{fit} evaluated at x , weighted appropriately by the error bars σ . Here σ refers to the standard deviation in the observed value y . The simulation parameters, slope m^{sim} and intercept c^{sim} , are minimized for $\sum_{i=1}^N \left(\frac{y_i^{sim} - y_{fit}^{sim}}{\sigma_i^{sim}} \right)^2$ and the experimental parameters slope m^{exp} and intercept c^{exp} minimized for $\sum_{i=1}^N \left(\frac{y_i^{exp} - y_{fit}^{exp}}{\sigma_i^{exp}} \right)^2$. N is the number of data points, in our case the number of chromosomes $N = 23$. The data $(y_i^{exp}, \sigma_i^{exp})$ and $(y_i^{sim}, \sigma_i^{sim})$ correspond to chromosome relative center of mass position and standard deviation, for experimental and simulation. We have

$$\chi^2 = \sum_{i=1}^N \frac{(y_{fit,i}^{sim} - y_{fit,i}^{exp})^2}{y_{fit,i}^{exp}} \quad (2.12)$$

After fitting the relative center of mass position simulation and experimental data to a straight line, we computed the χ^2 error from equation 2.12, using the fitted value $y_{fit,i}^{sim}$ and $y_{fit,i}^{exp}$. We then obtained a p-value from the χ^2 cumulative distribution function, using MATLAB chi2cdf function. We checked whether the systematic exclusion of sub-sets of chromosomes (say large vs small) might provide better fits to the relative center of mass position data in terms of the p-value.

2.6 Conclusion

Model of large-scale nuclear architecture are dependent upon the assumptions that go into the model as well as the numerical values of control parameters. This chapter

discussed a number of possibilities for these. It stresses the physical nature of the assumptions as well as the need to avoid excessive complexity in defining a large-scale nuclear architecture model. We optimized our parameters for the GM12878 cell type but recognize that all differences between cell types might not be contained in our simple way of assigning the cell-type specific properties solely based on levels of transcriptional activity. However, the analysis provided here should provide the simplest way of understanding the generic effects of activity in establishing patterns of nonequilibrium activity that reflect nuclear architecture across specific cell types. All the materials and methods presented in this chapter are used to generate the results of the next chapter.

Chapter 3

Results from a First-principles Approach to Large-scale Nuclear Architecture

Chromosomes are not distributed at random within the interphase nucleus, an observation that is central to our current understanding of large-scale nuclear architecture in the interphase nuclei of metazoans [Meaburn and Misteli, 2007, Cremer and Cremer, 2010, Bickmore and van Steensel, 2013]. Gene rich, more open, early-replicating euchromatin regions are typically distributed more centrally than gene-poor, relatively more compact, late-replicating heterochromatin [Cremer and Cremer, 2010]. Chromosomes are organised territorially, with each being segmented into relatively more (A) and less (B) active compartments that are then further subdivided into topologically associated domains [Lieberman-Aiden et al., 2009, Dixon et al., 2012, Fraser et al., 2015]. In humans, gene-rich chromosome 19, containing a large number of house-keeping genes, is distributed more centrally across several cell types than the similarly sized but gene-poor chromosome 18 [Croft et al., 1999, Boyle et al., 2001]. This observation generalises to a gene-density-based radial positioning

schema for all chromosomes [Takizawa et al., 2008].

Gene-rich regions within chromosomes tend to orient towards the nuclear centre, with expressed alleles often found further from the nuclear envelope than ones that are not expressed [Takizawa et al., 2008, Therizols et al., 2014]. In some human cell types, chromosomes appear to be positioned by size, with the centres of mass of smaller chromosomes disposed more centrally than those of larger ones [Sun et al., 2000, Bolzer et al., 2005, Kölbl et al., 2012]. In female cells, the two X chromosomes are differentially positioned, with the more compact, inactive X-chromosome found somewhat closer to the nuclear envelope than the active one [Dyer et al., 1989, Jégu et al., 2017]. Actively transcribed chromosomes tend to have rougher, more elliptical territories than less active ones [Eils et al., 1996, Berezney et al., 2005, Khalil et al., 2007, Sehgal et al., 2014, Jégu et al., 2017].

The probability with which two loci along individual chromosomes are found in proximity to each other in ligation assays follows a power-law $P(s) \sim 1/s^\alpha$ with $\alpha \simeq 1$ over an approximately 1 - 8 Mb range, consistent with a fractal globule picture of chromosome structure [Lieberman-Aiden et al., 2009, Mirny, 2011]. Currently, experiments suggest that such organization is cell-type dependent and that α ($1 \leq \alpha \leq 1.5$) also varies across chromosomes over a comparable range [Sanborn et al., 2015, Kang et al., 2015].

Most model approaches to nuclear architecture assume *a priori* that chromosomes are structured polymers in thermal equilibrium [Cook and Marenduzzo, 2009a, Tark-Dame et al., 2011, Marti-Renom and Mirny, 2011, Heermann et al., 2012, Vasquez and Bloom, 2014, Imakaev et al., 2015]. Some models ignore thermal fluctuations altogether in favour of incorporating loop structure as derived from the Hi-C data, while also requiring compatibility with physical restrictions on the overlaps of chromosomes [Imakaev et al., 2015, Amitai and Holcman, 2017, Tjong et al., 2016]. Others account for the domain structure of individual chromosomes [Odenheimer

et al., 2005, Jost et al., 2014, Jost et al., 2017, Chiariello et al., 2015, Haddad et al., 2017, Ghosh and Jost, 2017, Zhang and Wolynes, 2017, Tiana et al., 2016, Di Pierro et al., 2016, Di Pierro et al., 2017].

As summarized above and in previous chapters of this thesis, large-scale nuclear architecture exhibits generic features that are largely common across cell types. These should severely constrain potential models [Bickmore, 2013]. However, set against this stringent requirement, virtually all prior models for such architecture are incomplete: (i) these models fail to predict gene-density based or size-based positioning schemes; (ii) no simulations reproduce the chromosome-specific distribution functions for gene density or chromosome centre-of-mass that FISH-based experiments provide; (iii) the differential positioning of the active and inactive X chromosomes cannot be obtained using any model proposed so far and (iv), the spatial separation of heterochromatin and euchromatin, seen in interphase cell nuclei across multiple cell types, has not been reproduced in model calculations in which this information is not incorporated *a priori*. Understanding these discrepancies remains an outstanding problem.

In this chapter, extending these ideas, we describe our research on an *ab initio* biophysical approach to predicting both cell-type-specific and cell-type independent features of large-scale nuclear architecture, using data from RNA-seq experiments as a proxy for activity and a Hi-C-derived description of chromosome looping in each cell type. We work with three different models, which we call the gene density model, the gene expression model, and the combined model. These model provides a unified understanding of a number of common features of large-scale nuclear architecture observed across diverse cell-types.

3.1 Overview of models and parametrization

In the **gene density model**, the number of genes in 1 Mb interval of chromatin regions is counted from the GENCODE database mentioned in 2.1.1. We vary the number of active monomers (those that have a large number of genes in a 1 Mb chromatin region), to include 5%, 10%, and 20% of them by the number of genes they contain, and assign them an effective temperature of $T = 12$. We study two versions of the **gene expression model**. In the first version, we use gene expression data from the HeLa cell line, altering the fraction of active monomers and the temperature assignments to probe how our results depend on the details of the model. For this first version of our model, we consider two cases. In the first case, we vary the number of active monomers 5%, 10% and 20% and assign them an effective temperature of $T = 12$. This enables us to check the role of the fraction of monomers we take to be active. In the second case, we chose 5% (higher amount of gene expression in 1 Mb chromatin regions) active monomers and assign them different effective temperature $T = 6$, $T = 10$ or $T = 20$. This enables us to examine the role of the difference in effective temperatures between active and passive monomers.

In the second version of the gene expression model, we use the gene expression data of GM12878, IMR90, NHEK, HMEC, HUVEC cell types and define active monomers using the derivative cutoff method mentioned in 2.1.2.

For the **combined model**, we use the gene expression data as a base for initial active temperature assignments. To that, we add a further a fraction of the most active monomers (the top 5%) from the gene density calculation. These are assigned an effective temperature of $T = 12$ as mentioned in 2.1.3.

An overview of our models is supplied in Table 3.1.

Table 3.1: Overview of our models

Model	Cell type (or database)	Percentage of active monomers (or method)	Temperature of active monomers
Gene density model	GENCODE	Top 5%, 10% or 20%	$T = 12$
Gene expression model I	HeLa	Top 5%, 10%, 20% Top 5%	$T = 12$ $T = 6, T = 10, T = 20$
Gene expression model II	GM12878, IMR90, NHEK, HMEC, HUVEC	Derivative method	$T = 6 - 12$
Combined model	GM12878, IMR90, NHEK, HMEC, HUVEC	Derivative method + 5% from gene density	$T = 6 - 12$

3.2 Results from the Gene Density Model

We simulate 46 polymer chains representing chromosomes in a confined spherical nucleus using Langevin dynamics. We begin with the haploid case (23 chromosomes) in order to include random loops, ≈ 700 of them. We generate these by specifying a probability for two random monomers on the same chain to be in contact, hence this number fluctuates slightly between random realizations. The same set of loops are replicated for the homologous chromosomes. Thus, the number of loops indicated above is doubled for the diploid genome which we simulate.

The activity of monomers in the gene density model is defined using the GENCODE density. We show simulation outputs for three different input conditions where top 5% (red line), 10% (black line) or 20% (magenta line) monomers are assigned an active temperature of $T = 12$ and the remaining monomers are assigned an effective temperature of $T = 1$. Figure 3.1 shows $S(R)$, the radial distribution of the monomers associated with each chromosome, for the gene density model. The distribution function does not change by a substantial amount for most chromosomes

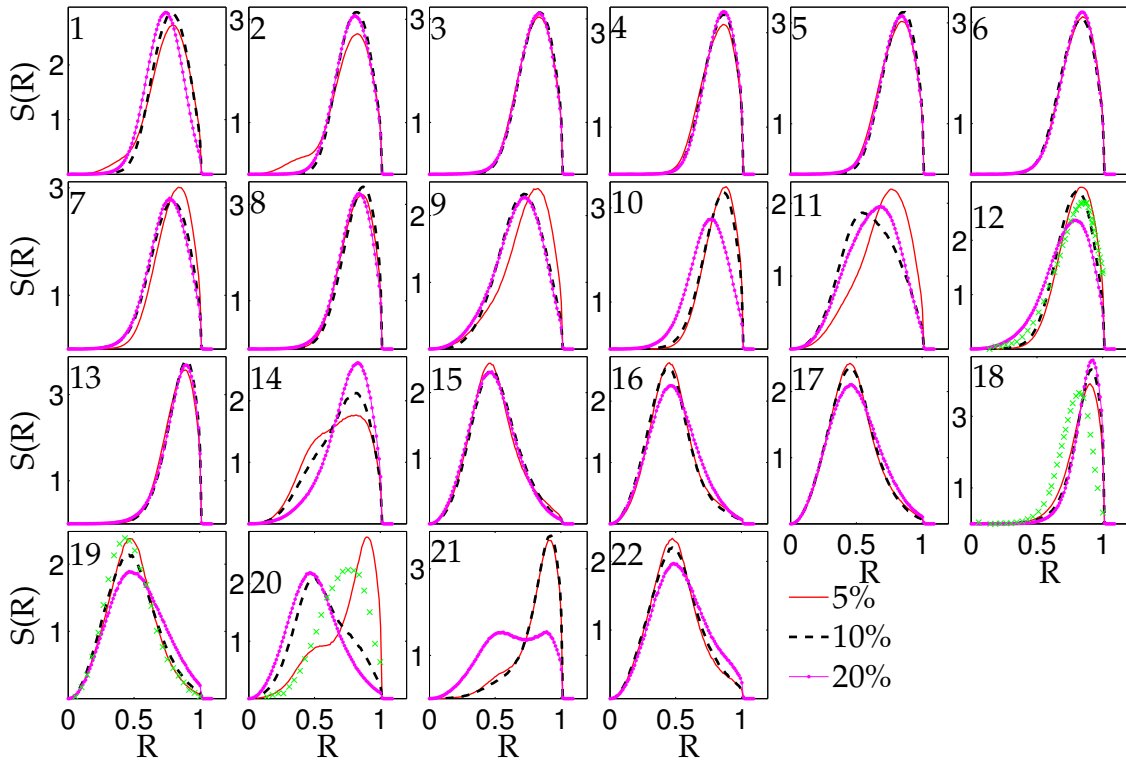


Figure 3.1: $S(R)$, the radial distribution of the monomer density associated to each chromosome, indicated by the chromosome number in the corner of each subplot. The top 5% (solid red line), 10% (black dashed line) and 20% (magenta dot-dashed line) of monomers by gene density are assigned an active temperature of 12 times the temperature assigned to inactive monomers, which is scaled to be the physiological temperature. We show experimental DNA density distribution for chromosomes 12, 20, 18 and 19 obtained from Ref. [Kreth et al., 2004].

as the fraction of active monomers is varied, with the exception of chromosomes 20 and 21. The experimental data agree reasonably with the distribution functions calculated for the 5% cutoff. The calculation captures, in particular, the very different distribution of chromosomes 18 and 19.

Figure 3.2 shows $S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome. The top 5% (red line), 10% (black line) and 20% (magenta line) of monomers by gene density are assigned an active temperature of $T = 12$ times the temperature assigned to inactive monomers, the physiological temperature. As the number of active monomers increases in the chromosome, their positioning shifts towards the interior of the nucleus. Experimental data for the gene-rich chromosome

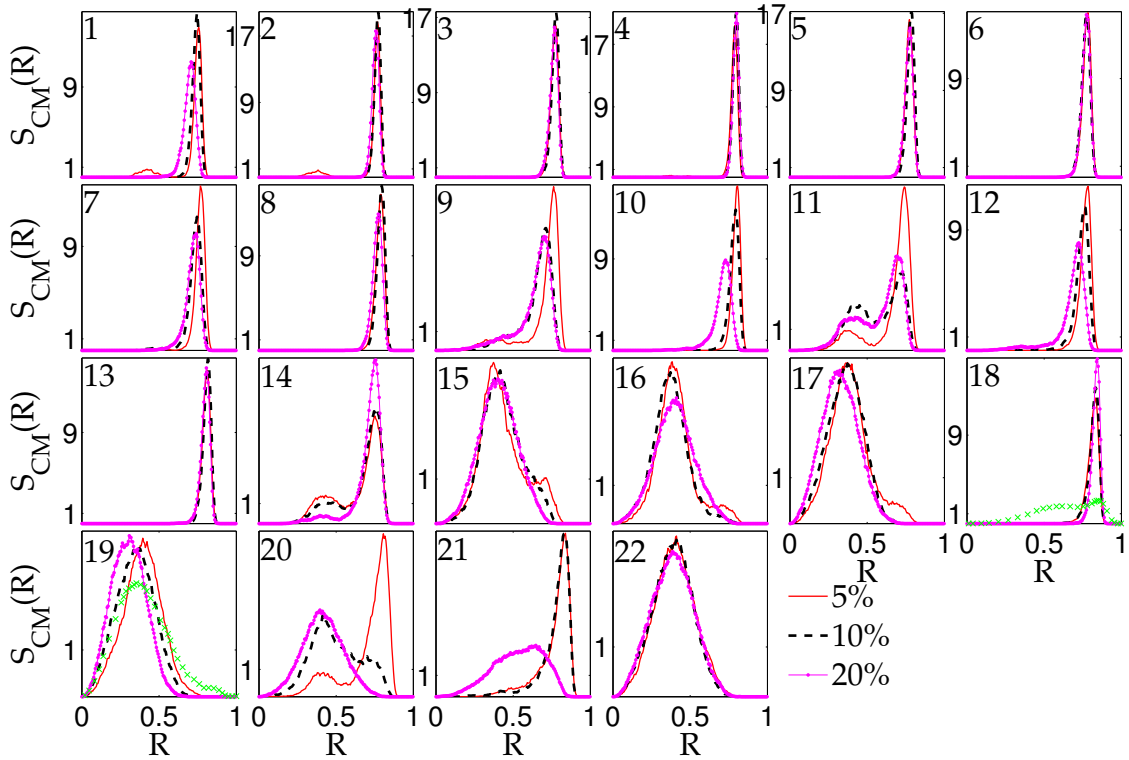


Figure 3.2: $S_{CM}(R)$, the radial distribution of the centre of mass of each **chromosome**, indicated by the chromosome number in the corner of each subplot. The top 5% (solid red line), 10% (black dashed line) and 20% (magenta dot-dashed line) of monomers by gene density are assigned an active temperature of 12 times the temperature assigned to inactive monomers, which is scaled to be the thermodynamic temperature. We show experimental relative centre of mass distribution data (green crosses) for chromosomes 18 and 19 obtained from Ref. [Kalhor et al., 2011].

19 matches well with simulation data. Together these data shows that activity is important for positioning.

Figure 3.3 shows the mean centre of mass positions of all chromosomes plotted against chromosome sizes in Mb, for the gene density model. The y-axis indicates the mean position of the centre of mass relative to the centre and the periphery of the nucleus. The top 5% (green circle), 10% (blue cross) and 20% (red star) of monomers by gene density are assigned an active temperature of $T = 12$.

From this plot, we observe that large chromosomes with both high and low gene density are generally found towards the periphery, indicating that chromosome size

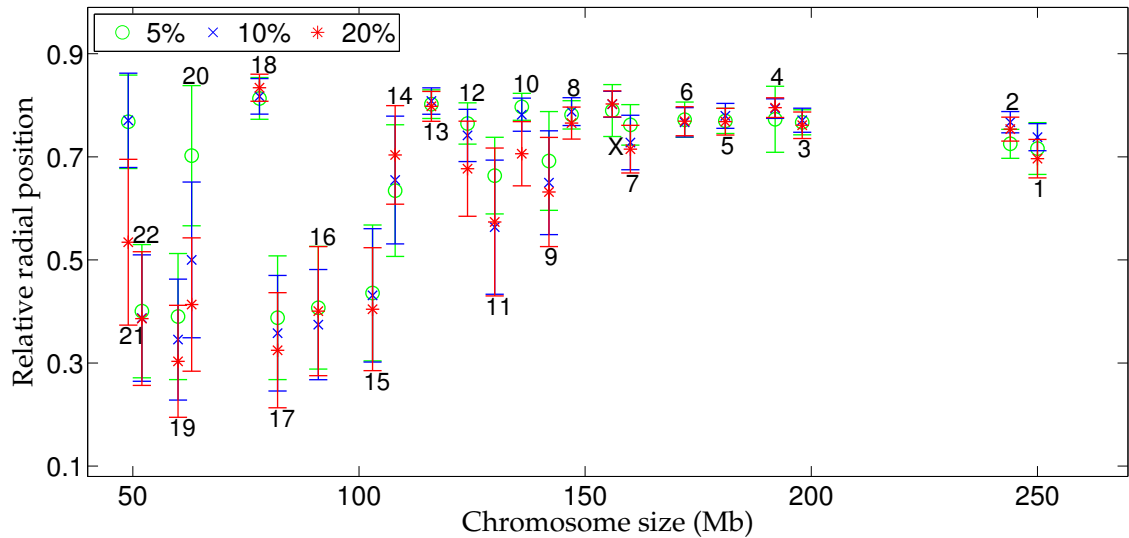


Figure 3.3: **Relative centre of mass position of all chromosomes plotted against their sizes in Mb on the x-axis.** The range in y-axis from 0 to 1 represents the radial distance from the centre of the spherical nucleus to its periphery. The top 5% (green circle), 10% (blue cross) and 20% (red star) of monomers by gene density are assigned an active temperature of $T = 12$ times the temperature assigned to inactive monomers, which is scaled to be the physiological temperature. The vertical line denotes the standard errorbars computed for each chromosome independently corresponding to a standard deviation above and below the mean value. The chromosome numbers are shown alongside.

may also play a role in determining their positions. Small chromosomes with high gene density are found more often towards the centre of the nucleus. Small chromosomes with low gene density can occupy positions intermediate between central and peripheral, suggesting that both size and gene density together determine their absolute positions. Overall as the number of active monomers increases, the positioning of the smaller chromosomes shows a shift towards the nuclear interior.

The relative centre of mass positioning data appears to have two branches, with a lower branch showing a rough linear dependence of location relative to the centre of the nucleus with chromosome size and an upper branch which includes the small gene-poor chromosome 18 which shows far less dispersion with chromosome size. The mean centre of mass position of chromosomes 21 and 20 show a large shift towards the centre of the nucleus as the overall percentage of active monomers is increased

from 10% to 20%. This indicates that activity strongly influences positioning but that size can also play a role in determining where chromosomes are positioned.

3.3 Results from the Gene Expression Model I

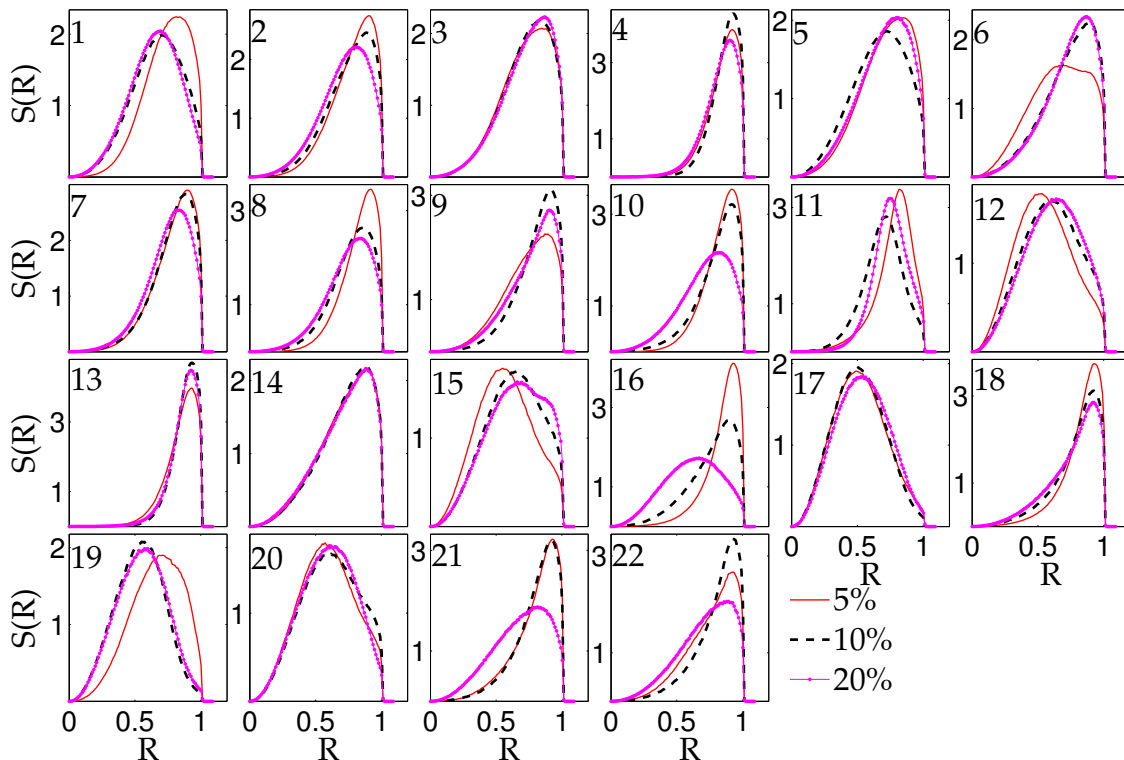


Figure 3.4: $S(R)$, the radial distribution of the monomer associated to each chromosome is plotted in each subfigure. The top 5% (red line), 10% (black line) and 20% (magenta line) of monomers in the HeLa cell line with the largest expression values are assigned an active temperature of $T = 12$ times the temperature assigned to the inactive monomers $T = 1$, which is scaled to be the physiological temperature.

The activity of monomers in this version of the gene expression model is defined using expressed genes from the HeLa cell line. We included 106 number of loops from Hi-C experiments and 120 number of random loops using the random loop model between monomers within each chromosome in the haploid cell nucleus; these numbers are doubled for the diploid case, which we simulate.

In Figure 3.4, we show $S(R)$, the radial distribution of the monomer density asso-

ciated with each chromosome. The top 5% (red), 10% (black) and 20% (magenta) monomers in the HeLa cell line with hog expression values are assigned an active temperature of $T = 12$. These distributions suggest as the number of active monomers increases, the positioning of active chromosomes shifts towards the interior of the nucleus.

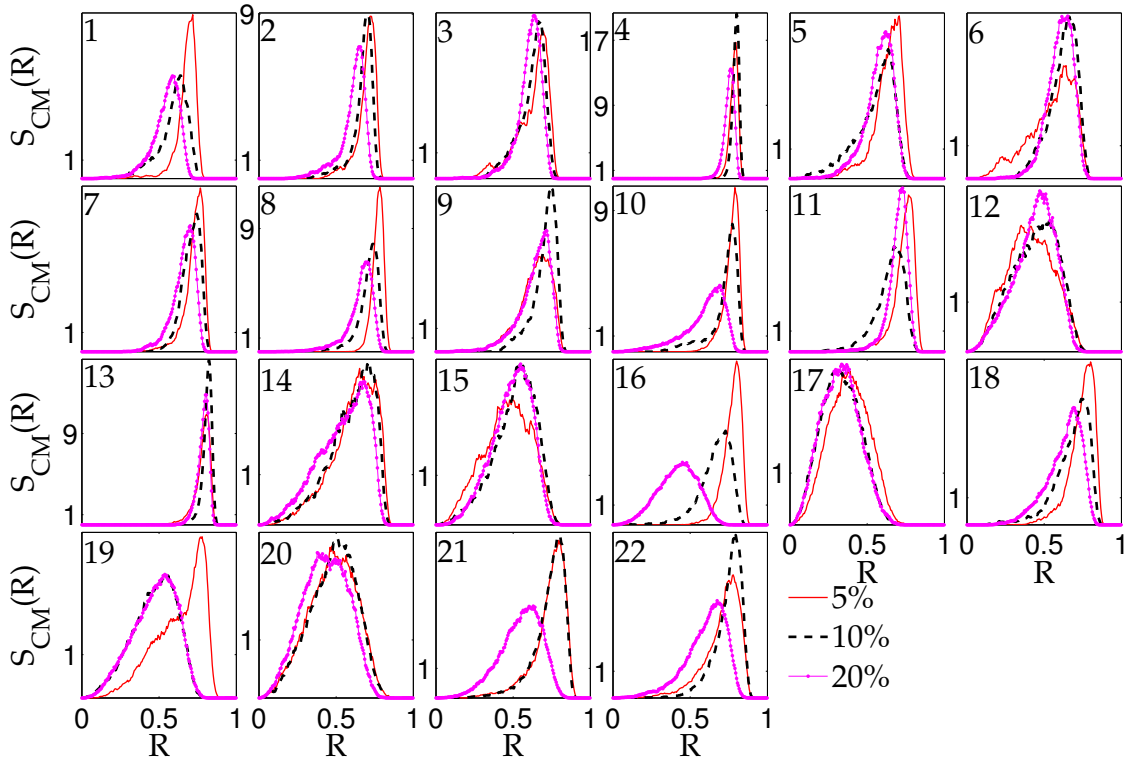


Figure 3.5: $S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome for HeLa cell line, indicated by the chromosome number in the corner of each subfigure. The top 5% (solid line), 10% (black line) and 20% (magenta line) of monomers by gene expression are assigned an active temperature of $T = 12$ times the temperature assigned to inactive monomers $T = 1$, which is scaled to be the physiological temperature.

Figure 3.5, shows $S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome. The top 5% (red), 10% (black) and 20% (magenta) high expressed monomers in the HeLa cell line are assigned a temperature of $T = 12$ times the temperature assigned to inactive monomers. Broadly, major features in the gene expression model are comparable to those in the gene density model. It can be seen that the width of the distribution and positioning of the distribution peak

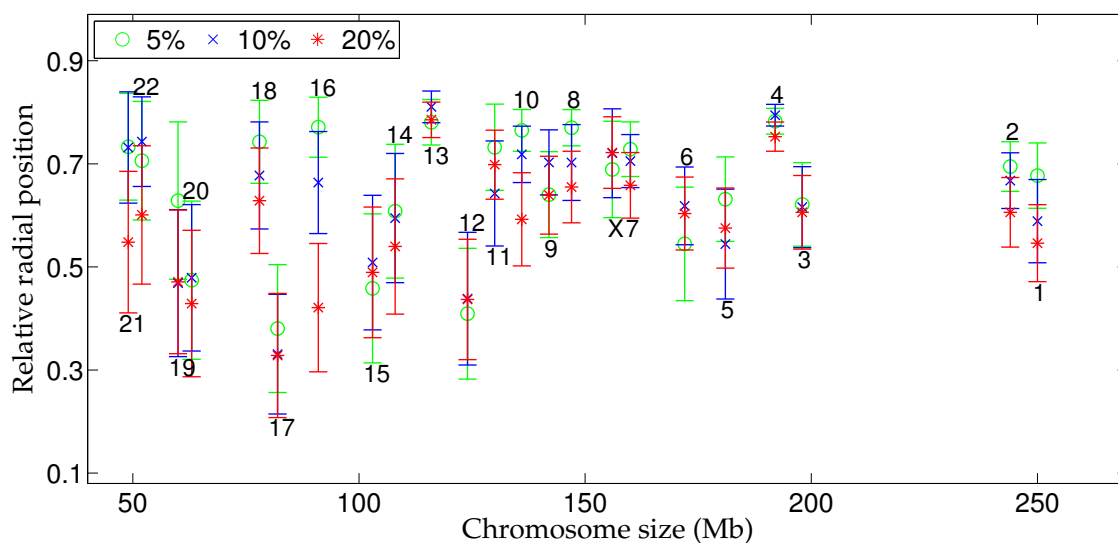


Figure 3.6: **Relative centre of mass position of all chromosomes plotted against their sizes with a vertical error-bar corresponding to standard deviation of their positioning.** The range in y-axis from 0 to 1 represents the radial distance from the centre of the spherical nucleus to its periphery. The top 5% (green circle), 10% (blue cross) and 20% (red star) of monomers by HeLa gene expression are assigned an active temperature of $T = 12$ times the temperature assigned to inactive monomers $T = 1$. The chromosome number is indicated in top or bottom of each errorbar.

for the larger chromosome are similar while it is different for smaller and gene-rich chromosome. The position of chromosome 12 differs in the gene density and the gene expression models. It is found close to the periphery in the gene density model (Figure 3.2) and towards the interior in the gene expression model (Figure 3.5). This ordering is reversed for chromosome 16 and 22 between the gene density and the gene expression models. Finally, chromosome 19 is shifted from a more interior position to a less interior position when going from gene density to gene expression models.

Figure 3.6 shows the centre of mass position of all chromosomes plotted against their sizes. Given the gene expression data, gene dense chromosomes tend to be found towards the nuclear periphery (see Figure 3.6), in comparison to their positioning in the gene density model (see Figure 3.3) while the larger and gene-poor chromosomes are more interior in the case of the gene expression model compared to the gene

density model. As the fraction of active monomers is increased, the smallest chromosomes 21 and 22 move inward towards the centre of the nucleus. The centre of mass positioning of chromosomes 20 and 21, in particular, are very sensitive to the fraction of active monomers, but less sensitive to the active temperature assigned to those monomers (see Figure 3.9). Chromosome 16 shows a larger shift in centre of mass position in the gene expression model in comparison to the gene density model as the fraction of active monomers is increased. Overall, the gene expression data leads to a larger variation in the position of the mean centre of mass as the fraction of active monomers is increased, in particular for chromosomes 12, 16, 19 and 22.

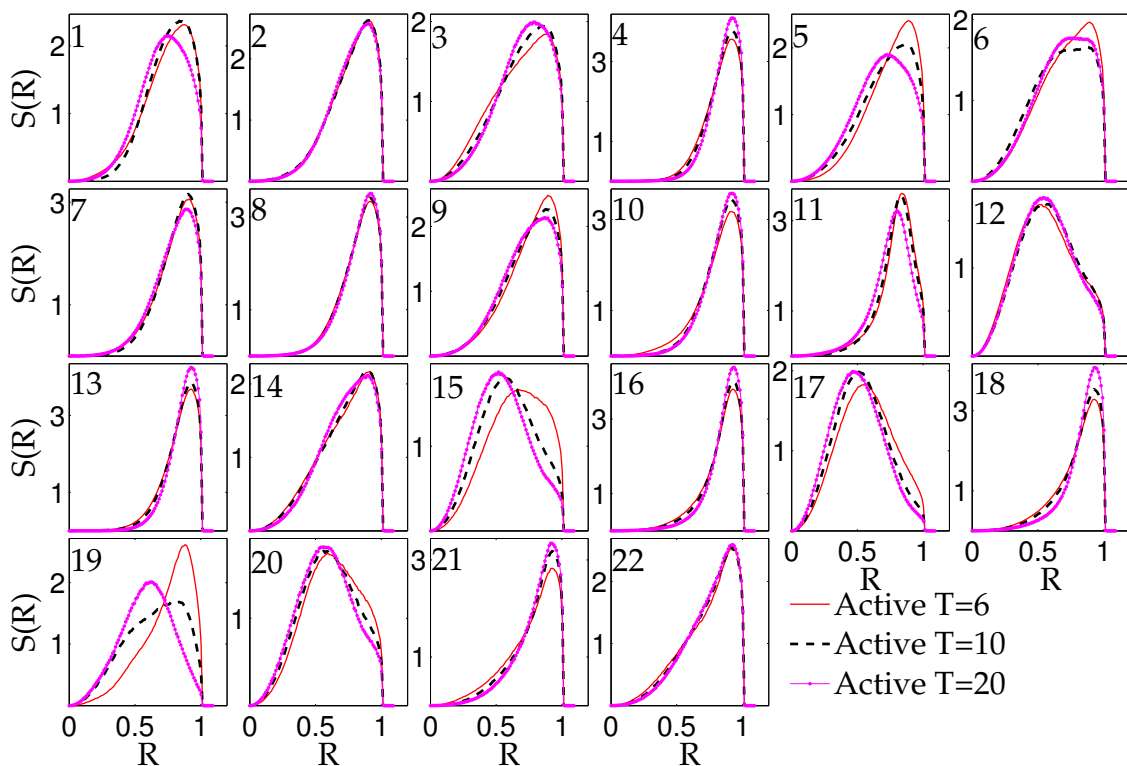


Figure 3.7: $S(\mathbf{R})$, the radial distribution of the monomer density associated to each chromosome for HeLa cell line, indicated by the chromosome number in the corner of each subfigure. The top 5% monomers contain high gene expression in HeLa cell line are assigned an active temperature of 6 (solid red line), 10 (black dashed line) and 20 (magenta dot-dashed line) times the temperature assigned to inactive monomers $T = 1$, which is scaled to be the physiological temperature.

