Probing structural organisation of chromatin during B cell development through mechanistic modelling

A thesis submitted for the degree of

Doctor of Philosophy

By

Anubhooti

(Reg. No. 13LAPH19)

Department of Animal Biology
School of Life Sciences
University of Hyderabad
Hyderabad - 500 046
India

December 2022

University of Hyderabad
School of Life Sciences
Hyderabad -500046, India

Jagan Pongubala, Ph.D.
Department of Animal Biology

Tel. No. 040-23134580
Email: jpsl@uohyd.ernet.in

Certificate

This is to certify that this thesis entitled **"Probing structural organisation of chromatin during B cell development through mechanistic modelling"** submitted by **Ms. Anubhooti** bearing registration number **13LAPH19** is a *bonafide* work carried out by her under my supervision and guidance and thus can be submitted in partial fulfilment of the requirement for the award of Doctor of Philosophy in the Department of Animal Biology, School of Life Sciences. I attest that her thesis is free from plagiarism and has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

Part of this thesis has been presented in the following conferences:

- **Anubhooti**, Wasim Abdul, Jagannath Mondal and Jagan M.R. Pongubala. 3D modelling of Chromatin Determines Dynamic Spatial & Structural Changes During B Lymphoid Differentiation. Presented Poster at International Conference on Virus Evolution, Infection and Disease Control-2022, December 15th-17th 2022, School of Life Sciences, University of Hyderabad, Hyderabad.
- **Anubhooti**, Wasim Abdul, Jagannath Mondal and Jagan M.R. Pongubala. 3D modelling of Chromatin Determines Dynamic Structural Changes During B Lymphoid Differentiation. Presented Poster at 48th Annual Conference of Indian Immunology Society, Infections, Vaccines & Immuno-Innovations for Human Health, July 8th-9th 2022 (Virtual), Department of Molecular and Human Genetics, Banaras Hindu University, Varanasi.
- **Anubhooti**, Anurupa Devi Yadavalli and Jagan M.R. Pongubala. 3D modelling of Chromatin Structure During B Lymphoid Differentiation. Presented Poster at 6th Meeting of the Asian Forum of Chromosome and Chromatin Biology, March 3rd-5th 2017, CSIR-Centre for Cellular and Molecular Biology, Hyderabad
- **Anubhooti**, Anurupa Devi Yadavalli and Jagan M.R. Pongubala. 3D modelling of Chromatin Structure During B Lymphoid Differentiation. Presented Poster at BioQuest-2015, September 23rd-24th 2015, School of Life Sciences, University of Hyderabad, Hyderabad

Under preparation:

- **Anubhooti**, Wasim A, Mondal J and Pongubala JMR. Discovery of dynamic changes in chromatin structure through polymer physics model.

Further, the student has passed the following courses towards fulfilment of the coursework requirement for Ph.D.

Course code	Name	Credits	Pass/Fail
AS 801	Analytical Techniques	4	Pass
AS 802	Research ethics, Data analysis and Biostatistics	3	Pass
AS 803	Lab seminar and Records	5	Pass

JMR 27/12/2022
Supervisor

Dr. Jagan Pongubala
Professor
Department of Animal Biology
School of Life Sciences
University of Hyderabad
P.O. Central University, Gachibowli
Hyderabad-500 046

K. Sreenivasan
अध्यक्ष / HEAD
जंतु जैविकी विभाग
Department of Animal Biology
Head
Dept. of Animal Biology

I/c Anuradha 28/12/22
Dean
School of Life Sciences

DEAN
School of Life Sciences
University of Hyderabad
Hyderabad-500 046.

University of Hyderabad
Hyderabad -500046, India

Declaration

I, **Anubhooti (13LAPH19)**, hereby declare that this thesis entitled **"Probing structural organisation of chromatin during B cell development through mechanistic modelling"** submitted by me under the guidance and supervision of **Prof. Jagan Pongubala** is an original and independent research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

Name: Anubhooti
Reg. No.: 13LAPH19

Date: 27.12.2022

Signature of the Student

Signature of the Supervisor

Dr. Jagan Pongubala
Professor
Department of Animal Biology
School of Life Sciences
University of Hyderabad
P.O. Central University, Gachibowli,
Hyderabad-500 046. T.S. INDIA.

Acknowledgements

I extend heartfelt thanks to my supervisor Prof. Jagan Pongubala for providing an opportunity to work on this topic and rendering necessary help, guidance and support to complete my Ph.D.

A word of special gratitude to my collaborator Dr. Jagannath Mondal from TIFR-H for his keen interest, insightful knowledge & expertise to steer the project and for his unwavering encouragement, motivation and eternal willingness to help me overcome all the hurdles during my entire collaboration with him. I also remain indebted to Mr. Wasim Abdul, Research Scholar, TIFR-H for our scientific discussions, his help and suggestions in computing & programming related issues which were forthcoming at any time of the day.

I convey sincere thanks to my doctoral committee members Prof. Anita Jagota and Dr. Nooruddin Khan for their patronage, encouragement and feedback.

I am thankful to Prof. Srinivasulu Kurukuti, Head, Department of Animal Biology, Prof. Y. Suresh, officiating Head of Department and the former Heads, Prof. Anita Jagota, Prof. Jagan Pongubala and Prof. B Senthilkumaran for their support and in allowing me to use departmental facilities.

I extend thanks to Prof. N. Siva Kumar, Dean, School of Life Sciences and the former Deans, Prof. S. Dayananda, Prof. P. Reddanna and Prof. K.V. Ramaiah for facilitating and making available all necessary School resources for my research work.

I offer my sincerest gratitude to our Vice-Chancellor Prof. B.J. Rao for his insightful advices, guidance and immense support so that I could finish my course on time.

I would like to thank all the previous and current non-teaching staff of Dept. of Animal Biology: Mrs. Vijaya Lakhsmi, Mr. Nikhil, Mr. Jagan, Mr. Gopi, Mr. Rangaswamy and Mr. Sreenu for their help in office and

administrative work.

I sincerely thank DBT-BINC for funding my doctoral studies and DST and DBT for providing financial assistance to the lab and the infrastructural facility to the School.

I am also thankful to Late Prof. Surajit Sengupta from TIFR-H for his guidance and everlasting concern during my initial days of collaboration with him.

I am thankful to Prof. Krishnamoorthy Kannan (Proprietor KK Biotech, former Vice Chancellor in Nagaland University) and Dr. Dinesh Kumar, Professor Biotechnology and Dean Academic, Central University of Haryana for their constant support and encouragement, continual concern and for recognizing my potential and motivating me to believe in myself during my toughest times in this journey.

I would like to thank my former lab members, Dr. Ravi Boya, Dr. Sameena Nikhat and Dr. Anurupa Devi Yadavalli for scientific discussions and helpful suggestions.

Special thanks to my very dear friends, Ms. Priyanka Kriti Narayan and Ms. Arpita Prusty, for the constant succour & encouragement received on all fronts and professional help to conclude my research goals. I am thankful to both of them for always believing in me, helping and pushing me up to the next higher pedestal. They both have always been there for me through thick and thin and provided me the strength in difficult times. I am also thankful to my friend Dr. Ranay Yadav for always being there and helping me and my son in any possible way. It would not have been possible to undertake this journey without the support system provided by the three of them. I am also grateful to Anant Ujagare for his support & concern at all times. Thanks is due to other members of my lab, Mr. Ashok Kumar, Ms. Nidhi Singh, Ms. Shravya Kurella and Ms. Apoorva Soni for their help and support.

I would like to extend my heartfelt thanks to Dr. Gaurav Bajpai, Postdoc, Weizmann Institute of Science, Israel for his ever enthusiastic nature to help me understand even a miniscule conceptual doubt. He had been a

great source of support, encouragement and patience who always motivated me to learn. He is the most friendly and helping person who would give his valuable time immediately for our innumerous insightful scientific discussions.

Special thanks to Dr. Pankaj Popli, Dr. Sneha Menon, Dr. Bhupendra Dandekar, Dr. Parswa Nath, Susmita Sarkar and the entire research team of Dr. Jagannath Mondal who constantly encouraged me to chase my research goals and helped me stay on track.

I would like to thank the day care facility provided by TIFR-H for my son, without whose cooperation it would have been impossible to carry out the research and successfully achieve my research goals.

I would like to extend thanks to my dear friends and erstwhile batchmates at IIIT-Hyderabad, Ms. Farheen Farzana, Dr. Navneet Bung, Dr. Koushik Kasavjhala, Ms. Pragya Saxena, Mr. Varun Raheja and Mr. Aditya Jitta from University of Helsinki, for endorsing me at all times in pursuit of my goals. The regular online sessions with all of them really pushed and motivated me to reach the finish line.

My family members have been my pillars of strength with their unflinching faith, patience and enormous encouragement. No words can suffice the painstaking endeavours and sacrifices made by my father Mr. Shekhar Srivastava (Ex Group General Manager, ONGC), my mother Mrs. Kranti Srivastava, my brother Siddharth (Senior Engineer at Spotify, London) and my beloved son Aarav in this entire journey, for which I shall ever remain indebted to each one of them. I take this opportunity to thank my parents for providing their unconditional support in every way possible, guiding me whenever I was in dilemma & to help me achieve what I am today. I thank my father for engaging with me even during the process of thesis writing. I am thankful to my mother for being so understanding & patient with me, for her endless sacrifices, care & affection and for being the person whom I could blindly fall back on for anything. I am thankful to my brother for standing by my side, supporting & guiding me during tough times. I thank him for extending his help in Python code and teaching me LaTeX. I remain thankful to my aunt Mrs. Kumkum

Sinha for providing the support to me and my son during our independent stay at Hyderabad. I am thankful to my beloved son for being my source of fulfilment & pure happiness, for going above & beyond his age, for pouring me with his unconditional love and bearing me on tough days with his innocent belief in his mother that helped me to start afresh after every failure. I will never be able to compensate for the times when he was always deprived of my attention for his entire childhood due to my research commitments.

Above all, I am thankful to the God Almighty for every opportunity, for every learning experience and for showing me the right perspective in unfavourable situations.

Contents

List of Figures

List of Tables

List of Abbreviations

3C Chromosome Conformation Capture

3D Three Dimensional

BACH Bayesian 3D constructor for Hi-C data

BACH-MIX Variant algorithm for structural variants in BACH

CG Coarse-Grained

ChIA-PET Chromosome Interaction Analysis with Paired-End Tag sequencing

ChIP-seq Chromatin Immuno Precipitation and Sequencing

CLP Common Lymphoid Progenitor

CMP Common Myeloid Progenitor

COM Center of Mass

CTCF CCCTC-binding Factor

DamID DNA Adenine Methyltransferase Identification

DL Dynamic Loop

EM Electron Microscopy

FISH Fluorescence *in situ* Hybridization

GAM Genome Architecture Mapping

GL Giant Loop

GMP Granulocyte-Macrophage Progenitor

Hi-C High-throughput chromosome conformation capture

HSC Haematopoietic Stem Cell

LAD Lamina Associated Domains

LMPP Lymphoid-primed Multi-Potent Progenitor

LSK Lin$^-$Sca-1$^+$c-Kit$^+$

LT-HSC Long Term-Haematopoietic Stem Cell

MCMC5C Markov Chain Monte Carlo 5C

MD Molecular Dynamics

MEP Megakaryocyte Erythrocyte Progenitor

MLS Multi-Loop/Subcompartment

MPP Multi-Potent Progenitor

NK-cells Natural Killer cells

PCA Principal Component Analysis

RBC Red Blood Cell

Rg Radius of Gyration

RL Random Loop

RT-PCR Real Time-Polymerase Chain Reaction

RW Random Walk

SAW Self-Avoiding Walk

SBS Strings and Binder Switch

scHi-C single-cell Hi-C

SPRITE Split-Pool Recognition of Interactions by Tag Extension

ST-HSC Short Term-Haematopoietic Stem Cell

SXT Soft X-ray Tomography

TAD Topologically Associated Domains

TCC Tethered Conformation Capture

TF Transcription Factors

VMD Visual Molecular Dynamics

WLC Worm-Like Chain

Abstract

Understanding the 3D organisation of the genome provides insights into the intricate relationship between chromatin architecture and its effects on the functional state of the cell. In coordination with the combinatorial activity of *cis*-regulome and the transcription factors (TFs), the cell type specific gene expression is governed via the organisation of genome in 3D that also plays a dominant role in cell differentiation and varied cellular functions. However, the concerted dynamics of the mega-size genomic regions and *cis*-regulome is unclear. Although chromosome conformation capture techniques have been pivotal in understanding chromatin organisation inside the cell nucleus, they are limited to only a static 2D representation of it. In order to quantitatively understand the structural alterations and dynamics of chromatin in 3D, we have developed a computational model that not only captures the hierarchical structural organisation but also provides mechanistic insights into the spatial rearrangements of chromatin during developing lymphoid lineage cells. From the combination of approaches of polymer physics representing chromatin as a homopolymeric chain and incorporation of the biological information of chromosomal interactome inferred from the Hi-C data, we generated a coarse grained bead-on-a-string polymer model of chromatin to comprehend the characteristics underlying the differential chromatin architecture. Our study showed that our simulated chromatin structure not only recapitulates the intrinsic features of chromatin organisation, including the fractal globule nature, compartmentalization, presence of topologically associating domains (TADs), phase separation and spatial preferences of genomic regions in chromosomal territories; but is also able to capture cell type specific compartmental switching and changes in the spatial positioning of lineage specific genomic regions upon comparative analyses of these simulated chromatin structures in differentiating B cell stages. Analysis of the compactness of the switched regions showed insights into consequential structural rearrangements & acquired open or closed states for gene regulation for aiding interactions with the *cis*-regulome and TFs, thereby orchestrating the cell fate. Further, we emphasise on the predictive potential of our model by identifying genes that demonstrated undergoing structural rearrangement in our simulated structures which were subsequently validated through their differential expression patterns *in vitro*. From these results we were able to apprehend the distinct structural changes of chromatin for its regulatory role in sustaining cell specificity.