Figure 3.7 shows $S(\mathbf{R})$, the radial distribution of the monomer density associated with each chromosome. The data shown here correspond to three different assign-

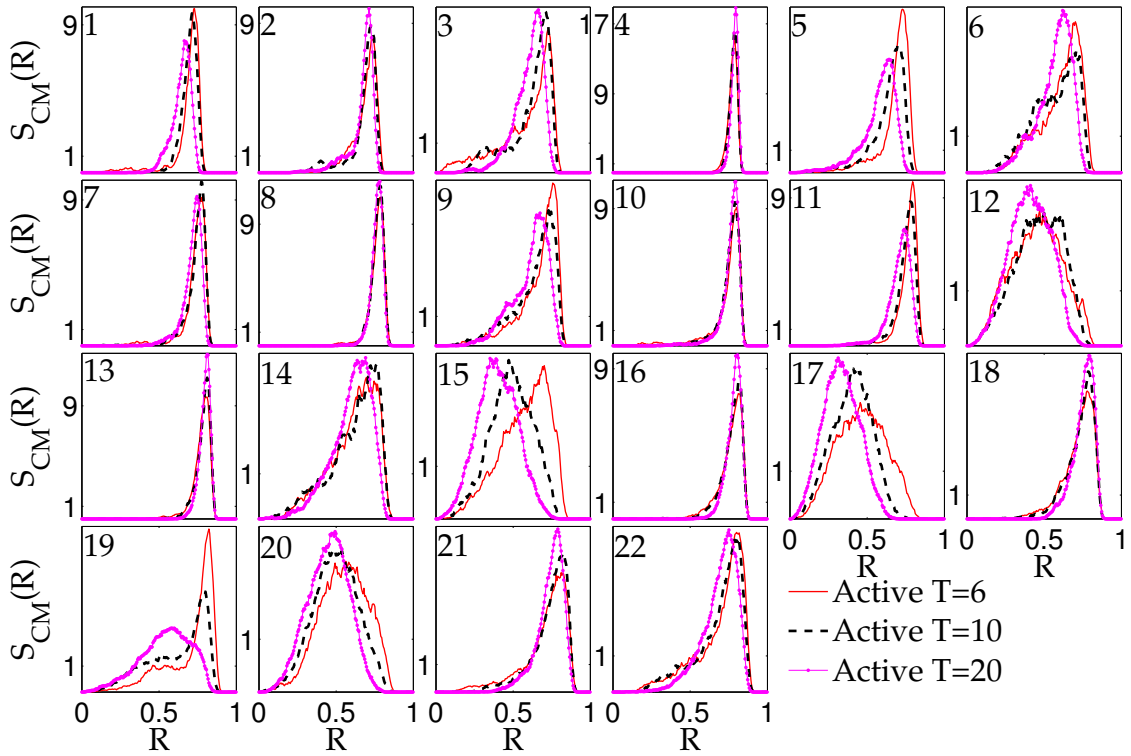


Figure 3.8: $S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome for HeLa cell line, indicated by the chromosome number in the corner of each subfigure. The top 5% monomers contain high gene expression in HeLa cell line are assigned an active temperature of 6 (solid red line), 10 (black dashed line) and 20 (magenta dot-dashed line) times the temperature assigned to the inactive monomers $T = 1$.

ments of active temperature $T = 6$ (solid red line), $T = 10$ (black dashed line) and $T = 20$ (magenta dot-dashed line), with a 5% fixed fraction of cutoff on the gene expression associated with active monomers. As the active temperature increases of the fragments of monomers for the same chromosome, the positioning of chromosomes shifts on average towards the interior of the nucleus.

Figure 3.8 shows $S_{CM}(R)$, the radial distribution of the centre of mass of each chromosome. There is some variation with the maximum temperature. For active temperatures $T = 6$ or $T = 10$, there are no significant changes in the peak of the distribution. Larger active temperatures ($T = 20$) tend to shift the peak values of the distribution to small values of R at least for gene-rich chromosomes. Chromosome 16 and 17 are both gene-rich chromosomes with similar sizes. They show an interesting

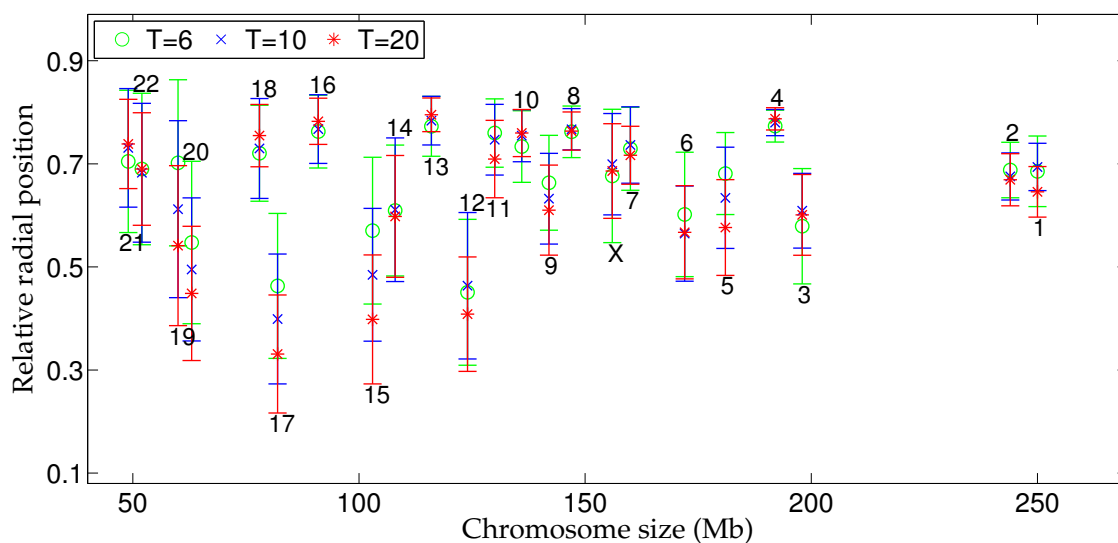


Figure 3.9: **Relative centre of mass position of all chromosomes plotted against their sizes for HeLa cell line**, each with an errorbar corresponding to standard deviation of the data. The range in y-axis from 0 to 1 represents the radial distance from the centre of the spherical nucleus to its periphery. The top 5% monomers contain high gene expression in HeLa cell line are assigned an active temperature of 6 (green circle), 10 (blue cross) and 20 (red star) times the temperature assigned to the inactive monomers. The chromosome number is indicated above or below each error-bar.

behaviour when their peak distribution is compared between variation of high active monomers and variation of larger active temperatures for the gene expression model. As the fraction of active monomers increases, the peak position shifts from the periphery to the interior in chromosome 16 (see Figure 3.5) but there is no difference of the peak position when the active temperature varies between $T = 6$ to $T = 20$ (see Figure 3.8). In case of chromosome 17, as the fraction of active monomers increase, the position of the peak does not change significantly (see Figure 3.5) but larger active temperatures shift the peak from the periphery ($T = 20$) to the interior ($T = 20$) (see Figure 3.8).

Figure 3.9 shows the centre of mass position of all chromosomes plotted against their sizes. The centre of the nucleus is at the bottom of the y-axis and the periphery of the nucleus is towards the top of the same axis. Chromosomes 13, 4, X, 2, 7, 16 and 14 do not show any changes in positioning due to variation of effective temperature.

The positioning of chromosome 16 is towards the periphery in the higher active temperature case (Figure 3.9) compared to Figure 3.6.

We conclude from the first version of the gene expression model for the HeLa cell line, that the effect of varying the temperature of active monomers at a fixed small fraction of active monomers appears to be smaller than that induced by purely changing the fraction of active monomers at a fixed active temperature. The number and nature of loops that we incorporate into our simulations are also important in determining the chromosome positioning, since they determine the overall compactness of the chromosome.

3.4 Results from the Gene Expression Model II

We study the second version of gene expression model thoroughly for the GM12878, HMEC, IMR90, HUVEC, and NHEK cell types in this and the following section, which discusses the combined model. The assignment of active temperature in this version of model is more complex than in the previous version of gene expression model. Here the gene expression curve is divided into three different regimes using the numerical derivative method explained in Chapter 2.1.2. Monomers in the lower part of the curve are assigned an inactive temperature of $T = 1$, those in the plateau part of the curve are assigned an active temperature of $T = 6$ and monomers in the upper part of the curve are assigned an interpolated active temperature between $T = 7$ to $T = 12$ (see Figure 2.3). We used experimental Hi-C long range loops for non-bonded monomers, i.e. loops whose length is more than 2 Mb. We leave out random loops.

Figure 3.10 shows the simulation predictions for the radial distribution function $S(R)$ of the monomer density associated with each chromosome from the GM12878 cell type in red. We also show the experimental data in magenta oval for chromosome 12,

20, 18 and 19 from Ref. [Kreth et al., 2004]. The experimental and simulation peak for chromosome 12 and 18 appear at the same location. Chromosome 19 shows a somewhat flattened distribution (red colour); this is improved in the combined model (blue). Chromosome 20 also shows a similarly flat distribution but so does the combined model. Chromosomes Xa and Xi are positioned differently because all the monomers of chromosome Xi are inactive. Here both chromosomes Xa and Xi have a similar number of Hi-C loops.

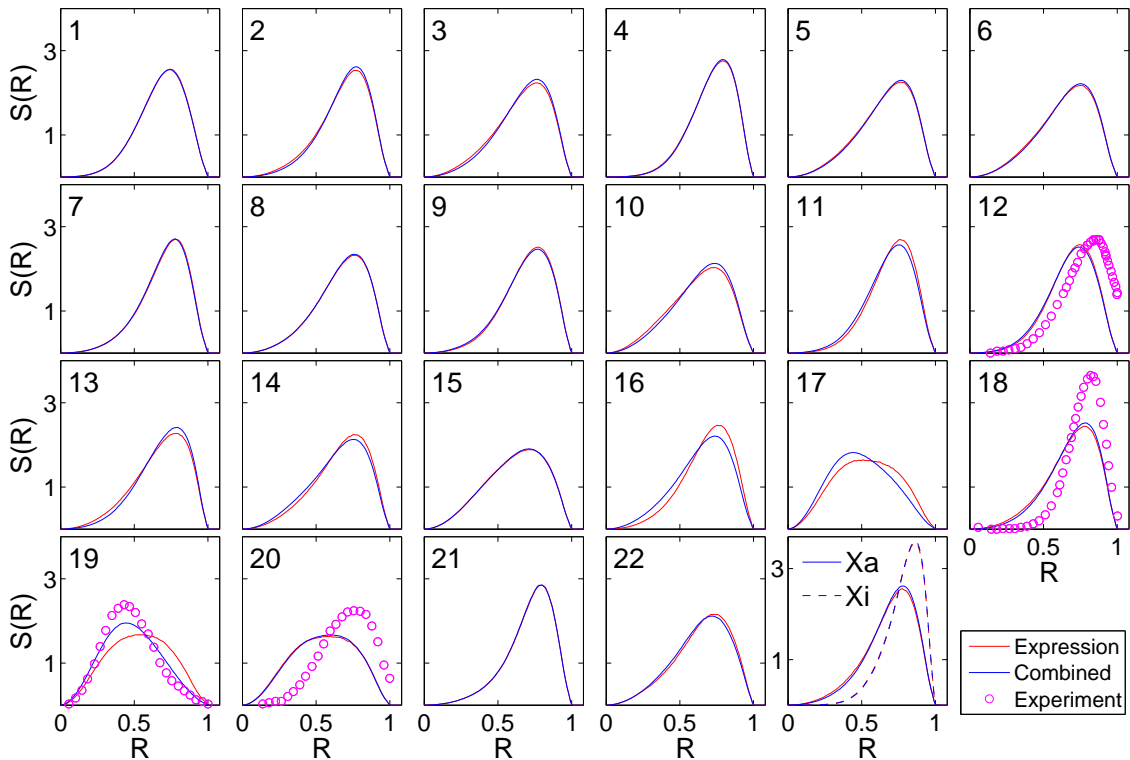


Figure 3.10: **The distribution functions $S(R)$ for all chromosomes, within the gene expression and the combined model for the GM12878 cell type.** $S(R)$ monomer distribution of each chromosome for gene expression (red) and combined model (blue) is shown. Experimental data from Ref. [Kreth et al., 2004] for Chromosomes 12, 18, 19 and 20 are shown in magenta.

Figure 3.11 shows predictions for the centre of mass distribution $S_{CM}(R)$ of all chromosomes for GM12878 cell type, in red. We also show the experimental data in magenta oval for chromosome 18 and 19 from Ref. [Kalhor et al., 2011]. Chromosomes Xa and Xi show differential positioning. The distribution of $S_{CM}(R)$ has narrower peaks than in $S(R)$. The location of the peak varies among chromosomes

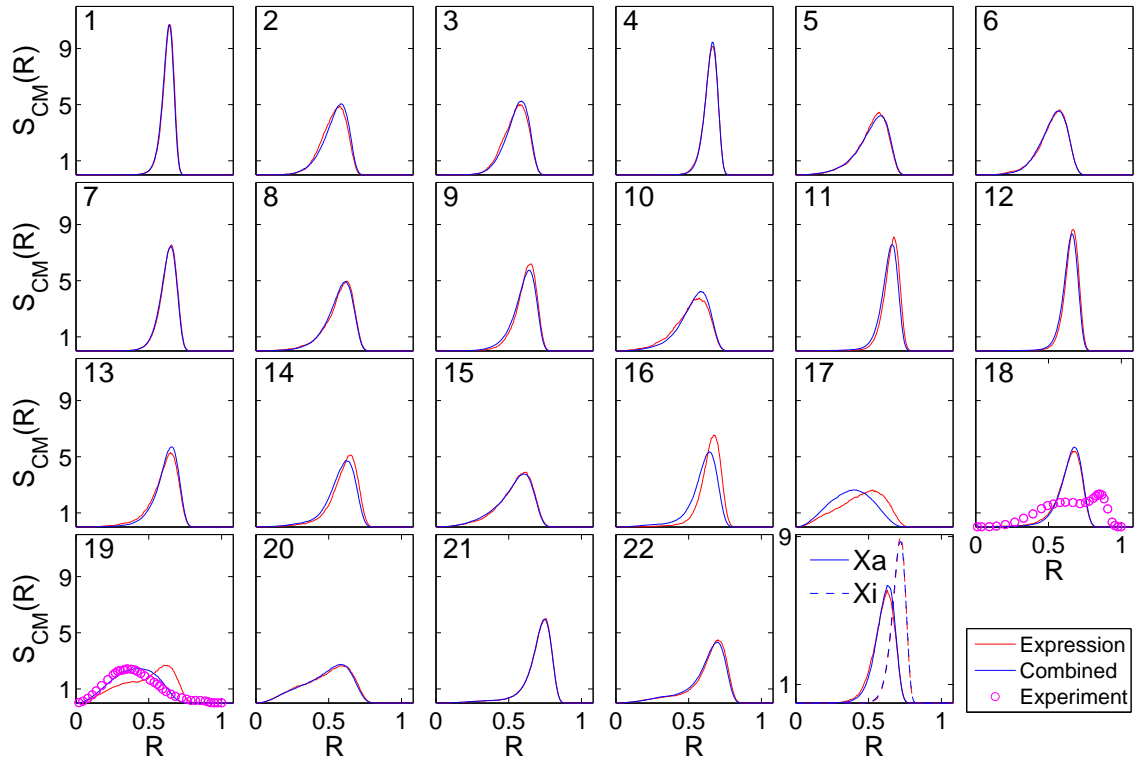


Figure 3.11: **The centre of mass distribution $S_{CM}(R)$ for all chromosomes, within the gene expression and combined model for the GM12878 cell type.** $S_{CM}(R)$ centre of mass distribution of each chromosome for gene expression (red) and combined model (blue) is shown. Experimental data from Ref. [Kalhor et al., 2011] for Chromosomes 18 and 19 are shown in magenta.

more for $S_{CM}(R)$ than it does in $S(R)$. Overall, apart from the gene-rich chromosomes, we do not see a substantial difference between the predictions of the gene expression and combined models.

3.5 Results from the Combined Model

The combined model bases itself largely on the second version of gene expression model, since it uses two cutoffs for defining active monomers but also assigns an additional 5% monomers that have high gene density an effective temperature of $T = 12$. Thus, the combined model has features of both gene density and gene expression model. It provides the most comprehensive fits to the experimental data. As in the earlier gene expression model, here we include only those Hi-C long range

loops from experiments whose length is more than 2 Mb. we use no random loops. A typical snapshot of assignment of different temperature to monomers of chromosome 12 is shown in Figure 3.12 for 5 cell types. This exhibits how our activity assignment changes for different cell type.

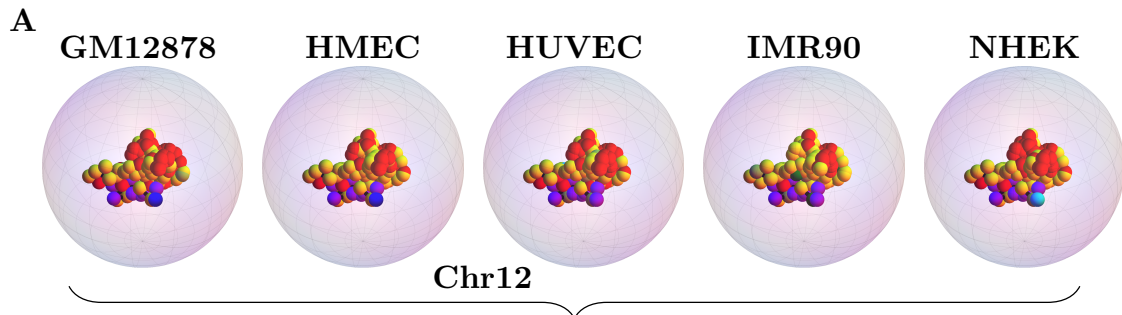


Figure 3.12: **Schematic of effective temperature assignment to each monomer for chromosome 12 for the combined model.** The red monomers are simulated at $T = 1$, yellow at $T = 6$, yellow-green at $T = 7$, green at $T = 8$, cyan at $T = 9$, blue at $T = 10$, indigo at $T = 11$ and violet at $T = 12$ times the physiological temperature.

3.5.1 $S(\mathbf{R})$ and $S_{CM}(\mathbf{R})$

Our computed $S(\mathbf{R})$ for each chromosomes in the 5 cell types GM12878 (blue), HMEC(green), HUVEC(black), IMR90(cyan) and NHEK(red) are shown in Figure 3.13. We compare our results for chromosomes 12, 20, 18 and 19 (shown in magenta ovals) with experimental results for the GM12878 cell type, extracted from Ref. [Kreth et al., 2004]. $S(\mathbf{R})$ for chromosomes 18 and 19 exhibit well-separated peaks, a feature that holds across cell types. Simulation data for the different cell types all yield fairly similar plots for $S(\mathbf{R})$, with the exception of chromosomes 17 and 20 in the GM12878 cell type where, although the simulation and experimental data peak at somewhat different locations, the overall shape of the curve is rendered accurately, including the relative shift in peak positions. Switching off activity completely in chromosome Xi compared to Xa leads to differential positioning, even those both of them have identical loop assignments.

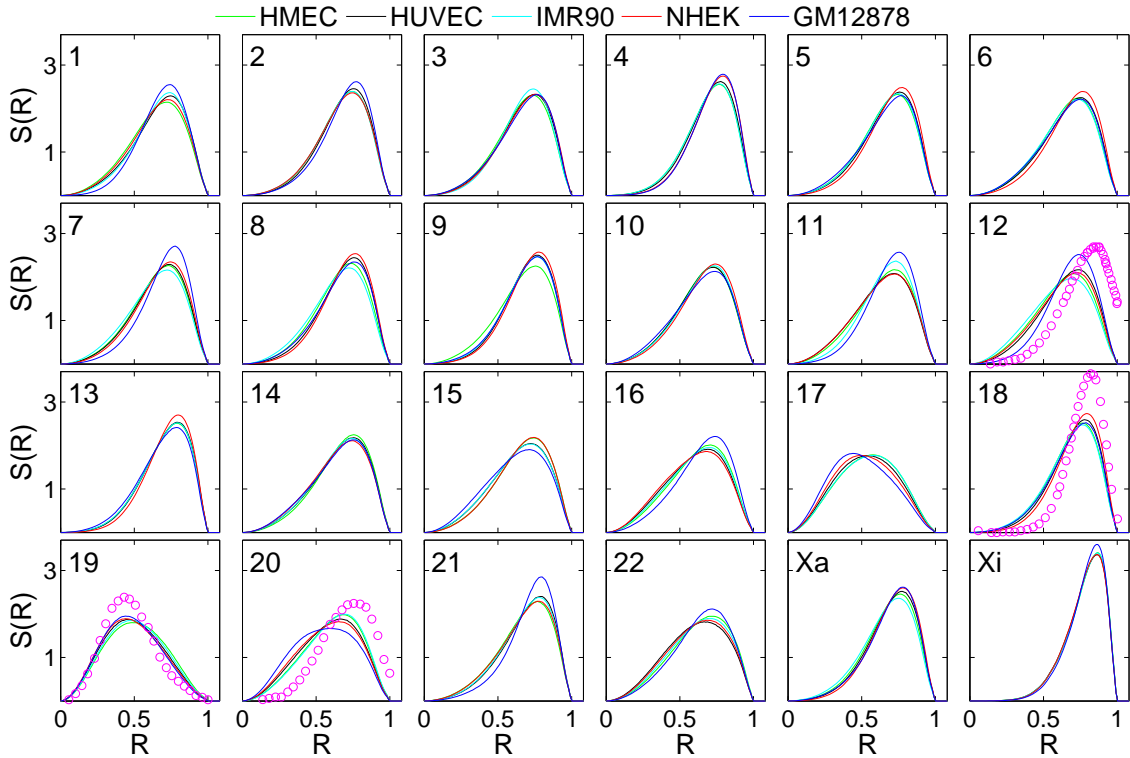


Figure 3.13: **Computed distribution functions $S(R)$ for all simulated chromosomes across 5 cell types for the combined model.** Distribution functions $S(R)$ for each simulated chromosomes for GM12878 (blue), HMEC(green), HUVEC(black), IMR90(cyan) and NHEK(red) is shown. Chromosome numbers are mentioned in the left upper corner of each subfigure. Experimental data for chromosomes 12, 20, 18 and 19 obtained from Ref [Kreth et al., 2004] for the GM12878 cell type is plotted in magenta ovals together with the simulation prediction.

Figure 3.14 shows the distribution of centre of mass $S_{CM}(R)$ of each chromosomes for 5 cell types GM12878 (blue), HMEC(green), HUVEC(black), IMR90(cyan) and NHEK(red). We compare our results for chromosomes 18 and 19 (shown in magenta ovals) with experimental data of GM12878 cell type extracted from Ref. [Kalhor et al., 2011]. The centre of mass distribution is captured with reasonable accuracy, especially for chromosome 19. The somewhat broader distribution of $S_{CM}(R)$ for chromosome 18 is also in agreement with the left tail of the experimental data, although the experimental data show a weaker and more outward shifted peak than the simulation prediction. Broadly, differences in positioning of chromosomes across cell types are more apparent in $S_{CM}(R)$ compared to $S(R)$. The largest variability across cell types is seen in chromosomes 1, 4, 7, 11, 12, 16, 21 and 22. $S_{CM}(R)$ for

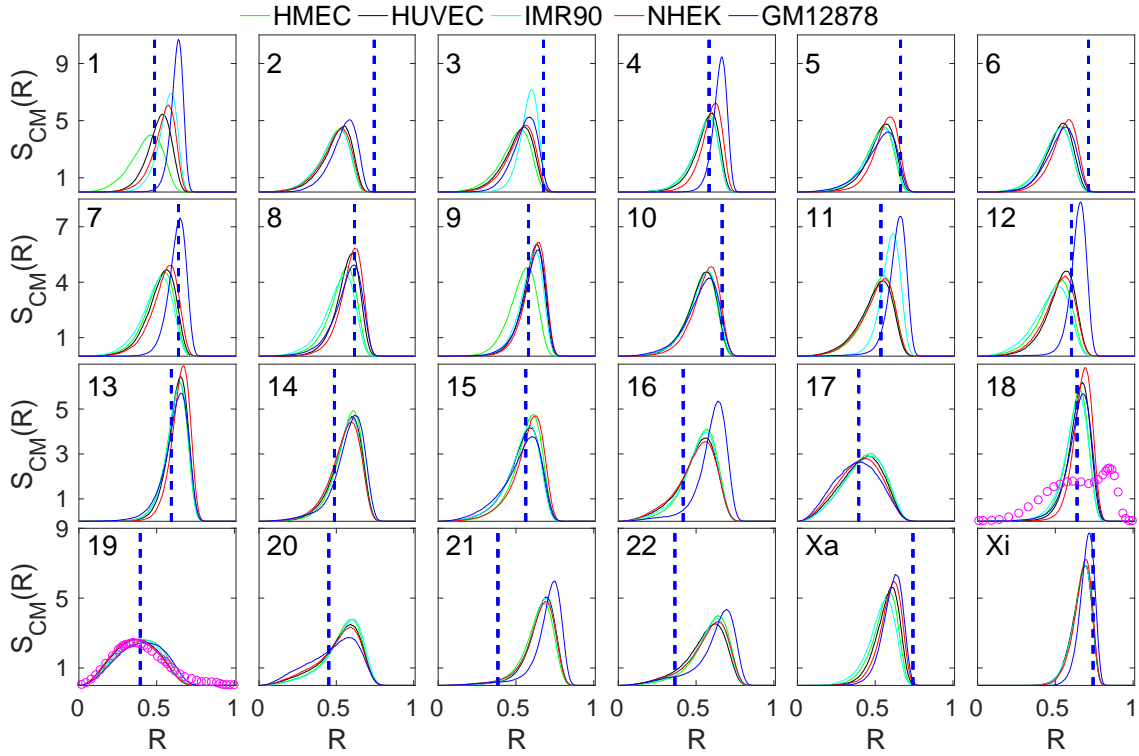


Figure 3.14: **Centre of mass distributions $S_{CM}(R)$ for all simulated chromosome across 5 cell types for the combined model.** The centre of mass distribution $S_{CM}(R)$ for each simulated chromosomes for GM12878(blue), HMEC(green), HUVEC(black), IMR90(cyan) and NHEK(red) is shown. Chromosome numbers are mentioned in the left upper corner of each subfigure. Experimental data of chromosomes 18 and 19 obtained from Ref. [Kalhor et al., 2011] for the GM12878 cell type are plotted in magenta ovals together with the simulation prediction. The vertical dashed line in each subplot refers experimental relative center of mass position of chromosomes in GM12878 cell type.

gene-poor chromosomes appears to be sharply peaked while gene-rich chromosomes have relatively broader distributions across all cell types.

Figure 3.15 shows how $S(R)$ for the combined model of GM12878 cell type varies when we include or exclude looping and activity. There are four possible way to make such a combination: (i) both loops and activities are present (Act:Y, Lps:Y), (ii) loops absent but activity present (Act:Y, Lps:N), (iii) loops present but activity absent (Act:N, Lps:Y) and (iv) both activity and loops are absent (Act:N, Lps:N). The predictions of the different models differ substantially for both the gene-rich chromosomes as well as the smallest chromosomes. In general, both looping and

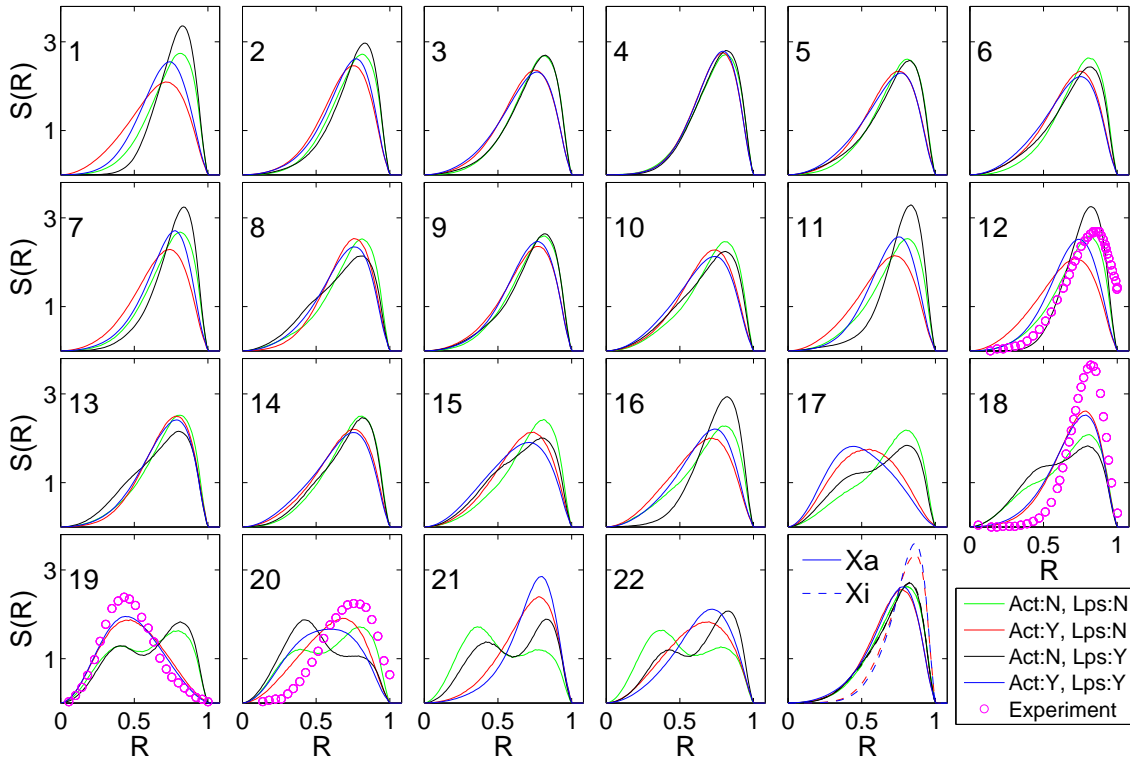


Figure 3.15: **Distribution functions $S(R)$ for all chromosomes with different combinations of activity and loops for the GM12878 cell type.** The $S(R)$ monomer density distribution for each chromosome is shown for the models mentioned below. All cases involving activity are shown for the combined model. **Act:N, Lps:N** Both activity and loops are switched off, with all monomers at the same effective temperature of $T = 1$, shown in green; **Act:Y, Lps:N** Activity is present, implying an inhomogeneous distribution of temperatures, but loops are switched off, shown in red; **Act:N, Lps:Y** Activity is absent but loops are present, shown in black; **Act:Y, Lps:Y** Both activity and loops are present, shown in blue colour. This is the original “combined model”, also shown in (Figures 3.10 and 3.13); The experimental data for chromosomes 12, 18, 19 and 20 are shown in magenta colour from Ref. [Kreth et al., 2004].

differential activity are needed to best represent available experimental data.

Similarly, Figure 3.16 shows, how different combinations of monomer activity and loops changes in $S_{CM}(R)$ for the combined model of GM18278 cell type. In the absence of activity smaller chromosomes shows bimodal type of distribution. There is not much difference in $S(R)$ or $S_{CM}(R)$ arising from the presence or absence of loops while the activity is present.

We plot results from the combined model, where both activity and loops are present,

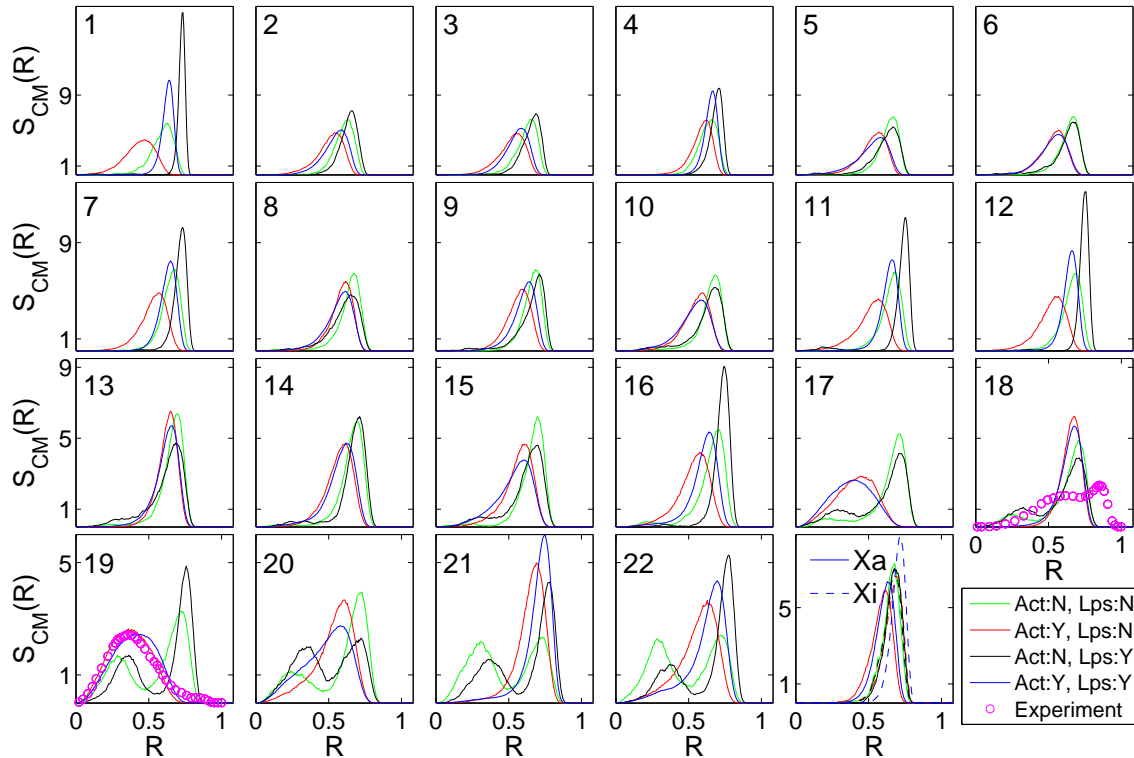


Figure 3.16: **Centre of mass distribution $S_{CM}(R)$ for all chromosomes with different combinations of activity and loops for the GM12878 cell type.** $S_{CM}(R)$ centre of mass distribution for each chromosome is shown for the models described below. **Act:N, Lps:N** Both activity and loops are switched off, with all monomers at the same effective temperature of $T = 1$, shown in green; **Act:Y, Lps:N** Activity is present but loops are switched off, shown in red; **Act:N, Lps:Y** Activity is absent but loops are present, shown in black; **Act:Y, Lps:Y** Both activity and loops are present, shown in blue colour. This is the original “combined model”, also shown in (Figures 3.14 and 3.15); The experimental data for chromosomes 18 and 19 are shown in magenta from Ref. [Kalhor et al., 2011].

alongside results from null model where both activity and loops are absent in Figure 3.17. The $S(R)$ and $S_{CM}(R)$ for small chromosomes shows a bimodal distribution in the absence of activity. We believe that this is likely because all the distributions are obtained as an average of homologous chromosomes. When the activity is absent, steric effects become important and the presence of one chromosome across a radial shell could potentially exclude the other from having a similar radial location. Incorporating activity nullifies this, since it reduces the barrier to crossing of chromosomes.