Chapter 1

Introduction

1.1 Why is studying 3D chromatin structural organisation important?

Our genome, together with the *cis*- and *trans*-regulatory elements, is responsible for expressing genes in a regulated manner across hundreds of different cell types. Though all the cells of an organism have identical genetic blueprint, yet it is specialised to perform remarkably diverse functions across cell types. This is principally because of the differential regulation & expression of genomic regions by the *cis*-regulome and various transcription factors (TFs) via their 3D organisation and positioning in the nuclear space. In essence, the structural arrangement of the genome facilitates interactions within distinct nuclear compartments and thus, plays a pivotal role in gene regulation, consequently governing the functional state and fate of the cell. The three-dimensional architecture of chromatin in space, therefore, represents a crucial link which maps the linear genomic information with its corresponding biological function [11]. Studying this causal relationship between genome function and its spatio-temporal organisation in the nucleus is, therefore, very exciting to understand a wide spread of cellular processes including differentiation, replication, repair, epigenetic modification and genomic stability etc. A comprehensive quantitative description elucidating in what way chromosomes fold & interact will provide further insights into the mechanism of cell functioning. It has been thought that 3D chromatin organisation, which is far from random, is governed by a set of principles that take into account the necessity of physically connecting remote functional regions of the genome, like regulatory elements, enhancers and promoters, in order to instigate specific transcriptional programmes [12]. Recent findings have really highlighted the importance of this structural order in controlling embryonic development and how its disruption may lead to human anomalies [13–16]. Unravelling the formation, maintenance or any perturbation of the chromatin spatial structure leading to consequential

effects in the downstream molecular processes would, therefore, certainly be pivotal in understanding common genetic diseases. Further, elaborative functional importance of 3D genome architecture is reviewed in [17–19].

1.2 The Chromatin Organization: from the 3D nucleus to the linear genome

To understand 3D chromatin organisation, it is important to first understand its components and their salient features. The eukaryotic genomes, specifically the mammalian genomes are folded inside a nucleus roughly five orders of magnitude smaller. For example, there are about 3.2 billion base pairs of nucleotides in the human genome, which is ∼2m in length when stretched out but is remarkably packed into a nucleus of diameter ∼10μm *in vivo*. Interestingly, the genome is nevertheless accessible to all essential cellular functions despite this tight folding. To comprehend how this exceptional folding is achieved in 3D, comprehensive understanding of the distinct levels of chromatin compaction and organisation is required.

Genome folding occurs at multiple scales ranging from whole chromosome structures to interactions across a few kilobases(kb), where each scale highlights an important interplay between structure and function [20]. This multi-scale organisation leads to a hierarchical architecture, where the histone-DNA interactions correspond to the smallest sub-nucleosomal scale, nucleosomes and regulatory loops to the nucleosomal scale, chromosome domains and compartments to the supra-nucleosomal scale or sub-chromosomal scale and finally chromosome territories to the largest nuclear scale as shown in Figure 1.1 and reviewed in [5, 6, 21, 22].

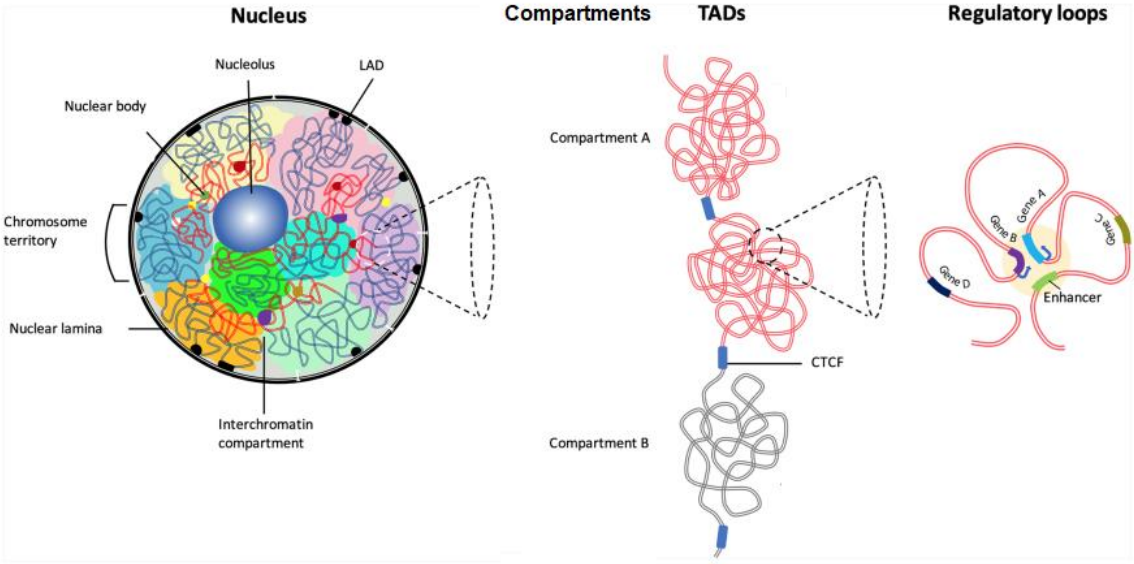


Figure 1.1: Levels of chromatin organisation. The multi-scale organisation leading to a hierarchical genome architecture with chromosome territory formation at the nuclear scale, formation of compartments at the chromosomal scale, TADs at sub-chromosomal scale and regulatory loops at sub-megabase scale. Adapted from [1]

1.2.1 Chromosome Territories

At the largest scale, all chromosomes position themselves in their specific domain or region known as 'chromosome territories' that results in minimal intermixing between individual chromosomes (leftmost image in Figure 1.1). Chromosome territories were first visualised and identified by FISH experiments utilising probes which specifically covered the entire chromosome [23] and with further advancement of chromosome conformation techniques, they were also observed in Hi-C maps [7] (as shown in Figure 1.2). Intuitively, chromosomes are expected to randomly intermingle and fill out the entire nuclear space due to the high orders of magnitude of compactness achieved by them, in order to fit in the much smaller dimensions of the nucleus. On the contrary, each chromosome occupies a non-random, largely non-overlapping space, defining its own territory inside the nucleus. Wherever these territories overlap, intermingling may take place, raising the possibility of functional connections across loci on dif-

ferent chromosomes [24]. However, it has been shown that such inter-chromosomal interactions (between different chromosomes) are quite less frequent than the intra-chromosomal interactions (on the same chromosome) [7]. The infrequent intermingling might also be a consequence of the crowding arising from the tight packaging of the chromosomes inside such a small volume of the nucleus. The positioning of these territories is shown to be cell specific and conserved between human and other primates, suggesting a functional role of specific chromosome organisation inside the nucleus [25–27].

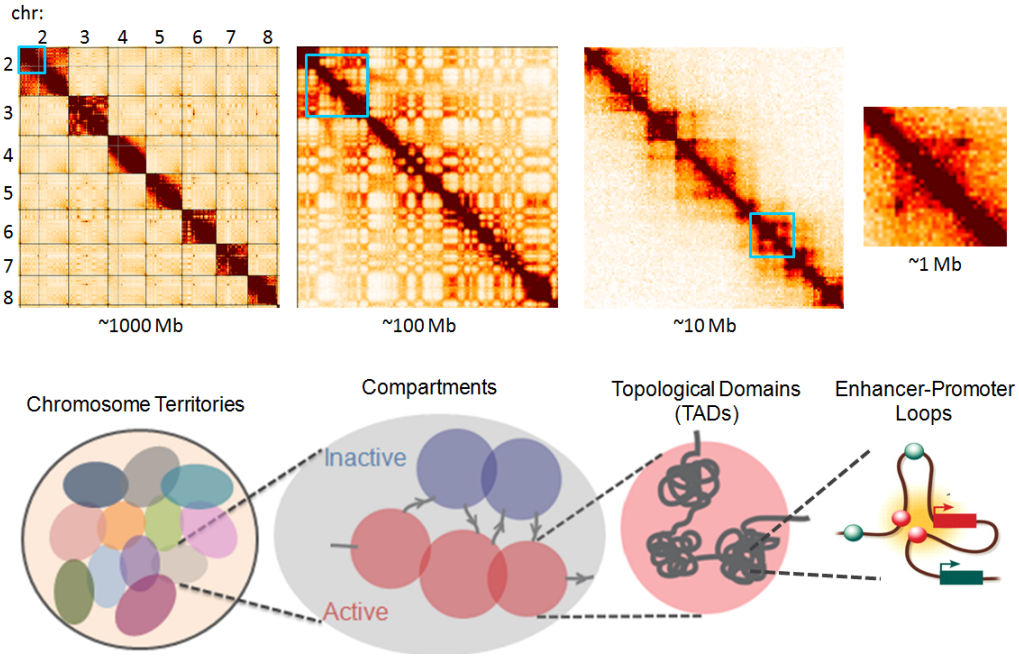


Figure 1.2: Hi-C represents chromatin organisation. Interpretation of Hi-C interactome data capturing different levels of chromatin organisation. Adapted from RPM-lab [2]

Within these chromosome territories, chromatin is known to be organised in a fractal structure [28], devoid of any knots, occupying the nuclear volume fraction of 0.1 [29, 30]. It has been shown that within the territories, the gene-rich regions are generally positioned at the periphery, a position favourable for easy accessibility to the transcriptional machinery regulating their expression. The gene-poor regions having less commonly expressed genes, such as the tissue specific or developmentally

regulated genes, are buried at the interior of chromosome territories and are expressed only after a conformational change [31–33]. It has been proposed that since the probability of chromosomal translocations increases with spatial proximity of loci or chromosomes, chromosome territories might play an important role in minimising such undesired inter-chromosomal rearrangements by acting as a barrier between chromosomes to some extent [24, 34]. Apart from this, not much of the functional relevance of chromosome territories is known, with few studies proposing that they also facilitate chromosome condensation prior to mitosis [35].

1.2.2 Chromatin Compartments

The next level at the sub-chromosomal scale is the formation of Chromatin Compartments (Figure 1.1), a consequence of the spatial segregation of euchromatin (composed of 'active' and 'open') region from heterochromatin (composed of 'inactive' and 'closed') region as defined cytogenetically for all chromosomes. Additionally, compartmentalisation of the genome into A and B compartments alternating along chromosomes, having a typical size of around ∼5 Mb each, was captured as checkerboard pattern in the Hi-C map (Figure 1.2) [7, 22]. On comparison of the epigenomic states of these two compartments with the ChIP-Seq (Chromatin Immuno-Precipitation and Sequencing) data, it was demonstrated that the A compartment strongly correlates with decondensed chromatin regions having histone marks that are involved in transcription, whereas the B compartment region correlates with dense chromatin regions and histone marks that are not involved in transcription. As a result, euchromatin and heterochromatin were given a new definition as A (*permissive*) and B (*repressive*) compartments based on the presence of largely active and inactive chromatin, respectively. Subsequent experiments with higher resolution further subdivided these compartments into A1-A2 and B1-B4 according to their chromatin signature as discussed in [36].

The regions of similar epigenetic states were observed to contact each other forming both intra- and inter-chromosomal contacts, giving rise to the plaid-pattern as indicated in the Hi-C map [7] (although, as mentioned in section 1.2.1, inter-chromosomal interactions are less frequent compared to intra-chromosomal interactions). They tend to indulge in homotypic (A-A or B-B) rather than heterotypic (A-B) contacts, i.e., in the genome, a locus from the compartment A interacts with other A compartment loci preferentially more often than it would with B compartment loci of the genome. Similarly, regions in B compartment tend to associate with other B compartment-associated regions than A compartment-associated regions.

Imaging experiments have also enhanced our understanding of the preferential positioning of these compartments [37, 38]. It was shown that the less compact A compartments containing gene-rich and active euchromatin regions position themselves at the centre of the nucleus, possibly for an easy access to the hub of transcriptional machinery, and also beneath the nuclear pores for faster transportation of the transcribed mRNA outside the nucleus for further translational processes. On the other hand, the denser B compartments composed of the gene-poor and inactive heterochromatin regions are generally positioned close to the nuclear envelope and surrounding the nucleoli. The B compartments have also demonstrated a strong link with the chromatin areas that are in contact with the nuclear envelope, also known as lamina-associated domains or LADs [38, 39]. Thus, the functional operations are influenced by the genome's division into compartments and LADs. These A/B compartments have also been found to be cell-type specific and are associated with distinct chromatin patterns [7, 36, 40].

1.2.3 Topologically Associated Domains (TADs)

At a further smaller scale within a compartment, chromatin is compacted and organised into multiple sub-megabase regions, known as topologically associated do-

mains or TADs (Figure 1.1), mainly formed by loop extrusion [41–43]. They are marked by their appearance along the diagonal of Hi-C maps in the form of contiguous square domains as seen in Figure 1.2 and discussed in [7]. These regions within the chromatin compartments exhibit a very high self-interaction frequency but they are comparatively separated from nearby domains due to the presence of boundary insulators [44, 45] (blue spheres shown in Figure 1.3). This is because of the fact that the physical interaction of DNA sequences within a TAD is much more frequent than with the sequences outside the TAD. TADs can be active or inactive, being smaller in size (median size of around 400-500kb, ∼900 kb in mice [44, 45]), contrary to the bigger A and B compartments which span many megabases and appear as alternate active and inactive sections throughout the chromosomes. The adjoining TADs are not necessarily of opposite chromatin status. However, there exists a preferential clustering where groups of adjacent TADs of same chromatin type can organise themselves into corresponding A or B compartments (detailed discussion in [46, 47]), e.g. broader and more active TADs would very likely result in the formation of A compartment domains [48, 49].