We conclude that there are subtle differences in the gene density distribution func-

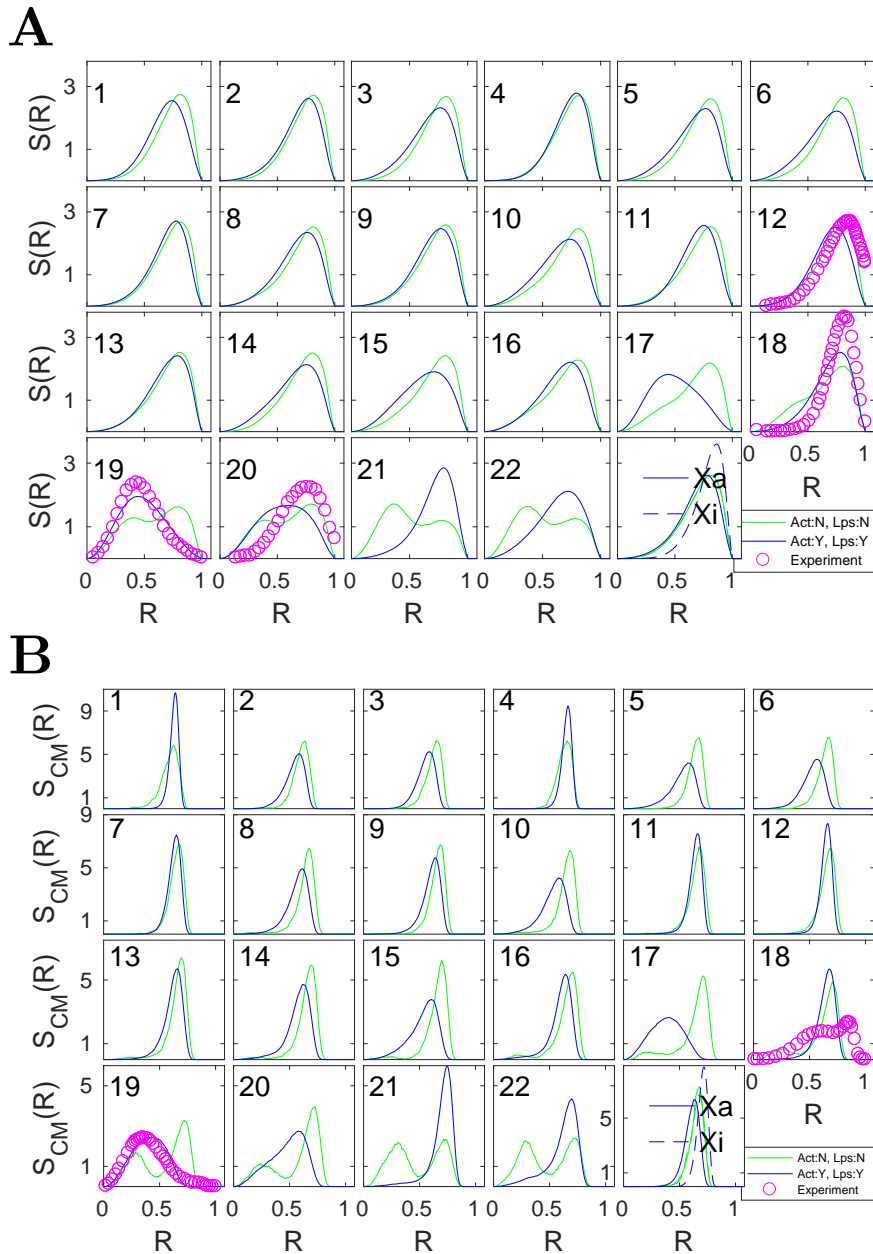


Figure 3.17: **Comparison of two models in presence of both activity and loops (Act:Y, Lps:Y) combined model and in absence of activity and loops (Act:N, Lps:N) for the GM12878 cell type.** Distribution functions $S(R)$ and centre of mass distribution $S_{CM}(R)$ for all chromosomes in Figure A and B respectively. The experimental data for chromosomes 18 and 19 are shown in magenta from Ref. [Kalhor et al., 2011].

tion $S(R)$ as well as in the mean centre of mass distribution $S_{CM}(R)$ of chromosomes across cell types. These originate both in differences in activity profiles across different cell types as well as variations in their loop content.

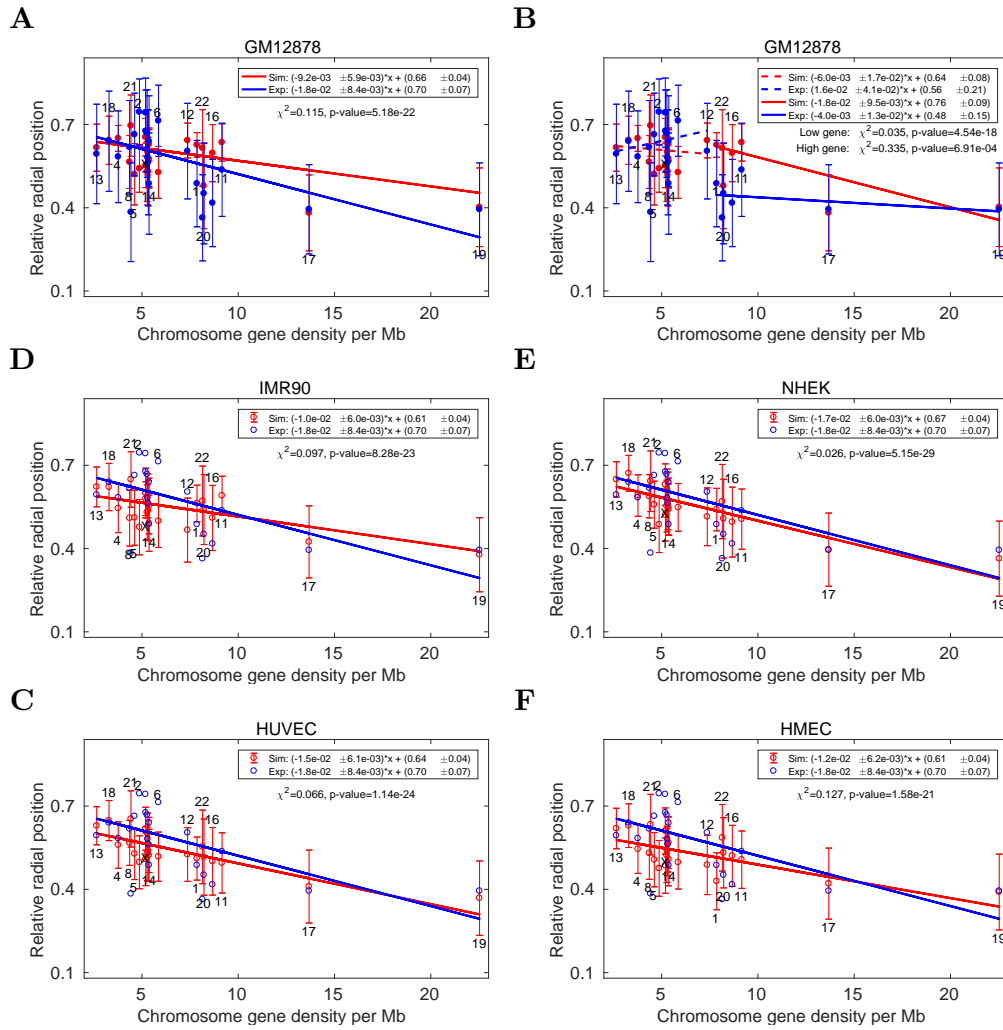


Figure 3.18: Predictions for the mean centre of mass location for each chromosome, computed for the GM12878, HMEC, IMR90, NHEK, and HUVEC cell types plotted as a function of chromosome gene density per Mb, compared to experimental data of GM12878 cell type extracted from Ref. [Kalhor et al., 2011]. Simulation and experimental points are shown using red and blue filled circles respectively in GM12878 cell type together with errorbars while in other cell types GM12878 experimental data shown with hollow blue ovals for illustrative purpose. All the simulation points of 5 cell types and experimental points of GM12878 cell type are fitted to a straight line using weighted least square method and their slope and intercept values are mentioned in the inset of each subfigure. For GM12878 cell type, whether the linear fits are dominated by high gene dense chromosome such as 19, 17 or all chromosomes follows the linear trends. We show two different linear fits in Figure B. Surprisingly, experimental data have positive slope for high gene dense chromosome [19, 17, 11, 16, 20, 22, 1] and negative slope for low gene dense chromosomes [13, 18, 4, 8, 21, 5, 2, X, 3, 10, 9, 15, 7, 14, 6, 12], while simulation data follow the linear fit with a positive slope for all chromosomes. The χ^2 statistics and p-values are also mentioned in each subfigure.

3.5.2 Relative Centre of Mass Positions

Figures 3.18 show our computation of the mean centre of mass of each chromosome within the combined model for the GM12878, IMR90, HUVEC, HMEC, and NHEK cell types against chromosome gene density per Mb. We also show the experimental data for the GM12878 cell type extracted from Ref. [Kalhor et al., 2011]. For other - IMR90, HUVEC, HMEC and NHEK - cell types, GM12878 experimental data is shown for illustrative purposes only. We show fits to straight lines as a function of chromosome gene density per Mb. To find whether the single straight line fits can be improved, we perform separate fits to high gene dense chromosomes [19, 17, 11, 16, 20, 22, 1] and low gene dense chromosomes [13, 18, 4, 8, 21, 5, 2, X, 3, 10, 9, 15, 7, 14, 6, 12] in subfigure 3.18B. We find that the simulation data is fit far better to a combination of two lines, one for chromosomes of low gene density and the other for chromosomes of large gene density. Note that the p-values depend upon the number of degrees of freedom, as mentioned in Table 3.3.

Table 3.2: χ^2 and p-value of least-square fits for the mean centre of mass location for each chromosome against chromosome sizes within the combined model for the GM12878, IMR90, HUVEC, HMEC, and NHEK cell types against chromosome sizes. When the experimental least square fits and simulation least square fits are exact similar then χ^2 value is close to zero and p-value is absolute zero. The least square fits are done for larger chromosomes [1, 2, 3, 4, 5, 6, 7, 8, 9, X] and smaller size chromosomes [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] separately. For smaller chromosomes we tried excluding 21 and 22 chromosomes also to see how the fits and χ^2 get improve significantly.

	Size dependent							
	All chromosomes		Larger chromo		Smaller chromo		Smaller - [21,22]	
Cell type	χ^2 -value	p-value	χ^2 -value	p-value	χ^2 -value	p-value	χ^2 -value	p-value
GM12878	0.297	1.7e-17	0.041	4.8e-10	0.394	6.8e-8	0.072	4.9e-10
HMEC	0.508	5.6e-15	0.327	4.9e-6	0.380	5.5e-8	0.110	3.9e-9
HUVEC	0.403	4.6e-16	0.205	6.2e-7	0.384	5.9e-8	0.105	3.2e-9
IMR90	0.314	3.1e-17	0.191	4.5e-7	0.330	2.4e-8	0.065	3.0e-10
NHEK	0.391	3.3e-16	0.142	1.2e-7	0.427	1.1e-7	0.151	1.9e-8

Figures 3.19 shows our computation of the mean centre of mass of each chromosome

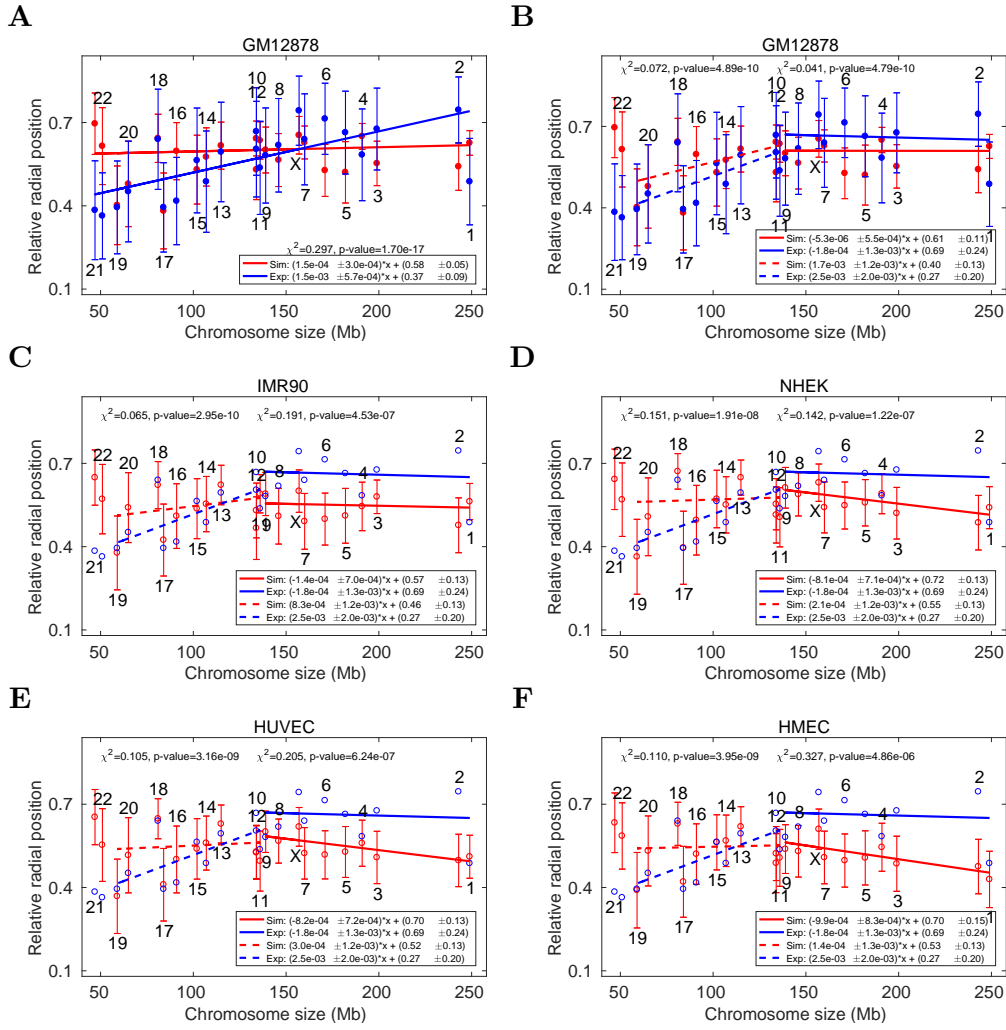


Figure 3.19: Predictions for the mean centre of mass location for each chromosome, computed for the GM12878, HMEC, IMR90, NHEK, and HUVEC cell types plotted as a function of chromosome sizes, compared to experimental data of GM12878 extracted from Ref. [Kalhor et al., 2011]. Simulation and experimental points are shown using red and blue filled circles respectively in GM12878 cell type together with errorbars while in other cell types due to unavailability of experimental data, again GM12878 experimental data shown with open blue ovals for illustrative purpose. The relative radial position 0 and 1 represent the centre and periphery of the nucleus. Chromosome numbers are indicated above or below of each errorbar. All the simulation points of 5 cell types and experimental points of GM12878 cell type are fitted with least square fit and their slope, intercept, χ^2 statistics and p-values are mentioned in the inset of each subfigure. The least square fits are shown for larger and smaller chromosome separately. And for smaller chromosomes the least square fits after dropping chromosome 21 and 22 is shown from subfigure B to F. The χ^2 error are p-value are compared between different models in Table 3.2.

within the combined model for the GM12878, IMR90, HUVEC, HMEC, and NHEK cell types against chromosome sizes. We also show the experimental data for the GM12878 cell type extracted from Ref. [Kalhor et al., 2011]. For other IMR90, HUVEC, HMEC and NHEK cell types, GM12878 experimental data is indicated for illustrative purpose only. In the initial fit to the data, all chromosomes are considered in the fits for both experiment (blue) and simulation (red) data in subfigure 3.19A. We tried a number of different fits, both excluding certain subsets of chromosomes as well as fitting two independent straight lines to different parts of the data fit to check whether the fits and χ^2 error could be improved. These investigations show that the slope of the line for chromosomes of a larger size [1, 2, 3, 4, 5, 6, 7, 8, 9, X] vs. those of a smaller size [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] can be different, cf. subfigures 3.19B to F. The χ^2 values for the fits are compared in all cases: inclusion of all chromosomes, only for larger chromosomes, only for smaller chromosomes, and only for smaller chromosomes with the exclusion of chromosomes 21 and 22 in Table 3.2. Both simulation and experiments least square fits show that the slope of the fit line for the larger chromosomes is slightly negative, while the slope for the smaller chromosomes are positive. When chromosomes 21 and 22 are dropped from the fit for smaller size chromosomes, the χ^2 value and fits improved significantly, as mentioned in last column of the Table 3.2. We do not know why chromosomes 21 and 22 do not follow the general slope trend for smaller chromosomes.

The simulations reproduce all experimental systematics, except for chromosomes 21 and 22. For the GM12878 cell type, the positions of virtually all chromosomes, with the exception of chromosome 21 and 22 lie within the error bars of the experimental data. It is important to note that the experimental and simulation data coincide more-or-less exactly for some chromosomes. The positions of chromosomes 7, 9, 13, 17, 18, 19 and 20 are very close to the experimental data, reproducing the unusual non-monotonicity in their positions.

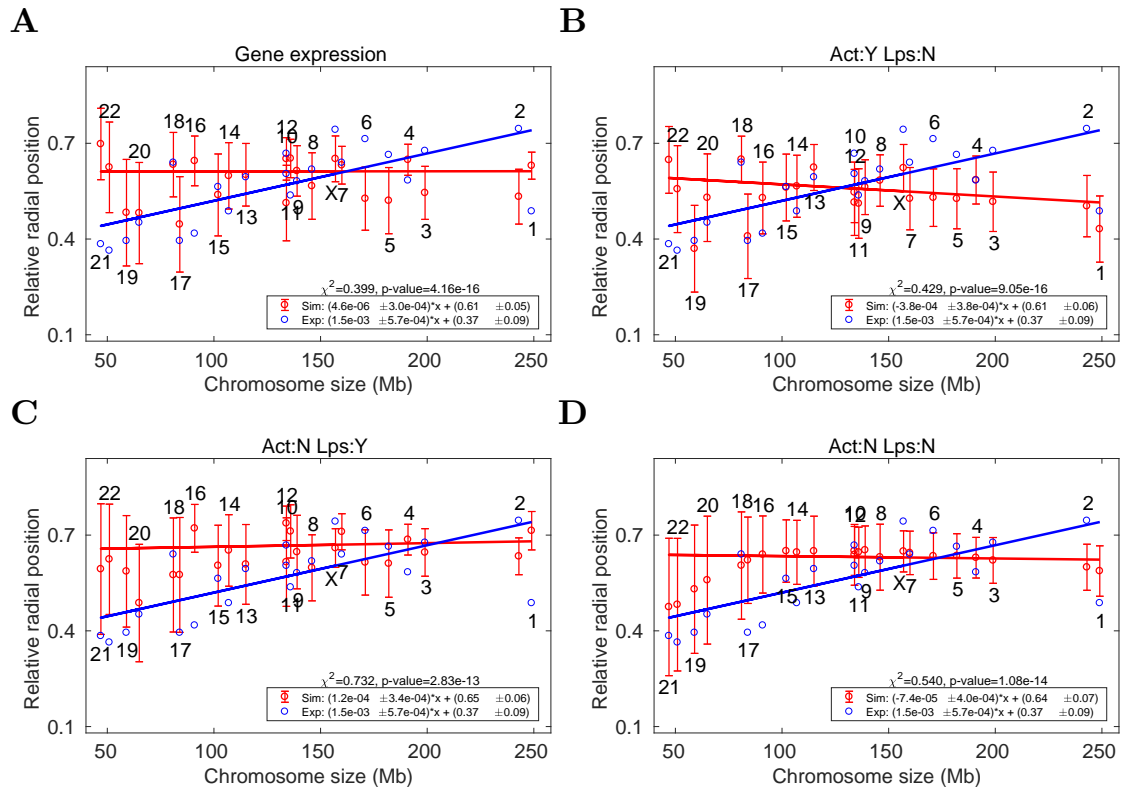


Figure 3.20: **The relative centre of mass position of each chromosome, in increasing order of their sizes, for different models.** (A) **Act:Y, Lps:Y** The effective temperature assignment of monomers is derived from the combined model of GM12878 cell type and actual loops refer from Hi-C experiments; (B) **Act:Y, Lps:N** The effective temperature assignment of monomers is taken from the combined model of GM12878 cell type but no loops are present; (C) **Act:N, Lps:Y** All monomers are at the same temperature (no activity), but loops inferred from Hi-C are present; (D) **Act:N, Lps:N** All monomers are at the same temperature (no activity) and the loops are also absent; Simulation data points (red) are shown together with the experimental data (blue) extracted from Ref. [Kalhor et al., 2011] along with respective errorbars. The chromosome number is mentioned at either the top or the bottom of the errorbars. The simulation and experimental points are fitted to a straight line for all chromosomes and slope, intercept, χ^2 statistics and p-values are mentioned in the inset of each subfigure.

If the two smallest chromosomes are excluded, an approximate size-dependence of chromosome positions relative to the nuclear centre is predicted. However, the activity associated with each individual chromosome also plays a role in determining its position. The mean centre of mass locations for chromosomes in different cell types are similar but not identical. Chromosomes 18 and 19, although similarly sized, have very different positions relative to the nuclear centre, as also seen in

the data of Figures 3.13 and 3.14. Chromosomes 21 and 22 in Figure 3.19A are positioned more towards the exterior of the nucleus in the simulations than in the experimental data. When chromosome centres of mass are plotted against gene densities the slope of the straight line is negative in all cell types (Figure 3.18). Thus, depending on the region that is fit, one can have reasonable fits to both size dependence and gene density dependence of chromosome centres of mass relative to the nuclear centre. The fact that the smallest chromosomes 21 and 22, lie outside of the fit to chromosome size may reflect aspects of their activity that our method does not resolve, as well as variations in loop assignments. On removing chromosome 21 and 22 from the fits in other cell types, Table 3.3 shows that χ^2 does not change significantly.

Figure 3.20 shows the mean centre of mass position as computed for the GM12878 cell type, across a variety of simulation conditions, including for the gene expression model as well as for the combined model with various choices for the incorporation of loops and activity. Figure 3.20A shows results for the gene expression model. In Figure 3.20B we show results for the case in which we allow differential activity but ignore looping.

In Figure 3.20C we show results for the case in which differential activity is absent but looping, as prescribed by the Hi-C data, is retained. All monomers then experience the same effective temperature, which we take to be the physiological temperature $T = 1$. Finally, in Figure 3.20D, we show results for the case where both looping and activity are absent, so this case corresponds to the case of chromosomes without loops at thermal equilibrium. It appears that in the absence of activity, relative centre of mass chromosome positioning correlates well to chromosome size, but comparisons to the $S(R)$ or $S_{CM}(R)$ distribution functions in the Figures 3.15 and 3.16 are not as good.

From these, we conclude that in the absence of both activity and looping, chro-

mosome positioning is only weakly structured. Our simulations indicate that $S(R)$ or $S_{CM}(R)$ distribution function is weakly size dependent or even independent of size in all conditions where activity is switched off. Allowing for loops induces some changes in positioning but these results do not match with experiment. Allowing for activity, but ignoring loops leads to a differential positioning for larger chromosomes.

Only models which incorporate both activity and looping are successful in reproducing, depending on the region that is fit, both fits to a size-dependence of chromosome positioning as well as a gene density dependence for all chromosomes, against relative chromosome centres of mass position. [A comparison of all the models for relative center of mass positions data are given in Table 3.3.](#) It is clear from [Table](#) that the combined model is the one which gives the best fit to relative center of mass positions data. Although, p-value in the model case with activity and no loops (Activity:Y, Loops:N) is similar to that for the combined model, this model does not represent the organization of chromatin loops so cannot be relevant biologically.

The model predicts the centre of mass positions of most chromosomes with reasonable accuracy, well within the error bars on the measurements for virtually all chromosomes. Finally, the fact that a number of broad features of the experiments are reproduced in the model suggests that the large-scale structure and positioning of individual chromosomes are principally determined by inhomogeneous activity across chromosomes, the presence of loops and confinement.

3.5.3 Distribution of Active and Inactive Monomers

It is observed from experiments that active genes are often located in the interior and inactive genes at the periphery [[Fedorova and Zink, 2009](#), [Therizols et al., 2014](#)].

Table 3.3: Different Model comparison. Chi-square goodness of statistics for different model of COM relative radial positions. If the prediction of experimental least square fits and simulation least square fit are equivalent that means p-value is 0.

	Size dependent				Density dependent	
	All except 21,22		All chromosomes		All chromosomes	
Model	χ^2 -value	p-value	χ^2 -value	p-value	χ^2 -value	p-value
Combined (GM12878)	0.098	2.1e-20	0.297	1.7e-17	0.115	5.2e-22
Gene expression			0.399	4.2e-16	0.325	4.4e-17
Activity:Y, Loops:N			0.429	9.0e-16	0.061	5.3e-25
Activity:N, Loops:Y			0.732	2.8e-13	0.984	6.6e-12
Activity:N, Loops:N			0.540	1.1e-14	0.382	2.6e-16
HMEC	0.349	6.2e-15	0.508	5.6e-15	0.127	1.6e-21
HUVEC	0.240	1.5e-16	0.403	4.6e-16	0.066	1.1e-24
IMR90	0.181	9.4e-18	0.314	3.1e-17	0.097	8.3e-23
NHEK	0.209	3.9e-17	0.391	3.3e-16	0.026	5.2e-29

To show how active monomers are distributed in our simulated nuclei, we divide the monomers into active (whose effective temperature range is between $T = 6$ to $T = 12$) and inactive (whose effective temperature is $T = 1$) for each chromosome. Figure 3.21 depicts the partial distribution functions $S(R)$ for inactive (blue) and active (red) monomers in the GM12878 cell type. The distribution for active monomers is shifted towards the nuclear centre whereas, for the inactive monomers, it is seen to be shifted towards the nuclear periphery. These results relate to the experimental observation that active alleles are positioned more towards the interior of the nucleus, an effect strong enough to be apparent in our simulations

3.5.4 Monomer Distribution across Cell types

According to experiments, active alleles tend to be positioned more towards the interior of the nucleus compared to inactive ones [Takizawa et al., 2008]. We examine how individual monomer positioning changes due to active temperature in our simulated nuclei across different cell types. We show monomer-specific distribution functions $S_M(R)$ for 6 tagged monomers across chromosomes 1, 2, 6, 7 and

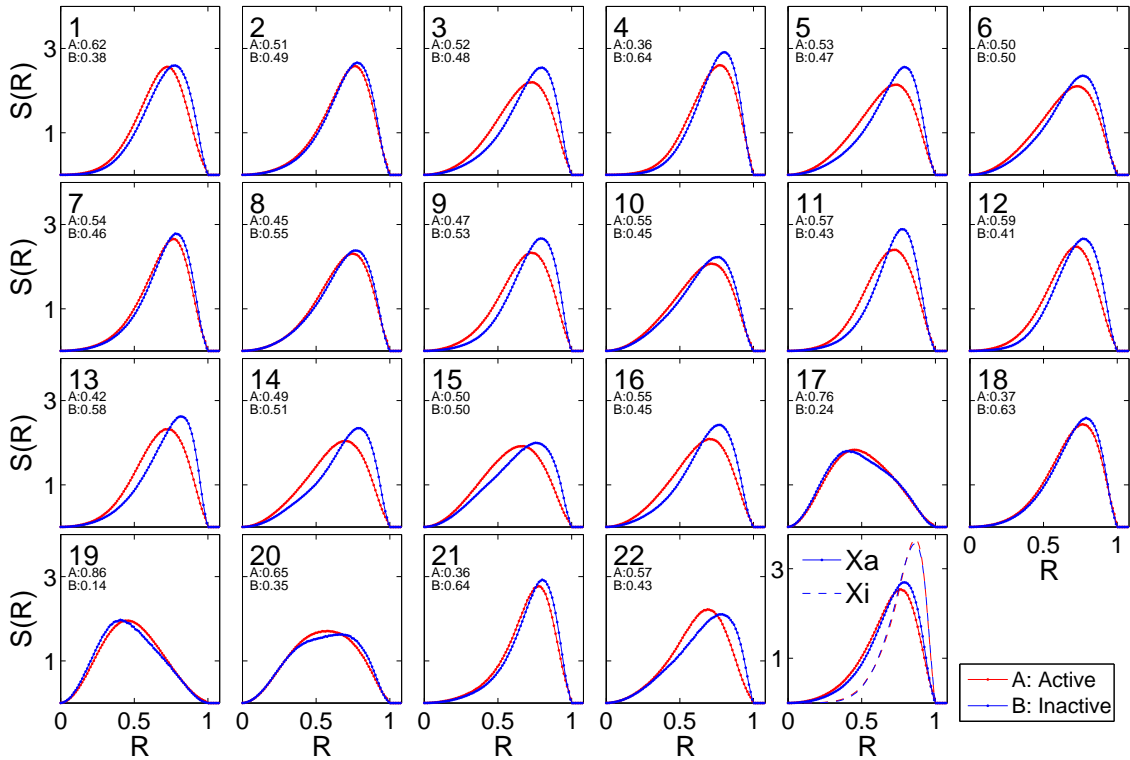


Figure 3.21: **Density distribution $S(R)$ of active (red) and inactive (blue) monomers of all chromosomes for the GM12878 cell type.** The distribution of active monomers is more interior with respect to inactive monomers. Here, inactive monomers refer to those monomers assigned a temperature of $T = 1$; remaining monomers are active. For each chromosome fraction of active (A) and inactive (B) monomers are mentioned in the upper left corner below the chromosome index in each subfigure.

15 for GM12878, HMEC, IMR90, NHEK, and HUVEC cell types in Figure 3.22. These monomers contain multiple loci and typically show differential activity across the cell types. Such monomer specific distributions are not identical but depend on both active temperatures as well as the overall activity and loop content of the chromosomes. These results suggest that signatures of activity in some cases can be prominent at the level of individual monomers or loci, but overall they are less prominent in chromosome-specific monomer density distributions or chromosome centres of mass distributions. These subtle differences originate both in differences of activity profiles as well as variations in their loop content across different cell types.

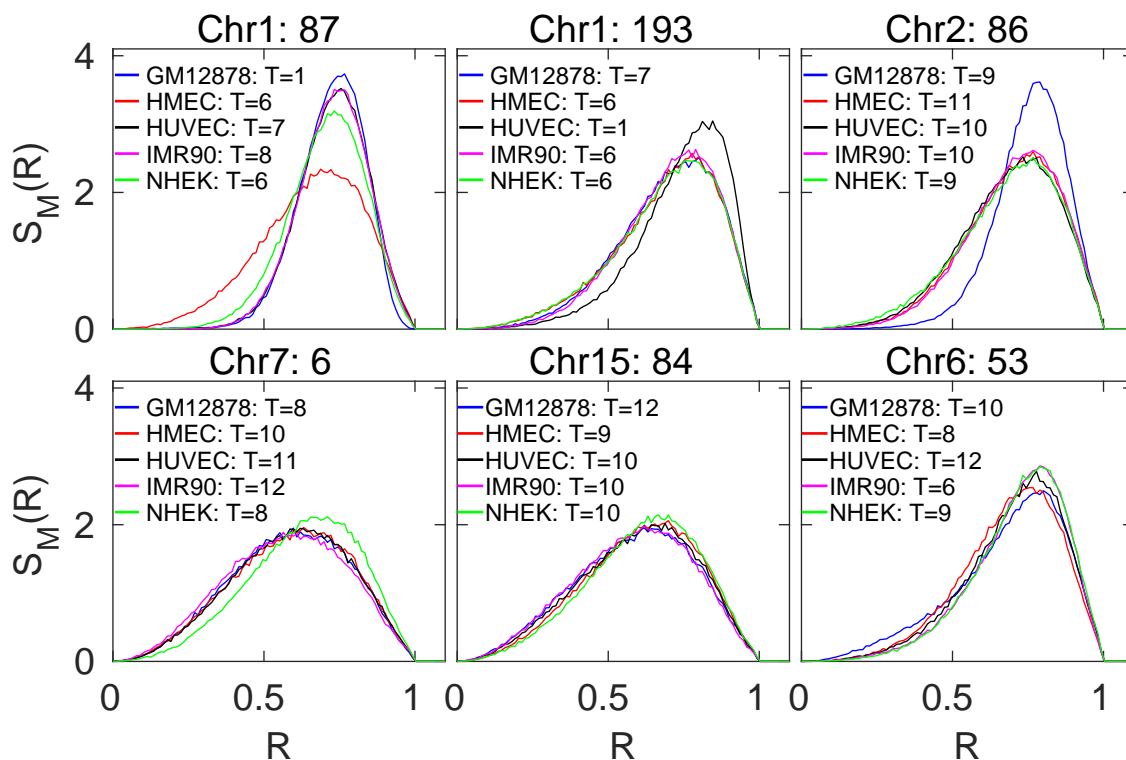


Figure 3.22: **Density distribution $S_M(\mathbf{R})$ of specific monomers as indicated, on chromosomes 1, 2, 7, 15, and 6, plotted for 5 cell types studied here.** These monomer-specific distributions can differ depending on cell type, suggesting that loci associated to these monomers can be positioned differently depending on their levels of activity, but also on the levels of inhomogeneous activity of the chromosome they belong to.