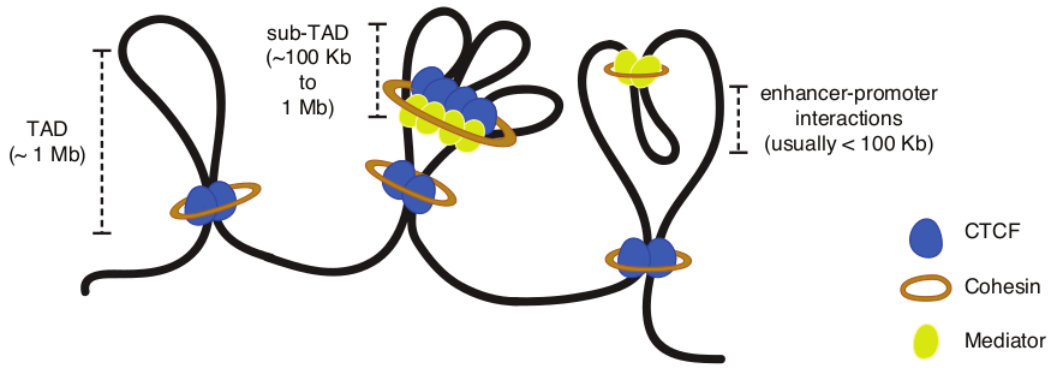


Figure 1.3: TADs, Sub-TADs and Loops. Representation of organisation of topologically associated domains (TADs), sub-TADs and regulatory loops such as the enhancer-promoter loops. Adapted from [3]

The insulation at the TAD boundary regions, which is often linked to housekeep-

ing genes (in ∼34% of TAD boundaries), repetitive elements, tRNA and most significantly, the CCCTC-binding protein (CTCF) [50] (detected at ∼76% of all boundaries with converging CTCF-motifs, i.e. motifs oriented in a way that they approach each other), results in the demarcation of these TADs from the neighbouring regions. The significance of these border-elements as CTCF-motifs was studied in the loop-extrusion model mentioned earlier, where interphase chromatin is extruded by the ring-shaped cohesin complex until it encounters the chromatin-bound convergent CTCFs. A '*stripe*' is produced at the TAD border as a result of the unidirectional extrusion of chromatin and the landing of Cohesin close to a CTCF site. Hence, CTCF plays a key role in nuclear organisation and appears to be a major player in chromatin structure formation in general.

In contrast to the A/B compartments which are tissue-specific, correlating with cell-type specific gene expression patterns as mentioned in 1.2.2, the TADs are positioned in a way that their boundaries are secured [51] and mainly conserved across species and amongst different cell types, i.e most of the TADs are tissue-invariant, but not all of them [44, 45], although their epigenetic state and compaction can vary significantly. Although recently, studies have shown that TADs are dynamic and can undergo conformational changes during cell differentiation, which suggests their capability to not only constrain but to also facilitate the most significant enhancer-promoter interactions during cellular development [52, 53]. Overall, TADs are shown to be the universal building blocks of chromosomes [47] as domains displaying high regulatory potential [54]. Each of human and mouse genomes are known to exhibit regions of over 2,000 TADs, with over 90% of their genomes sharing 50-70% of TAD boundaries between them [51].

Further classification of TADs can be made into smaller sub-TADs (Figure 1.3), exhibiting a lesser degree of conservation across tissue types and seemingly related to cell type-specific gene expression [55, 56].

1.2.4 Chromatin Loops

The organisation of the above domains is guided by specific DNA contacts: DNA-DNA or 'chromatin loops' which form the next level of organisation. Chromatin loops, roughly less than 1Mb [36], facilitate the actual interactions between distant genomic loci causing the activation or the repression of genes [25]. The genome adopts such loop conformations to be able to achieve structure-mediated regulatory effect on genes which are governed by a large set of *cis*-balancing elements such as enhancers, insulators or repressors, often located at a considerable distance from the target genes (Figure 1.3). Thus, the functional relevance of chromatin loops is to bring together two elements to a spatial proximity to elicit a regulatory response.

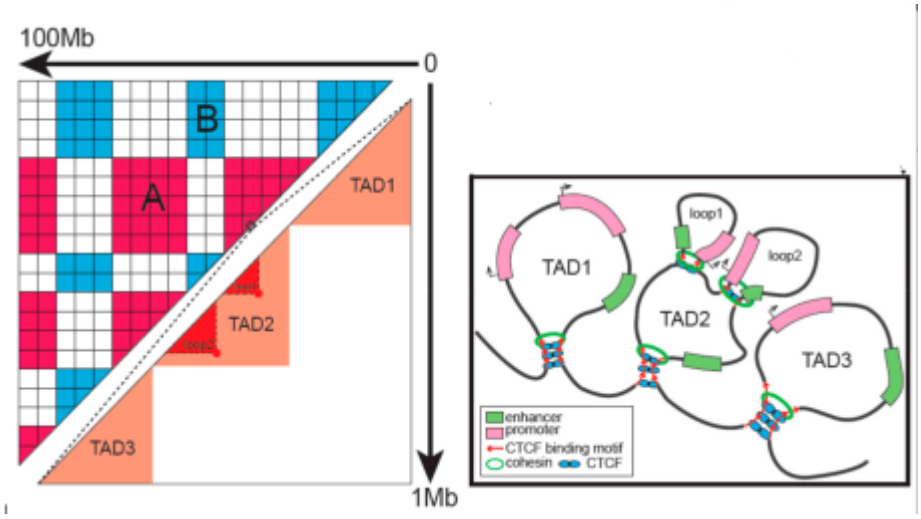


Figure 1.4: Loop formation. (*right*) Existence of loops demarcating TAD boundaries and also within TADs. (*left*) TADs and loops been represented in the Hi-C data matrix. Adapted from [4]

CTCF binding not only intervenes most of the chromatin loops [36] but also frequently demarcates TAD boundaries, although the chromatin loops are also positioned within them (Figure 1.4). Similar to TADs, most CTCF-mediated loops have a constitutive character and are highly conserved across cell types and during differentiation. Contrarily, interactions that are directly associated to the tran-

scriptional process, such as enhancer-promoter or Polycomb-mediated contacts, are linked to CTCF-dependent loops that are more dynamic [36, 52, 57]. In both situations, chromatin loops enable the further division of TADs into more dynamic, nested substructures known as sub-TADs, which have more cell-to-cell variability than TADs [16, 36, 55, 58–60].

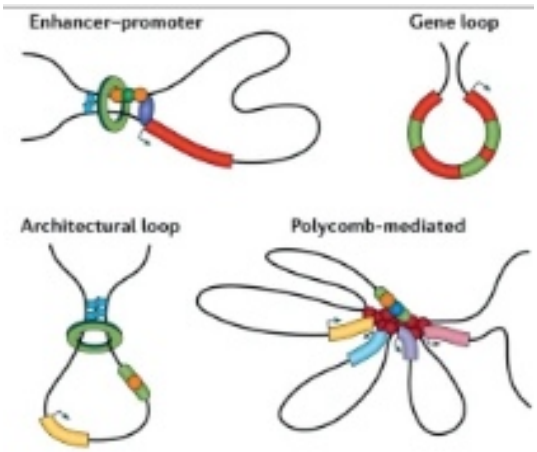


Figure 1.5: Types of loops within TAD domain. Enhancer-promoter loop, Polycomb-mediated loop, gene loop, and architectural loop are a few examples of chromatin loops that may exist inside a domain. Adapted from [5]

Enhancer-promoter loop, Polycomb-mediated loop, gene loop, and architectural loop are a few examples of chromatin loops that may exist inside a domain [5] where the gene regulatory elements use this looping process as a tool to control genes over enormous genomic lengths (Figure 1.5).

1.3 Background on Techniques and Approaches to study Chromatin Organization

The interpretation of genomic function and evaluation of its role in cellular processes is dependent on our knowledge of the scrupulous interaction between the different hierarchical orders of chromatin organisation. Three main categories of the tech-

niques to study this chromatin organisation are broadly available as: (i) image based (ii) biochemical or genomics-based and (iii) computational & modelling techniques (representative Figure 1.6)

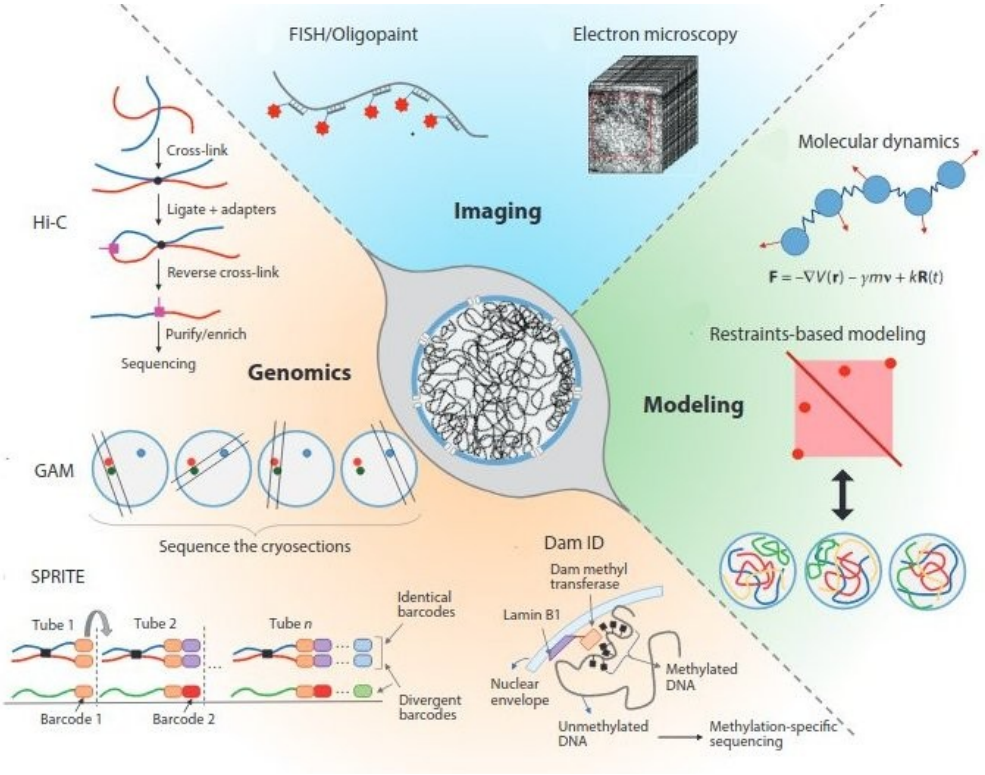


Figure 1.6: Overview of techniques to study chromatin organisation. Three categories to group methodologies for analysing chromatin organisation: (i) Imaging (ii) Biochemical or Genomics and (iii) Computational & Modelling techniques. Adapted from [6]

1.3.1 Imaging Techniques

The initial approaches for studying the 3D genome organisation were predominantly 'image-based' methods, including FISH, EM, X-Ray, Super-resolution light microscopy as the most popular ones.

- DNA FISH (Fluorescence *in situ* Hybridization) allows the localisation of intended chromatin domains or entire chromosome in a cell by hybridising fluorescently tagged probes to DNA after fixation, and then visualise the labels

under light microscopy [61,62]. Although done at a lower resolution with probe size around 40kb, FISH had the capabilities for live cell imaging during dynamic chromatin movements and interactions and in measurement of distances between two or more loci in space. [63,64].

- Electron Microscopy (EM) imaging allowed visualisations of individual cell nuclei's sections of condensed versus decondensed chromatin in exquisite detail and even the *in situ* analysis of chromatin fibre structure [65], but mostly constrained to accentuate specific DNA sequences. Electron microscopy, when combined with DNA-specific labelling, could unveil 3D chromatin structures of nanometer-scale in frozen samples [66].

- Super-resolution light microscopy (resolution up to 20nm), combining the traditional light microscopy and electron microscopy offered some important novel approaches to the imaging of genome architecture including the use of patterned excitation on the same sample and combinatorial labelling techniques to concurrently expose several distinct loci.

- Soft X-ray Tomography (SXT) is used for imaging chromatin organisation, distribution and biophysical properties during neurogenesis. It offered mesoscale resolution upto 20-50nm in intact, unprocessed cells [67].

1.3.2 Biochemical Techniques

Despite the advances, analysis of only a small number of loci could be possible by these imaging methods. The advent of novel 'genomics-based' experimental or 'biochemical approaches', as detailed below, could overcome this limitation:

- Chromosome Conformation Capture (3C), a technique that enables the identification and measurement of the frequency of physical interactions between DNA

segments [68]. This approach was further augmented to enable the analysis of contacts across the entire genome (known as Hi-C), in turn providing enormous quantitative data in regard to genome architecture at the level of large cell populations [7].

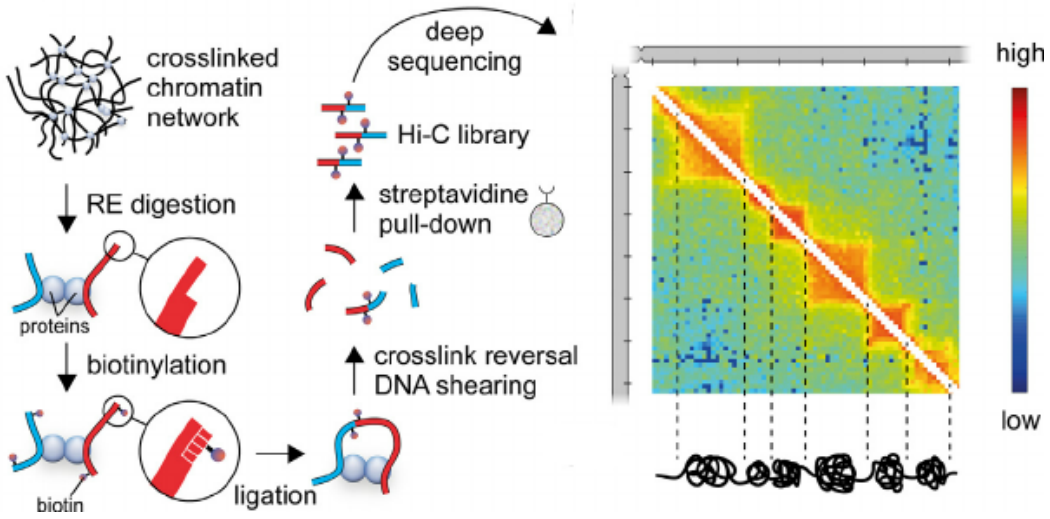


Figure 1.7: Schematic representation of Hi-C protocol Steps involved in Hi-C method include: digestion, biotinylation, ligation, cross-linking, pull-down and deep sequencing. Detailed description in [7]. Adapted from [8]

In Hi-C, different chromatin regions that are adjacent to one another spatially, undergo the protocol of cross-linking, fragmentation, ligation, and are then tagged with adapters (see Figure 1.7). Following reverse cross-linking, purification, sequencing, and mapping of the fragments to their genomic sites, genome-wide contact frequency matrices are produced. Apart from 3C (one-to-one) and Hi-C (all-to-all), other chromosome conformation capture based approaches include 4C (one-to-all), 5C (many-to-many), ChIA-PET (Chromatin Interaction Analysis with Paired-End Tag sequencing), Capture 3C, Capture Hi-C, TCC (Tethered Conformation Capture), Single-cell Hi-C (scHi-C), Dilution Hi-C and DNase Hi-C.