3.5.5 Ellipticity and Regularity in Two-dimensional Projection

In Figure 3.23, we show comparisons between 2d FISH data for chromosome regularity and ellipticity on WI38 cells, for which data is available [Sehgal et al., 2014], to predictions from our simulations for the GM12878 and IMR90 cell types. Both IMR90 and WI387 are lung fibroblast cell lines. Chromosomes are indexed along the X-axis, in order of their gene density and Xi chromosome in GM12878 cell type is simulated with superloops. The simulation results and experimental data appear to follow each other, with the simulations finding the same dip and subsequent rise of both ellipticity and regularity around chromosome 22. Both ellipticity and regularity

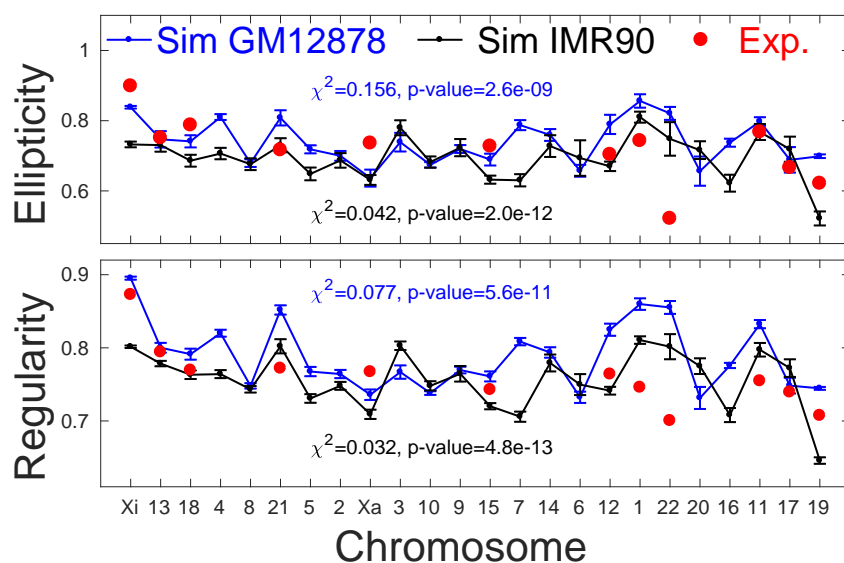


Figure 3.23: **Ellipticity and Regularity for each chromosome as predicted by the combined model** and obtained from simulations representing the GM12878 (blue) and IMR90 (black) cell types. These are compared to experimental data (red oval symbols) from 2d FISH experiments Ref. [Sehgal et al., 2014] for a cell type closely related to the IMR90 cell type. Ellipticity values of 1 represent a perfect elliptical chromosome and regularity values of 1 refer to a perfectly regular chromosome, without roughness. The X-axis is plotted in order of increasing gene density.

peak for chromosome 11, a feature both of the simulations and of the experiments. The ellipticity and regularity also appear to decrease weakly with increasing gene density, although individual chromosomes may deviate from this general trend.

3.5.6 Three-dimensional Volume and Surface Area of Chromosome

Grid method

Figure 3.24A and Figure 3.24B shows our predictions for the fractional volume and fractional surface area of each chromosome in three dimensions for the GM12878 and IMR90 cell types, as indicated in the figure with blue oval and red triangle respectively using the grid method. These quantities are plotted against chromosome

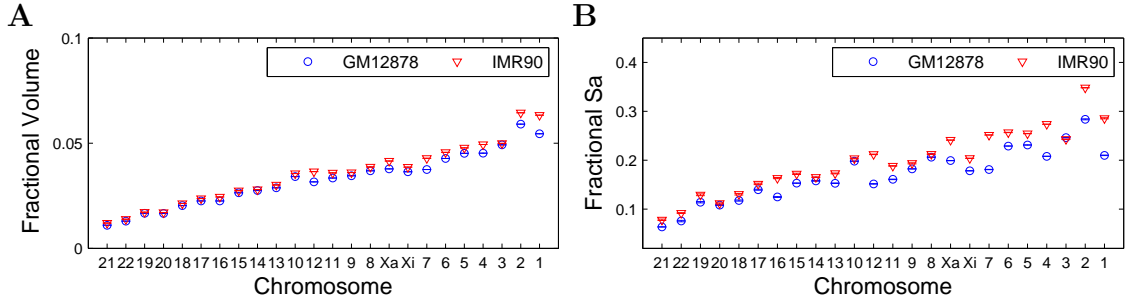


Figure 3.24: **Fractional volume and fractional surface area of each chromosome predicted by grid method.** (A) The fractional volume of each chromosome, normalised by the overall nuclear volume, in order of increasing chromosome size for GM12878 (blue oval) and IMR90 (red triangle). (B) The fractional surface area of each chromosome, normalised by the nuclear surface area, in order of increasing chromosome size for GM12878 (blue oval) and IMR90 (red triangle) cell. The trend of both cell type is more similar for volume than surface area.

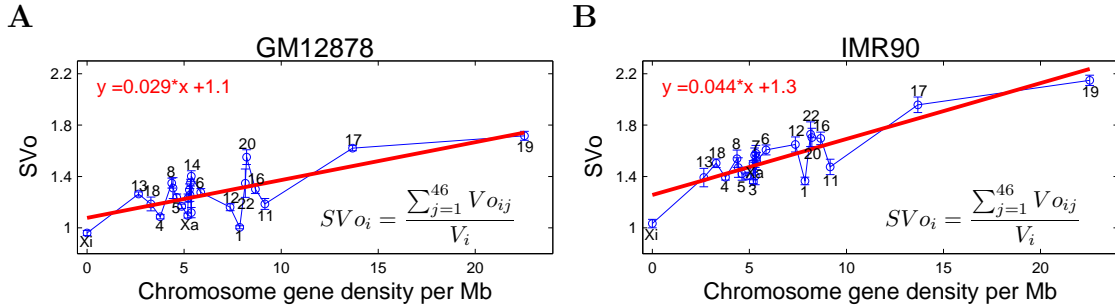


Figure 3.25: **Summed volume overlap (SVo) of chromosomes in GM12878 and IMR90 cell types as predicted by grid method,** with the X-axis plotted in order of increasing gene density per chromosome. There is a weak increase with gene density in both cell types, shown as the solid line, representing the best linear fit to the data. The IMR90 cell shares more volume overlaps with other chromosomes compared to the GM12878 cell type. The volume overlap for the self chromosome is considered to be 0.

sizes. Larger chromosomes have a larger fractional volume and smaller chromosomes have the smallest fractional volume. There is a weak cell-type dependence but data for the two cell types appear to track each other closely overall. Similar statements hold for the fractional surface area in Figure 3.24B, for these cell types.

Figures 3.25A and 3.25B show the summed volume overlap (SVo), sometimes referred to as the intermingling and used to understand chromosome-chromosome interactions in trans, of different chromosomes in our model. This is computed in the following way. The summed volume overlap of chromosome i is the sum of the overlap volume of chromosome i with all other chromosomes j (V_o) divided by the

actual volume (V) of chromosome i . Here the volume overlap of each chromosome with itself is ignored. The ordering of chromosomes according to their gene density per chromosome as shown on the X-axis. The largest overlap is for the most gene-rich chromosome. There are perceptible differences in the overlaps of chromosomes in the GM12878 and the IMR90 cell types.

In summary, the simulations reproduce broad features of individual chromosome territories. More active chromosomes appear to deviate more from a spherical shape and tend to have rougher territories [Berezney et al., 2005]. The summed volume overlap appears to increase approximately linearly with chromosome gene density, with the Xi being an exception to this trend. Activity and looping tend to have countervailing trends since activity expands chromosome territories while looping contracts them.

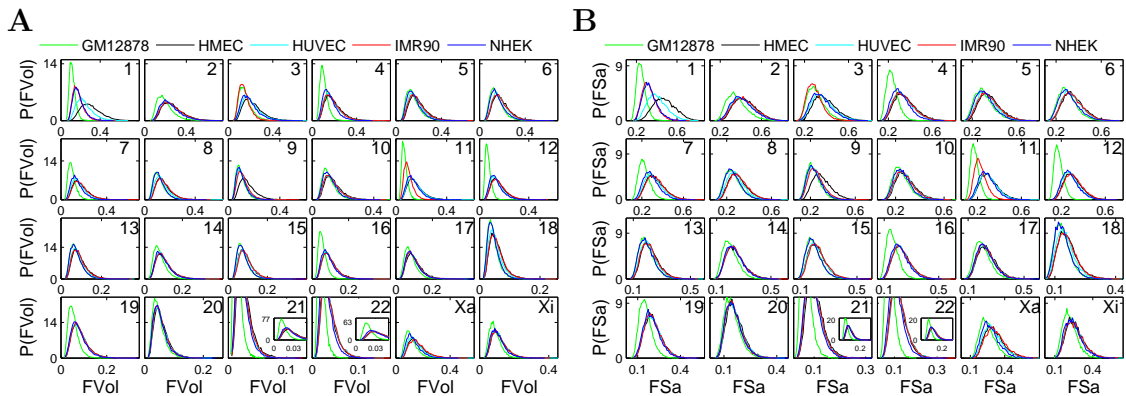


Figure 3.26: **Histogram of (A) fractional volume (FVol) and (B) fractional surface area (FSa) obtained using 3d ellipsoid fit method for GM12878 (green), IMR90 (red), NHEK (blue), HMEC (black) and HUVEC (magenta) cell types. The territory associated with each chromosome is fit to the smallest 3d ellipsoid which contains the chromosome territory. Fractional volume and fractional surface area of individual CT is the actual volume and surface area of 3d ellipsoid divided by total volume and total surface area of nuclei.**

3d ellipsoid fit method

Figure 3.26A shows the histogram of fractional volumes calculated for all chromosomes in the GM12878, IMR90, HMEC, HUVEC, and NHEK cell types, for the

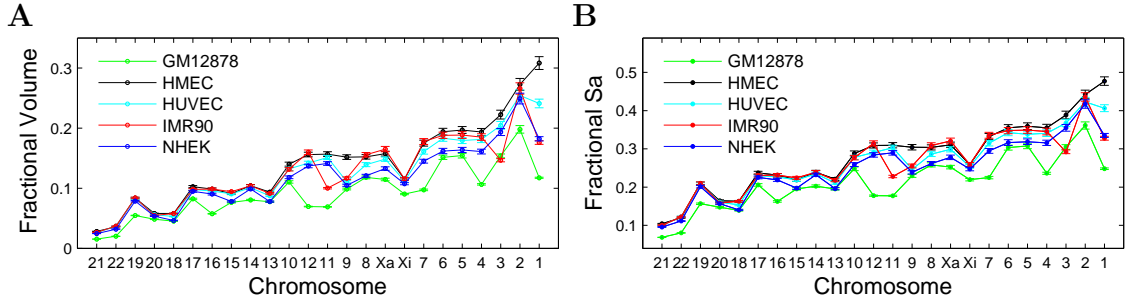


Figure 3.27: Computed (A) fractional volume and (B) fractional surface area for each chromosome, calculated for GM12878 (green), HMEC (black), HUVEC (magenta), IMR90 (red) and NHEK (blue) cell types by ellipsoid fit method. From the distributions shown in Figs. 3.26A and 3.26B, the average fractional volume and average fractional surface area are plotted in order of chromosome sizes. The total volume of 46 chromosomes including the overlap volume between chromosomes using ellipsoid fit method is 4.2 in GM12878 cell and 5.8 in IMR90 cell which is much more higher than our grid method which found only 1.5 for GM12878 cell and 1.6 for IMR90 cell with respect to nucleus volume. This shows that ellipsoid fit method has much more error than grid method. Similarly the linear volume to surface area trend is also visible for each chromosome in both ellipsoid and grid fitting methods.

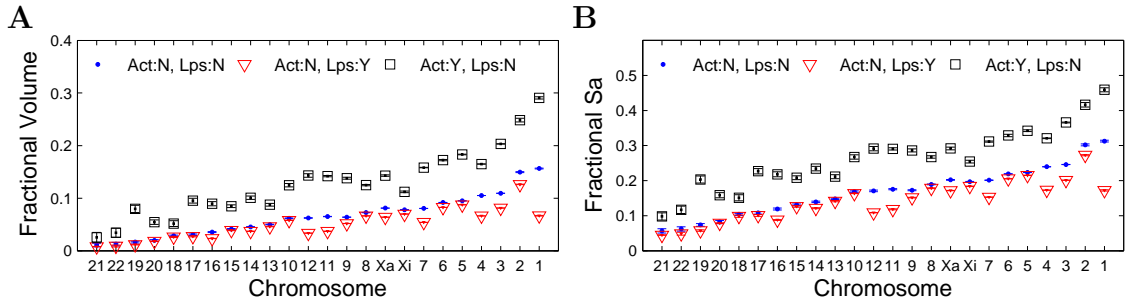


Figure 3.28: (A) Fractional volume and (B) fractional surface area shown for (Act:N, Lps:N), (Act:N, Lps:Y) and (Act:Y, Lps:N) cases with blue dot, red triangle and black square using the ellipsoid fit method. Act:N, Lps:N in which activity is absent and no permanent loops are present; Act:N, Lps:Y in which activity is absent, but loops are present; Act:Y, Lps:N in which effective activity is taken from the combined model appropriate to GM12878 cell type but loops are absent.

combined model, using the ellipsoidal fit described in the 2.5.2. Figure 3.26B shows the analogous plot for the fractional surface area. Figure 3.27A shows the fractional volume computed from these distribution functions, while Figure 3.27B shows the fractional surface area. There is a strong dependence of fractional volume and fractional surface area on chromosome size for the smallest chromosomes but this dependence is weaker for the larger chromosomes. The inactive X chromosome splits

off from this general trend. Figure 3.28A shows the fractional volume for different model of activity and looping. Activity increases the volume while loops decrease the volume. It clearly visible that compactness of chromosome increases with the decrease of activity and increase of looping. Figure 3.28B shows the fractional surface area (S_a) for different model of activity and looping. A trend similar to that for the volume is also seen here but these trends are mostly affected for larger chromosomes.

3.5.7 The Differential Positioning of the Xa and Xi Chromosome in the Presence of Superloops in Xi

Experiments investigating the positioning of X chromosomes in female mammalian cell within interphase have consistently found that their active (Xa) and inactive (Xi) homologs are differentially positioned. The Xi is silenced in such a way that the whole chromosome become transcriptionally inactive and it is located most often towards the periphery of the nucleus [Jégu et al., 2017]. This contrasts with the more central disposition of the Xa, which is larger and more extensively transcribed than more compact Xi.

Superloops are only found in the inactive X chromosome and unfortunately are only available for GM12878 cell type. This is processed using HiCCUPS method and stored in GEO (accession GSE63525) from Ref [Rao et al., 2014]. Again, we ignore superloops of smaller than 2 Mb and the remaining leftover loops are assigned in our simulation using permanent FENE bond.

Recently, Hi-C experiments have revealed that the Xi chromosome has more large-scale loops than Xa. This provides a further level of compaction in Xi. Such loops were termed as superloops [Rao et al., 2014, Darrow et al., 2016]. The experiments observed 27 large superloops each spanning between 7 and 74 Mb, present in the GM12878 cell type.

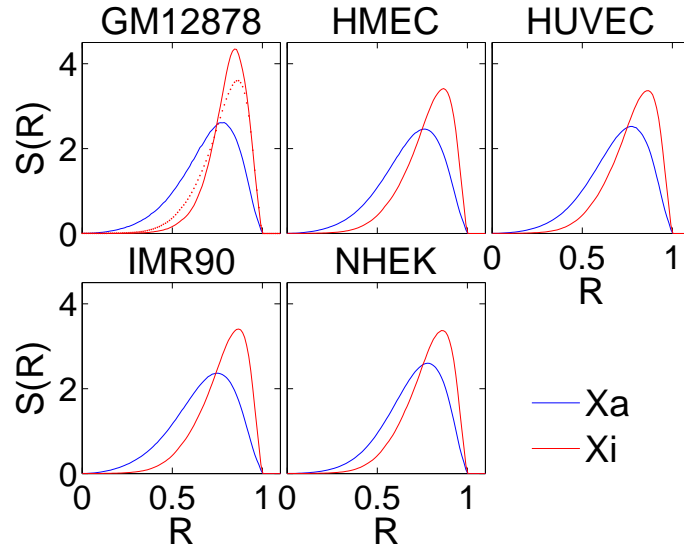


Figure 3.29: **Density distribution $S(R)$, for the Xi and Xa chromosome as obtained from simulations across 5 cell types**, named in the header to each subfigure. The inactive X chromosome, Xi, is shown in red and the active X chromosome, Xa, is shown in blue. Loops on the Xi in the GM12878 cell type can include (red solid line) or exclude (red dashed line) “superloops” as seen in recent experiments Ref. [Rao et al., 2014, Darrow et al., 2016]

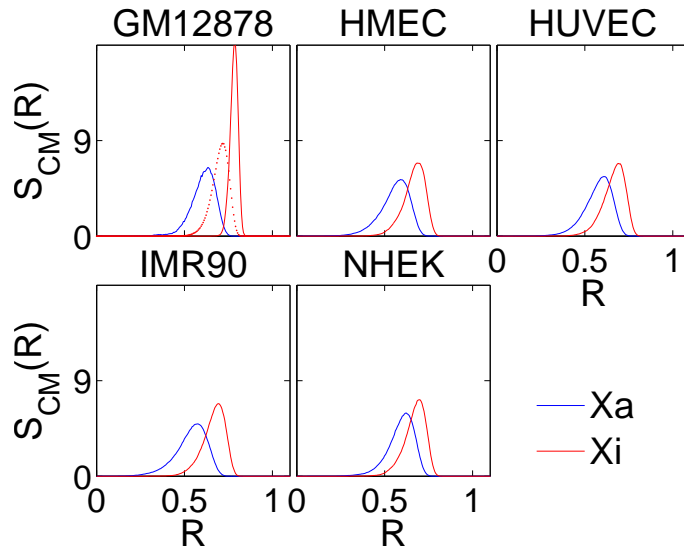


Figure 3.30: **Distribution of the location of the centre of mass $S_{CM}(R)$ of the Xi and Xa chromosome as obtained from simulations across 5 cell types**, named in the header to each subfigure. The inactive X chromosome, Xi, is shown in red and the active X chromosome, Xa, is shown in blue. Superloops on the Xi in the GM12878 cell type can be included (red solid line) or excluded (red dashed line).

Figure 3.29 shows our predictions for how Xa and Xi chromosome are differentially positioned across all the 5 cell types we study through $S(R)$. We calculate $S(R)$ for

the Xi in two ways for GM12878 cell type: First, we ignore the presence of superloops, as shown by the red dashed line. Second, we account for such superloops, shown using a red solid line. $S(R)$ of Xi is sharply peaked close to the nuclear periphery, but accounting for superloops leads to a narrower $S(R)$ distribution. Although Xa chromosome has a peak at a comparable location, its distribution has a long tail towards the nuclear centre. We do not have information about the presence of superloops in other cell types, so we ignored this feature in Xi for HMEC, HUVEC, IMR90 and NHEK cell type. Figure 3.30 shows the $S_{CM}(R)$ for 5 cell types, verifying this essential distinction. Here again, for the GM12878 cell type, the red dashed line is the case without superloops, while the red solid line is for the simulations that include them. The distinction between the distribution of Xa (blue solid line) and Xi (red solid line) suggest that different positioning is more clearly visible in $S_{CM}(R)$ than in $S(R)$. Thus our predictions for the positioning of Xa and Xi chromosomes, which emphasize the role and importance of activity, loops and superloops, yield predictions that other models cannot.

We compute the contact probabilities $P(s)$ by applying a cutoff to the monomer-monomer distance distributions obtained in our simulation, averaging across a large number of simulation configurations. Figure 3.31 shows our computation of the contact probability $P(s)$ for both Xa and Xi, across the GM12878, HMEC, HUVEC, IMR90, and NHEK cell types. The active X chromosome shows more prominent power-law scaling of the contact probability than the inactive X chromosome, where any fit to a power law can only be over a far shorter genomic scale. Exponents for the power-law scaling of $P(s)$ range from 1.11 - 1.24, with the smallest values obtained for the GM12878 cell type. For the Xi chromosome, accounting for superloops leads to a comparable scaling. However, in other cell types, such superloop information is unavailable. Accounting for loops as obtained through conventional Hi-C leads to the power-law exponent obtained over a limited range varying from 1.52 - 1.72. The variation in the scaling of $P(s)$ between Xa and Xi should be accessible experi-

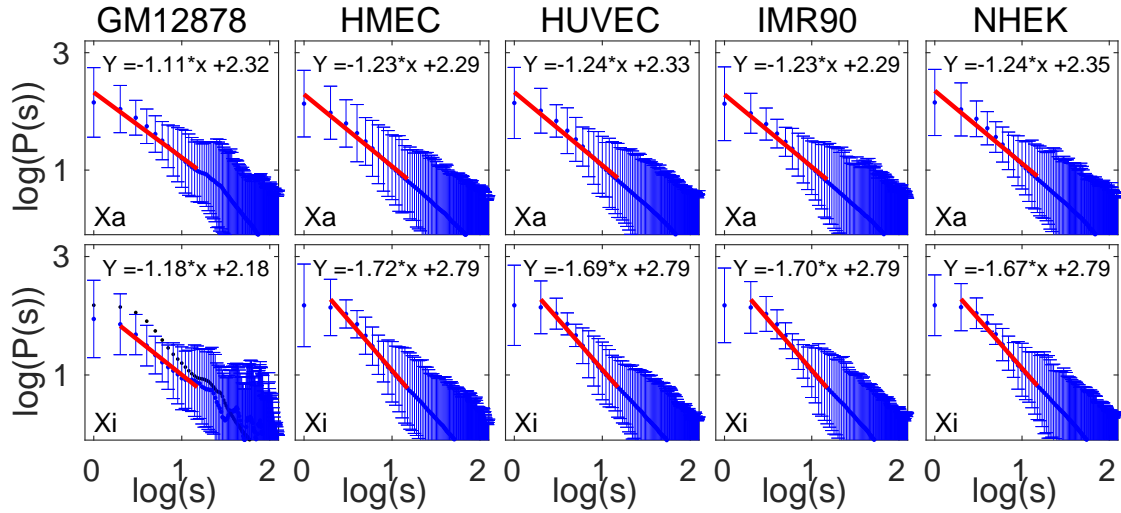


Figure 3.31: **Contact probability $P(s)$ vs s , for the active (top row) and inactive (bottom row) X chromosomes**, computed for 5 cell types within our simulations. The Xa chromosome exhibits a reasonable power-law decay of $P(s)$ with an exponent α between 1.1 and 1.25. The Xi chromosome shows a reduced region of power-law scaling, with an exponent across this reduced range which is between 1.5 and 1.7. Red lines show the power-law fit in both cases, with the fit parameters indicated within each sub-figure. In the absence of superloops on the Xi in the GM12878 cell type leads to fit for black dots $\alpha = 1.52$ (fitted line is not shown) while the fit to blue dots in the presence of superloops reduced the $\alpha = 1.18$, which bring this exponent close to Xa fitted value. For the remaining cell types, superloop information is not available for the Xi chromosome.

mentally, but the presence of superloops in other cell types as well might lead to a smaller divergence between these cases.

3.5.8 Contact Probability and Spatial Distribution

We compute the contact probability $P(s)$, for the chromosome, by applying a cutoff to distance distributions. Figure 3.32 exhibits simulation results for $P(s)$ of chromosome 1, across the five different cell types we study here, as well as the predictions of the effects of varying both activities and looping in the combined model. The data for small s show a power-law $P(s) \sim 1/s^\alpha$ behaviour over approximately a decade, as represented in the red straight line. Fitting an exponent α directly to GM12878 cell type data, in this range yields $\alpha \approx 1.06$. For larger s , $P(s)$ saturates. This value

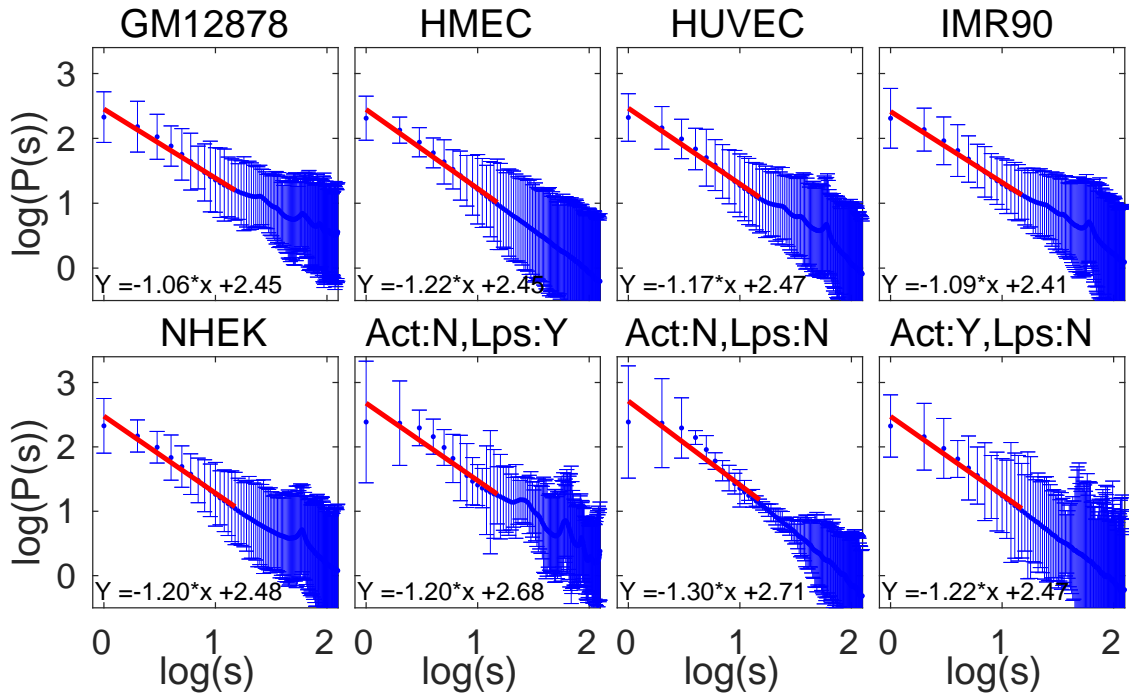


Figure 3.32: **Contact probability $P(s)$ as a function of genomic distance for chromosome 1, fit to a power law in the range of 1-15 MB** plotted for GM12878, HMEC, HUVEC, IMR90, NHEK cell types and different combinations (Act:N, Lps:Y), (Act:N, Lps:N), (Act:Y, Lps:N) of presence or absence of loops and activity. Simulation data is plotted with blue dots displayed with errorbars. Depending on the region that is fit, a power law scaling is obtained with an exponent between roughly 1.17 and 1.22. These fits are shown with red colours and the coefficient of fits are mentioned in each subfigure.

is very close to that obtained experimentally across the same region of genomic separation [Lieberman-Aiden et al., 2009, Sanborn et al., 2015]. Values of α for all other cell types are consistently larger, with the exception of the IMR90 cell type. Overall, fitting α directly to the data across cell types yields $0.97 \leq \alpha \leq 1.27$. We see $P(s) \sim 1/s^\alpha$ with $\alpha \simeq 1$ over a 1 – 10 Mb range, as predicted by the fractal globule model, even though our model lacks virtually all the requisite ingredients for this model. All we require is that activity is differentially distributed along the chromosome, that we account for looping as drawn from the Hi-C data, and that we account for crowding by other chromosomes, all features that previous work elides. Our model specification can be relaxed in several ways so that we can examine and quantify independent contributions to this behaviour. In the absence of both ac-

tivity and loops (Act:N, Lps:N), the exponent is highest. Adding loops or activity reduces this exponent. However, only the combined model, which includes both activity and looping obtains α values closest to those in experiments.

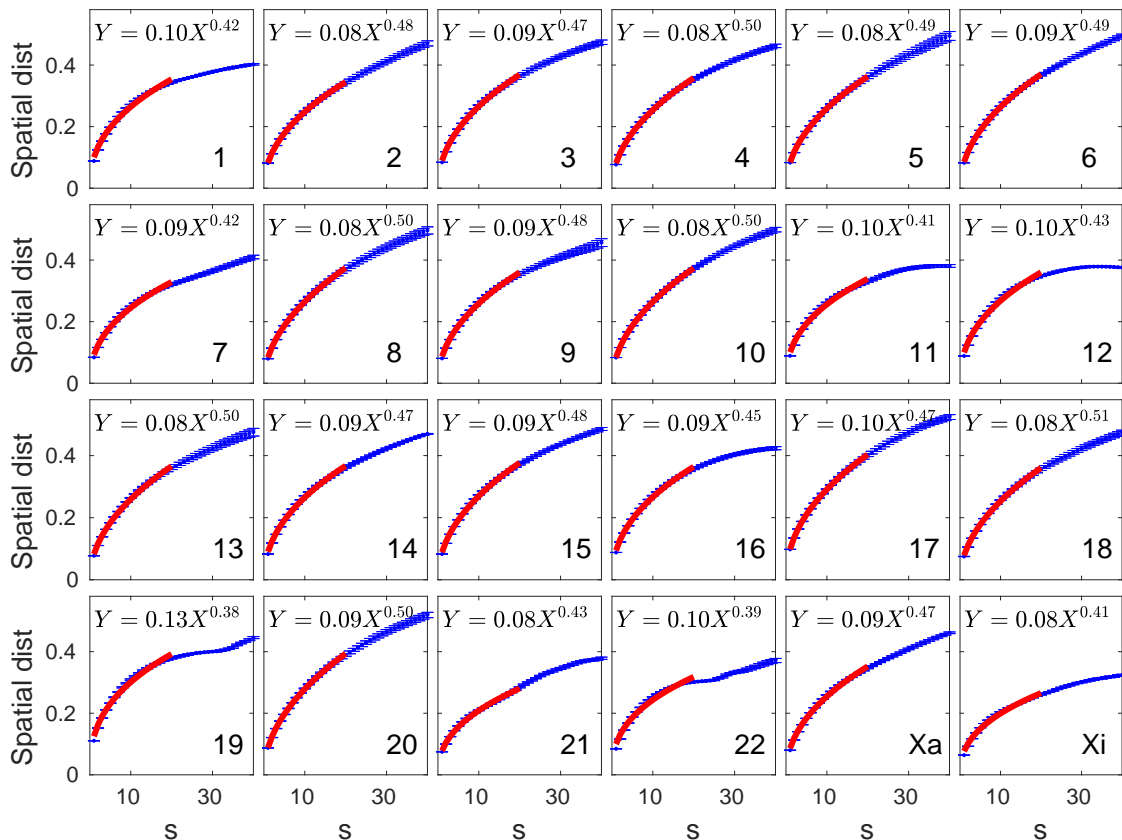


Figure 3.33: **Spatial distance between monomers of each chromosome for GM12878 cell type.** The scaling of contacts in the initial 40 Mb region for each chromosome is shown with blue dots with blue errorbars. The best power-law fit to the data in the initial 20 Mb region is shown in red colour. The chromosome number is mentioned in the bottom corner of each subfigure. The range of exponents is 0.37–0.50 in GM12878 and cell type. Here superloops are included in Xi chromosome.

We can quantify the saturation by plotting the mean three-dimensional spatial separation of monomers as a function of their internal distance in Figure 3.33. As has been noted previously, such data show an initial power-law rise accompanied by a saturation, indicative of the compactness of individual chromosome configurations at large scales. The exponent associated with this power-law varies between 0.37 and 0.50. These are consistent with earlier observations based on FISH measurements, which found values in this range as well [Wang et al., 2016].

3.5.9 Asphericity and Prolatensess

In Figure 3.34, we show the spread of the asphericity parameter Δ and the shape parameter Σ , across chromosomes in GM12878, HMEC, HUVEC, IMR90, NHEK cell types and different combinations (Act:N, Lps:Y), (Act:N, Lps:N), (Act:Y, Lps:N) of presence or absence of loops and activity. The simulations yield a linear relationship between Δ and Σ . Larger chromosomes have a smaller value of Σ and Δ while the smaller chromosomes have a larger value of Σ and Δ , implying that larger chromosomes are more close to spherical, while smaller chromosomes prefer a more prolate, rod shape nature. From these figures, we observed that all chromosomes are prolate ellipsoids in the absence of activity. Activity brings larger chromosomes towards an oblate shape. We see that larger chromosomes are more spherical and that smaller chromosomes are rougher and more rod shaped. The regularity and ellipticity indices calculated for the 2-d projections are in reasonable agreement with experimental trends Figure 3.23. However, we predict that the asphericity and prolateness of the Xi chromosome should provide an exception to the general trend for other chromosomes. We find that the data appears to fall into two classes, one a more compact set corresponding to all chromosomes with the exception of chromosomes 1, 21 and Xi, contained within an elliptical domain as shown in Figure 3.34. Values of Δ and Σ for these special three chromosomes appear to be somewhat displaced from the locations for the other chromosomes, falling approximately onto the periphery of a larger ellipse in all 5 cell types. In the absence of activity, both if loops are present or absent, the Δ and Σ values for these chromosomes falls within the inner elliptical region.

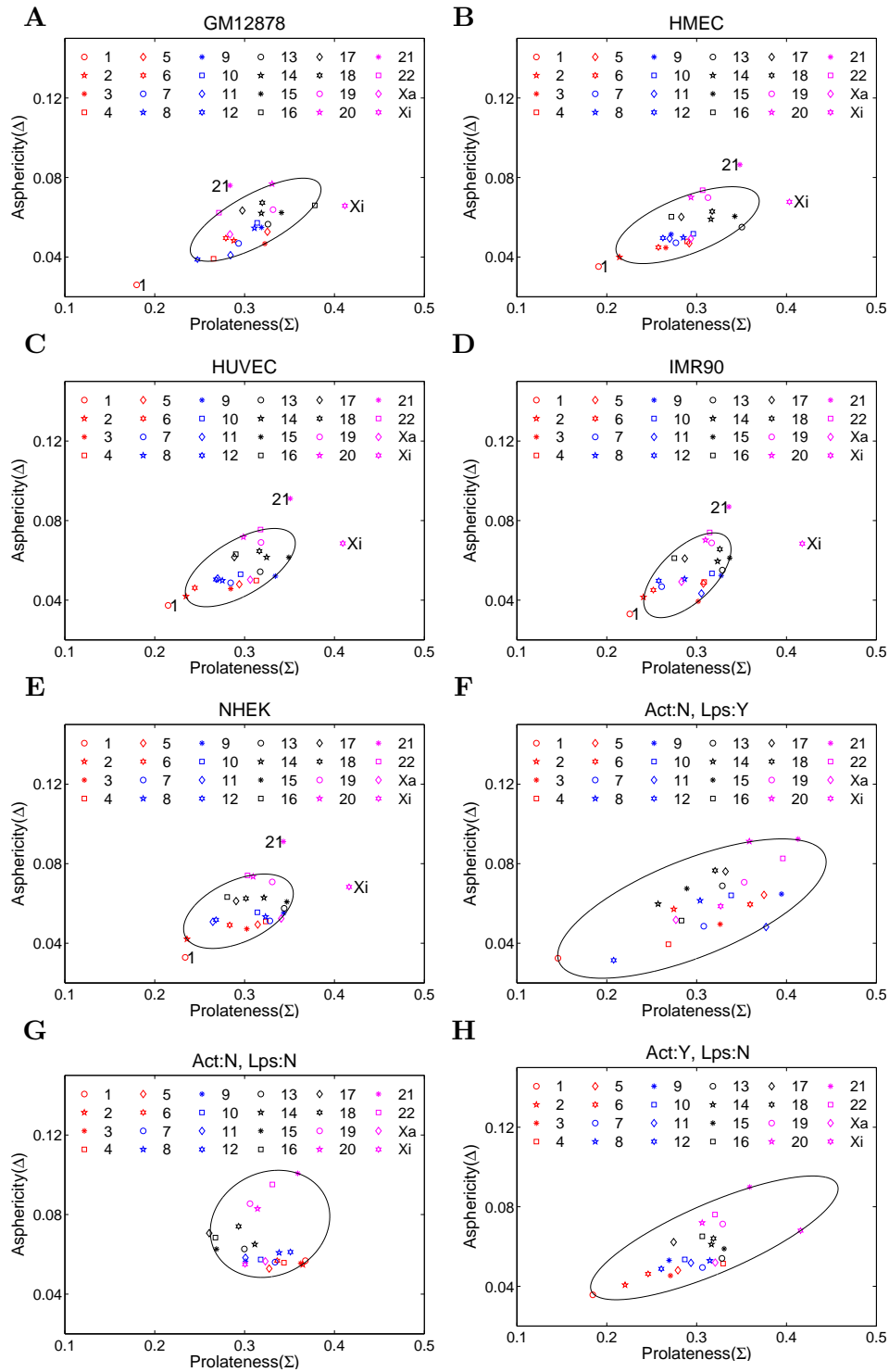


Figure 3.34: Calculated average values of the prolateness parameter Σ versus the asphericity parameter Δ for each chromosome across GM12878, HMEC, HUVEC, IMR90, NHEK cell types and different combinations (Act:N, Lps:Y), (Act:N, Lps:N), (Act:Y, Lps:N) of presence or absence of loops and activity. Each data point corresponds to chromosomes mentioned in the legend. The data suggests from figure A to E that values of Σ and Δ for chromosomes 1, 21 and Xi take more extremal values than for the other chromosomes, as shown by the ellipse plotted together with the data while in Fig F to H such extremal values are not there.

3.5.10 Distance Maps and Contact Maps

Figures 3.35A-E shows a heat map of monomer distances of chromosomes, indexed in increasing order of gene density for the GM12878, HMEC, HUVEC, IMR90, and NHEK cell types. One feature of the data is that the more active chromosomes show smaller values of inter-chromosomal distance, likely reflecting the fact that more active regions are enriched towards the nuclear centre. In Figures 3.35F-J, we show the enlarged distance maps for chromosome 1. Applying a cutoff to such data, we can derive the likelihood of contacts arising from intra-chromosomal interactions, yielding $P(s)$. Solid lines outside the figure body indicate those permanent attachments between different monomers that the Hi-C data provides. Note that regions connected by such loops exhibit a larger overlap. Figures 3.35K-O shows the contact maps inferred after applying a cutoff to the corresponding distance map. The borders of the axes show, in black and green, the active temperatures associated to specific monomers belonging to those chromosomes. The black colour refers to the most active monomers, with an effective temperature of 12 in units of the physiological temperatures whereas the green colour shows monomers with an effective temperature in the range 6 – 11. Monomers with a lower effective temperature are not shown. Regions with the same high effective temperature appear to contact each other more, but these are further modulated by the presence of internal loops. Note the presence of a dark banded region towards the centre of chromosome 1, associated to a large inactive region on this chromosome. This is a prominent feature of the experimental data, also seen in other cell types [Rao et al., 2014].

To summarise, our model yields structural information for chromosome structures and shapes that are broadly in agreement with available data. Our simulated distance maps lack the fine detail of distance maps computed in Hi-C experiments, which provide data for contacts at the smaller scales of 10 - 100 kB, but nevertheless are relevant to experiments that probe large-scale structuring. Our computed

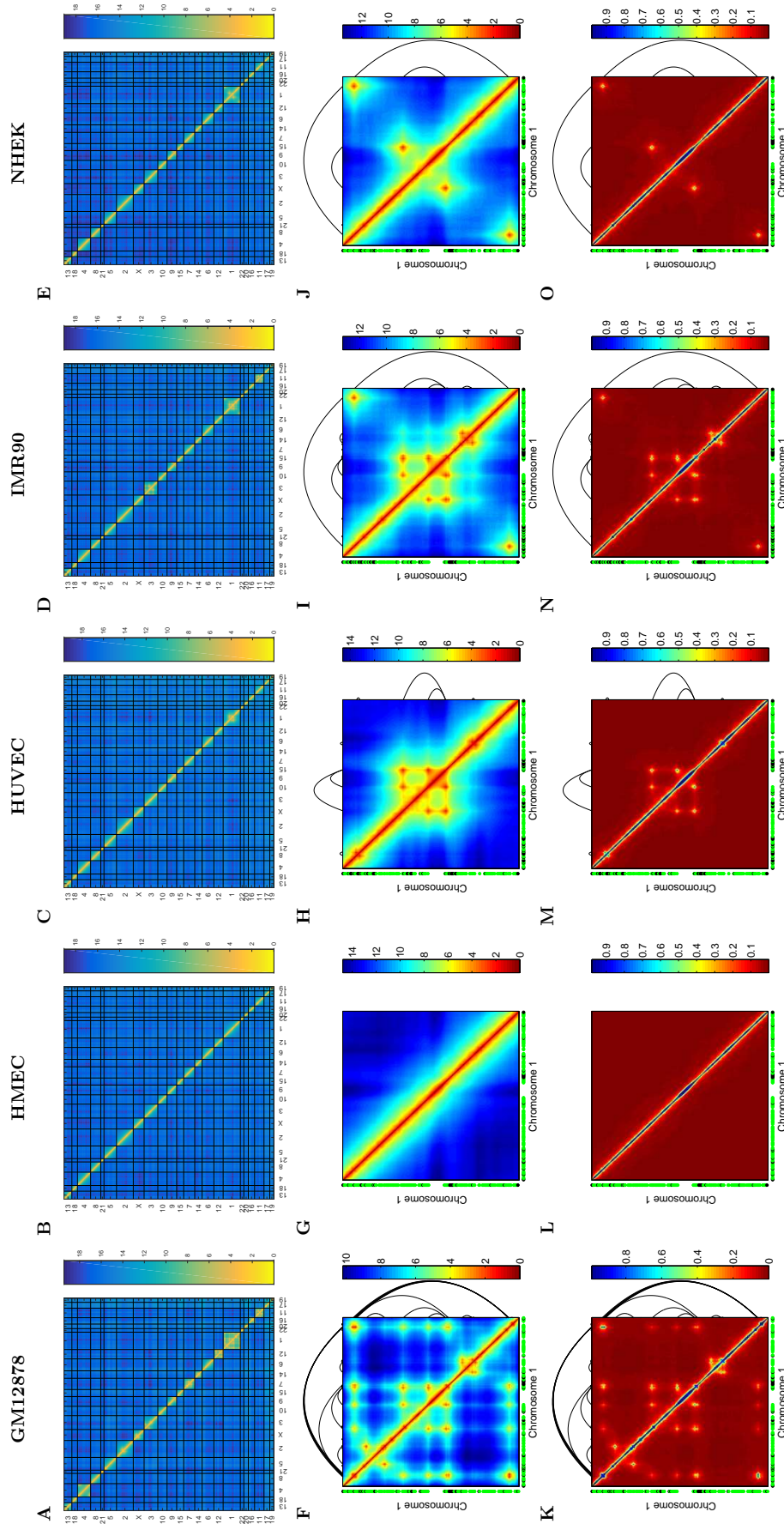


Figure 3.35: Heatmap of mean distances between monomers, the distance map, and contact maps for GM12878, HMEC, HUVEC, IMR90, and NHEK cell types. A-E. Heatmap of mean distances between monomers, the distance map for all monomers in which chromosomes are ordered by their gene density. F-J. Heatmap of the mean distance between monomers for chromosome 1, expanded out from Figure A to E. The locations of the permanent loops inferred from the Hi-C data are plotted with the black curve. Individual monomers at $T = 6$ and $7 \leq T \leq 12$ are shown in green and black, adjacent to the X and Y axis, respectively. K-O. Contact map inferred from the distance matrix for chromosome 1, Figure F to J. The locations of the permanent loops inferred from the Hi-C data are plotted with the black curve. Individual monomers at $T = 6$ and $7 \leq T \leq 12$ are shown in green and black, adjacent to the X and Y axis, respectively. Colorbars of distance and contact maps are shown in respective sub-figures.

$P(s)$ contains about a decade or so of power-law decays, with exponents that are comparable to those seen in experiments. Our model-based predictions for trends in the asphericity and prolateness of chromosomes with chromosome size and gene density are testable.

3.6 Conclusions

Model descriptions of chromosomes must bridge multiple scales, ranging from microscopic length-scales of a few angstroms to scales of microns, of order the nuclear size. For now, brute-force atomistic simulations of the 23 pairs of chromosomes in human nuclei contained within the densely crowded, fluid and confined environment of the nucleoplasm are impossible. They are likely to remain so at least for the foreseeable future. Understanding how microscopic descriptions connect to macroscopic ones thus requires intuition for the processes that act to couple these scales, so that model building, which is as much about what to leave out as it is about what to leave in, can proceed.

The principal results of this chapter are the following:

1. Reproducing trends from experiments on large-scale nuclear architecture across a number of human cell types (and presumably for all higher eukaryotes) requires that we include both inhomogeneous activity and the looping of chromosomes in a computational model. Models that lack these two ingredients simply cannot hope to provide an explanation for the full variety of experimentally available data, although they may provide reasonable fits to one or two specific properties, provided they are tuned to do so. This is the central message of our work, together with a numerical implementation of this idea that yields results that can be compared to experiments.

2. The results in this chapter are for a set of models whose parameter values were finally converged on through a combination of biophysical reasoning and extensive simulations. (Indeed, the work that went into the formulation of the three different version of the models we study here - gene density, gene expression and the combined model - required extensive preparatory analysis and simulations, but these results are left out here for compactness, since they were only a means to an end.) The central question is how to associate measures of energy-consuming non-equilibrium processes acting on chromatin with a local active temperature. We explored a number of different ways of doing so, but a combination of gene-expression-based and gene-density-based methodologies seems to work best.
3. There are subtle differences in large-scale nuclear architecture across cell types. Our model suggests where these might originate, as well as provides a reasonable explanation for why a number of general trends in such architecture tend to be similar across different cell types.
4. The model contains, within itself, the seeds of further generalization e.g. to include the effects of lamin-associated domains, incorporating the nucleolus as an additional nuclear landmark, as well as the computation of dynamic properties.

The model described in this chapter stresses a specific biophysical effect, ignored in previous work, of relevance to the modelling of chromosomes in living cells. We began by emphasising the relevance of non-equilibrium effects arising from local transcriptional activity for descriptions of nuclear architecture [Chu et al., 2017, Almassalha et al., 2017]. We proposed that the intensity of active processes should increase with increased transcription levels. We mapped a reasonable measure of local transcriptional activity, inferred from combining population-level measures of local RNA-output with estimates of the local gene density, into an effective temper-

ature seen by each monomeric unit in our polymer model of chromosomes. We then performed simulations of these confined polymers, with properties chosen to reflect generic biophysical aspects of chromosomes. The monomers in our simulation represented 1Mb sections of chromosomes, although we could have defined our model at the smaller scales of 0.1 or even 0.01 Mb. However, the averaging inherent in summing transcriptional output over a 1Mb scale renders the model relatively less sensitive to errors and noise in this input. Further, the 1Mb scale is believed to be an appropriate building block for chromosome territories.

A more detailed and explicit model for non-equilibrium activity and its consequences for an active temperature description would be useful, but the form such a model ought to take is presently unclear and best left to more extensive investigations. Irrespective of potential quantitative improvements on the model front, the broad trends we describe here should be largely robust.

Chapter 4

Motif Identification Through Clustering of ChIP-Seq Data

Three-dimensional interactions of chromatin and transcription factors (TFs) constitute a primary mechanism for regulating transcription in mammalian genomes [Li et al., 2010]. TFs bind to regulatory sequences known as transcription factor binding sites (TFBSs) in order to up or down-regulate gene expression. Mutations in TFBS positions often lead to genetic diseases [Lee and Young, 2013]. Many TFs show a highly cell type-specific binding pattern, which is due to the combinatorial action of TFs with cofactors, and chromatin accessibility of DNA-binding sites [Spitz and Furlong, 2012]. High throughput experimental methods determine TF binding regions of 100-1000 bp while the functional TFBS is very short, typically 6-25 bp within that region. It is difficult to identify which of these TFBS are real and functional in gene regulation, and which are non-functional [Li et al., 2010]. Some of the experimental *in vitro* methods include Electro-Mobility Shift Assay (EMSA), DNase I footprinting/protection assay, Systematic Evolution of Ligands by EXponential enrichment (SELEX) and *in vivo* methods include ChIP-chip (chromatin immunoprecipitation with DNA microarray), ChIP-Seq (chromatin immunoprecipitation with

high throughput sequencing) and ChIP-exo (chromatin immunoprecipitation with exonuclease digestion) [Jayaram et al., 2016]. Most of the computational approaches to TFBS recognition are based on input experimental data and each one has its own pros and cons. Key experimental and computational methods are described in the next section.

4.1 Identification of Transcription Factor Binding Sites (TFBS)

In this section, first, we describe the ChIP-Seq and ChIP-exo methods for genome-wide identification of TF binding regions, then computational predictions of TFBS within those regions.

4.1.1 Experimental Approaches

Chromatin immunoprecipitation (ChIP) related techniques revolutionized the study of *in vivo* TF-DNA binding interaction by enabling the genome-wide identification of regions occupied by TFs of interest. ChIP-Seq is a method widely used for *in vivo* genome-wide identification of TFBS [Johnson et al., 2007]. In this method, DNA-protein complexes are crosslinked using formaldehyde, sonicated to break the DNA, and treated with a TF-specific antibody to precipitate the protein of interest. By then reversing the crosslinks, sequencing the DNA fragments and mapping them to a reference genome, a genome-wide map of TFBS with a resolution of 100-200 bp can be obtained. A typical ChIP-Seq workflow method is shown in Figure 4.1. ChIP-Seq has higher resolution, fewer artefacts, greater coverage, and a larger dynamic range than ChIP-chip and older methods. It provides a more precise mapping of protein-binding sites that allows for a more accurate list of targets for TFs and enhancers

[Mchaourab et al., 2018, Park, 2009].

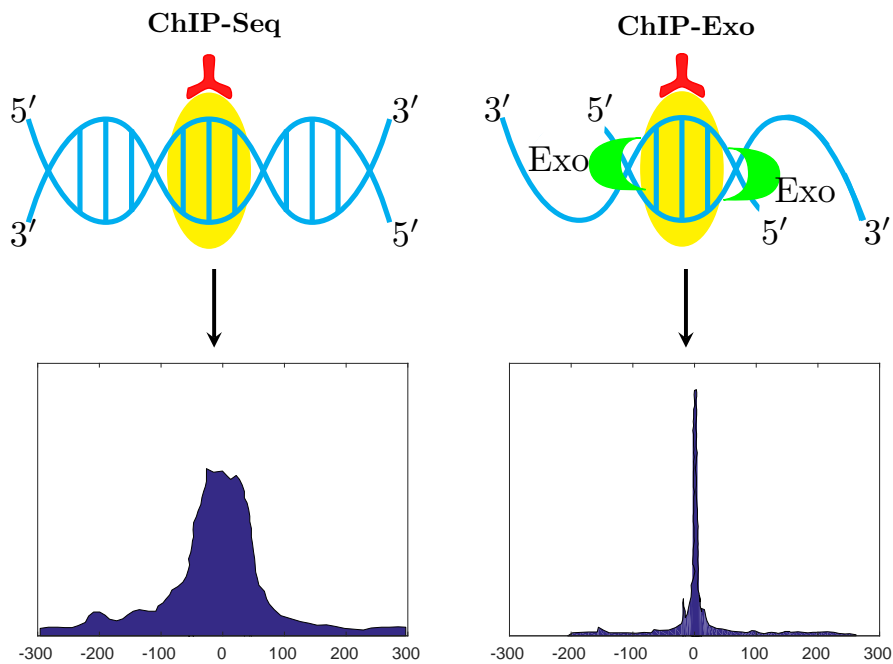


Figure 4.1: **Difference between ChIP-seq and ChIP-exo workflow.** Cells are cultured using standard conditions and harvested for ChIP-Seq and ChIP-exo. ChIP-Seq reports on the sonication borders of ChIP-enriched DNA fragments, wherein the location of the protein-DNA crosslink is deduced. In contrast, ChIP-exo, 5'-3' exonuclease is employed to trim the DNA sequences on one strand to within a few bp of the crosslinking point.

A recently developed method, ChIP-exo (ChIP with exonuclease digestion) improves upon ChIP-Seq by providing near base pair mapping resolution for protein-DNA interactions [Rhee and Pugh, 2011]. ChIP-Seq has a limitation that some DNA not bound by the protein of interest contaminates the sequencing library, resulting in high false positives rate [Stower, 2011]. In ChIP-exo, an exonuclease step is introduced after proteins are crosslinked to DNA. This removes DNA flanking the crosslinked site and DNA contaminants [Stower, 2011]. It can identify low-occupancy binding sites at a higher resolution than ChIP-Seq. ChIP-exo methodology incorporates lambda exonuclease digestion in the library preparation workflow to effectively footprint the left and right 5' DNA borders of the protein-DNA crosslink site. Thus, rather than sequencing from the distal sonication borders as in ChIP-seq, ChIP-exo enriched DNA fragments are sequenced from the left and right 5' DNA

borders of the protein-DNA crosslink site, shown in Figure 4.1.

It is very common for TFs to interact with DNA via co-factors and indirectly, which means a mixture of different motifs might be found in the ChIP-seq data. Detection of such mixtures of motifs corresponding to known TFBS presents a challenge to traditional computational motif-finders tools.

4.1.2 Computational Approaches

TFBS are generally characterized by short conserved patterns or motifs. These are short, usually 6-20 bp, and somewhat variable. At a basic level, they can be represented by strings, with variable nucleotides represented by IUPAC symbols: for example, R (puRine) for A or G; S (Strong) for C or G; and so on (figure 4.2). Such a representation turns out to be rather restrictive in describing the complexity of actual TFBS. Instead, binding motifs are commonly represented by position-weight-matrices (PWMs), a probabilistic representation where each position within a binding site is described by an independent categorical distribution over the four nucleotides (A, C, G, T). At each base position of a TFBS, for each nucleotide, the PWM provides a score that is proportional (when normalized, equal) to the probability that it occurs at that position. Multiplying these probabilities for each base of sequence yields a likelihood for observing that sequence under a given PWM model. PWMs are conveniently visualised using sequence logos. A PWM of a given TF is often used to scan regulatory sequences to identify potential TF binding sites. Over the last decade, an unprecedented wealth of data on TF-DNA interactions has been catalogued and used as motif collections in databases such as TRANSFAC [Matys et al., 2006], JASPAR [Sandelin et al., 2004, Mathelier et al., 2016, Khan et al., 2017], Factorbook [Wang et al., 2012], HT-SELEX [Jolma et al., 2013], UniPROBE [Hume et al., 2014], and CisBP [Weirauch et al., 2014].

In many cases, sequence logos reflect strong preference to one or a small number of related sequences, or weak base preference in spite of contributing to binding. Sometimes PWMs fail to detect true binding sites due to dependencies among base positions, for example in case of multiple binding modes, DNA shape or deformability, cooperative interactions, DNA methylation which can impact binding. To incorporate such complexities, more complicated models like dinucleotides and higher order k-mers have been proposed. However, the improvement is minor or even undetectable, especially when comparing across different datasets, and the PWM remains the most commonly used model for analysis of TF binding [Lambert et al., 2018].

Finding statistically enriched motifs in a set of regulatory sequences is commonly known as the motif-discovery problem. In the last 2 decades, numerous tools both simple and sophisticated have become available for motif discovery task. More complex models are better in describing the data they were derived from, but on the other hand simpler models are easier to evaluate and to scrutinize for artifactual features such as noise or experimental bias.

PWMs can be visualised using sequence logos [Schneider and Stephens, 1990, Crooks et al., 2004]. In each position of a sequence logo, nucleotides are stacked on top of each other, sorted according to their frequencies, and the height of each letter is proportional to its frequency. The information content (IC) at position i of the motif is given by

$$IC(i) = \log_2(N_\alpha) + \sum_{\alpha} p_{i\alpha} \log_2(p_{i\alpha}) = 2 - \text{entropy}(i) \quad (4.1)$$

The IC is measured in bits and for DNA, N_α is 4. At a given position in the motif, if all nucleotides occur with equal probability, the IC is 0 bits, while if only a single nucleotide occurs, then IC is 2 bits. An example of a PWM and its sequence logo is shown in Figure 4.3.

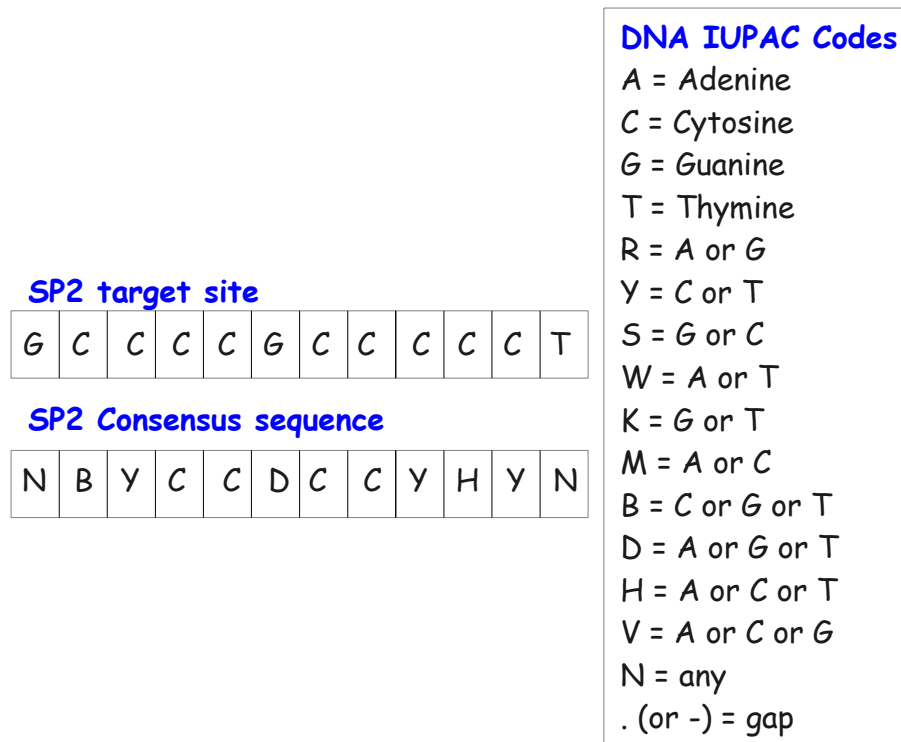


Figure 4.2: **Consensus model of TF-DNA binding.** A single SP2 target site or IUPAC degenerate consensus sequence. The box inset displays all possible degenerate IUPAC bases for the different DNA bases [Cornish-Bowden, 1985].

A TF binds its specific binding motifs with a higher affinity than other genomic sequences of the same length. PWM-based models assume each nucleotide at each position contributes independently from other positions. They are easy to implement, easy to visualize using sequence logos, have small number of parameters and provide useful approximation of binding sites for the majority of studied TFs. However, PWM-based model fail to capture if there is an interactions between nucleotides, which can lead to inaccurate predictions, particularly for low-affinity sites. This approximation can be refined by including contributions of higher order sequence features, such as dinucleotides or longer k -mers.

The algorithms for the motif discovery are categorized based upon (i) probabilistic methods where the model parameters are estimated using maximum likelihood principle or Bayesian inference, (ii) regular-expression or string based methods which mostly rely on exhaustive enumeration, ie, counting and comparing nucleotide k -mer

Frequency Matrix of SP2

A	[160	39	52	0	0	174	0	0	0	206	115	187	226	228	203]
C	[253	811	1386	1679	1643	0	1686	1656	1108	1196	971	554	717	794	745]
G	[1028	257	1	7	0	1280	0	0	0	0	99	325	401	408	422]
T	[245	579	247	0	43	232	0	30	578	284	501	620	342	256	316]

Sequence logo of SP2

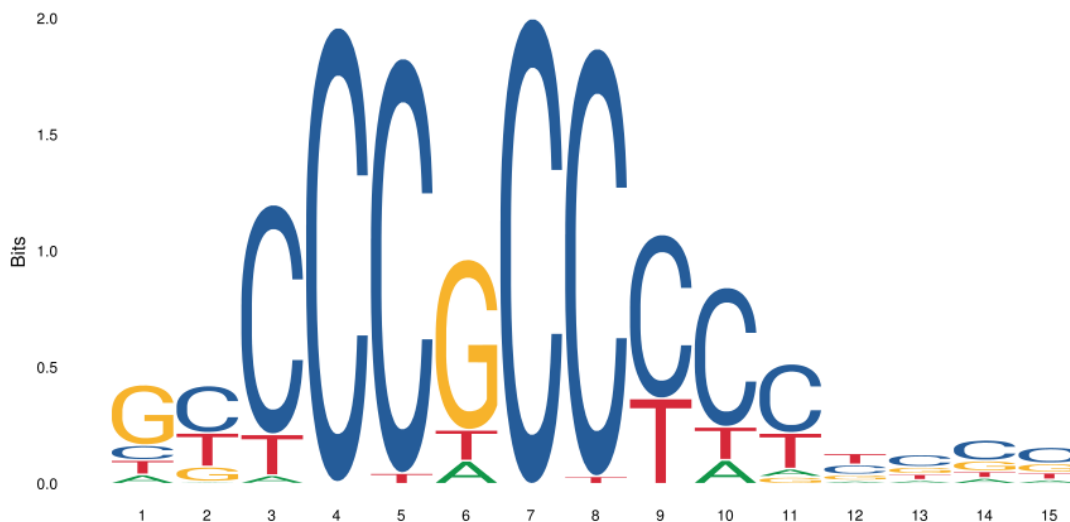


Figure 4.3: **Frequency matrix of SP2 TF and its sequence logo representation is shown.** It is downloaded from JASPAR database (MA0516.1). In the frequency matrix at position 7, frequency of nucleotides other than C are 0. So, at 7th position in sequence logo diagram, information content of nucleotide C is 2 bits.

frequencies and (iii) based upon other methods such as machine learning etc.

Regular-expression methods are based on counting matching patterns with a certain maximum number of mismatches. They learn motifs from the sequences using overrepresent k-mers. A typical tool for this purpose is Weeder [Zambelli et al., 2014] which uses suffix-trees to hold data and enforces some constraints on locations where mismatches are allowed.

Probabilistic-based algorithms perform heuristic searches by iteratively optimizing an initial PWM. These methods select positions from the input data, align their associated sequences, build a PWM and score the obtained model. For example ‘Multiple Expectation Maximization for Motif Elicitation’ (MEME) [Bailey et al., 1994], begins with separate profiles for for each input k-mer, then selects the current best profile to optimized deterministically in further ‘expectation maximization’

(EM) steps. MEME does not allow gaps so it cannot discover the motifs in the sequences which exhibit insertions and deletions. The Gibbs sampler [Lawrence et al., 1993] is another method that can be seen as MEME's stochastic counterpart. Unlike MEME, it overcomes the generation of too many initial profiles by building only one random initial profile that is subsequently improved. Both algorithms have drawbacks such as: they assume the presence of a motif in each input sequence; they may prematurely end in local optima; they are not suitable for analysis of large input data such as genome-wide ChIP-seq peaks.

Another algorithm proposed by Siddharthan et al [Siddharthan et al., 2005] PhyloGibbs, that explicitly accounts for the phylogenetic relationship between the species in the alignment and then uses Gibbs sampling to rigorously assign posterior probabilities to all binding sites that it reports. This algorithm performed significantly better than MEME and Gibbs sampler. A nice review of *de novo* motif discovery tools before ChIP-Seq and after ChIP-Seq era can be viewed in following references [Zambelli et al., 2012, Lihu and Holban, 2015].

The goal of existing *de novo* motif discovery program is to find motifs that are statistically over-represented in the entire dataset, and more suited for finding common patterns in data. Most existing *ab initio* motif finders do not scale to large datasets, or fail to report motifs associated with cofactors which may be present only in a small fraction of sequences.

4.2 THiCweed: Introduction

We present THiCweed (**T**op-down **H**ierarchical **C**lustering to **w**eed out the signals in ChIP-Seq peaks), a new approach to analyzing TF binding data from high-throughput ChIP-Seq experiments. THiCweed clusters bound regions based on sequence similarity using a divisive hierarchical clustering approach based on sequence

similarity within sliding windows, while exploring both strands. ThiCweed is specially geared towards data containing mixtures of motifs, which present a challenge to traditional motif-finders. Our implementation is significantly faster than standard motif-finding programs. On synthetic data containing mixtures of motifs it is as accurate or more accurate than other tested programs.

THiCweed, offers both speed and accuracy in finding multiple motifs in large datasets. It does not require prior information on the number of motifs or the lengths of the motif, since its approach is based on clustering rather than traditional motif-finding, and the clustering is based on stringent statistical criteria. On synthetic data, it outperforms all current alternatives greatly on speed and is close to the best current alternative in terms of accuracy. On real genomic data, it reveals an unusual complexity in the structure of sequence motifs, in particular in internal dependencies and in flanking sequence extending far beyond the core motif.