- Genome Architecture Mapping (GAM), a procedure that sequences DNA after cryosectioning nuclei into thin slices [69].

- Split-Pool Recognition of Interactions by Tag Extension (SPRITE) in which DNA and RNA fragments are sequentially barcoded, allowing to identify and sequence both DNA-DNA and DNA-RNA interactions [70].

- DNA Adenine Methyltransferase IDentification (DamID) whereby the Dam methyltransferase marks chromatin sites near the nuclear lamina and genome-wide sequencing is used to map them [39].

1.3.3 Computational Techniques

Although advances in both 'imaging' and 'experimental' methods discussed above have been impressive, they also have their limitations and therefore, in order to better understand the genome architecture, computational models have proved to be the indispensable tool of choice. [22]. Techniques for nuclear architecture imaging have advanced significantly, although they still suffer from on screen limited throughput choices for visualising chromatin interactions and has its own artefacts such as milder fixation and structure degradation. 3C-based experimental techniques too, had limitations of unclear inefficiencies and certain biases. In fact, the current experimental techniques prevent a direct quantitative description of the folding, movement and interaction of chromosomes within the nucleus. Moreover, unravelling the processes that control genome architecture or predicting how it will evolve in various environments or organisms is not viable by mere descriptions, however accurate. These gaps may be filled using the computational methods that have the potential to answer these unresolved questions. Consequently, the interest had shifted from generating more and more collections of detailed and quantitative data sets from experimental approaches, to developing tools for data analysis and its interpretation, and also developing high predictive powers of computational methods. This led to the development of computational models which can be broadly classified into:

polymer models or ***direct models*** or ***thermodynamic-based approaches*** or

top-down approaches and ***data-driven models*** or ***inverse models*** or ***restraint-based approach*** or ***bottom-up approaches*** based on the key strategies used to build these models.

consensus structure models and ***ensemble models*** based on whether the model can generate a single consensus structure or ensemble of configurations.

optimization-based methods and ***probabilistic methods*** based on the different mathematical approaches to operate on spatial distances between genomic loci.

1.3.3.1 Polymer Models and Data-Driven Models

As the name suggests, '*polymer models*' are based on the fact that DNA can be considered as an extremely long semi-flexible polymer chain obeying the laws of physics (detailed review in [71–73]). To elaborate, polymer models use fundamentals of polymer physics and folding principles of polymers to explain vital features of the large-scale architecture of chromosomes. The behaviour of chromosomes, is mostly inferred 'directly' from the principles of polymer physics and thermodynamics, depending upon on a relatively small set of factors such as the persistence length of chromatin, or its looping probabilities etc and a very limited physical assumptions among others. Due to this *de novo* approach, these models have a particularly strong potential to provide quantitative, predicted mechanistic insights into the chromosomal architecture that has correlated well with the experimental observations. Therefore, even after a limited set of inputs to start with and while not being very informative about possible mechanisms of folding, these models do provide significant explanations to the chromosomal organisation. Well known polymer models from the literature include random walk (RW) model [74,75], self-avoiding walk (SAW) model [76], random loop (RL) model [77], dynamic loop (DL) model [78, 79], strings and binder switch (SBS) model [80], Worm-like chain (WLC) model [81], Rouse model, entropy driven thermodynamic model [82] and more complex equilibrium globule [83] and fractal

globule [7, 28, 84, 85] model.

- The RW model was one of the initial looped-polymer models (driven by *cis*-interactions) for individual chromosomes that relied on loops as a potential explanation for how chromosomal territories are predicted. The existence of giant loops of chromatin emerging from an underlying backbone were described by a generalisation of the polymer description as a random walk also known as the Giant Loop or GL model. The parameters of this model were determined by fitting the analytical demonstration to the FISH data on human chromosome 4 [75].

- Contrary to the RW or GL (and another Multi-Loop/Subcompartment or MLS) models, which assumed that loops were effectively 'frozen' and would originate only at specific chromosomal loci that arise at typical genomic sizes (∼1 Mbp for RW or GL, ∼120 kbp for MLS) without any change in their positions during the course of simulations, another straightforward RL model proposed that chromatin loops are 'annealed', that is, they arise and vanish at random loci and at all genomic sizes in order to account for the observed folding of the chromatin fibre inside the nucleus [77]. The model was eventually developed into the so-called DL model, in which the chromosomal fibre was modelled as a self-avoiding chain and the looping-related protein-chromatin interactions were incorporated by a probabilistic and dynamic method employing MC simulations to produce starting configurations for mitotic rather than interphase chromosomes [79]. The looping probability was the main model parameter thus giving rise to different loops of varied size. The RL or random coil is considered the simplest model characterised by non-interacting monomers as opposed to the self-avoiding chains, which are characterised by non-overlapping monomers that display excluded volume interactions giving rise to an increased effective volume (due to self-avoidance effects) compared to the random coil.

- Another SAW polymer is one that folds spontaneously in a random conformation and thus, there is entropy induced intermingling within the polymer. However, a model that was developed in 2004 suggested that polymers also have a high self-attraction force that can give rise to separate chromosomal territories [86].

- The WLC or Kratky-Porod model adds intrinsic stiffness, a characteristic of semi-flexible polymers like double-stranded DNA, by coupling chain bending with an energy cost. It is a bit complex model as it takes persistence length into account.

- Unlike the RW/GL, MLS, and RL models that considered looping interactions implicitly while completely ignoring the role of diffusible chromatin-binding proteins in the nucleoplasm that mediate such looping interactions, the SBS model precisely modelled the action of these diffusible binding proteins [80]. The concentration of these binding proteins was shown to significantly affect chromosome looping and hence folding by adopting to many different states. It is a unique variation of the DL model in which the polymer fiber is modelled as a self-avoiding chain and the binding molecules are modelled as a certain concentration of Brownian particles. According to the SBS model, chromatin resides inside nuclei as a complex mixture of differently folded areas that are subject to local specific stimuli and are capable of self-organizing across spatial scales through general physical mechanisms. In fact, using the SBS model, it is possible to create more intricate organisational structures with multiple nested layers. One of the many conceivable states obtained by the SBS model turns out to be the so-called crumpled globule model or fractal globule model in particular.

- The fractal globule model describes the compact polymer state that results from polymer condensation with 'topological constraints' in the process [84]. Given

that it exhibits the same scaling behaviour of the contact probability as a function of genomic distance, s, at a scale of ∼1-10 Mbp [85], it is said to be in agreement with the original Hi-C data of the human genome [7]. In particular, intrachromosomal (*cis*) contacts appeared to degenerate as a power law $\sim s^{-1}$ of the genomic distance s, as opposed to the expected power law decay of $\sim s^{-3/2}$ for confined polymers at equilibrium (the equilibrium globule model). It was suggested that the fractal polymer that underwent fast confinement prevented chain relaxation and self-entanglements, resulting in a knot-free polymer that could unfold easily without being constrained by entanglements. This is an appealing property that could facilitate local decondensation of chromatin for gene regulation. This was in contrast to the equilibrium globule which underwent slow confinement resulting in highly tangled polymer. The role of entropy comes into play as a guiding force in territorial separation by restraining the intermingling between different chromosomes [35, 87, 88]. Entropy-based models, however, are unable to account for the wide range of specific, functional contacts made possible by chromatin looping (such as enhancer-promoter interactions), the organisation of its domains (LADs, TADs), and, in particular, the behaviour of the contact probability between genomic loci as revealed by Hi-C experiments.

To summarise, *direct* models essentially do not consider biological information (such as DNA sequence-related fibre heterogeneity, chemical modifications to DNA or histones, the presence of particular chromatin binding proteins, etc) and only account the physical assumptions. Despite this extreme simplification, these polymer physics models are actually able to explain a vast array of qualitative characteristics, such as the spatial segregation of chromosomes into distinct territories, as well as quantitative information, such as the scaling laws of average contact frequencies with genomic distance. However, several elements of large experimental data sets, including the

apparent segregation of higher eukaryotic chromosomes into topological domains, are not fully explained by these models.

These led to the development of so called '*data-driven models*' which can completely assimilate a wealth of experimental data sets, like genome-wide contact frequencies. Unlike the *direct* models, these methods use the experimentally generated contact map as input to reiterate the genome's underlying 3D structure, thus, also obtaining the name '*inverse-models*' or 3D reconstruction methods. These *indirect* models rely primarily on translating known data into conformations that are essentially based on an implicit relationship between the contact frequencies and spatial distances between genomic regions. The goal of these 'bottom-up' methods is to 'reconstruct' the conformations by satisfying the 'spatial restraints' derived from the progressively larger data sets produced by imaging methods, and particularly the high-throughput 3C-based methods. Both these models, in practice, can be differentiated by the amount of experimental data used: the *inverse* models often use thousands or more values from massive, often genome-wide data as opposed to the *direct* models, which typically use a small number of parameters. Popular 3D reconstruction models include:

(i) reconstructing the yeast chromosome from 3C data [68] where the authors tallied the number of cross-linking occurrences as contact frequencies converted into the corresponding spatial separations between 13 loci that are spread out over the ∼320 kbp long chromosome 3 giving rise to 78 pairs of measurements,

(ii) reconstruction of the ∼2.5 Mbp long immunoglobulin heavy-chain locus in B lymphocytes based on single-cell FISH experiments [89] theorising that the spatial configuration of this chromosome region makes way for genomic reshuffling from the earlier Pre-Pro-B stage to the Pro-B cell stage,

(iii) reconstruction of all the 16 chromosomes in budding yeast genome from Hi-C data at a model resolution of 10kb for each chromosomal segment [90] (utilising an op-

timised strategy in moving the positions of the beads in such a way, that their relative distances are a close match to the estimated distances from the contact frequencies. It is achieved by keeping the sum of the squared differences between modelled distances and desired distances to be minimum, while still satisfying the other constraints) thereby giving rise to a 'water-lily' model that displayed all 16 centromeres grouped together close to the same nuclear pole, chromosomal arms reaching out from the pole, and rDNA occupying the opposite pole,

(iv) reconstructing the fission yeast from Hi-C data [91] which totals onto a $\sim$14 Mbp region and comprises only three chromosomes at a model resolution of 20kb, utilising the FISH data fitted to measured contact frequencies obtained in Hi-C for seven interchromosomal pairs among the 18 different pairs of loci in 100 or more cells for each pair, which particularly reveals that co-regulated genes frequently lie in close vicinity,

(v) reconstruction of a 500kb long α-globin locus on human chromosome 16 from 5C data to calculate the contact frequencies among 70 fragments scattered over this domain [92] with the two important model parameters of equilibrium length and stiffness computed as Z-scores from 5C giving rise to an ensemble of structures,

(vi) a bacterial genome's reconstruction from 5C data comprising a single circular $\sim$4 Mbp long chromosome for 339 fragments at a genomic resolution of 12kb done by Bau *et al* [93]

(vii) reconstruction using probabilistic methods called MCMC5C [94] of the 142 kbp region of chromosome 7 containing a cluster of *Hox* genes [95] based on previously acquired 5C data. This region was shown to play a crucial role in development and cell differentiation revealing statistically significant difference in the chromosomal region in differentiated cells which was more compact than in undifferentiated cells. The same method was applied to Hi-C data previously acquired for a human chromosome 14 arm measuring 88.4 Mbp [7] which exhibits fairly good correlation of

FISH data with the distances predicted by their model for three pairs of loci. It was unfortunately not feasible to extend the analysis to the entire human genome because the reliable sampling consumes excessive computation time. In a recent Hi-C data set on mouse embryonic stem cells [44], a similar MCMC-based sampling technique termed BACH-MIX [96] was utilised to distinguish chromosome topologies within ∼Mb long topological domains.

(viii) reconstruction of a population of genome structures where a total of 2 x 428 spheres represented the whole diploid human genome, each of which was referable to a distinct chromosomal region [97]. The experimental input was taken from human lymphoblasts retrieved using an improved variant of 3C-based methods called TCC or tethered chromosome conformation method which utilises a procedure for cross-linking in which cross-linked fragments were tethered to beads. It was possible to predict the preferential bearings of chromosome territories with this model with respect to the nuclear centre or periphery, which were in consonance with those that FISH already identified.

(ix) recent reconstruction methods including ShRec3D [98] and ChromSDE [99] which use shortest path reconstruction in 3D and semi-definite programming methods respectively.

(x) a novel study that uses a neural network to infer the relationship between the genomic compartment in which a locus is located as determined by DNA-DNA proximity ligation (Hi-C) and the epigenetic information acquired from ChIP-Seq for that locus [100, 101].

With such a large amount of work, the *inverse* models in comparison to *direct* models, lag behind in terms of relative paucity of their predictive power. *Inverse* models fail to predict, for example, the change in chromosome models due to translocation or a change in gene expression, since as an input for reconstruction, new data from such experiments would be required.

Certain limitations of both *direct* and *indirect* models led to the development of a hybrid model i.e. between polymer models and data-based reconstruction models [102, 103]. In this model, chromosomes were modelled as chains of beads, each with a diameter of 3 kbp, simulated using Brownian dynamics undergoing motions with an assumption of persistence length and other additional forces. In yet another study, a least-biased effective energy landscape for the chromosome was derived using a maximum entropy method and the chromosome conformation capture data [104]. There is scope of improvement in this field owing to its advantages over both polymer-based and reconstruction-based approaches.

1.3.3.2 Consensus Structure Models and Ensemble Models

The models discussed above can be divided into *consensus structure models* and *ensemble models* based on whether a single consensus structure or ensemble of configurations is generated from the model. There are both advantages and disadvantages for consensus as well as ensemble methods. Since ensemble approaches take into account the fact that Hi-C data is collected from an ensemble of conformations, they are more acceptable in a biological context. However, it is not simple to investigate a group of inferred ensemble of 3D structures. One of the options is to characterise the ensemble average [97], while another alternative is to pick a few structures that reflect the diversity of the ensemble [94]. In contrast, the consensus methods, generate a single structure which is easier to analyse and can be thought of as a visualisation of the contact map. Many of the models in the foregoing description generated a single 3D structure of one chromosome [68, 89] or multiple chromosomes [90, 91]. However, the chromosomal polymers' dynamics is not accounted for by the consensus methods and distinct configurations may exist in different cells of a population even in the absence of the chromosomal dynamics. Ensemble methods are much more arduous in terms of computation than consensus methods because they require sampling of candidate

3D structures from a very large dimensional space .

1.3.3.3 Optimisation-Based Models and Probabilistic Models

Based on the underlying strategy to develop the model, another important categorisation of the models discussed in section 1.3.3.1 can also be done into *optimisation-based approaches* and *probabilistic approaches*. In the former approach, an objective function is minimised to fulfil all the set of constraints in order to build the model [7, 35, 68, 80, 87–93, 97–99] while in the latter strategy, a probability distribution, such as Gaussian distribution (MCMC5C) or Poisson distribution (BACH-MIX, PASTIS) of structures from contact frequency data is followed from which it is possible to derive 3D structures [94, 96, 105]. Apart from MCMC5C and BACH-MIX discussed in 1.3.3.1 to reconstruct the 3D structure, PASTIS employs calculation of maximum likelihood of the model parameters, with the highest likelihood given the observed contact data.

Various optimisation methods such as numerical optimisation or gradient-descent optimisation are used to minimise the objective function in optimised-based methods. The optimization algorithms may fetch altogether different results, based on the chosen initial configuration, raising scepticism on the explanation of the particular configurations derived from a single optimization run. Additionally, it was not possible to immediately interpret the variability seen after the output of various initial random structures was clustered (partially reflected by the multiplicity of clusters), either in terms of the biological variability of structures within a cell population or as statistical errors when building a model from necessarily small amounts of data. Despite the fact that most *inverse* models employ optimization techniques, it seems more assuring to use probabilistic sampling approaches which can determine uncertainties in reconstructed models and their parameters. This approach can offer a thorough and less biassed perspective of the ensemble of models that seem to be

compatible with the experimental results. The potential capacity of such approaches to focus on the limitations of existing experimental data sets affecting the model uncertainties, is particularly interesting. The implication could be useful in developing future experimental techniques, with an aim to minimise such uncertainties to the extent possible.

1.4 Interest and Focus of the study

This research work focuses on understanding the dynamics of chromatin organisation and transitional changes in developmental cell stages during cell differentiation of the hematopoietic system. The study was motivated from the previous work of high-throughput Hi-C analyses with epigenetic landscapes and genome-wide expression profiles done on progenitor (Pre-Pro-B) and committed (Pro-B) cell stages by our lab [53]. It had provided significant insights on the architecture of genome in the context of 2D information obtained from Hi-C. Here, in this work, we improve our analysis of studying the genomic architecture of lymphoid lineage developmental cells using the 2D information of the Hi-C data and building upon it to first generate a prototype model structure of the two cell stages using the approach and underlying principles of polymer physics. We were specifically interested in investigating their 3D structures and organisation of chromatin in these cell stages by performing a comparative analysis and studying their transitional dynamics through Molecular Dynamics (MD) simulations.

To begin with the most fundamental concepts of our research on blood cell development, we brief here about the Haematopoietic system. The broader background of the study is discussed in section 1.4.1 followed by the specific research focus in section 1.4.2. The approach we followed to carry out our study is further discussed in section 1.5.

1.4.1 Interest: Blood Cell Development (The Haematopoietic System)

Haematopoiesis is described as the process of development of all the cellular components of blood and immune system. Haematopoietic stem cells (HSCs), which are responsible for maintaining and producing a variety of cells that make up an organism's blood and immune system, are the initial point of haematopoiesis (Figure 1.8). Hence, it forms an excellent model to study changes occurring at the chromatin architectural level.

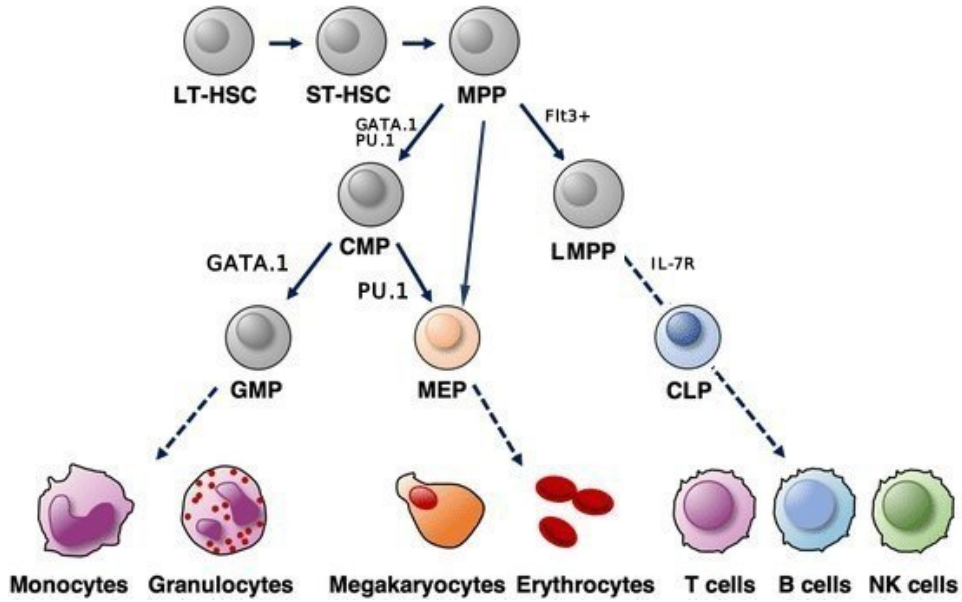


Figure 1.8: Schematic representation of Hematopoiesis Development of different blood cells from HSC. Adapted from [9]

Two essential properties, *self-renewal* and *multipotent* differentiation are the basis of definition of HSCs, which are capable of producing cells of all blood lineages: from erythrocytes (or RBCs that transport oxygen) and megakaryocytes (which produce platelets that control blood clotting) to both innate and adaptive immune cells (leukocytes) which fight infections. The most primitive self-renewing HSCs with long-term reconstituting activity (LT-HSCs) & also the short-term (ST)-HSCs were found in the

mouse bone marrow population as part of the LSK population ($Lin^{-}Sca\text{-}1^{+}c\text{-}Kit^{+}$). This is reviewed in detail in [106].

ST-HSCs generate multipotent progenitors, MPPs, defined by the absence of self-renewal and restricted lineage differentiation capacities. The MPP population is heterogeneous and includes progenitor subgroups dedicated to *myelo-erythroid* (megakaryocytes, erythrocytes, granulocytes, mast cells, dendritic cells and only monocyte-macrophage cells among agranulocytes) or *myelo-lymphoid* (granulocytes, mast cells, dendritic cells and both monocyte-macrophage & lymphocytes among agranulocytes) lineages. As a result, the myelo-erythroid subgroup of MPPs has the ability to directly develop into either common-myeloid progenitors (CMPs) or megakaryocyte-erythrocyte progenitors (MEPs). MEPs differentiate into megakaryocytes/platelets and erythrocytes while the CMPs ultimately bring about granulocytes, mast cells, dendritic cells and macrophages via granulocyte–macrophage progenitors (GMPs). The CMP can give rise to all types of myeloid colonies, while the MEP or the GMP generates only megakaryocyte-erythrocyte (ME) or granulocyte-macrophage (GM) lineage cells, respectively, indicating that the CMP retains the potential to differentiate into MEP along with GMP (Figure 1.8 and detailed review in [106]). It is because of the co-expression of PU.1 and GATA-1 that MPPs first become committed to CMPs; nonetheless, their mutual exclusion is essential for the differentiation of CMPs into either megakaryocytic-erythroid or granulocytic-monocytic progenitors. CMPs must express GATA-1 in order to differentiate into MEPs and express PU.1 in order to differentiate into GMPs.

On the other hand, for the myelo-lymphoid subset of the MPPs, the MPPs can attain $Flt3^{+}$ to develop into lymphoid primed MPPs or the LMPPs which can produce all granulocyte/macrophage progenitors (GMPs) as well as lymphocytes but no longer have the ability to self-renew or differentiate into megakaryocytes or erythrocytes. As LMPPs acquire IL-7R, they even loose the myeloid developmental potential and

become solely committed to the lymphoid lineage i.e. common lymphoid progenitors (CLPs) which can further differentiate to produce natural killer (NK) cells, dendritic cells, and the B and T lymphocytes (Figure 1.8). Two important TFs that promote initiation of lymphocyte development from LMPPs are Ikaros and the E-protein TF, E2A. *Rag1*, *Rag2*, *Dntt*, and the cytokine receptor IL-7R are among the lymphocyte-specific genes that E2A activates to cause lymphocyte-specific priming.

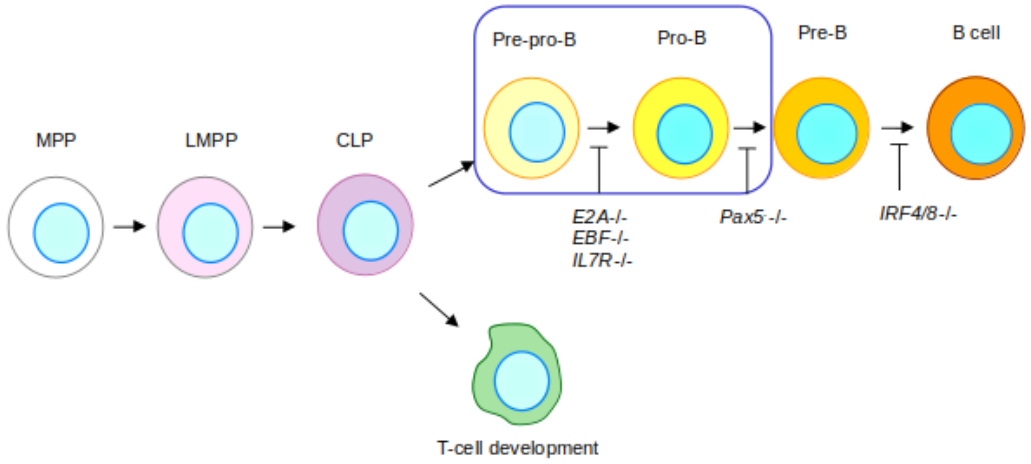


Figure 1.9: B-cell development Developmental stages during B cell differentiation. The blue box represents the focus of this study.

The CLPs, usually regarded as the branch-point for the formation of B- and T-cells, determine the fate of lymphocytes after the LMPPs (Figure 1.9). The CLPs loose myeloid potential once they acquire IL-7R and progress towards lymphoid lineage. B-cell lymphopoiesis requires a number of TFs, including PU.1, E2A, Ikaros, and FOXO1, which 'prime' the genomic *cis*-regulatory areas. The B lineage-specific TFs EBF1 and PAX-5 are responsible to activate this 'priming' process (Figure 1.9). Additionally, E2A cooperates with FOXO1 to activate EBF1, a crucial B-lineage determinant. The TF Runx1 has been shown to be responsible for activating EBF1 expression in addition to E2A and FOXO1. [107]. B-cell development is caused by the interaction of all these factors. On the other hand, Notch-DLL4 signalling, which

promotes the differentiation of progenitors towards T-lineage, causes the CLPs to mostly differentiate towards T-lineage upon entering the thymus [108].

1.4.2 Focus: Chromatin Structural Dynamics during B-Cell Commitment

The induction as well as maintenance of lineage-affiliated genetic programs is brought out by lineage specification and commitment. This incorporates two aspects: the expression of lineage-specific genes as well as the repression of alternate-lineage genes in order to establish the lineage identity. Complex cellular dynamics plays a role to achieve this by involving spatial & structural rearrangement of genome architecture in order to integrate the lineage-specific transcription factors & cytokine signals along with several other epigenetic mechanisms. As seen above in section 1.4.1, a lot has been known in terms of the role of *cis*-regulome in cell type specific gene regulation; however the structural changes that enable interactions between regions of the genome in order to achieve lineage-specific gene regulation are still unknown. Therefore, this dynamics, involving spatial re-organization of chromatin during the developmental transitions of B-cells, is the focus of this study in order to understand the 'lineage specific chromatin organisation'. To do so, we are focusing on two cell stages during the B cell development: Pre-Pro-B cell stage and Pro-B cell stage. The blue rectangular box in Figure 1.9 highlights our focus of this research. The Pre-Pro-B cells are undifferentiated cells arrested at a multipotent cell stage and maintained such that they still have the potential to differentiate into both B and T cells. On the other hand, the Pro-B cells are differentiated cells committed towards B cell development and have lost the differential potential to alternate lineages. In this study, we elucidate the principles underlying the intra-chromosomal structural dynamics and investigate its role in cell-type specific gene expression patterns, an area that has been underappreciated in theoretical and computational studies. In particular, we

are trying to determine the transitional structural variations in the two cell stages that maintain cell identity and orchestrate B cell commitment.