4.3 THiCweed: Methods

There are two components to our approach:

- First is an efficient method of divisive hierarchical clustering. Starting with one large cluster, we split it in two clusters (or three, the third consisting of poor matches to either cluster). The scoring is described below, and is based on the likelihood ratio of a sequence belonging to one or the other cluster, done iteratively starting from an initial heuristic split. We then split each new cluster into two (or three) further clusters; and proceed until no further splits are possible. For each split we apply stringent statistical criteria to accept or reject the split. Further optimizations are described in subsection Algorithm.
- During this clustering process, we include shifts and reverse complements of

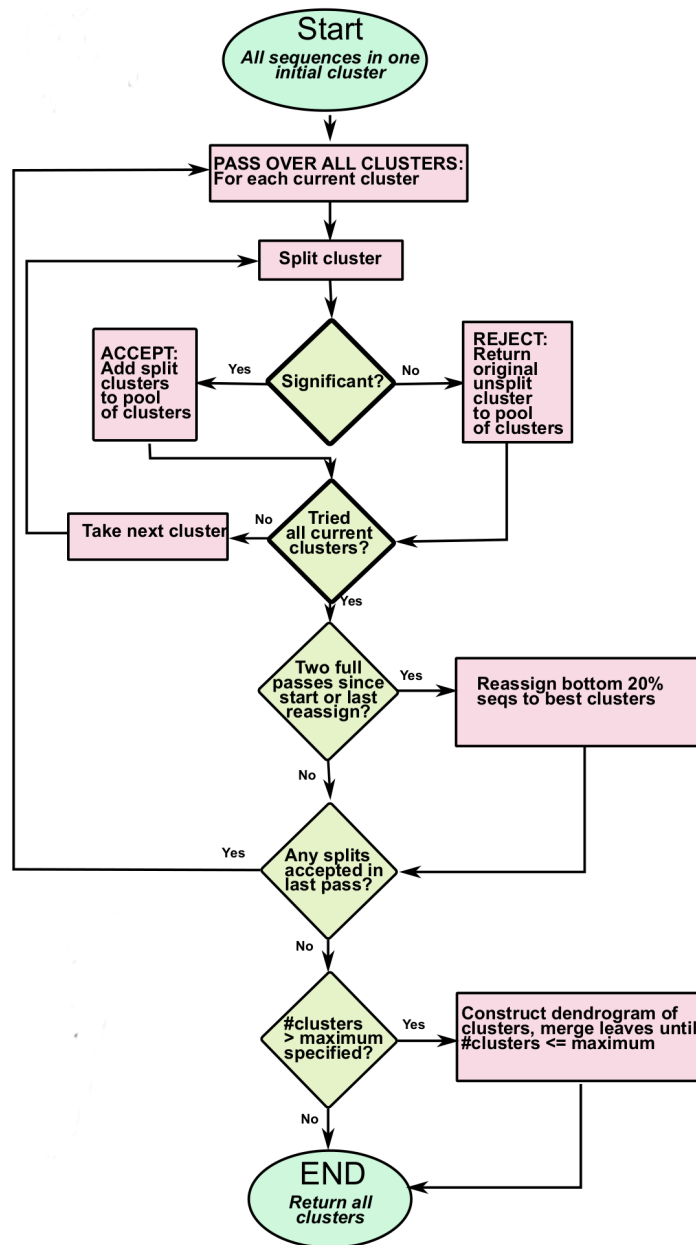


Figure 4.4: **Flowchart for the hierarchical clustering algorithm.** Flowchart for the hierarchical clustering algorithm. The initialization is with all sequences in one cluster. At every pass, an attempt is made to split every current cluster. Splits are accepted or rejected based on significance. Every two passes, a reassignment of low-scoring sequences to the best available cluster is made. When a pass has ended with no splits being made, the program terminates returning the current clusters.

individual sequences to find optimal clusters. This is implemented by considering fixed-sized “windows” of length W , one window within each sequence. Sequences may have variable length; we permit up to half the window to lie

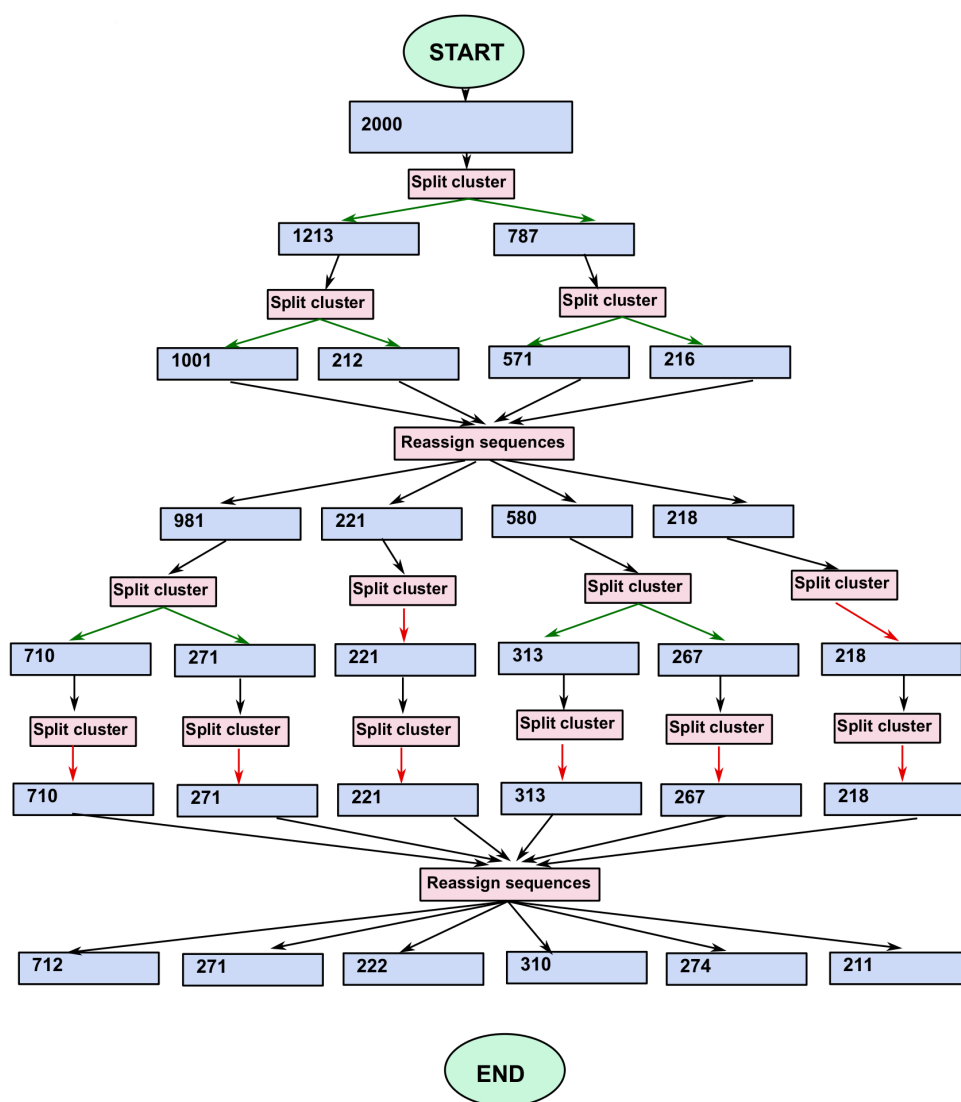


Figure 4.5: **An example of the possible run of hierarchical clustering algorithm.** A possible run for an input of 2000 sequences. The blue boxes represent cluster sizes, green arrows from “Split Cluster” boxes indicate successful splits and red arrows indicate unsuccessful splits. Each horizontal row of “split cluster” boxes represents one pass.

outside the sequence, with the missing nucleotides scored as N’s, so that for each sequence of length L , $2L$ configurations (L window positions and two orientations) are considered and the optimal window chosen. The default choice of W is one-third the median sequence length, that is, much longer than a typical TF motif. whose positioning and orientation is sampled. This, it turns out, constitutes an effective and fast implementation of an *ab initio* motif

finder on large ChIP-seq data sets, in addition to detecting the variations in motif and sequence context alluded to in the previous point.

THiCweed can also be used on sequences that have been previously aligned by a “feature” (motif) to discover additional motifs/complexities, by disabling shifts and reverse complements, similar to the program NPLB [Narlikar, 2014, Mitra and Narlikar, 2015].

Our divisive clustering is in contrast to typical (agglomerative) hierarchical clustering, where individual data points are formed into clusters, requiring $O(N^3)$ or at best $O(N^2 \log N)$ time for N data points.

4.3.1 Algorithm

Top-down hierarchical clustering

The algorithm and a typical run through it are portrayed in Figure 4.4 and Figure 4.5 and described below. We first take the simpler case of input data that has been pre-aligned with all sequences of the same length, where we don’t consider shifts and reverse-complements of sequences. The steps are as follows:

1. Initialize with one cluster containing all sequences.
2. Split every current cluster C (initially just one cluster), into two clusters C_1 and C_2 , using scoring and significance criteria described below. Sequences not consistently clustering with either C_1 or C_2 (as described below) are concatenated into a third cluster C_p . In each round, all these unclustered sequences from each division are concatenated into one cluster.
3. After every two iterations of step (2), if the current state has more than two clusters, reassign the poor-scoring sequences (sequences whose likelihoods in their current cluster are low) to the “best” available cluster.

4. Repeat from (2), until no new clusters are formed and no reassignments are made.

The user may specify a maximum number of desired clusters, and if the number of clusters at the end is greater than this, a dendrogram of current clusters is constructed and closest leaves are joined until the number of clusters is sufficiently reduced.

Scoring

Only windowed portions of sequences are scored. Let the window length be W . Consider a cluster C with N sequence windows in it, S^1, S^2, \dots, S^N . The probability of seeing this data if all these windows were drawn from the same PWM model is

$$P(C) = \prod_{i=1}^W \frac{\prod_{\alpha} \Gamma(n_{i\alpha} + c) \Gamma(4c)}{\Gamma(\sum_{\alpha} n_{i\alpha} + 4c) \Gamma(c)^4} \quad (4.2)$$

where $n_{i\alpha}$ is the number of times nucleotide α appears in column i , and c is a pseudocount (0.5 by default). If the cluster contains a single sequence, this expression reduces to $(\frac{1}{4})^W$.

The likelihood that a sequence window S is sampled from the same PWM as sequences in a cluster C that contains N seqs is

$$P(S|C) = \frac{P(S.C)}{P(C)} = \prod_{i=1}^W \frac{n_{iS_i} + c}{N + 4c} \quad (4.3)$$

where S_i indicates the i 'th nucleotide in sequence window S , and n_{iS_i} is the number of occurrences of that nucleotide at position i in the cluster.

When splitting a cluster, an initial split is made by ranking each sequence by its likelihood of belonging to that cluster, and moving the “best” 25% to another cluster.

Then sequences are selected in random order, removed from their current cluster and re-assigned to the more likely cluster, considering all possible window choices (position and orientation) within the sequence during the reassignment, until no further reassignments are made.

Finding Optimal Clusters

The significance of the split is assessed using two criteria. First, we demand that the ratio of the likelihoods of the two clusters, to the likelihood of the unsplit cluster, as calculated from equation 4.2, exceed a threshold, calculated from the LLR of two columns being cleanly separated in nucleotide composition. That is, suppose the two clusters consisted of random sequences, and were split on a single position – say, one cluster contained only A or C in that position, the other only G or T – while the nucleotides at all other positions are evenly distributed. This is not a significant split (it is always possible to do this, or better, for any cluster). Call the log likelihood ratio in this case L_1 . However, if the clusters differed in this manner in two positions – one cluster contained only A or C in those two positions, the other only G or T – this would be significant. Call the log likelihood ratio of this split L_2 . We demand the LLR of the split performed be equal to at least $L_T = L_1 + T(L_2 - L_1)$ where T is a parameter set to 0.4 by default and L_1 and L_2 can be calculated quickly using equation 4.2. The measure the significance of split cluster using score L_T depends upon adjustable parameter T . We predict the clusters of motifs with varying values of T in synthetic data of known motifs. Then we computed the accuracy of predicted clusters using adjustable Rand index (ARI). We got high ARI value for $T = 0.4$, so we fix this value as default parameter in ThiCweed.

Second, we demand that the splits be reproducible. using the following approach: we perform the split four times with four random initializations. With the resulting four pairs of clusters, we demand that at least three of the six pairwise cluster

comparisons that result have an adjusted Rand index (“ARI”) [Hubert and Arabie, 1985] greater than a threshold r (by default 0.2). An ARI of 1.0 indicates perfect agreement while random clusterings would have ARIs close to zero. If the three pairwise comparisons between the first three splits each exceed r , the fourth split is not performed. If the split is accepted, the three pairs of clusters resulting from the three splits are identified based on majority membership, and sequences that failed to be consistently clustered by this criterion (that is, did not cluster in the same way according to this association) are put in a third cluster.

Splits that fail one of these two significant criteria are rejected: the split clusters are joined again and returned to the pool. Figure 4.6 shows results on the synthetic data (1000bp set described in section 4.3.2) of THiCweed runs with various choices of T and r . While some extreme choices give poor performance presumably because they encourage excessive insignificant splitting or discourage valid splitting, the overall performance of clusters is not extremely sensitive to the choices of T and r . Based on this synthetic data, the choice of $T = 0.4$, $r = 0.2$ are set as defaults for the program.

When reassigning sequences (step 3 of the algorithm), we consider the poorest 20% of the sequences (measured by their likelihoods in their current clusters). For each sequence S , we first remove it from its current cluster, then calculate $P(C')$ for each available cluster C where $C' = C + S$, using the above formula, and add it to the best cluster. In practice, on average 4% and at most about 10% of the sequences considered in this step get reassigned.

4.3.2 Benchmarking: Synthetic Data

We generated synthetic datasets consisting of sequences of length 100bp each, with motifs drawn from random PWMs placed within the central 40bp of these sequences,

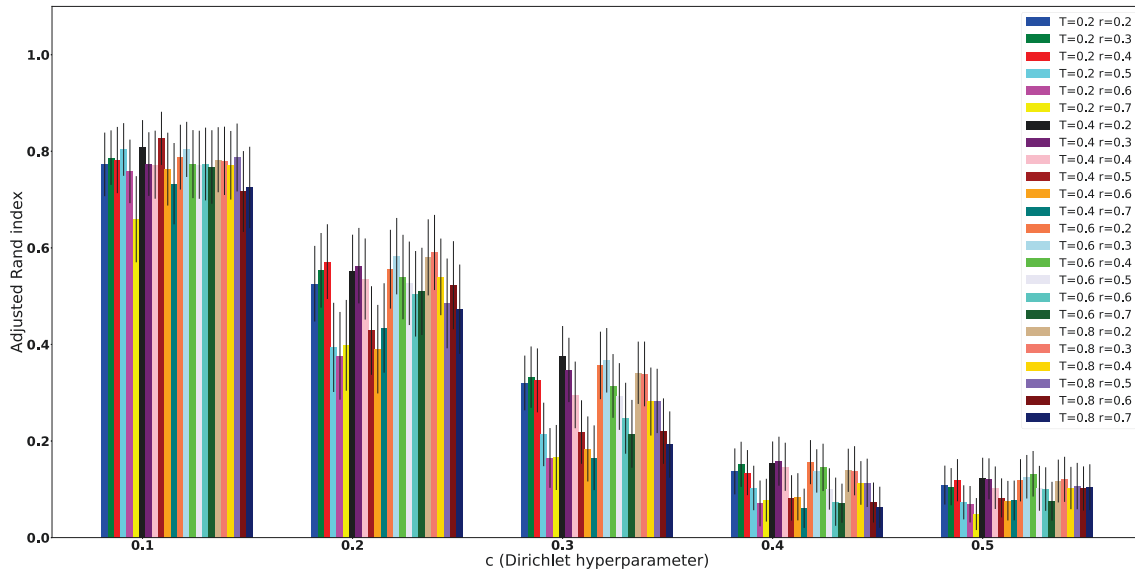


Figure 4.6: **Performance of THiCweed on synthetic datasets of size 1000 sequences/dataset for various value of T and r .** Accuracy of predicted clustering to known clustering given by ARI (value =1 is perfect). Synthetic data containing motifs drawn from PWMs sampled column wise from Dirichlet distributions with hyperparameter c . Error bars shown with thin vertical lines from 20 synthetic datasets.

and otherwise random (each nucleotide having probability 0.25). The PWMs had columns sampled from Dirichlet distributions with uniform hyperparameter c (ie, each column \mathbf{v} denoting the probability distribution over the four bases A, C, G, and T, was independently sampled from the distribution $P(\mathbf{v}) \propto v_{\alpha}^{c-1}$). Drawing from a Dirichlet distribution with a low value of c is more likely to result in a probability distribution that is highly skewed, ie is different from a uniform 0.25 probability per base. This skewness reduces with increase in c , a high value of c making the motif less distinguishable from background. Five datasets were generated with $c = 0.1, 0.2, 0.3, 0.4, 0.5$. Each dataset consisted of 20 files, with each file having sequences containing between 2 and 5 distinct motifs (one motif per sequence), the motifs drawn from PWMs of a “core” width of 5–10 bp and a tapering “flank” to a full width of 10–20bp (to reflect what is often in real data, as described below). The core positions were drawn from Dirichlet distributions with the hyperparameter c as described above, while the flanks tapered off rapidly from the core c to a

hyperparameter of 20 (essentially a uniformly random vector). The performance of the programs and therefore the conclusions do not change when the flank is omitted.

Each sequence contained one motif, and each dataset contained motifs drawn from a small number of PWMs. The number and lengths of PWMs were varied across datasets for each c , but the distribution of numbers and lengths was the same for different c 's. Figure 4.7 shows synthetic motifs for $c = 0.1, 0.3$ and 0.5 , all with a core width of 6bp and a full width of 20bp.

THiCweed and five other programs (Peak-Motifs, MuMoD, Chipmunk, Meme-Chip, Weeder2) were run on these sets, in multiple-motif ZOOPS mode (zero or one occurrences of a motif per sequence). The “known” clustering of the set was the assignment of sequences to PWMs, and the “predicted” clustering for each program was the assignment of sequences to predicted motifs. The known and predicted clusters were compared using the adjusted Rand index, and the results plotted as a function of c . Higher ARI indicates a better match between the clusterings, with 1.0 indicating perfect agreement and 0.0 being the value expected by chance.

Two such datasets are shown here, with dataset 1 containing 1000 sequences per file, and dataset 2 containing 5000 sequences per file. The ARIs are averaged over all 20 files for each value of c in each dataset.

Despite the “filter” keyword used in the command line, Chipmunk sometimes predicts multiple motifs per sequence because it searches for matches for predicted motifs in all sequences. For computing the adjusted Rand index, each sequence was classified to the best-matching motif, as per the score reported by Chipmunk. The same was done for Peak-Motifs. In addition, sequences where no motifs were reported were assigned to an additional cluster.

Table 4.1: Commandline options for various tools

Tools	Options
THiCweed	No additional parameters
MuMoD	Default parameters were used for the curves marked MuMoD. For MuMoD(i) the true number of motifs was specified.
ChIPMunk	In all runs, the correct number of motifs was specified . The length of the motif was given as 7:20.
Weeder2	Used with default options, but with a background frequency model derived from synthetic data.
MEME-ChIP (MEME)	Dreme was disabled with ‘-dreme-m 0’, and the known number of motifs specified with ‘-meme-nmotifs’, with default parameters otherwise.
MEME-ChIP (DREME)	meme was disabled with ‘-meme-nmotifs 0’
Peak-motifs	Default parameters were used

4.3.3 ENCODE Data

Here we used data from the ENCODE project [Consortium et al., 2012, Landt et al., 2012, Sloan et al., 2016], consisting of ChIP-seq peaks. NarrowPeak files were downloaded from the ENCODE website. 75bp flanking sequence was taken about each peak location, and repetitive regions (lowercase sequence in chromosome files downloaded from the UCSC Genome Browser [Karolchik et al., 2003], identified using RepeatMasker and Tandem Repeat Finder with period of 12 or less) were rejected for the purposes of this work. The cell types, ENCODE accession numbers for various factors and THiCweed output on ENCODE factors is available on the website (<https://www.imsc.res.in/~rsidd/thicweed/encodePredictions/>).

The ZNF143 clusters were compared with nucleosome positioning data in the same cell-type (GM12878) from ENCODE and PhastCons [Hubisz et al., 2010] phylogenetic conservation data (with other primates) from the UCSC genome site [Karolchik et al., 2003], distances from nearest transcriptional start sites (TSS), and DNase-seq values from ENCODE, using custom python scripts. For TSS we used the refGene data from the hg19 release on the UCSC genome browser site.

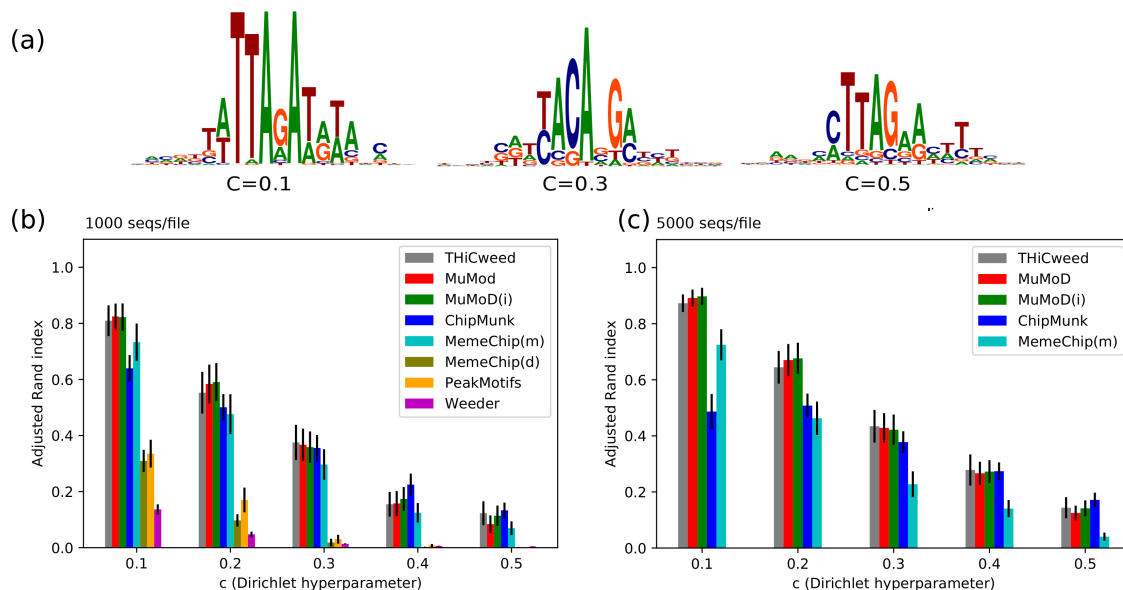


Figure 4.7: **Synthetic motifs and comparison of performance with other tools.** (a) Examples of embedded synthetic motifs. In this case all these have core widths of 6bp and full widths of 20bp, which are common to corresponding files in all datasets. The PWMs are sampled from different values of c , which varies from the indicated value in the core to a large value of 20 at the periphery. This is intended to model the appearance of motifs observed in real data. (b) and (c): Adjusted Rand index (higher is better) of predicted clustering to known clustering of synthetic data sets, containing motifs drawn from PWMs sampled columnwise from Dirichlet distributions with hyperparameter c . Error bars in black (standard error from 20 datasets). (b) In the case of 1000 seqs/file, THiCweed is competitive but somewhat inferior on this metric to MuMoD and ChipMunk, and somewhat superior to MemeChip (meme mode). (c) With 5000 seqs/file, comparing the better-performing programs from the previous figure, THiCweed is very close to MuMoD in performance.

4.4 Results

4.4.1 Synthetic Data

Results for the two datasets described in Methods are plotted in Figure 4.7 parts A,B,C for $c = 0.1, 0.2, 0.3, 0.4, 0.5$ (smaller value of c corresponds to sharper motifs).

In all cases THiCweed was run with default parameters, and in particular, a “window size” of 33bp or one-third the median input sequence length. As noted, it is designed to be run with large window sizes on real genomic data. Also, the stringent

criteria for splitting a cluster ensure that spurious clusters are unlikely, so setting the maximum number of clusters helps only marginally (not shown). Since clusters are split according to significance criteria, there is no option to set a minimal amount of clusters.

MuMoD was run both with default parameters (“MuMoD”) and with the additional information of number of motifs (“MuMoD(i)”); the latter provides only marginal improvement. Chipmunk (in ChipHorde mode) requires the exact number of motifs to be told to it, which was done in these cases, and the range of lengths of the motif was given. Meme-Chip with its default options run the MEME motif-finder on a random subset of the input data, with inferior results. Forcing MEME for the full set improved the results, at a significant cost in running time. For comparison, we also disabled MEME entirely in favour of DREME, a heuristic approach based on regular expressions rather than PWMs. Weeder2 was run with default options but a background model derived from synthetic data, as described in Methods. With 1000 seqs/set, THiCweed is competitive with MuMoD and ChipMunk on this metric.

Only the best performers were tested with 5000 seqs/set. All programs show improved performance here, because the motif strength is maintained the same but background “noise” reduces as $N^{-\frac{1}{2}}$ with increasing number of sequences N . But THiCweed’s improvement is sharper: it catches up with MuMoD and is largely superior to ChipMunk.

The reason for poor performance of Peak-Motifs seems to be its prediction of a very large number of motifs that are minor variations of one another. While it is hard to judge the relevance of this for real data, in the case of synthetic data these are certainly spurious, and THiCweed’s statistical criteria for splitting help it avoid this problem.

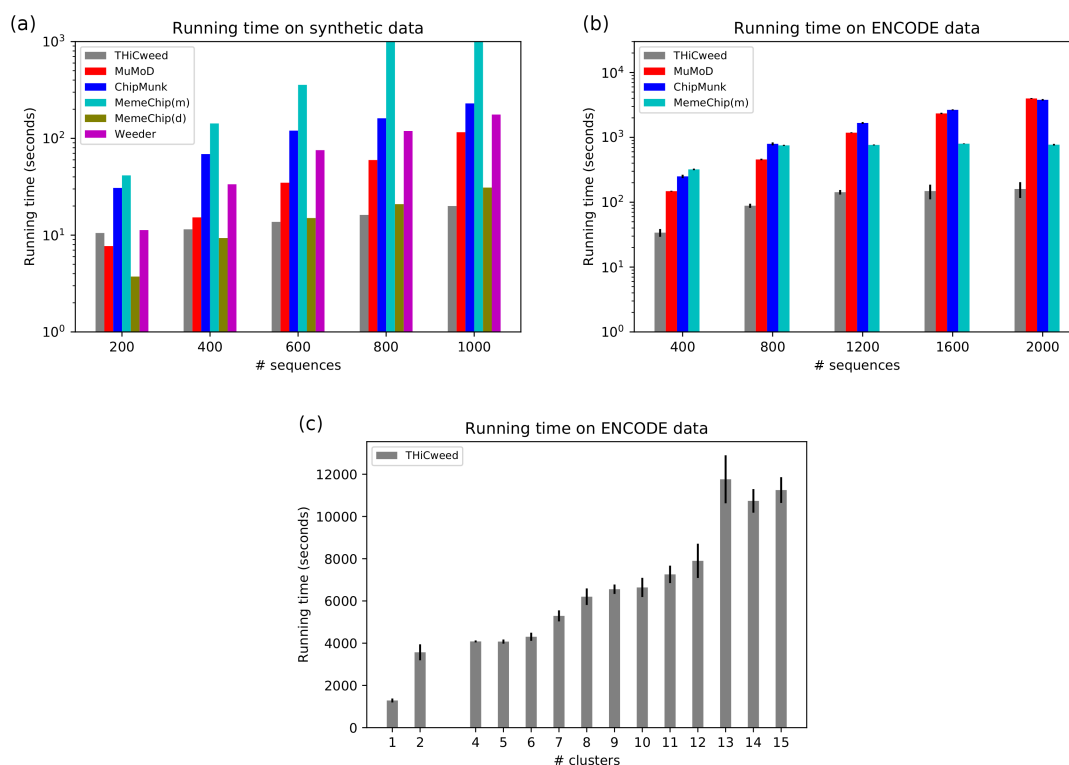


Figure 4.8: **Running time of various programs.** This is on synthetic data and THiCweed’s performance on real data varies significantly with the complexity of the sequence features; nevertheless, it remains on average much faster than other programs (Peak-Motifs was not tested but it is the fastest in this comparison).

4.4.2 Running Times: Synthetic Data

Figure 4.8 (a) shows running times of all the programs tested, except Peak-Motifs, for synthetic input data consisting of 200, 400, 600, 800 and 1000 sequences, each 1000bp long and containing two different motifs, each of length 10 sampled with Dirichlet parameter 0.2, in 60:40 proportion. Meme-Chip in MEME mode is an outlier: though its performance in accuracy is not very far behind other programs (Figure 4.7, its running time would seem to disqualify it from realistic datasets (and indeed it disables MEME by default for sequence sets larger than about 600×100 bp). It appears that, of the other programs, Chipmunk and Meme-Chip (Dreme mode) have runtimes increasing roughly linearly with data size; MuMoD and Weeder running times increase superlinearly; and THiCweed’s increase is somewhat sublinear.

4.4.3 Running Times: ENCODE data

Figure 4.8 (b) shows the results of THiCweed, MuMoD, ChipMunk and Meme-Chip (MEME mode) on real ENCODE data, consisting of 400–2000 random samples from a set of CTCF ChIP-seq peaks (dataset ENCFF001USS). The results are similar to on the synthetic data, except that, somewhat surprisingly, Meme-Chip is faster than MuMoD and ChipMunk on larger datasets.

Figure 4.8 (c) shows the running time of THiCweed as a function of the number of clusters found, on 92 ChIP-seq datasets each consisting of 27000–33000 peaks, across multiple TFs and cell lines. The running time increases with the number of clusters, but somewhat sublinearly. On such realistic ChIP-seq datasets, THiCweed’s running time is about two orders of magnitude less than MuMoD, which can take days, and is also much faster than all other programs tested. Meme-Chip uses the MEME step on only a small fraction of the input sequences; and Weeder2 learns motifs from a small fraction of the sequences and uses those to analyse the rest [Zambelli et al., 2014]. THiCweed processes the majority of files of this size in under two hours, with interesting and biologically relevant results.

4.4.4 ChIP-Seq Data from the ENCODE Project

Running on actual genomic data yields a variety of different results depending on the factor being examined and the size of the dataset.

THiCweed has no prior knowledge of the number of different motif clusters, but by default reports a maximum of 15. In some cases far fewer are reported. Because of the statistical criteria on splitting clusters that we use, described in Methods, we believe that large numbers of clusters, if produced, are statistically significant, but THiCweed can recluster the output into smaller numbers of clusters for ease of visualization, and this is done in some cases here. Also, it works with window sizes

much larger than typical motif lengths that one considers; here we used 50bp. We compare the discovered motifs to previously reported motifs from JASPAR [Sandelin et al., 2004, Mathelier et al., 2016]; the THiCweed website also includes comparisons to motifs from HocoMoco [Kulakovskiy et al., 2012] and FactorBook [Wang et al., 2013].

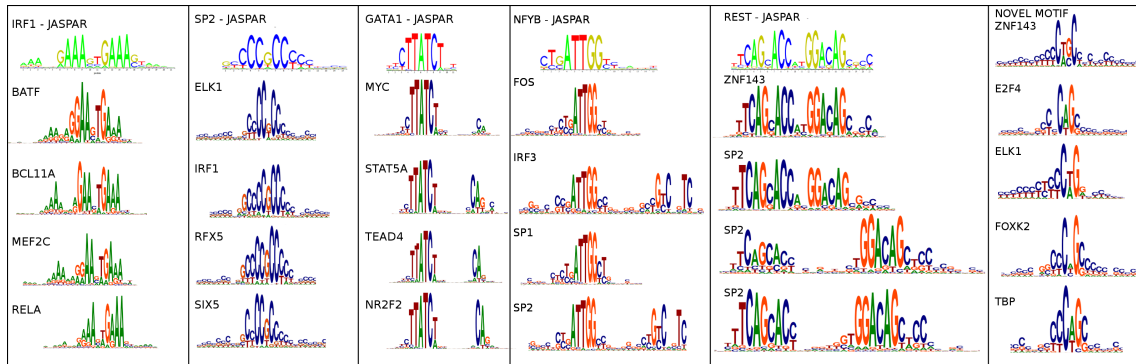


Figure 4.9: **Motifs that occur across multiple chip-seq datasets**, in addition to zinger motifs identified in [Hunt and Wasserman, 2014]. The factor for which the motif is the canonical motif according to JASPAR is indicated at the top of each column, together with the JASPAR sequence logo. Below are datasets for various other TFs where THiCweed finds the same motif.

Ubiquitous “zinger” motifs

Hunt and Wasserman [Hunt and Wasserman, 2014] observed that certain TF motifs occur repeatedly in different ChIP-seq datasets, which they termed “zingers”. In particular they identified CTCF-like, JUN-like, ETS-like and THAP11-like motifs in multiple datasets. We see all of these in our analysis of ENCODE data too (for example, the THAP11-like and CTCF motifs occur in Figure 4.12, but several other motifs appear across multiple experiments. Figure 4.9 shows examples that resemble IRF1, SP2, GATA1, NFYB, REST, and a novel motif that we could not identify. Of these, SP2 and the novel motif are roughly as ubiquitous as CTCF. Both frequently co-occur with CTCF and the SP2-like motif tends to be concentrated near TSS (an example is in Figure 4.12). We suspect a role for these in chromatin organization, a topic to be explored in future work.

Also noteworthy is the appearance of a secondary motif in multiple cases for the GATA-like and NFYB-like motifs; and the variable spacing of the REST-like motif. The canonical motif has two halves, TCAGCACC and GGACAG, separated by two nucleotides. But we pick up variants, previously described in [Otto et al., 2007], with longer spacing (8 and 9 bp here). Such widely spaced motifs cause problems for conventional motif-finders, but are readily picked up in our approach.

Examples of THiCweed output

Figure 4.10 shows four examples of motif output. In some cases the output has been reclustered and filtered for compactness of viewing; complete results for these and many more factors are available on the THiCweed website.

We make the following observations:

- Zinger motifs are widespread here. The SP1-like motif that we documented above occurs in IRF1 and NFYA. The unidentified motif in the previous section appears in REST and FOXA1. CTCF occurs in NFYA and FOXA1. ETS-like occurs in IRF1.
- The canonical motif for IRF1 occurs in two clusters, one of which has an additional poly-T tail.
- Similarly, the canonical motif for NFYA appears in three clusters, one of which also exhibits a weak secondary motif to the left.
- The canonical REST motif occurs as a closely-spaced dimer (4th cluster), partial closely-spaced dimer (5th cluster), monomer (3rd cluster) and a widely-spaced dimer (2nd cluster). All of these variants also occur in THiCweed output for SP2 (Figure 4.11) suggesting an interaction between SP2 and REST. The widely-spaced dimer is not picked up by other motif finders.

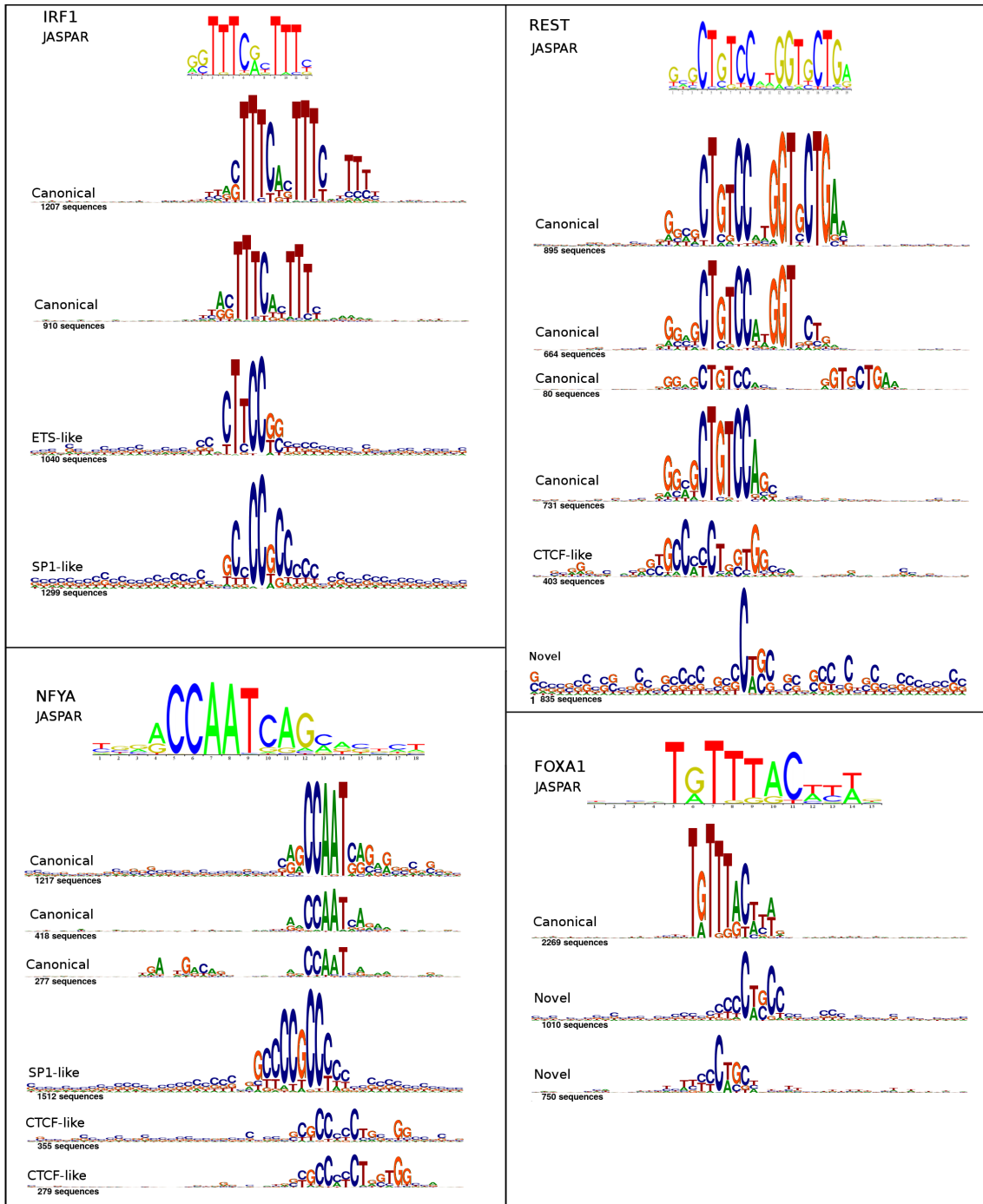


Figure 4.10: Sample THiCweed output on four CHIP-seq datasets: IRF1 (5543 peaks), NFYA (4497 peaks), REST (3998 peaks), FOXA1 (4029 peaks). Not all output clusters are shown here. The full output is available on the THiCweed website.