1.5 Our Approach: Our Hybrid Model Method and its Advantages

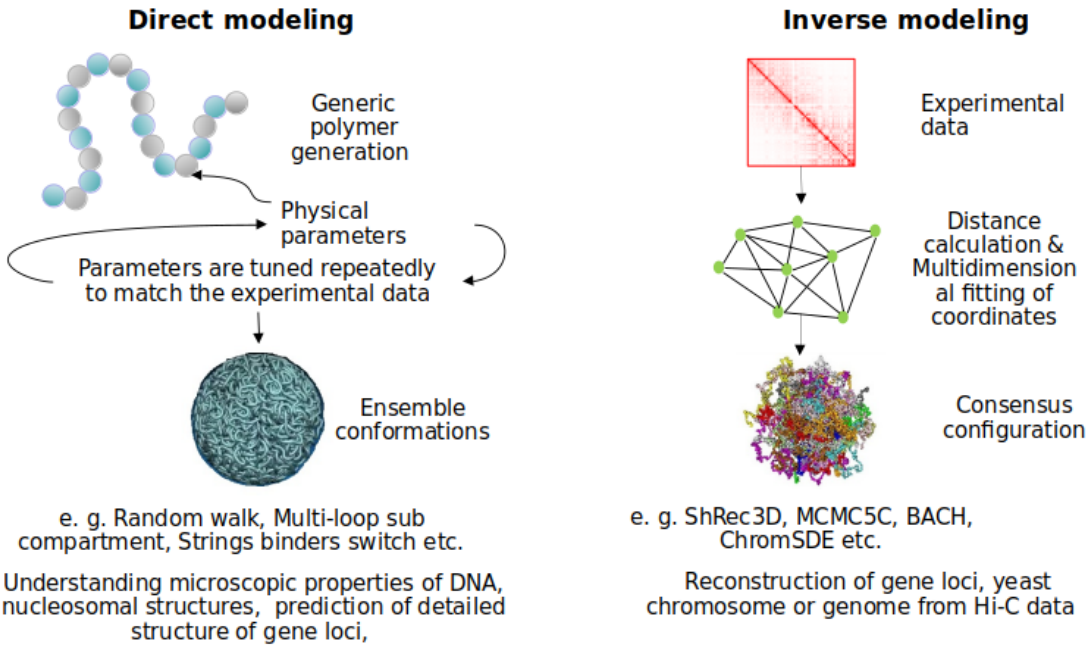


Figure 1.10: Direct versus Inverse models Illustration of comparison between *direct* and *inverse* models.

As we have already discussed in detail in section 1.3.3, there have been various computational methods developed using polymer physics simulations, which are solely guided by a limited number of physical assumptions and parameters. Although, these *direct* models helped in understanding qualitative & quantitative properties, they essentially did not consider any biological informations but only retained simplified physical assumptions failing to take into account every aspect of the extensive experimental data sets. On the other hand, the rich experimental data sets, such genome-wide contact frequencies, are fully incorporated into the *indirect* or *inverse*

models as input to rebuild the underlying 3D structure of a genome. Such models lack in their predictive power, for example, it is impossible to foresee the effects of a translocation or a change in gene expression through these models since new data set would be required from such experiments as an input for the reconstruction. A pictorial representation of both approahes is highlighted in Figure 1.10. Limitations of both *direct* and *indirect* models led us to the development of our hybrid model where we generated a physical coarse-grained bead-on-a-string polymer model and incorporated the experimental datasets as input to it and let this hybrid model evolve with time. From this time-evolved trajectories, we study the dynamic changes occurring in the two systems representing the two cell stages. By doing so, we introduce a predictive computational model, with minimal biological information to begin with, in order to study the cell-type specific 3D chromatin folding. Through this combinatorial approach, we were able to utilise the concepts of the polymer physics along with the relevant biological information to be able to derive a fundamental relationship between genome organisation & cell type-specific gene expression and also provide mechanistic insights to its regulation. We show that our polymer model is a powerful tool for investigating structural rearrangements and predicting consequential gene expression patterns upon cell differentiation. We have studied the organisation of a murine chromosome that shows crucial changes during B cell development. In order to study its dynamics, we have modelled a self-avoiding polymer chain with harmonic bonds between consecutive beads and incorporated Hi-C information as weak harmonic bonds and performed Langevin dynamic simulations, the details of which are discussed in Chapter 2.

1.6 Objectives: Overall and Specific

The overall objective of this study is to determine the structural alterations during the developmental transitions of B-cells with a focus to understand the lineage specific chromatin organization.

This was then classified into the following specific objectives:

1. Develop and validate a robust prototype structural model of a chromosome using high-throughput chromatin interactome by employing mechanistic modelling.

2. Comparative structural characterization to identify cell type specific chromatin architecture at different levels of chromatin organization using the simulated models.

3. Determine novel spatial rearrangements leading to differential changes and correlate with gene expression patterns.

Chapter 2

Methodology

2.1 Data Acquisition

The experimental data used in the present study was obtained from the in-house experiments of high-throughput Hi-C sequencing data published online (GSE85858) [53]. The Ebf1$^{-/-}$ cells indicate the Pre-Pro-B stage while the Pro-B cell stage is represented by Rag2$^{-/-}$ cells.

2.2 Model Generation

We have computationally modelled the chromosome 11 of mouse genome as a beads-on-a-string homopolymeric chain consisting of spherical non-overlapping beads (self-avoiding walk polymer) of defined diameter σ, connected by a spring. Each of the beads in the present model maps genomic region of size 40kb, which is same as the Hi-C matrix resolution as reported in [53]. This leads to the total number of beads as 3053 (size of chromosome/resolution = 122082543bp/40x10^{3}bp) in our model system as the length of chromosome 11 of the mouse genome is 122082543bp (in mm10 genome). 200 such initial self-avoiding polymer chains with 3053 beads each, were generated where each polymer chain was defined in a confinement of radius (r_{conf}) 0.986μm in real units. The confinement depicts the chromosomal territory (as shown in Figure 2.1), where the radius of the confinement is calculated proportional to the known genomic volume fraction of an eukaryotic cell (explained in section 2.3.2). Based on the volume fraction of 0.1 for a eukaryotic genome (refer section 1.2.1), the diameter of the 40kb sized spherical bead was calculated to be 63.13nm (i.e. $\sigma = 63.13$nm) and the radius of confinement to be 15.6 times larger than the bead diameter (i.e. $r_{\text{conf}} = 15.6\sigma$). The calculations for determining the bead diameter σ and the radius of confinement r_{conf} are discussed in section 2.3. The contact information from the Hi-C data was then integrated to these initial random generated structures, the details of

which are discussed in section 2.4.

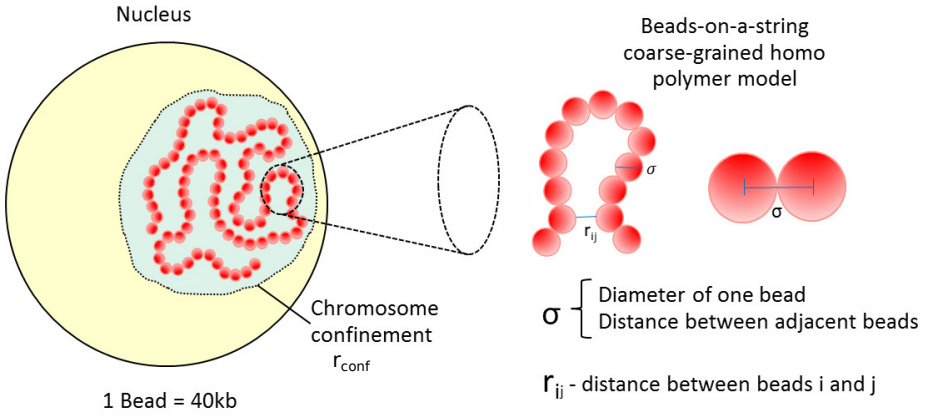


Figure 2.1: Prototype of initial configuration. Prototype representing the bead-on-a-string homopolymeric chromatin model with bead size (σ) and radius of confinement (r_{conf}) shown. The size of one bead is considered to be 1σ or 40kb in genomic units.

2.3 Determining Bead Size and the Radius of Confinement of the Polymer

2.3.1 Determining Bead Size, σ

To determine the bead size, we assume that the volume of chromosome 11 having L basepairs (V_{L}) *in vivo* is equal to the volume of the modelled polymer for chromosome 11 *in silico*. Volume of the polymer *in silico* is computed as the volume of one bead $\times$ total number of beads in the polymer. Therefore, if $4/3\pi(\sigma/2)^3$ is the volume of one bead with diameter σ and N as the total number of beads in the polymer, V_{L} *in*

silico can then be written as

$$\text{Vol. of one bead} \times \text{No. of beads} = \text{Vol. of chromosome of length L bp}$$

$$i.e.\ \frac{4}{3}\pi\left(\frac{\sigma}{2}\right)^3 \times N = V_{\mathrm{L}} \tag{2.1}$$

Now, to compute the volume V_{L} *in vivo*, we first compute the volume of 1 bp. V_{L} can then be derived as volume of 1 bp $\times$ L bp, with an assumption of uniform volume of each basepair. Since the entire genome of total length G bp occupies a volume denoted by V_{genome}, we assume that 1 bp will effectively occupy a volume V_{genome}/G and therefore, chromosome with L bp will occupy a volume V_{genome}/G $\times$ L i.e. V_{L} *in vivo* can be written as

$$V_{\mathrm{L}} = \frac{V_{\mathrm{genome}}}{G} \times L \tag{2.2}$$

Here, if 0.1 is the volume fraction where volume of the genome occupies 10% of the nuclear volume (V_{nucleus}) as discussed in section 1.2.1 and also in [29, 30], then, we can derive V_{genome} as $V_{\mathrm{genome}} = 0.1 \times V_{\mathrm{nucleus}}$ and substitute it in equation 2.2 as

$$V_{\mathrm{L}} = \frac{\left(0.1 \times V_{\mathrm{nucleus}}\right)}{G} \times L \tag{2.3}$$

Here, $V_{\mathrm{nucleus}} = 4/3\pi(d_{\mathrm{nucleus}}/2)^3$ where the nuclear diameter, d_{nucleus} =$\sim$7μm for

lymphocytes since in normal situations, the coarse, dense nucleus of a lymphocyte is approximately about 7μm in diameter [109]. Substituting V_{nucleus} in equation 2.3, we obtain V_{L} *in vivo* as

$$V_{\text{L}} = \frac{\left(0.1 \times \frac{4}{3}\pi\left(\frac{d_{\text{nucleus}}}{2}\right)^3\right)}{G} \times L \tag{2.4}$$

Equating the V_{L} *in vivo* from equation 2.4 and V_{L} *in silico* from equation 2.1, we get

$$\frac{4}{3}\pi\left(\frac{\sigma}{2}\right)^3 \times N = \frac{\left(0.1 \times \frac{4}{3}\pi\left(\frac{d_{nucleus}}{2}\right)^3\right)}{G} \times L$$

Simplifying that gives us,

$$\boxed{\sigma = d_{nucleus}\left(\frac{0.1 \times L}{G \times N}\right)^{1/3}} \tag{2.5}$$

For $d_{\text{nucleus}} = 7\mu$m,

L = 122082543 bp,

G = 2 × haploid = 2 × 2725521370 bp and

N = L/40kbp = 3053 beads, we obtain

σ = **63.13 nm**

2.3.2 Determining Radius of Confinement, $r_{\mathbf{conf}}$

Based on the fact that the total genome has a volume fraction of 0.1 within its entire nuclear volume or the confining volume enveloping that genome, i.e $V_{\text{genome}} = 0.1 \times V_{\text{nucleus}}$; we assume that V_{L} *in silico* would also occupy a volume fraction of 0.1 within its chromosomal territory defined as $V_{\text{confinement-for-L}}$ *in silico*. That is to say, *in silico*

$$V_{\text{L}} = 0.1 \times V_{\text{confinement-for-L}} \tag{2.6}$$

where

$$V_{\text{confinement-for-L}} = \frac{4}{3}\pi\left(\frac{d_{\text{confinement-for-L}}}{2}\right)^3$$

Substituting $V_{\text{confinement-for-L}}$ in equation 2.6 and then equating the resulting V_{L} *in silico* to V_{L} *in vivo* from equation 2.4, we obtain

$$0.1 \times \frac{4}{3}\pi\left(\frac{d_{\text{confinement-for-L}}}{2}\right)^3 = \frac{\left(0.1 \times \frac{4}{3}\pi\left(\frac{d_{\text{nucleus}}}{2}\right)^3\right)}{G} \times L \tag{2.7}$$

Simplifying it, we get

$$\left(d_{\text{confinement-for-L}}\right)^3 = \frac{\left(d_{\text{nucleus}}\right)^3}{G} \times L$$

$$d_{\text{confinement-for-L}} = d_{\text{nucleus}} \left(\frac{L}{G}\right)^{1/3}$$

$$\boxed{r_{\text{confinement-for-L}} = \frac{d_{\text{nucleus}}}{2} \left(\frac{L}{G}\right)^{1/3}} \tag{2.8}$$

With values of d_{nucleus}, L for chromosome 11 and G mentioned in section 2.3.1, we calculated $r_{\text{confinement-for-L}}$ or $r_{\mathbf{conf}} = \mathbf{986.57nm}$ i.e. 986.57nm/63.13nm = 15.6σ. Therefore, $r_{\mathbf{conf}} = \mathbf{15.6}\sigma$

2.4 Incorporation of Hi-C Data

Once the polymer chains are generated with model parameters, σ, and the radius of confinement r_{conf}, the biological information from the Hi-C interactome data is incorporated, thus obtaining the current hybrid model. The intra-chromosomal interaction data for chromosome 11 was obtained from the in-house generated genome-wide Hi-C for Pre-Pro-B cells as well as the Pro-B cells [53] and was integrated into all the 200 initial polymer models as weak harmonic bonds. For this, we first extracted the N×N intra-chromosomal matrices from the normalised genome-wide Hi-C contact frequency matrices for both the Pre-Pro-B and Pro-B cell stages. The steps to process the raw reads of the Hi-C data have already been discussed in [53] wherein the Iterative mapping module of hiclib (`https://github.com/mirnylab/hiclib-legacy` by Mirny lab) was used. After ICE (iterative correction and eigenvector decomposition) normalisation, the corrected contact frequency matrix was converted to contact

probability matrix using the method previously employed in [104], i.e.

$$P_{\mathrm{ij}} = min\Big(1, \frac{c_{\mathrm{ij}}}{min\big(n_{\mathrm{i}}, n_{\mathrm{j}}\big)}\Big) \tag{2.9}$$

$$where, n_{\mathrm{k}} = max\big(n_{\mathrm{k\text{-}4,k}}, ..., n_{\mathrm{k\text{-}1,k}}, n_{\mathrm{k+1,k}}, ..., n_{\mathrm{k+3,k}}\big)$$

where c_{ij} is the contact frequency and P_{ij} is the contact probability between regions or beads i and j. These experimentally derived Hi-C contact probability maps were integrated in the current model as harmonically restrained bonds between two given beads representing the corresponding 40kb sized genomic region in the Hi-C contact frequency matrix. However, unlike bonds between consecutive beads, these 'Hi-C bonds' between non-consecutive beads are restrained by contact probability-dependent distances and distance-dependent force constants. In particular, if the Hi-C pair contact probability, P_{ij} between i & j of the N×N contact probability matrix is such that, $P_{\mathrm{ij}} \geq P_{\mathrm{c}}$ where P_{c} is the probability cut-off at 0.04, then the corresponding 'Hi-C bond' is modelled via a harmonic restraint of spring constant, $k_{\mathrm{Hi\text{-}C}}$, defined as

$$\boxed{k_{\mathrm{Hi\text{-}C}} = \frac{k_0}{r_{\mathrm{ij}}}} \tag{2.10}$$

where, $r_{\mathrm{ij}} = \sigma / P_{\mathrm{ij}}$, $k_0 = 2.0$ kJmol^{-1}nm^{-2}. Here, the amplitude term k_0 establishes the maximum limit to the force constant for the Hi-C bond. For longer distances, this function effectively assigns lesser values of the force constant. The threshold probability cut-off, P_{c}, was chosen in order to consider only the minimal set of Hi-C

data above the selected threshold of contact probability. An overview of the complete workflow is depicted in Figure 2.2.