Comparison with other programs

Figure 4.11 compares the output of THiCweed with three other programs. All programs pick up the main motif (though with varying numbers of instances). All

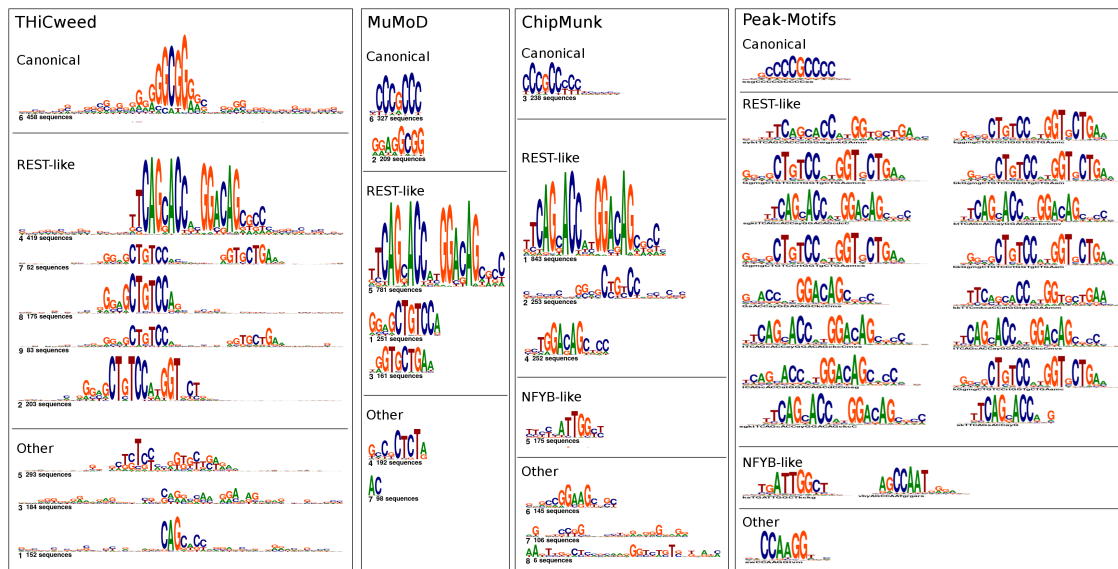


Figure 4.11: **Comparison of clustering of 2,019 peaks** for SP2 by THiCweed, with motifs found by MuMoD, ChIPMunk and Peak-motifs programs.

also pick up the REST motif, but only THiCweed picks up the widely-spaced version in one piece. THiCweed also seems to reveal a larger surrounding-sequence context in many cases, notably for the SP1-like motif which generally occurs in a CG-rich background. Peak-Motifs identifies a very large number of motifs, most of which appear to be minor variations of the main motif. This may explain the poor performance of Peak-Motifs on our synthetic benchmark: the adjusted Rand index would penalize breaking up clusters into smaller clusters.

Biological relevance of these clusters

We typically find several different motifs, variants of a motif, and a few apparently uninformative clusters in THiCweed runs. Biological significance to these are suggested on comparing other genomic features such as phylogenetic conservation (via PhastCons scores [Hubisz et al., 2010] from the UCSC Genome Browser [Karolchik et al., 2003]) and nucleosome occupancy and DNase-seq data (from ENCODE [Consortium et al., 2012]). Figure 4.12 compares each of 8 clusters for ZNF143 with a plot of conservation, nucleosome occupancy (in an extended region of 1000bp on

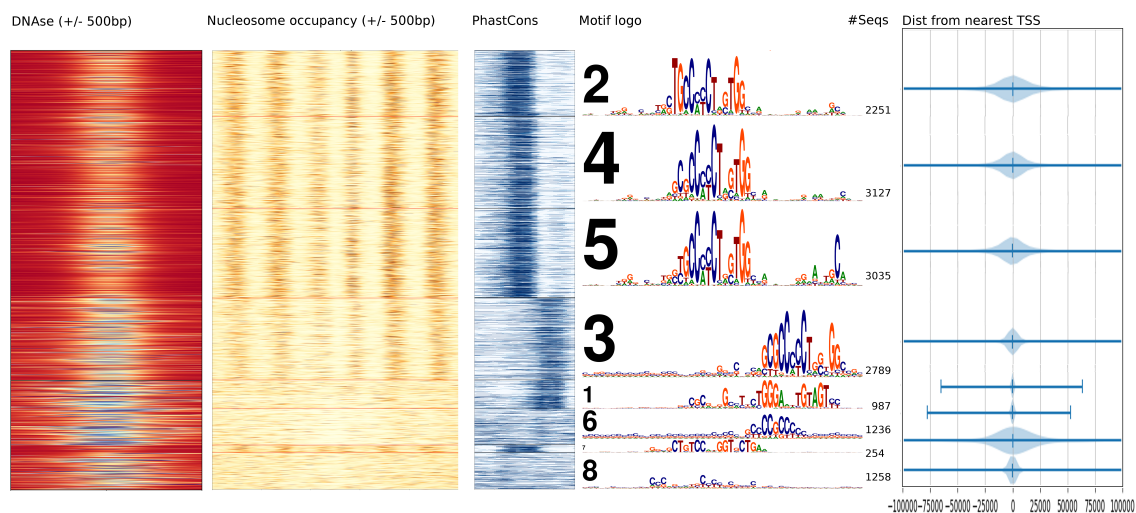


Figure 4.12: **Biological relevance of sequence clusters.** Comparison of sequence clusters of 14,937 ZNF143 ChIP-seq peaks with DNase-seq values (colour scale: blue=open, red=closed), nucleosome occupancy (colour scale: white = 0, brown = 5+), PhastCons conservation score (colour scale:white=0, dark blue=1), and distance from nearest TSS, suggesting connections between the motif structure in different sequence clusters, biological function, evolutionary conservation pressure, nucleosome positioning and open/closed chromatin.

each side), distance to the nearest TSS, and DNase-seq values. Cluster 6 (SP2-like motif) tends to be concentrated close to TSSs (mostly within 1000bp – a pattern we see consistently), shows little phylogenetic conservation, and no sign of nucleosome positioning. Cluster 1 (a motif resembling THAP11, identified in [Hunt and Wasserman, 2014] as a zinger motif), too, is concentrated near TSSs; it too shows little effect in nucleosome positioning, but is strongly conserved. Cluster 7, resembling the REST motif, is spread away from TSSs, is phylogenetically conserved, and has an effect on nucleosome positioning (which we observe in other datasets where this motif occurs).

Cluster 8 seems uninformative, but it appears concentrated near the TSSs (within about 5000bp), which would likely not happen if it consisted only of random un-clusterable sequences left over from the other clusters.

The remaining clusters are variants of the CTCF motif; cluster 5 includes the previously documented “M2” motif. Cluster 3 appears different from other CTCF clusters

in that it occurs in a GC-rich background, is more concentrated near TSS (mostly within about 10000bp), appears a little less conserved and a little less effective at nucleosome positioning, with more open chromatin as shown by DNase.

4.5 Discussion

Motif-finding in large datasets produced by ChIP-seq and similar experiments is a qualitatively different problem in complexity from what traditional motif-finders are used to handle. Additionally, one could liken the problem of finding rarely-occurring motifs to finding a needle in a haystack. We view THiCweed’s approach as “sequence feature analysis” (over large windows) rather than “motif-finding” (detection of short patterns). Our novel clustering algorithm can comfortably handle tens of thousands of sequences at a time, and with significant heterogeneity in motif content. It successfully picks up biologically relevant motifs even when they occur in fewer than 5% of the input sequences, such as the REST-like motif in ZNF143 (cluster 7 in Figure 4.12). Its large window size enables it to also pick up secondary motifs like the M2 CTCF motif in the ZNF143 data (Figure 4.12), the widely-spaced dimer in SP2 and REST (Figures 4.10 and 4.11), and peripheral features such as an overall CG-richness in some motifs (eg CTCF-like cluster 3 in Figure 4.12). The significance criterion used for splitting, and the differences in biological parameters in Figure 4.12, suggest that these differences are important and are not artifacts.

Uniquely among the programs we have tested, THiCweed achieves its combination of speed and accuracy without resorting to heuristics in scoring (as DREME and Chipmunk do, using regular expressions and “seeding” respectively) and without resorting to training on a small subset of the sequences (as Weeder does). THiCweed’s clustering algorithm is stochastic, but is essentially similar to an iterated K -means clustering with $K = 2$, with significance criteria to avoid spurious splits. Instead of

invoking pairwise distances and calculating a centroid, however, we calculate multinomial likelihoods correctly within the limitations of the PWM assumption. The clustering algorithm and wide-window approach ensures that little or no prior information is required to run the program: significant short motifs can be found inside longer windows by eyeballing, but other relevant sequence features can be picked up too.

A possible shortcoming is that within THiCweed’s framework, only one motif occurrence per sequence will be detected (unless two motifs co-occur with a restricted spacing, as in the extended REST motif and the secondary M2 CTCF motif). Sequences that match no dominant motif may end up in a relatively uninformative cluster such as cluster 8 in Figure 4.12. One may ask whether, in clusters that do not match the canonical motif, the motif nevertheless occurs elsewhere in some of the peaks in addition to the non-canonical motif in the cluster. It is possible that for a given input sequence, the canonical motif occurs more than once. But THiCweed would show only one occurrence. So, to check this possibility, we ran FIMO [Grant et al., 2011], with the canonical motif on input fasta files and a q -value threshold of 10^{-3} . We count what proportion of sequences that were not clustered with a recognizable motif, a motif match was found by FIMO anyway. The proportion was quite small: only 25 factors out of 93 showed any occurrence, mostly in much fewer than 1% of such sequences.

These too showed matches in very few cases; the exceptions were SP2 and EGR1, both of which have GC rich canonical motifs, which reported matches in about 15% and 14%, respectively, of such sequences. It would therefore seem that the “missing” of canonical motifs because of occurrence of other strong motifs within ChIP-seq peaks is not a common concern in practice.

In cases where there is a profusion of similar but slightly different motif patterns as well as an occurrence of many different motifs (as in the ZNF143/CTCF case), it

appears that the differences may have biological significance, as reflected by nucleosome positioning and phylogenetic conservation. We plan to explore this, and the significance of some of the novel zinger motifs, further in a future work.

Chapter 5

Discussion and Conclusion

In this thesis, we have described one possible origin of large-scale nuclear structuring. Organelles such as the nucleus in eukaryotic cells are membrane-bound and thus explicitly compartmentalized. However, compartmentalization can be an emergent phenomenon, deriving from multiple interactions in a complex system. We propose that it is natural to identify the hierarchical structuring of the human cell nucleus at its large-scale as an emergent property associated with activity. We might then reasonably expect to be able to understand a number of generic properties of large-scale nuclear architecture using simpler polymer models that describe this activity explicitly while omitting other details.

The model for large-scale nuclear architecture described in this thesis stresses a specific biophysical effect, relevant to the modelling of chromosomes in living cells. Our central assumption is that a connection between levels of inhomogeneous activity arises from local transcriptional activity across different regions of chromatin and large-scale properties of nuclear architecture [Chu et al., 2017, Almassalha et al., 2017]. We propose that the intensity of active processes should increase with increased transcriptional output. We map a reasonable measure of local transcriptional activity, inferred from combining population-level measures of local RNA-

output with estimates of the local gene density, into an effective temperature seen by each monomeric unit in our polymer model of chromosomes. This inhomogeneous activity is associated with non-equilibrium, ATP-consuming processes acting locally on chromatin. The fact that a number of broad features of the experiments are reproduced in our model suggests that the large-scale structure and positioning of individual chromosomes are principally determined by inhomogeneous activity across chromosomes, the presence of loops and confinement.

Generally, the active beads in chromosomes appear in the form of clusters; they rarely appear alone. The average length of a continuous stretch of inactive beads varies from a minimum of 2 in chromosome 10 to a maximum of 7 in chromosome 22 for the GM12878 cell type. The average cluster size of active beads varies between 1.5 in chromosome 21 to 9 in chromosome 19. For other cell types, the value of the average cluster size of active beads and inactive beads varies but are broadly consistent with that of the GM12878 cell type. In the case of chromosome 22, both continuous long stretches of inactive beads, as well as of stretches of active beads are found consistently. This is possibly the reason that the statistical properties of chromosome 22 are special and differ from those of the rest of the chromosomes.

Our biophysical description of chromosomes and their structuring, given our coarse-graining to the 1Mb scale, reproduces the different spatial distributions of chromosomes, a feature seen across multiple cell types. A central consequence of our model is that gene expression should correlate to a larger strength of mechanical fluctuations, i.e. activity, and that the radial distribution of chromosomes should be attenuated towards the boundaries of the nucleus. This is an emergent property, arising from the combination of differential activity and confinement, that could not have been inferred from how the model was constructed. We ignore hydrodynamical couplings between different sections of chromosomes mediated by the intervening nucleoplasm, on the grounds that they can be neglected in a highly confined sys-

tem where the (inverse) system size effectively cuts off this interaction and where a number of other constituents, not modelled explicitly, are available to take up momentum [Bruinsma et al., 2014]. In view of the large-scale separation of our basic units, the monomers, vis a vis these microscopic force units, an effective description in terms of uncorrelated Gaussian noise should be appropriate at the monomer scale.

Our model recovers the territorial nature of chromosomes. Such territoriality stems from two sources. First, the compactness of individual chromosomes is assured through looping, with the mean distance between two points on a chromosome saturating as the contour length between them is increased. Making chromosomes compact automatically ensures some degree of territoriality. However, territoriality is further enhanced through segregation by differential activity, which ensures that individual chromosomes are well separated, since levels of activity are not constant across chromosomes. Crucially, this segregation based on differential activity also produces the experimentally seen distinctions in the positioning of more active and less active chromosomes. Much work on the large-scale architecture of chromosomes has described individual chromosomes, ignoring the role of packing within the nucleus. Our model, in contrast, describes all chromosomes at an equivalent level and chromosome confinement is a crucial aspect of our model. Varying the active temperature assigned to monomers, either by changing the absolute scale of active temperatures or by changing the relative proportion of active to inactive monomers, induces changes in the positioning of chromosome centres-of-mass while leaving the distribution of gene density largely invariant; it is important to note that these quantities are not equivalent. The variation across experiments might also reflect genuine variations in activity distributions in inequivalent conditions [Küpper et al., 2007].

The distribution functions we calculate, $S(R)$ certainly depend on confinement, as implemented using the size of the enclosing volume. If we increase the radius of

the nucleus keeping all else the same, these distribution functions become largely featureless. There is a dependence on topology as well, since increased compactness of chromosomes decreases the importance of confinement.

We made some approximations in our polymer model of large-scale nuclear architecture and here we describe the physical rationale for them. The **first** approximation is that we ignore the effects of self-avoidance, working with self-repelling but not fully self-avoiding polymer models for chromosomes. The fractal globule model suggests that intermediate configurations between open and collapsed self-avoiding polymer configurations are relevant to the understanding of the structure of individual chromosomes. There are no biological reasons to assume that chromosome conformations in interphase arise from anything like polymer collapse. In polymer collapse, first, small crumples are folded, leading to formation of an effectively thicker polymer-of-crumples, which next forms large crumples itself. Since interphase chromatin is less condensed than chromatin within or exiting mitosis. A more profound difficulty for the model is that the fractal globule state appears to be only a metastable one [Mirny, 2011]. We ignore self-avoidance in our model for the following reason. The cell utilizes a large number of enzymes that change the structure of DNA through active, energy-consuming processes. Due to this, *in vivo* biological systems are far from the ones encountered in equilibrium soft matter systems, and time-scales for topology changes far exceed any relevant experimental time-scale.

Our **second** approximation relates to the unit of coarse-graining. The monomers in our simulation are represented by 1Mb sections of chromosomes, although we could have defined our model at smaller scales of 100 kb or even less. However, the averaging inherent in summing transcriptional output over a 1Mb scale renders the model relatively less sensitive to errors and noise in this input. Further, the 1Mb scale is believed to be an appropriate building block for chromosome territories [Lieberman-Aiden et al., 2009]. A more detailed and explicit model for non-equilibrium activity

and its consequences for an active temperature description would be useful, but the form such a model ought to take is presently unclear and best left to more extensive investigations. Irrespective of potential quantitative improvements on the model front, the broad trends we describe here should be largely robust.

A **third** approximation relates to the use of an effective temperature for active regions of chromosomes. The utilization of an active temperature is a convenience as opposed to a necessity since it is used to depict the net impact of dynamic mechanical fluctuations, which we expect to be uncorrelated from monomer to monomer. Our estimates for the scale of the effective temperature are acquired from the following biological arguments. ATP-dependent chromatin remodelling enzymes, present in huge numbers in the cell nucleus, can surmount hindrances an order of magnitude larger than energy scales associated with physiological temperatures while positioning nucleosomes [Hargreaves and Crabtree, 2011]. Therefore fluctuations of individual chromosome loci, which raise the effective temperature in the cell, might be used to mechanically regulate **in vivo** gene expression [Weber et al., 2012]. Since non-equilibrium energy input occurs through the hydrolysis of ATP which releases an energy of approximately $20k_B T$, the active temperature should be bound by less than $20k_B T$. In our simulations, we have varied the temperature of active monomers within a range of 6-20 times the effective temperature in physiological conditions. The variation in active temperatures gives roughly similar results. As we have pointed out, the scale of active temperature seems less important than the fraction of monomers which are assumed to be active.

The spatiotemporal organization of transcription is actively regulated and maintained by the nucleus. This regulated activity can be achieved by transcriptional machinery, which targets specific gene loci at precise times. To complete this process, the physical environment of the nucleus must follow these two opposing conditions. First, the nuclear medium must be rigid enough to maintain the 3D positioning

of the gene loci. Second, it must be fluid enough to allow for specific transcription agents to be driven towards their gene targets at specific times [Hameed et al., 2012].

In general TF-DNA interaction can be either direct, or indirect through contact with other proteins. This protein-DNA interaction occurs via different mode of interactions. The types of mode in direct DNA-binding are monomer binding, homodimer binding, and heterodimer binding; and indirect DNA-binding are piggyback binding and multi-TF binding. Most of the motif discovery tools don't report minority motifs, and they provide little indication of explanation in full dataset. The computational methods discussed in section 4.1.2 report highly similar motifs in content, mostly built from slightly different subsets of the bound sequences. As a consequence, the decision of choosing the meaningful motifs is left for biologists [Narlikar, 2013].

The different types of binding modes in ChIP-Seq data are difficult to distinguish by traditional methods. These data may contain binding sites for canonical TF only, non-canonical TF only, or both. Secondary motifs found in our THiCweed results are often very different from the canonical motif of the TF being assayed. There are many possible reasons for which secondary motifs appear in the results: First, cobinding i.e. two TFs binding to neighbouring sites as heterodimer, physical or cooperative interaction; Second, one TF binding to another that, in turn, binds to DNA through tethered binding; Third, Cohesin/polycomb and TFs of secondary motifs participate in demarcation and stabilization of inter-segment interactions of DNA at which primary TFs bind [Hunt and Wasserman, 2014]. In the case where sites of non-canonical motifs are frequently found to be in the same ChIP-Seq peaks as canonical motif sites, then two TFs are likely to interact at the protein level and influence each other in binding to their DNA sites. This is hard to find by any computational method. Contrary, if the majority of the peaks contain only sites for non-canonical motifs, then tethered binding is a more possible scenario. SP1 (or

SP2) and NFY (heterodimer of NFYA and NFYB) prefer to cobind neighbouring sites in the genome [Wang et al., 2012]. SP1 and SP2 TFs mostly bind to common sites in the genome. Similarly, YY1 interacts with MYC, ESRRA interacts with HNF4, and NHKB interacts with SPI1. Another secondary motif USF consistently occurs in all MAX or MYC datasets which suggests they compete for sharing binding sites [Wang et al., 2012].

Combinatorial regulation by TFs that do not bind DNA directly is an example of tethered binding. For example, canonical motif of ATF3 is CREB motif. In ChIP-Seq peaks of ATF3, almost half of the peaks contain USF sites but not CREB sites, suggests that ATF3 tethers to USF, which binds DNA directly [Wang et al., 2012]. Other examples include SP1 tethering to HNF4, STAT3 tethering to CEBPB, TCF12 tethering to FOXA and HNF4, IRF1 tethering to NFY, SREBF1 tethering to RFX5, and SIX5 tethering to ZNF143 [Wang et al., 2012].

THiCweed reports many examples of secondary motifs but not all have been verified yet. (i) SP2, JUN, TFE3, TFEB, MAX, ZNF263, CTCF, and ARNTL motifs found in USF; (ii) SP2, MNT, CTCF, ZNF263, and MYC motifs found in MAX; (iii) ZNF263, CTCF, SP2, JUN, EGR1, GATA4, CEBPA, and SPIC motifs found in MYC; (iv) CTCF, NFY, and REST motifs found in SP1; (v) SP2, CTCF, and ZNF263 motifs found in NFY; (vi) ZNF263, CTCF, IRF1, SP2, FOXP1, FOXJ3, and HNF4G motifs found in YY1; (vii) JUN, CTCF, TFEB, USF, SP2, TFE3, HNF4G, and ZNF263 motifs found in ATF3; (viii) SPIC, IRF1, and ZNF263 motifs found in SPI1; (ix) MAFG, NFE2, ZNF263, PRDM1, SP2, and ELK4 motifs found in IRF1; (x) IRF1, JUN, CEBPA, ZNF263, SP2, CTCF, and FOS motifs found in STAT3; (xi) ATF4, JUN, IRF1, CTCF, ZNF263, and SP2 motifs found in CEBPB. These results suggest that TFs possibly participate in co-binding events. Some of which are reported in earlier literature also. Motifs for SP2 and ZNF263 frequently occur in peaks for many other TFs, suggesting that they may participate in some

kind of genome organization or are related with nucleosome positioning. We will explore the significance of some of these novel zinger motifs, and the absence of canonical motifs due to the occurrence of other strong motifs within ChIP-Seq peaks, further in future.

5.1 Future Directions

The core of all model is a set of approximations made to render a calculation persuadable. Biophysical models for nuclear architecture must walk a delicate line between incorporating the requisite biological complexity on the one hand and a preference for simplicity and generality on the other hand [Cook and Marenduzzo, 2009b, Rosa and Everaers, 2008, Dorier and Stasiak, 2010, Bohn and Heermann, 2010, Barbieri et al., 2013, Pombo and Nicodemi, 2014]. The work we present in this thesis surrenders our ability to model chromatin behaviour at scales shorter than 1 Mb in order to make specific, testable predictions for chromosomes at large-scale [Ganai et al., 2014, Agrawal et al., 2017, Agrawal et al., 2018a]. However, we incorporate the precise details of the system (the chromatin inside the nucleus) in several ways, primarily through assigning profiles of activity to each chromosome, reflecting either their gene density or their gene expression. In addition, we incorporate chromosome looping as inferred from 3C related experimental data. The advantage of the large-scale nuclear architecture model which we described is that it can be improved easily by adding more relevant biological input.

We can extend our model with several choices. First, we can incorporate role of lamin proteins in anchoring specific lamin-associated domains (LADs) to the nuclear lamina, as well as the interactions of specific gene loci with nuclear pore complexes [Mattout et al., 2015]. While we omit the effect of lamins in the current work, the omission can at least be qualitatively justified by the biophysical intuition that the

activity-based physical segregation of chromosomes is a bulk or volume effect that should dominate, at the simplest level of description, over surface effects arising from interactions with the nuclear envelope. Thus, modelling the effects of interactions of LADs with the nuclear lamina by introducing weak monomer-specific interactions with the inner surface of the confining sphere in our simulations might be expected to modify the results we present here for specific chromosomes, but hopefully in a controlled manner.

Second, we can include nucleoli in our model, formed around nucleolar organizer regions containing multiple copies of rRNA genes, with such regions located on the short arms of the acrocentric chromosomes 13,14,15, 21 and 22 [Németh and Längst, 2011]. We can account at least qualitatively for the presence of the nucleolus, a relatively large and dense nuclear landmark, by excluding a pre-decided subvolume of space within the simulated nucleus from being occupied by other chromosomes and adding a weak attractive attraction that favours association to monomers associated to the p-arms of the acrocentric chromosomes.

Third, we simulate the nucleus as a spherical shell containing our model chromosomes, although nuclear shapes exhibit considerable variability and much of the experimental data comes from experiments on the relatively flattened nuclei of fibroblasts [Bolzer et al., 2005]. Our model could be generalised to account for the effects of variable nuclear shapes.

Fourth, we ignore the potential interactions of looping across chromosomes. Such interactions could potentially arise from the looping out of loci on different chromosomes to interact at transcription factories [Maharana et al., 2016]. We could account for this by making designated monomers on different chromosomes sticky with respect to each other, thus coupling regions of different chromosomes that are known to physically localise together when co-transcribed.

Fifth, while using RNA-seq data as a proxy for activity, we largely considering

steady-state gene expression only. Inferring activity from GRO-Seq (Global Run-On Sequencing), which also extracts nascent and rapidly degraded transcripts, may help to provide a more accurate view of transcription-coupled activity.

Last, the role of nuclear actin and associated motors remains unclear, although they could potentially contribute additional sources of non-equilibrium noise [de Lanerolle, 2012]. Indeed, all the possible improvements on our model that we list above could be incorporated, but only at the expense of putting more details in the model and with a number of further assumptions. We choose to leave these questions for future work.

In first-principles approaches, a small set of initial model assumptions, argued for on general grounds, must yield consistent explanations and descriptions for all data, not just those the model abstracts in its construction. The advantage of simple models is that they enable us to concentrate on underlying principles that are often obscured by the complexity of real data, including intrinsic heterogeneities across cell populations, varied experimental and analysis procedures and the lack of sufficient statistics in some cases. Prior models for nuclear architecture in mammalian cells fail to reproduce many general attributes of nuclear architecture known from experiment. Certainly, these properties emerge in our calculations, since they were not directly encoded in our model specification. This suggests that our methodologies provide as it yet unavailable biophysical insights of large-scale nuclear architecture in metazoans.

Approaches such as those we describe here are possibly the only ones that can provide first-principles-based answers to the following questions, since at their core they all relate to the biophysical principles that underly how macromolecules can relocate across micron scales in a statistically reproducible manner: Cancer cells show altered transcription patterns, are typically aneuploid and exhibit characteristic translocations [Lengauer et al., 1998, Roukos and Misteli, 2014, Ranade et al.,

2017]. They also display changes in chromosome positioning relative to normal cells, but how such changes correlate to the altered transcriptome is not understood [Marella et al., 2009]. Chromosome territories alter their shapes and positions following DNA damage, but also relax to their unperturbed positions as DNA repair proceeds [Mehta et al., 2013, Dabin et al., 2016]. What determines these large-scale positional shifts, and how they are modulated by energy-consuming DNA repair processes acting on chromatin, is unclear [Kruhlik et al., 2006, Ioannou et al., 2015]. Our understanding of how the unique properties of the stem cell transcriptome contribute to the biophysical properties of stem cell chromatin, including its fluidity, remains limited [Talwar et al., 2013, Pajeroski et al., 2007]. *Answering such questions requires that we understand, at a minimum, the coupling between the positioning of individual chromosomes and their transcription levels.* Our model describes a biophysically-motivated way of describing this coupling, yielding predictions that compare favourably to published experimental data while also providing benchmarks upon which more detailed studies can build. The first-principles approach we propose here connects cell-specific gene expression patterns to large-scale nuclear architecture, suggesting how the problems listed above might be fruitfully addressed.

TFBS are identified using patterns of conserved sequences. However, using only sequence information to find TFBS is oversimplified. The 3d ‘ of DNA, which reflects the physicochemical and conformational properties, is critical for the packaging and regulation of DNA in the cell. The protein-DNA binding interaction is a three-dimensional interactions, so finding the structure of DNA is important to understand the correct mechanism of protein-DNA binding. However, the relationship between TFs and corresponding DNA structural properties remains to be elucidated. It is known that a considerable number of TFs showed distinct DNA structural preferences. These structural features also show positional preferences in TFBS [Dai et al., 2015].

Chromatin interactions play a critical role and serve to regulate gene expression. The current model for the identification of TFBS uses genomic regions around CHIP-Seq peaks. This data provides only linear information of TFBS along the chromosome, is unable to determine the target genes of distal TFBS and suffers from high genomic background noise i.e. false positives. Information on spatial chromatin interaction can give clues on positioning the regulatory elements close to their target genes, and provide novel insights into the study of transcription regulation. Such data can be provided by, genome-wide, high-throughput methods such as Hi-C and ChIA-PET. ChIA-PET can be used for studying long-range chromatin interactions in a three-dimensional manner, as well as for determining TFBS for particular protein of interest. In future, we will explore how the binding sites for same or different TFs co-regulating in the context of spatial chromatin interactions.

Bibliography

- [Agrawal et al., 2017] Agrawal, A., Ganai, N., Sengupta, S., and Menon, G. I. (2017). Chromatin as active matter. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(1):014001.
- [Agrawal et al., 2018a] Agrawal, A., Ganai, N., Sengupta, S., and Menon, G. I. (2018a). A first-principles approach to large-scale nuclear architecture. *bioRxiv*, page 315812.
- [Agrawal et al., 2018b] Agrawal, A., Sambare, S. V., Narlikar, L., and Siddharthan, R. (2018b). Thicweed: fast, sensitive detection of sequence features by clustering big datasets. *Nucleic acids research*, 46(15):e29.
- [Almassalha et al., 2017] Almassalha, L. M., Bauer, G. M., Wu, W., Cherkezyan, L., Zhang, D., Kendra, A., Gladstein, S., Chandler, J. E., VanDerway, D., Seagle, B.-L. L., et al. (2017). Macrogenomic engineering via modulation of the scaling of chromatin packing density. *Nature biomedical engineering*, 1(11):902.
- [Amitai and Holcman, 2017] Amitai, A. and Holcman, D. (2017). Polymer physics of nuclear organization and function. *Physics Reports*, 678:1–83.
- [Armstrong, 2013] Armstrong, L. (2013). *Epigenetics*. Garland Science.
- [Ay et al., 2014] Ay, F., Bailey, T. L., and Noble, W. S. (2014). Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, 24(6):999–1011.

- [Bailey et al., 1994] Bailey, T. L., Elkan, C., et al. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc Int Conf Intell Syst Mol Biol*.
- [Barbieri et al., 2012] Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., and Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences*, 109(40):16173–16178.
- [Barbieri et al., 2013] Barbieri, M., Scialdone, A., Gamba, A., Pombo, A., and Nicodemi, M. (2013). Polymer physics, scaling and heterogeneity in the spatial organisation of chromosomes in the cell nucleus. *Soft Matter*, 9(36):8631–8635.
- [Bártová et al., 2008] Bártová, E., Krejčí, J., Harničarová, A., and Kozubek, S. (2008). Differentiation of human embryonic stem cells induces condensation of chromosome territories and formation of heterochromatin protein 1 foci. *Differentiation*, 76(1):24–32.
- [Bellomo et al., 2007] Bellomo, N., Bellouquid, A., and Herrero, M. A. (2007). From microscopic to macroscopic description of multicellular systems and biological growing tissues. *Computers & Mathematics with Applications*, 53(3-4):647–663.
- [Belmont, 2002] Belmont, A. S. (2002). Mitotic chromosome scaffold structure: new approaches to an old controversy. *Proceedings of the National Academy of Sciences*, 99(25):15855–15857.
- [Benedetti et al., 2013] Benedetti, F., Dorier, J., Burnier, Y., and Stasiak, A. (2013). Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic acids research*, 42(5):2848–2855.

- [Berezney et al., 2000] Berezney, R., Dubey, D. D., and Huberman, J. A. (2000). Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma*, 108(8):471–484.
- [Berezney et al., 2005] Berezney, R., Malyavantham, K. S., Pliss, A., Bhattacharya, S., and Acharya, R. (2005). Spatio-temporal dynamics of genomic organization and function in the mammalian cell nucleus. *Advances in enzyme regulation*, 45(1):17–26.
- [Bickmore, 2013] Bickmore, W. A. (2013). The spatial organization of the human genome. *Annual Review of Genomics and Human Genetics*, 14(1):67–84. PMID: 23875797.
- [Bickmore and van Steensel, 2013] Bickmore, W. A. and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. *Cell*, 152.
- [Bohn and Heermann, 2010] Bohn, M. and Heermann, D. W. (2010). Diffusion-driven looping provides a consistent framework for chromatin organization. *PLoS one*, 5(8):e12218.
- [Bolzer et al., 2005] Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. R., et al. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5):e157.
- [Boyle et al., 2001] Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A., and Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum Mol Genet*, 10.
- [Branco and Pombo, 2007] Branco, M. R. and Pombo, A. (2007). Chromosome organization: new facts, new models. *Trends in cell biology*, 17(3):127–134.