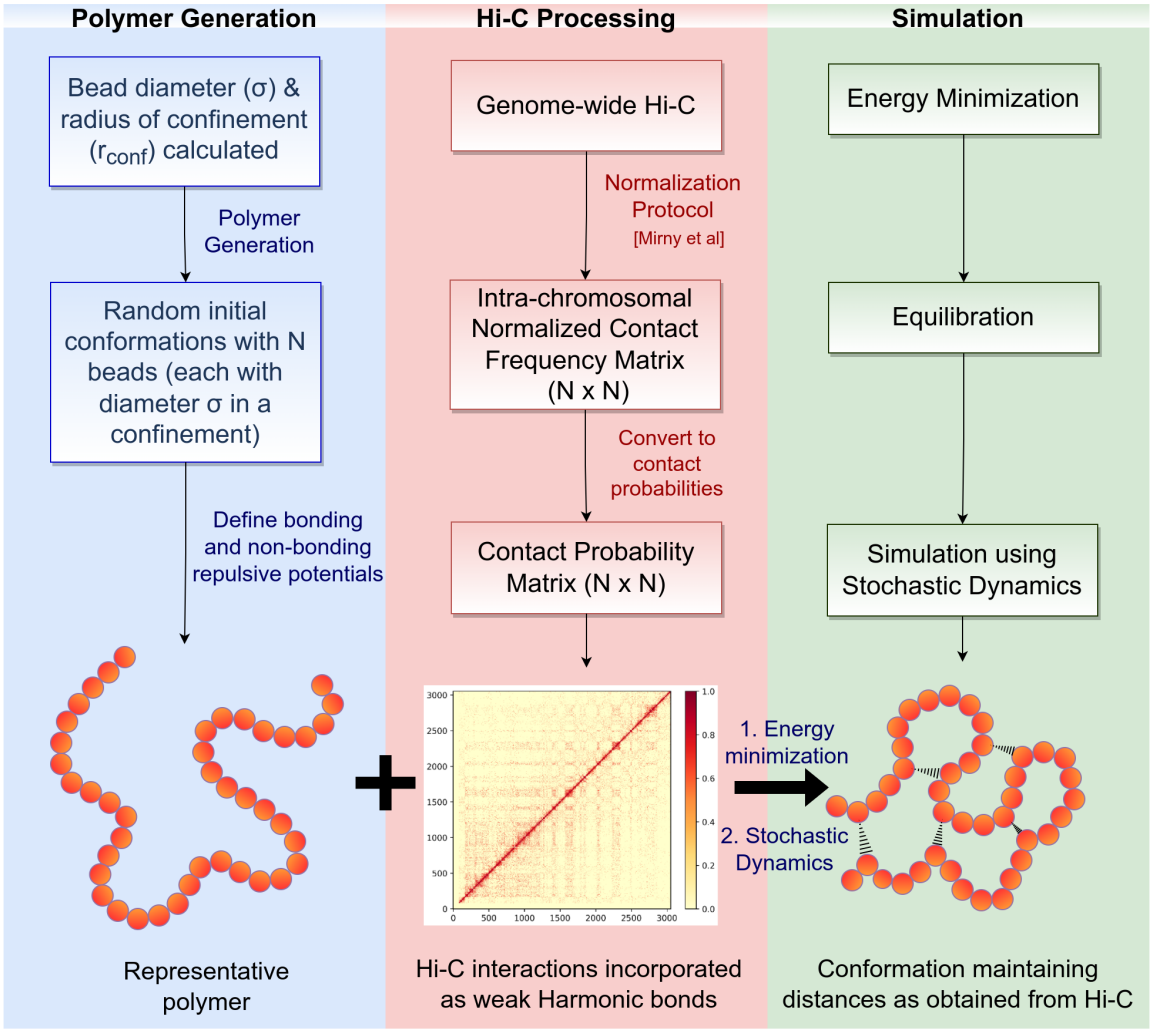


Figure 2.2: Schematic overview of the approach. The flow chart explains the workflow of the approach followed and the processes involved in each step

.

2.5 Defining the Force Field

The force fields were defined by defining the following potentials:

(i) The bonded interaction potential $V_b(r_{ij})$ between consecutive beads i & j separated by distance r_{ij} was defined as strong harmonic springs given by the equation

$$V_{\mathrm{b}}(r_{\mathrm{ij}}) = \frac{1}{2}k_{\mathrm{b}}\big(r_{\mathrm{ij}} - r_0\big)$$

with the equilibrium bond length, $r_0 = 1\sigma$ and a strong bond constant $k_{\mathrm{b}} = 300\mathrm{kJmol}^{-1}\sigma^{-2}$.

(ii) The angular potential (U_{angle}) between three consecutive beads was defined as

$$U_{\mathrm{angle}} = K_{\mathrm{a}}\big[1 - cos\big(\theta - \theta_0\big)\big]$$

where $K_{\mathrm{a}} = 2.0$, $\theta_0 = 180$, in order to provide rigidity to the polymer and reduce the possibilities of unwanted bending that can give rise to overlaps between beads.

(iii) Since we assume that Hi-C interactions take care of the attractive interactions, the other non-bonded interactions are repulsive that were defined as the repulsive term of the Lennard-Jones potential,

$$V_{\mathrm{LJ}}(r) = \frac{c^{12}}{r^{12}} - \frac{c^6}{r^6}$$

with the attractive term, $c^6 = 0.0$ and the repulsive term, $c^{12} = 1.0$ kJmol^{-1}. The interactions between all the non-consecutive or the non-adjacent bead pairs have been permitted purely via this repulsive potential.

2.6 Simulation Details

To perform the simulations, we have used GROMACS 5.0.7 [110] which is a popular open source programme. We energy minimised the 200 polymer configurations, generated using the method described in the previous sections, followed by 2×10^6 steps of Langevin Dynamics simulation for each of these configurations. A Langevin thermostat set to 310K and a coupling constant of 1 ps was used to maintain the tem-

perature of our system. All the 200 simulations were run for a total time of 2.372s in real units, 1 timestep(ts)=0.002τ (refer section 2.6.2 for derivation of time-scales) and at equal intervals of every 250 timesteps, we had recorded the coordinates of the system. Thus, the number of configurations saved for each of the 200 simulations will be total timesteps/250, giving rise to 8000 simulation frames for each polymer.

2.6.1 Langevin Dynamics for Molecular Simulations

In molecular dynamics simulations, for each atom or a coarse-grained (CG) unit in a system, Newton's second law of motion, also known as the equation of motion (F =ma) is solved. The full solvent features can be ignored in order to simplify the simulations and to do so, a popular method to accomplish this is *Langevin dynamics*. In our model simulations, we have used Langevin dynamics in order to mimic the effect of solvent and the real world scenario. In Langevin dynamics, the influence of the solvent can be roughly described by two extra force terms in the equation of motion. Hence, the resultant equation of motion or the Langevin equation for an atom or CG entity (i.e. bead i in our case) becomes

$$m\frac{d^2r_{\mathrm{i}}}{dt^2} = F_{\mathrm{i}} - \xi\frac{du}{dt} + \sqrt{6\xi k_{\mathrm{B}}T}\eta_{\mathrm{i}}(t)$$

where r_{i} is the vector position of bead i. The force i.e. mass times the acceleration shown on the left-hand side of the equation, experienced by the bead i is calculated from the three terms on the right of the equation. This equation is solved numerically during a simulation run wherein time is assumed to evolve discretely in the form of timesteps. These three terms are used to compute the force acting on the bead at one time point, which is used to determine how the force affects the bead's velocity and location at the next time point. This equation is basically solved for all of the beads in a CG system by the simulation software program GROMACS mentioned above.

The three force terms on the right of the equation can be explained as:

(i) The force that a bead experiences as a result of interactions with all the other beads in the system is represented by the first term, F_{i}. It is essentially a set of simplified potentials as mentioned in section 2.5 that takes into account the important aspect of the beads' connectivity in a polymeric chain via the bonded potential, avoiding beads from overlapping with each other in space via the non-bonded repulsive potential, and taking into account the polymer bending stiffness via the angular potential.

The next two terms help in approximating the effects of the solvent on the model system where

(ii) The term -$\xi \frac{du}{dt}$ refers to the frictional or viscous drag that the bead experiences as it moves inside the solvent. Here, ξ is the friction parameter associated to the solvent's viscosity and it follows an inverse relation to determine the movement of the bead in the solvent i.e. the bead's velocity is inversely proportional to this frictional force or drag due to the viscosity of the solvent.

(iii) To account for the random interactions within the solvent molecules, the second solvent term approximately represents them as the thermal jostling. The term $\eta_{\mathrm{i}}(\mathrm{t})$ denotes the random force that the solvent molecules exert on the bead i at a time point t. At each time step, it generates a random number in order to incorporate *stochasticity* in the simulations, while on the other hand, this *noise* has a well-defined mathematical description. The $\sqrt{6\xi k_B T}$ factor guarantees that the equation complies with the fluctuation-dissipation theorem, which describes the link between an object's diffusive motion (fluctuations) and the viscous drag it experiences while being pushed through a fluid (dissipation).

2.6.2 Derivation of Simulation Time Scales

Simulations were run for the total number of timesteps = 2×10^6 timesteps where 1 timestep (ts) = 0.002τ. The value of τ is calculated as $\tau = 3\pi\eta\sigma^3/k_{\mathrm{B}}\mathrm{T}$

where η (viscosity of water) = 10^{-3}Pa sec,

$k_{\mathrm{B}}T = 4\times10^{-21}$ J and

$\sigma = 63.13$nm

which results in $\tau = 0.593$ms. Therefore, the total time for which the simulation ran for each of the configuration was: total number of timesteps $\times$ 1 timestep = $(2\times10^6) \times (0.002 \times 0.593 \times 10^{-3}$ secs) = 2.372s

2.7 Calculation of Simulated Contact Probability Matrix

The simulations were done independently for both scenarios, the Pre-Pro-B cell stage and the Pro-B cell stage. Last 2000 frames of each of these 200 GROMACS trajectories were employed for any production analyses as these were energy minimzed and had attained equilibrium. Using the last 2000 frames in each simulated trajectory, all inter-particle distances were calculated in order to generate a distance matrix for each frame. For a single trajectory k, a final distance matrix D_{k} was then obtained by averaging over the 2000 total number of frames. From this distance D_{k}, we also generated probability matrix P_{k} for each of the k trajectories such that $P_{\mathrm{k}} = \sigma/D_{\mathrm{k}}$. All the analysis was done by converting the distances to contact probabilities using MDAnalysis Python Package. Further, a final simulation derived contact probability matrix was generated by averaging P_{k} over all the k trajectories. Therefore, the simulation-derived contact probability matrix ($\mathrm{P_{sim}}$PPB for Pre-Pro-B and $\mathrm{P_{sim}}$PB for Pro-B) is essentially averaged over 200(trajectories)$\times$2000(frames) independent

conformations. Here, each frame corresponds to a microstate of the ensemble of chromatin conformations whose average is the experimental Hi-C matrix. Since a set of 2000 frames belong to a particular initial conformation, instead of generating $200 \times 2000 = 4 \times 10^5$ probability matrices and performing an average over them, we average over 200 matrices, which have already been averaged frame-wise.

We further filtered this simulated contact probability matrix by replacing elements in the matrix with zeros which were also zero in the experimental matrix to avoid any misinterpretations from our simulated matrix. To render the representative 3D conformations of the chromosome model, we have used the open-source package Visual Molecular Dynamics (VMD) [111].

2.8 Principal Component Analysis

A potent tool for studying multivariate or data with many dimensions is principal component analysis (PCA). The fundamental principle behind PCA is to redefine the coordinate system such that the data may be described using as few dimensions as feasible. This is a form of clustering or dimensionality reduction method. The first component can be used to describe as much of the system variance as feasible, the second component can be used to describe as much of the remaining variance as possible, and so on. These axes of the coordinate system are known as the principal components. The data can then be viewed more simply by taking into account each region in relation to its values along the first two primary components.

PCA was first used on Hi-C data by [7] for the prediction of A and B compartments. In this setting, each region along the chromosome represents a dimension in the analysis. The first eigenvector or Principal Component 1 (PC1) of the correlation matrix was then used to calculate the compartment score, and genomic regions with positive or negative compartment scores were categorised as belonging to compart-

ments A or B, respectively. We have used the same principle in our study where we perform PCA on the simulation-derived contact probabilities matrices in order to identify A/B compartments. We have performed PCA analysis using the Python package.