- [Briand et al., 2018] Briand, N., Cahyani, I., Madsen-Østerbye, J., Paulsen, J., Rønningen, T., Sørensen, A. L., and Collas, P. (2018). Lamin a, chromatin and fp1d2: not just a peripheral ménage-à-trois. *Frontiers in cell and developmental biology*, 6.
- [Bridger et al., 2000] Bridger, J. M., Boyle, S., Kill, I. R., and Bickmore, W. A. (2000). Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts. *Curr Biol*, 10.
- [Bruinsma et al., 2014] Bruinsma, R., Grosberg, A. Y., Rabin, Y., and Zidovska, A. (2014). Chromatin hydrodynamics. *Biophysical journal*, 106(9):1871–1881.
- [Bystricky et al., 2004] Bystricky, K., Heun, P., Gehlen, L., Langowski, J., and Gasser, S. M. (2004). Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proceedings of the National Academy of Sciences*, 101(47):16495–16500.
- [Chiariello et al., 2015] Chiariello, A. M., Bianco, S., Piccolo, A., Annunziatella, C., Barbieri, M., Pombo, A., and Nicodemi, M. (2015). Polymer models of the organization of chromosomes in the nucleus of cells. *Modern Physics Letters B*, 29(09):1530003.
- [Chu et al., 2017] Chu, F.-Y., Haley, S. C., and Zidovska, A. (2017). On the origin of shape fluctuations of the cell nucleus. *Proceedings of the National Academy of Sciences*, 114(39):10338–10343.
- [Chubb et al., 2006] Chubb, J. R., Treck, T., Shenoy, S. M., and Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. *Current biology*, 16(10):1018–1025.
- [Consortium et al., 2012] Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57.

- [Cook and Marenduzzo, 2009a] Cook, P. R. and Marenduzzo, D. (2009a). Entropic organization of interphase chromosomes. *The Journal of cell biology*, 186(6):825–834.
- [Cook and Marenduzzo, 2009b] Cook, P. R. and Marenduzzo, D. (2009b). Entropic organization of interphase chromosomes. *The Journal of Cell Biology*, 186(6):825–834.
- [Cornish-Bowden, 1985] Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic acids research*, 13(9):3021.
- [Cremer and Cremer, 2001] Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, 2.
- [Cremer and Cremer, 2010] Cremer, T. and Cremer, M. (2010). Chromosome territories. *Cold Spring Harb Perspect Biol*, 2.
- [Cremer et al., 2018] Cremer, T., Cremer, M., and Cremer, C. (2018). The 4d nucleome: Genome compartmentalization in an evolutionary context. *Biochemistry (Moscow)*, 83(4):313–325.
- [Croft et al., 1999] Croft, J. A., Bridger, J. M., Boyle, S., Perry, P., Teague, P., and Bickmore, W. A. (1999). Differences in the localization and morphology of chromosomes in the human nucleus. *J Cell Biol*, 145.
- [Crooks et al., 2004] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190.
- [Dabin et al., 2016] Dabin, J., Fortuny, A., and Polo, S. E. (2016). Epigenome maintenance in response to dna damage. *Molecular cell*, 62(5):712–727.

- [Dai et al., 2015] Dai, Z., Guo, D., Dai, X., and Xiong, Y. (2015). Genome-wide analysis of transcription factor binding sites and their characteristic dna structures. In *BMC genomics*, volume 16, page S8. BioMed Central.
- [Dans et al., 2016] Dans, P. D., Walther, J., Gómez, H., and Orozco, M. (2016). Multiscale simulation of dna. *Current opinion in structural biology*, 37:29–45.
- [Darrow et al., 2016] Darrow, E. M., Huntley, M. H., Dudchenko, O., Stamenova, E. K., Durand, N. C., Sun, Z., Huang, S.-C., Sanborn, A. L., Machol, I., Shamim, M., et al. (2016). Deletion of dxz4 on the human inactive x chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31):E4504–E4512.
- [de Lanerolle, 2012] de Lanerolle, P. (2012). Nuclear actin and myosins at a glance. *Journal of Cell Science*, 125(21):4945–4949.
- [Dekker, 2014] Dekker, J. (2014). Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture. *Epigenetics Chromatin.*, 7.
- [Dekker et al., 2013] Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, 14.
- [Dekker and Mirny, 2013] Dekker, J. and Mirny, L. (2013). Biological techniques: chromosomes captured one by one. *Nature*, 502(7469):45–46.
- [Di Pierro et al., 2017] Di Pierro, M., Cheng, R. R., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, page 201714980.

- [Di Pierro et al., 2018] Di Pierro, M., Potoyan, D. A., Wolynes, P. G., and Onuchic, J. N. (2018). Anomalous diffusion, spatial coherence, and viscoelasticity from the energy landscape of human chromosomes. *Proceedings of the National Academy of Sciences*.
- [Di Pierro et al., 2016] Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2016). Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences*, 113(43):12168–12173.
- [Dixon et al., 2016] Dixon, J. R., Gorkin, D. U., and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. *Molecular cell*, 62(5):668–680.
- [Dixon et al., 2012] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485.
- [Dorier and Stasiak, 2010] Dorier, J. and Stasiak, A. (2010). The role of transcription factories-mediated interchromosomal contacts in the organization of nuclear architecture. *Nucleic acids research*, 38(21):7410–7421.
- [Doyle et al., 2014] Doyle, B., Fudenberg, G., Imakaev, M., and Mirny, L. A. (2014). Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS computational biology*, 10(10):e1003867.
- [Duan et al., 2010] Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., and Lee, C. (2010). A three-dimensional model of the yeast genome. *Nature*, 465.
- [Dyer et al., 1989] Dyer, K., Canfield, T., and Gartler, S. (1989). Molecular cytological differentiation of active from inactive x domains in interphase: implications for x chromosome inactivation. *Cytogenetic and Genome Research*, 50(2-3):116–120.

- [Eils et al., 1996] Eils, R., Dietzel, S., Bertin, E., Schröck, E., Speicher, M. R., Ried, T., Robert-Nicoud, M., Cremer, C., and Cremer, T. (1996). Three-dimensional reconstruction of painted human interphase chromosomes: active and inactive x chromosome territories have similar volumes but differ in shape and surface structure. *The Journal of cell biology*, 135(6):1427–1440.
- [Essers et al., 2005] Essers, J., van Cappellen, W. A., Theil, A. F., van Drunen, E., Jaspers, N. G., Hoeijmakers, J. H., Wyman, C., Vermeulen, W., and Kanaar, R. (2005). Dynamics of relative chromosome position during the cell cycle. *Molecular biology of the cell*, 16(2):769–775.
- [Farré and Emberly, 2018] Farré, P. and Emberly, E. (2018). A maximum-entropy model for predicting chromatin contacts. *PLoS computational biology*, 14(2):e1005956.
- [Fedorova and Zink, 2009] Fedorova, E. and Zink, D. (2009). Nuclear genome organization: common themes and individual patterns. *Current opinion in genetics & development*, 19(2):166–171.
- [Flaus and Owen-Hughes, 2011] Flaus, A. and Owen-Hughes, T. (2011). Mechanisms for atp-dependent chromatin remodelling: the means to the end. *The FEBS journal*, 278(19):3579–3595.
- [Flyamer et al., 2017] Flyamer, I. M., Gassler, J., Imakaev, M., Brandão, H. B., Ulianov, S. V., Abdennur, N., Razin, S. V., Mirny, L. A., and Tachibana-Konwalski, K. (2017). Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*.
- [Fodor et al., 2015] Fodor, É., Guo, M., Gov, N., Visco, P., Weitz, D., and van Wijland, F. (2015). Activity-driven fluctuations in living cells. *EPL (Europhysics Letters)*, 110(4):48005.

- [Fraser et al., 2015] Fraser, J., Williamson, I., Bickmore, W. A., and Dostie, J. (2015). An overview of genome organization and how we got there: from fish to hi-c. *Microbiology and Molecular Biology Reviews*, 79(3):347–372.
- [Fraser and Bickmore, 2007] Fraser, P. and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447.
- [Fritz et al., 2016] 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., and Stein, G. S. (2016). Chromosomes at work: organization of chromosome territories in the interphase nucleus. *Journal of cellular biochemistry*, 117(1):9–19.
- [Fudenberg et al., 2016] Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell reports*, 15(9):2038–2049.
- [Fullwood et al., 2009] Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., et al. (2009). An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 462(7269):58.
- [Ganai et al., 2014] Ganai, N., Sengupta, S., and Menon, G. I. (2014). Chromosome positioning from activity-based segregation. *Nucleic Acids Research*, 42(7):4145–4159.
- [Ghosh and Jost, 2017] Ghosh, S. K. and Jost, D. (2017). How epigenome drives chromatin folding and dynamics, insights from efficient coarse-grained models of chromosomes. *bioRxiv*, page 200584.
- [Gibcus and Dekker, 2013] Gibcus, J. and Dekker, J. (2013). The hierarchy of the 3d genome. *Molecular Cell*, 49(5):773 – 782.
- [Giorgetti et al., 2014] Giorgetti, L., Galupa, R., Nora, E., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive polymer modeling re-

veals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950 – 963.

[Giorgetti and Heard, 2016] Giorgetti, L. and Heard, E. (2016). Closing the loop: 3c versus dna fish. *Genome biology*, 17(1):215.

[Grant et al., 2011] Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

[Guelen et al., 2008] Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., and Talhout, W. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453.

[Haddad et al., 2017] Haddad, N., Jost, D., and Vaillant, C. (2017). Perspectives: using polymer modeling to understand the formation and function of nuclear compartments. *Chromosome Research*, 25(1):35–50.

[Halverson et al., 2014] Halverson, J. D., Smrek, J., Kremer, K., and Grosberg, A. Y. (2014). From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Reports on Progress in Physics*, 77(2):022601.

[Hameed et al., 2012] Hameed, F. M., Rao, M., and Shivashankar, G. (2012). Dynamics of passive and active particles in the cell nucleus. *PLoS One*, 7(10):e45843.

[Hargreaves and Crabtree, 2011] Hargreaves, D. C. and Crabtree, G. R. (2011). Atp-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell research*, 21(3):396.

[Harrow et al., 2012] Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774.

- [Heermann et al., 2012] Heermann, D. W., Jerabek, H., Liu, L., and Li, Y. (2012). A model for the 3d chromatin architecture of pro and eukaryotes. *Methods*, 58(3):307–314.
- [Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.
- [Hubisz et al., 2010] Hubisz, M. J., Pollard, K. S., and Siepel, A. (2010). Phast and rphast: phylogenetic analysis with space/time models. *Briefings in bioinformatics*, 12(1):41–51.
- [Hume et al., 2014] Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L. (2014). Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein–dna interactions. *Nucleic acids research*, 43(D1):D117–D122.
- [Hunt and Wasserman, 2014] Hunt, R. W. and Wasserman, W. W. (2014). Non-targeted transcription factors motifs are a systemic component of chip-seq datasets. *Genome biology*, 15(7):412.
- [Imakaev et al., 2015] Imakaev, M. V., Fudenberg, G., and Mirny, L. A. (2015). Modeling chromosomes: Beyond pretty pictures. *FEBS Letters*, 589(20):3031 – 3036. 3D Genome structure.
- [Ioannou et al., 2015] Ioannou, D., Kandukuri, L., Quadri, A., Becerra, V., Simpson, J. L., and Tempest, H. G. (2015). Spatial positioning of all 24 chromosomes in the lymphocytes of six subjects: evidence of reproducible positioning and spatial repositioning following dna damage with hydrogen peroxide and ultraviolet b. *PloS one*, 10(3):e0118886.
- [Jackson and Pombo, 1998] Jackson, D. A. and Pombo, A. (1998). Replicon clusters are stable units of chromosome structure: evidence that nuclear organization

- contributes to the efficient activation and propagation of s phase in human cells. *The Journal of cell biology*, 140(6):1285–1295.
- [Jayaram et al., 2016] Jayaram, N., Usvyat, D., and Martin, A. C. (2016). Evaluating tools for transcription factor binding site prediction. *BMC bioinformatics*, page 1.
- [Jégu et al., 2017] Jégu, T., Aeby, E., and Lee, J. T. (2017). The x chromosome in space. *Nature Reviews Genetics*, 18(6):377–389.
- [Jerabek and Heermann, 2012] Jerabek, H. and Heermann, D. W. (2012). Expression-dependent folding of interphase chromatin. *PloS one*, 7(5):e37525.
- [Johnson et al., 2007] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502.
- [Jolma et al., 2013] Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339.
- [Jost et al., 2014] Jost, D., Carrivain, P., Cavalli, G., and Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic acids research*, 42(15):9553–9561.
- [Jost et al., 2017] Jost, D., Vaillant, C., and Meister, P. (2017). Coupling 1d modifications and 3d nuclear organization: data, models and function. *Current opinion in cell biology*, 44:20–27.
- [Junier et al., 2010] Junier, I., Martin, O., and Képès, F. (2010). Spatial and topological organization of dna chains induced by gene co-localization. *PLoS computational biology*, 6(2):e1000678.

- [Junier et al., 2015] Junier, I., Spill, Y. G., Marti-Renom, M. A., Beato, M., and le Dily, F. (2015). On the demultiplexing of chromosome capture conformation data. *FEBS Letters*, 589(20):3005 – 3013. 3D Genome structure.
- [Kalhor et al., 2011] Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, 30.
- [Kalmárová et al., 2007] Kalmárová, M., Smirnov, E., Mašata, M., Koberna, K., Ligasová, A., Popov, A., and Raška, I. (2007). Positioning of nors and nor-bearing chromosomes in relation to nucleoli. *Journal of structural biology*, 160(1):49–56.
- [Kang et al., 2015] Kang, H., Yoon, Y.-G., Thirumalai, D., and Hyeon, C. (2015). Confinement-induced glassy dynamics in a model for chromosome organization. *Physical review letters*, 115(19):198102.
- [Karolchik et al., 2003] Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., et al. (2003). The ucsc genome browser database. *Nucleic acids research*, 31(1):51–54.
- [Khalil et al., 2007] Khalil, A., Grant, J., Caddle, L., Atzema, E., Mills, K., and Arnéodo, A. (2007). Chromosome territories have a highly nonspherical morphology and nonrandom positioning. *Chromosome research*, 15(7):899–916.
- [Khan et al., 2017] Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S. R., Tan, G., et al. (2017). Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, 46(D1):D260–D266.
- [Kölbl et al., 2012] Kölbl, A. C., Weigl, D., Mulaw, M., Thormeyer, T., Bohlander, S. K., Cremer, T., and Dietzel, S. (2012). The radial nuclear positioning of

genes correlates with features of megabase-sized chromatin domains. *Chromosome Research*, 20(6):735–752.

[Kreth et al., 2004] Kreth, G., Finsterle, J., Von Hase, J., Cremer, M., and Cremer, C. (2004). Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. *Biophysical journal*, 86(5):2803–2812.

[Kruhlak et al., 2006] Kruhlak, M. J., Celeste, A., Dellaire, G., Fernandez-Capetillo, O., Müller, W. G., McNally, J. G., Bazett-Jones, D. P., and Nussenzweig, A. (2006). Changes in chromatin structure and mobility in living cells at sites of dna double-strand breaks. *J Cell Biol*, 172(6):823–834.

[Kulakovskiy et al., 2012] Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J. (2012). Ho-comoco: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, 41(D1):D195–D202.

[Küpper et al., 2007] Küpper, K., Kölbl, A., Biener, D., Dittrich, S., von Hase, J., Thormeyer, T., Fiegler, H., Carter, N. P., Speicher, M. R., Cremer, T., et al. (2007). Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma*, 116(3):285–306.

[Lambert et al., 2018] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4):650–665.

[Landt et al., 2012] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., et al. (2012). Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831.

- [Lawrence et al., 1993] Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *science*, 262(5131):208–214.
- [Lee and Young, 2013] Lee, T. I. and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251.
- [Lengauer et al., 1998] Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature*, 396(6712):643.
- [Li et al., 2010] Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., Ariyaratne, P. N., Mohamed, Y. B., Ooi, H.-S., Tennakoon, C., et al. (2010). Chia-pet tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology*, 11(2):R22.
- [Lieberman-Aiden et al., 2009] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragozy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326.
- [Lihu and Holban, 2015] Lihu, A. and Holban, Ş. (2015). A review of ensemble methods for de novo motif discovery in chip-seq data. *Briefings in bioinformatics*, 16(6):964–973.
- [Loi et al., 2011] Loi, D., Mossa, S., and Cugliandolo, L. F. (2011). Non-conservative forces and effective temperatures in active polymers. *Soft Matter*, 7:10193–10209.
- [Louis et al., 2000] Louis, A., Bolhuis, P., and Hansen, J. (2000). Mean-field fluid behavior of the gaussian core model. *Physical Review E*, 62(6):7961.

- [Maeshima et al., 2016] Maeshima, K., Ide, S., Hibino, K., and Sasai, M. (2016). Liquid-like behavior of chromatin. *Current opinion in genetics & development*, 37:36–45.
- [Maharana et al., 2016] Maharana, S., Iyer, K. V., Jain, N., Nagarajan, M., Wang, Y., and Shivashankar, G. (2016). Chromosome intermingling the physical basis of chromosome organization in differentiated cells. *Nucleic acids research*, 44(11):5148–5160.
- [Malyavantham et al., 2008] Malyavantham, K. S., Bhattacharya, S., Alonso, W. D., Acharya, R., and Berezney, R. (2008). Spatio-temporal dynamics of replication and transcription sites in the mammalian cell nucleus. *Chromosoma*, 117(6):553–567.
- [Marella et al., 2009] Marella, N. V., Bhattacharya, S., Mukherjee, L., Xu, J., and Berezney, R. (2009). Cell type specific chromosome territory organization in the interphase nucleus of normal and cancer cells. *Journal of Cellular Physiology*, 221(1):130–138.
- [Marti-Renom and Mirny, 2011] Marti-Renom, M. A. and Mirny, L. A. (2011). Bridging the resolution gap in structural modeling of 3d genome organization. *PLoS Comput Biol*, 7.
- [Maston et al., 2006] Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59.
- [Mateos-Langerak et al., 2009] Mateos-Langerak, J., Bohn, M., de Leeuw, W., Giro-mus, O., Manders, E. M., Verschure, P. J., Indemans, M. H., Gierman, H. J., Heermann, D. W., Van Driel, R., et al. (2009). Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences*, 106(10):3812–3817.

- [Mathelier et al., 2016] Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 44(D1):D110–D115.
- [Mattout et al., 2015] Mattout, A., Cabianca, D. S., and Gasser, S. M. (2015). Chromatin states and nuclear organization in development a view from the nuclear lamina. *Genome Biology*, 16(1):1–15.
- [Matys et al., 2006] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenov, D., Krull, M., Hornischer, K., et al. (2006). Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl_1):D108–D110.
- [Mchaourab et al., 2018] Mchaourab, Z. F., Perreault, A. A., and Venters, B. J. (2018). Chip-seq and chip-exo profiling of pol ii, h2a. z, and h3k4me3 in human k562 cells. *Scientific data*, 5:180030.
- [Meaburn and Misteli, 2007] Meaburn, K. J. and Misteli, T. (2007). Cell biology: chromosome territories. *Nature*, 445.
- [Mehta et al., 2013] Mehta, I. S., Kulashreshtha, M., Chakraborty, S., Kolthur-Seetharam, U., and Rao, B. J. (2013). Chromosome territories reposition during dna damage-repair response. *Genome biology*, 14(12):R135.
- [Menon, 2010] Menon, G. I. (2010). Active matter. In *Rheology of complex Fluids*, pages 193–218. Springer.
- [Meshorer and Misteli, 2006] Meshorer, E. and Misteli, T. (2006). Chromatin in pluripotent embryonic stem cells and differentiation. *Nature reviews Molecular cell biology*, 7(7):540.

- [Millett et al., 2009] Millett, K. C., Plunkett, P., Piatek, M., Rawdon, E. J., and Stasiak, A. (2009). Effect of knotting on polymer shapes and their enveloping ellipsoids. *The Journal of chemical physics*, 130(16):04B623.
- [Mirny, 2011] Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res*, 19.
- [Mitra and Narlikar, 2015] Mitra, S. and Narlikar, L. (2015). No promoter left behind (nplb): learn de novo promoter architectures from genome-wide transcription start sites. *Bioinformatics*, 32(5):779–781.
- [Mukhopadhyay et al., 2011] Mukhopadhyay, S., Schedl, P., Studitsky, V. M., and Sengupta, A. M. (2011). Theoretical analysis of the role of chromatin interactions in long-range action of enhancers and insulators. *Proceedings of the National Academy of Sciences*, 108(50):19919–19924.
- [Murmamann et al., 2005] Murmann, A. E., Gao, J., Encinosa, M., Gautier, M., Peter, M. E., Eils, R., Lichter, P., and Rowley, J. D. (2005). Local gene density predicts the spatial position of genetic loci in the interphase nucleus. *Experimental cell research*, 311(1):14–26.
- [Nagano et al., 2013] Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., and Dean, W. (2013). Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502.
- [Narlikar, 2013] Narlikar, L. (2013). Mumod: a bayesian approach to detect multiple modes of proteindna binding from genome-wide chip data. *Nucleic Acids Research*, 41(1):21–32.
- [Narlikar, 2014] Narlikar, L. (2014). Multiple novel promoter-architectures revealed by decoding the hidden heterogeneity within the genome. *Nucleic acids research*, 42(20):12388–12403.

- [Naumova et al., 2013] Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., and Dekker, J. (2013). Organization of the mitotic chromosome. *Science*, 342.
- [Németh et al., 2010] Németh, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Pterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Lngst, G. (2010). Initial genomics of the human nucleolus. *PLOS Genetics*, 6(3):1–11.
- [Németh and Längst, 2011] Németh, A. and Längst, G. (2011). Genome organization in and around the nucleolus. *Trends in genetics*, 27(4):149–156.
- [Nora et al., 2012] Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., and Servant, N. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485.
- [Odenheimer et al., 2005] Odenheimer, J., Kreth, G., and Heermann, D. W. (2005). Dynamic simulation of active/inactive chromatin domains. *Journal of biological physics*, 31(3):351–363.
- [Osborne et al., 2004] Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J. A., Lopes, S., Reik, W., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10):1065.
- [Otto et al., 2007] Otto, S. J., McCorkle, S. R., Hover, J., Conaco, C., Han, J.-J., Impey, S., Yochum, G. S., Dunn, J. J., Goodman, R. H., and Mandel, G. (2007). A new binding motif for the transcriptional repressor rest uncovers large gene networks devoted to neuronal functions. *Journal of Neuroscience*, 27(25):6729–6739.
- [Pajerowski et al., 2007] Pajerowski, J. D., Dahl, K. N., Zhong, F. L., Sammak, P. J., and Discher, D. E. (2007). Physical plasticity of the nucleus in stem cell

- differentiation. *Proceedings of the National Academy of Sciences*, 104(40):15619–15624.
- [Park, 2009] Park, P. J. (2009). Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669.
- [Plimpton et al., 2007] Plimpton, S., Crozier, P., and Thompson, A. (2007). Lammgs-large-scale atomic/molecular massively parallel simulator. *Sandia National Laboratories*, 18.
- [Pombo and Nicodemi, 2014] Pombo, A. and Nicodemi, M. (2014). Physical mechanisms behind the large scale features of chromatin organization. *Transcription*, 5(2):e28447. PMID: 25764220.
- [Prestipino et al., 2005] Prestipino, S., Saija, F., and Giaquinta, P. V. (2005). Phase diagram of the gaussian-core model. *Physical Review E*, 71(5):050102.
- [Ranade et al., 2017] Ranade, D., Koul, S., Thompson, J., Prasad, K. B., and Sengupta, K. (2017). Chromosomal aneuploidies induced upon lamin b2 depletion are mislocalized in the interphase nucleus. *Chromosoma*, 126(2):223–244.
- [Rao et al., 2014] Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., and Robinson, J. T. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 59.
- [Rawdon et al., 2008] Rawdon, E. J., Kern, J. C., Piatek, M., Plunkett, P., Stasiak, A., and Millett, K. C. (2008). Effect of knotting on the shape of polymers. *Macromolecules*, 41(21):8281–8287.
- [Razin and Gavrilov, 2018] Razin, S. and Gavrilov, A. (2018). Structural–functional domains of the eukaryotic genome. *Biochemistry (Moscow)*, 83(4):302–312.

- [Rhee and Pugh, 2011] Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.
- [Rosa and Everaers, 2008] Rosa, A. and Everaers, R. (2008). Structure and dynamics of interphase chromosomes. *PLoS computational biology*, 4(8):e1000153.
- [Rouillard et al., 2016] Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., and Maayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016.
- [Roukos and Misteli, 2014] Roukos, V. and Misteli, T. (2014). The biogenesis of chromosome translocations. *Nature cell biology*, 16(4):293.
- [Sanborn et al., 2015] 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., et al. (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.
- [Sandelin et al., 2004] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1):D91–D94.
- [Sati and Cavalli, 2017] Sati, S. and Cavalli, G. (2017). Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, 126(1):33–44.
- [Sazer and Schiessel, 2018] Sazer, S. and Schiessel, H. (2018). The biology and polymer physics underlying large-scale chromosome organization. *Traffic*, 19(2):87–104.

- [Schlick, 2009] Schlick, T. (2009). From macroscopic to mesoscopic models of chromatin folding. *Bridging The Scales in Science in Engineering*, pages 514–535.
- [Schlick, 2010] Schlick, T. (2010). *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*, volume 21. Springer Science & Business Media.
- [Schneider and Stephens, 1990] Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100.
- [Schwarzer et al., 2017] Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51.
- [Sehgal et al., 2014] Sehgal, N., Fritz, A. J., Morris, K., Torres, I., Chen, Z., Xu, J., and Berezney, R. (2014). Gene density and chromosome territory shape. *Chromosoma*, 123(5):499–513.
- [Sexton et al., 2012] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148.
- [Sha and Boyer, 2009] Sha, K. and Boyer, L. A. (2009). The chromatin signature of pluripotent cells. *Stem Book. Cambridge: Harvard Stem Cell Institute*.
- [Shi et al., 2018] Shi, G., Liu, L., Hyeon, C., and Thirumalai, D. (2018). Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nature communications*, 9(1):3161.

- [Siddharthan et al., 2005] Siddharthan, R., Siggia, E. D., and Van Nimwegen, E. (2005). Phylogibbs: a gibbs sampling motif finder that incorporates phylogeny. *PLoS computational biology*, 1(7):e67.
- [Sloan et al., 2016] Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Stratton, J. S., Hitz, B. C., Gabdank, I., Narayanan, A. K., Ho, M., Lee, B. T., et al. (2016). Encode data at the encode portal. *Nucleic acids research*, 44(D1):D726–D732.
- [Smith et al., 2016] Smith, E. M., Lajoie, B. R., Jain, G., and Dekker, J. (2016). Invariant tad boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the cftr locus. *The American Journal of Human Genetics*, 98(1):185–201.
- [Solovei et al., 2009] Solovei, I., Kreysing, M., Lanctt, C., Ksem, S., Peichl, L., Cremer, T., Guck, J., and Joffe, B. (2009). Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell*, 137(2):356 – 368.
- [Spielmann et al., 2018] Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural variation in the 3d genome. *Nature Reviews Genetics*, page 1.
- [Spitz and Furlong, 2012] Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9):613.
- [Stillinger, 1976] Stillinger, F. H. (1976). Phase transitions in the gaussian core system. *The Journal of Chemical Physics*, 65(10):3968–3974.
- [Stower, 2011] Stower, H. (2011). Gene regulation: Resolving transcription factor binding. *Nature Reviews Genetics*, 13(2):71.
- [Straub, 2003] Straub, T. (2003). Heterochromatin dynamics. *PLoS biology*, 1(1):e14.

- [Streubel and Bracken, 2015] Streubel, G. and Bracken, A. P. (2015). Med23: a new mediator of h2b monoubiquitylation. *The EMBO journal*, 34(23):2863–2864.
- [Sun et al., 2000] Sun, H. B., Shen, J., and Yokota, H. (2000). Size-dependent positioning of human chromosomes in interphase nuclei. *Biophysical journal*, 79(1):184–190.
- [Takizawa et al., 2008] Takizawa, T., Meaburn, K. J., and Misteli, T. (2008). The meaning of gene positioning. *Cell*, 135(1):9–13.
- [Talwar et al., 2013] Talwar, S., Kumar, A., Rao, M., Menon, G. I., and Shivashankar, G. (2013). Correlated spatio-temporal fluctuations in chromatin compaction states characterize stem cells. *Biophysical journal*, 104(3):553–564.
- [Tanabe et al., 2002] Tanabe, H., Muller, S., Neusser, M., von Hase, J., Calcagno, E., Cremer, M., Solovei, I., Cremer, C., and Cremer, T. (2002). Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc Natl Acad Sci USA*, 99.
- [Tark-Dame et al., 2014] Tark-Dame, M., Jerabek, H., Manders, E. M., Heermann, D. W., and van Driel, R. (2014). Depletion of the chromatin looping proteins ctfc and cohesin causes chromatin compaction: insight into chromatin folding by polymer modelling. *PLoS computational biology*, 10(10):e1003877.
- [Tark-Dame et al., 2011] Tark-Dame, M., van Driel, R., and Heermann, D. W. (2011). Chromatin folding—from biology to polymer models and back. *J Cell Sci*, 124(6):839–845.
- [Therizols et al., 2014] Therizols, P., Illingworth, R. S., Courilleau, C., Boyle, S., Wood, A. J., and Bickmore, W. A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science.*, 346.

- [Tiana et al., 2016] Tiana, G., Amitai, A., Pollex, T., Piolot, T., Holcman, D., Heard, E., and Giorgetti, L. (2016). Structural fluctuations of the chromatin fiber within topologically associating domains. *Biophysical journal*, 110(6):1234–1245.
- [Tjong et al., 2016] Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., Le Gros, M. A., et al. (2016). Population-based 3d genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences*, 113(12):E1663–E1672.
- [Todd and Yildirim, 2007] Todd, M. J. and Yildirim, E. A. (2007). On khachiyan’s algorithm for the computation of minimum-volume enclosing ellipsoids. *Discrete Applied Mathematics*, 155(13):1731–1744.
- [Trapnell et al., 2010] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511.
- [Tuğrul et al., 2015] Tuğrul, M., Paixao, T., Barton, N. H., and Tkačik, G. (2015). Dynamics of transcription factor binding site evolution. *PLoS genetics*, 11(11):e1005639.
- [Uhler and Shivashankar, 2016] Uhler, C. and Shivashankar, G. (2016). Geometric control and modeling of genome reprogramming. *BioArchitecture*, 6(4):76–84. PMID: 27434579.
- [van Steensel and Belmont, 2017] van Steensel, B. and Belmont, A. S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5):780–791.
- [Vasquez and Bloom, 2014] Vasquez, P. A. and Bloom, K. (2014). Polymer models of interphase chromosomes. *Nucleus*, 5(5):376–390.

- [Vignali et al., 2000] Vignali, M., Hassan, A. H., Neely, K. E., and Workman, J. L. (2000). Atp-dependent chromatin-remodeling complexes. *Molecular and cellular biology*, 20(6):1899–1910.
- [Wang et al., 2012] Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9):1798–1812.
- [Wang et al., 2013] Wang, J., Zhuang, J., Iyer, S., Lin, X.-Y., Greven, M. C., Kim, B.-H., Moore, J., Pierce, B. G., Dong, X., Virgil, D., et al. (2013). Factorbook.org: a wiki-based database for transcription factor-binding data generated by the encode consortium. *Nucleic acids research*, 41(D1):D171–D176.
- [Wang et al., 2016] Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., Wu, C.-t., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602.
- [Wang and Wolynes, 2011] Wang, S. and Wolynes, P. G. (2011). Communication: Effective temperature and glassy dynamics of active matter.
- [Wang et al., 2017] Wang, Y., Nagarajan, M., Uhler, C., and Shivashankar, G. (2017). Orientation and repositioning of chromosomes correlate with cell geometry-dependent gene expression. *Molecular Biology of the Cell*, 28:1997–2009.
- [Weber et al., 2015] Weber, J. K., Shukla, D., and Pande, V. S. (2015). Heat dissipation guides activation in signaling proteins. *Proceedings of the National Academy of Sciences*, 112(33):10377–10382.
- [Weber et al., 2012] Weber, S. C., Spakowitz, A. J., and Theriot, J. A. (2012). Non-thermal atp-dependent fluctuations contribute to the in vivo motion of chromosomal loci. *Proceedings of the National Academy of Sciences*, 109(19):7338–7343.

- [Weirauch et al., 2014] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443.
- [Williamson et al., 2014] Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R. S., Paquette, D., Dostie, J., and Bickmore, W. A. (2014). Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes & development*, 28(24):2778–2791.
- [Wu et al., 2007] Wu, C., Bassett, A., and Travers, A. (2007). A variable topology for the 30-nm chromatin fibre. *EMBO reports*, 8(12):1129–1134.
- [Zambelli et al., 2012] Zambelli, F., Pesole, G., and Pavesi, G. (2012). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, 14(2):225–237.
- [Zambelli et al., 2014] Zambelli, F., Pesole, G., and Pavesi, G. (2014). Using weeder, pscan, and pscanchip for the discovery of enriched transcription factor binding site motifs in nucleotide sequences. *Current protocols in bioinformatics*, 47(1):2–11.
- [Zhang and Wolynes, 2015] Zhang, B. and Wolynes, P. G. (2015). Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19):6062–6067.
- [Zhang and Wolynes, 2017] Zhang, B. and Wolynes, P. G. (2017). Genomic energy landscapes. *Biophysical journal*, 112(3):427–433.
- [Zhu et al., 2018] Zhu, G., Deng, W., Hu, H., Ma, R., Zhang, S., Yang, J., Peng, J., Kaplan, T., and Zeng, J. (2018). Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic acids research*, 46(8):e50–e50.

- [Zidovska et al., 2013] Zidovska, A., Weitz, D. A., and Mitchison, T. J. (2013). Micron-scale coherence in interphase chromatin dynamics. *Proceedings of the National Academy of Sciences*, 110(39):15555–15560.
- [Žunić and Žunić, 2013] Žunić, D. and Žunić, J. (2013). Shape ellipticity based on the first hu moment invariant. *Information Processing Letters*, 113(19):807–810.
- [Zwanzig, 2001] Zwanzig, R. (2001). *Nonequilibrium statistical mechanics*. Oxford University Press.