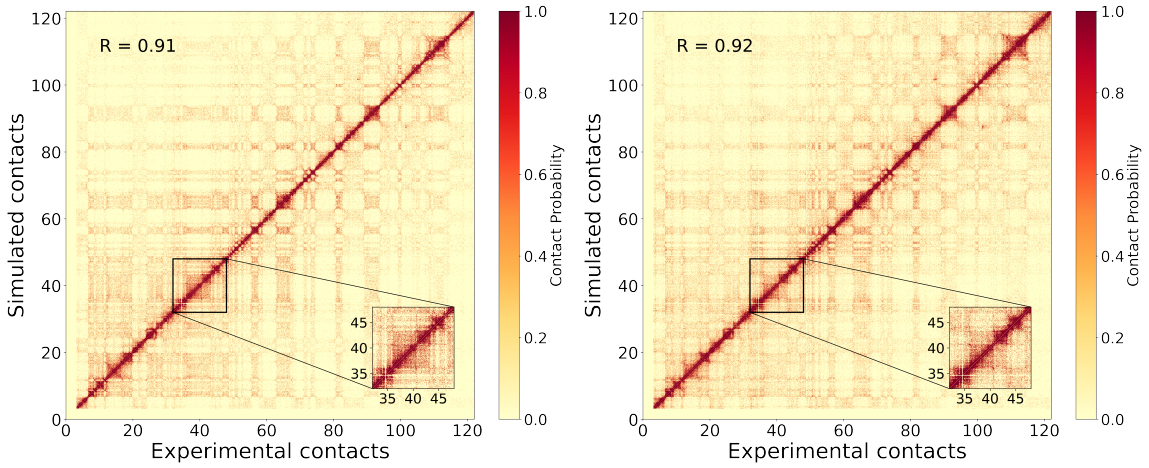
Chapter 3

Results

3.1 Polymer-based model recapitulates chromosome-conformation capture data

To begin with, we have modelled chromosome 11 of mouse for two reasons: first, it harbours crucial factors responsible for B cell development and second, this chromosome has a genomic length of 122kb that is intermediate in size. Hence, it is optimal in terms of handling complexity in a computationally affordable model. As described in chapter 2 in section 2.2, we have modelled the chromosome as a beads-on-a-string homopolymer consisting of identical monomers or beads at 40kb resolution in our study. The energy function incorporates experimentally rendered Hi-C probability matrix, excluded volume interaction and the resultant polymer model is constrained in a confinement (refer section 2.3.2) that is commensurate with its chromosomal territory (Figure 2.1).

To validate our proposed computational model, we first computed the simulation contact probability matrix by averaging over the ensemble of conformation simulated across multiple configurations (averaged over 200×2000 independent conformations) and compared it with the experimental contact matrix obtained from the Hi-C data (refer section 2.7 for details of simulation-derived contact probability calculation). Figure 3.1a and 3.1b compares the heatmap between simulation derived contact probability matrix and the experimental Hi-C contact probability matrix of chromosome 11 for Pre-Pro-B and Pro-B cell stage, respectively. With a Pearson correlation coefficient of 0.91 and 0.92 between corresponding experimental and simulated contact probability matrices of Pre-Pro-B and Pro-B, respectively, our model clearly indicates a very good agreement between simulations and experimental data for both cell stages. This is also evident through remarkably similar checker-board patterns of the corresponding matrices in both cell types. The results also show that inspite of considering only a small percentage of experimental interactions in our simulations above

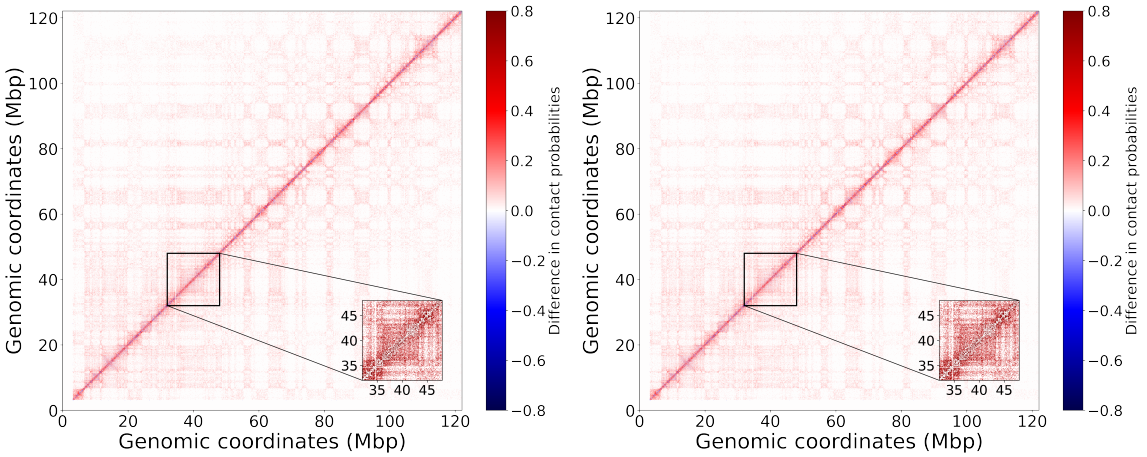


(a) Pre-Pro-B **(b)** Pro-B

Figure 3.1: Comparison of experimental versus simulation-derived contact probabilities The heatmap shows the comparison between experimental and simulated contact probability maps of chromosome 11 at 40kb resolution for (a) Pre-Pro-B and (b) Pro-B cells.

a threshold, P_c, such that $P_{ij} \geq P_c$ where $P_c = 0.04$, our model faithfully reproduces not only the considered Hi-C interactions but also those experimental Hi-C interactions which weren't included in the initial incorporation while generating the model. This result contributes to the model's efficiency and performance only on limited input information. In the heatmap generated in Figure 3.1, we also observed intense diagonal regions which indicate smaller distances having a higher contact probability between neighbouring chromosomal regions. This is justified since the proximal regions represented by the diagonal, tend to exhibit higher contact probabilities than the distal genomic regions that exhibit relatively smaller contact probabilities unless there is a possibility of formation of highly interacting regions such as TADs. We talk about the presence and predictability of these regions further in section 3.2.4.

For a more rigorous assessment of the computed simulation-derived matrix, we plotted a heatmap of the difference matrix, calculated as difference between the simulation-derived contact probability matrices and the experimental contact probability matrices for both Pre-Pro-B and Pro-B (Figure 3.2a for Pre-Pro-B and 3.2b



(a) Pre-Pro-B **(b)** Pro-B

Figure 3.2: Difference plot between simulation-derived and the experimental contact probabilities The heatmap shows the difference between experimental and simulated contact probability maps of chromosome 11 at 40kb resolution for (a) Pre-Pro-B and (b) Pro-B cells. The blue and red colours in the colour bar indicate higher contact probability in the experimental data and simulation, respectively.

for Pro-B) cell stage. We observe that this difference for any bead i and j is minimal in both cell types (the white regions in the difference plots), except in regions near the diagonal. In the diagonal region, simulation contact probability is estimated to be higher than the corresponding experimental probabilities. This could arise due to the high interaction frequencies between consecutive beads of the polymer owing to their physical proximity that accounts for the over estimation of contact probabilities. Even in the absence of any contact information, the simulation contact probabilities tend to be greater due to the closely packed adjacent beads that lead to the observed difference. Thus, the results from the difference heatmap indicate that for longer genomic distances, the simulation-derived contact probabilities are in agreement and exhibit least difference with the experimental contact probabilities.

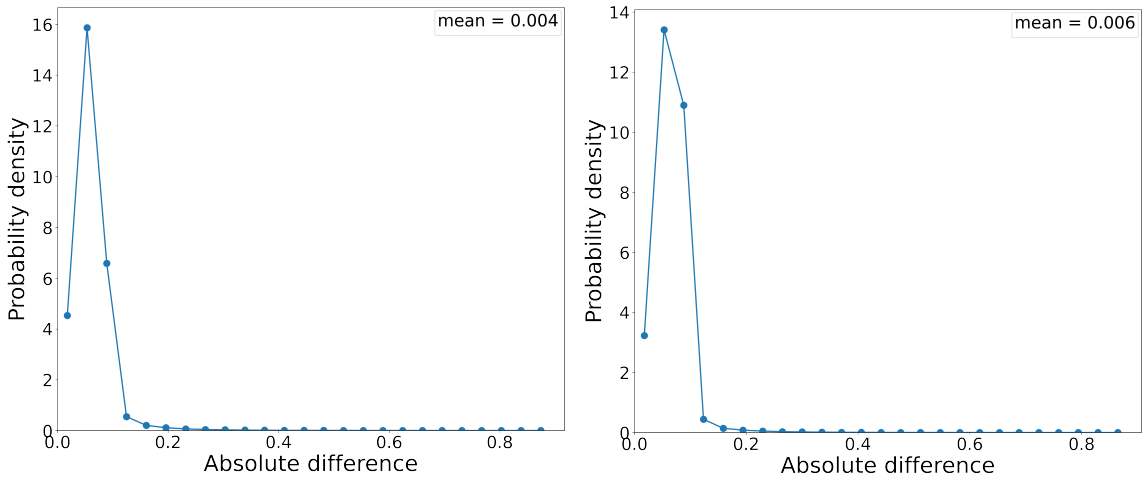
Further, in order to quantify this difference, we generated the probability density plot of absolute differences between the experimental and simulation contact probabilities for both cell types (Figure 3.3), after excluding the noise near the diagonal regions that was due to the higher differences in the contact probabilities, as observed

in Figure 3.2 above. In the distribution of absolute values of the difference in contact probabilities between experiment and simulation data obtained from the difference heatmap, we see that the discrepancy or the difference between the simulation and experiment contact probability is as tiny as <0.1, indicating that there is a fair amount of agreement between the simulation and experiment contact probabilities. We observe that above 90% of the contact probabilities show the absolute difference close to 0, indicating almost negligible difference between the experimental and simulation probabilities, thereby, suggesting our model's conformity with experiments. We observe the maximum difference value to be as low as 0.004 and 0.006 between experimental and simulation-derived probabilities for Pre-Pro-B (Figure 3.3a) & Pro-B cells (Figure 3.3b), respectively. Due to these small scale differences, we demonstrate that our model is sufficiently robust to be investigated for the analysis and prediction of key chromosomal properties. We discuss about them in section 3.2 where we further test our model for results independent of any experimental inputs.

Finally, we also show a conformation representative from the ensemble of conformations for the structure of chromosome 11 obtained via simulations for Pre-Pro-B (Figure 3.4a) and Pro-B (Figure 3.4b). The snapshots were generated using VMD software and rendered for image quality purposes. The chromosomal regions have been coloured with respect to their genomic location for both Pre-Pro-B and Pro-B chromosome models.

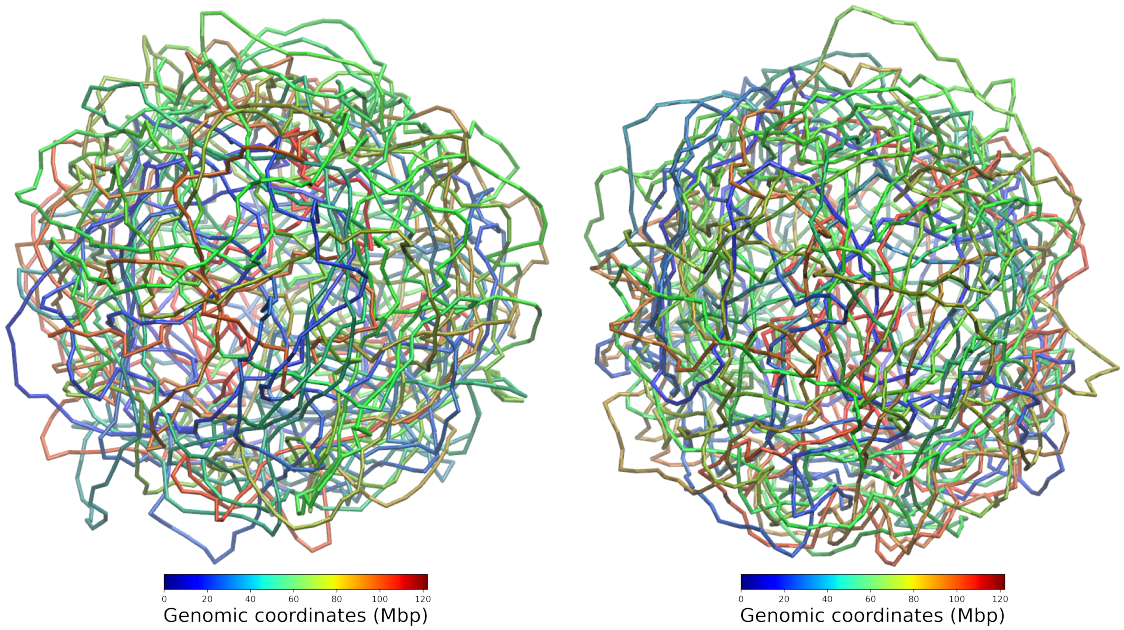
3.2 Model independently demonstrates intrinsic structure and folding properties of chromatin

The above results were obtained based on the initial input of the biological data provided during the model generation and incorporation of Hi-C step (discussed in section 2.4). In order to test the predictability, reliability and behaviour of our simu-



(a) Pre-Pro-B **(b)** Pro-B

Figure 3.3: Absolute difference plot between experiment and simulation-derived contact probability matrices The absolute difference plot shows maximum difference value of 0.004 and 0.006 between experimental and simulation-derived probabilities for (a) Pre-Pro-B and (b) Pro-B cells, respectively.



(a) Pre-Pro-B **(b)** Pro-B

Figure 3.4: Representative snapshot of chromosome 11 The regions of the chromosome have been coloured with respect to their genomic location for both (a) Pre-Pro-B and (b) Pro-B chromosome models.

lated model structures, we extended our investigation to explore some of the intrinsic properties of the chromatin which were independent of any implicit or explicit experimental inputs other than the Hi-C data used during the generation of the model.

3.2.1 Chromatin folding

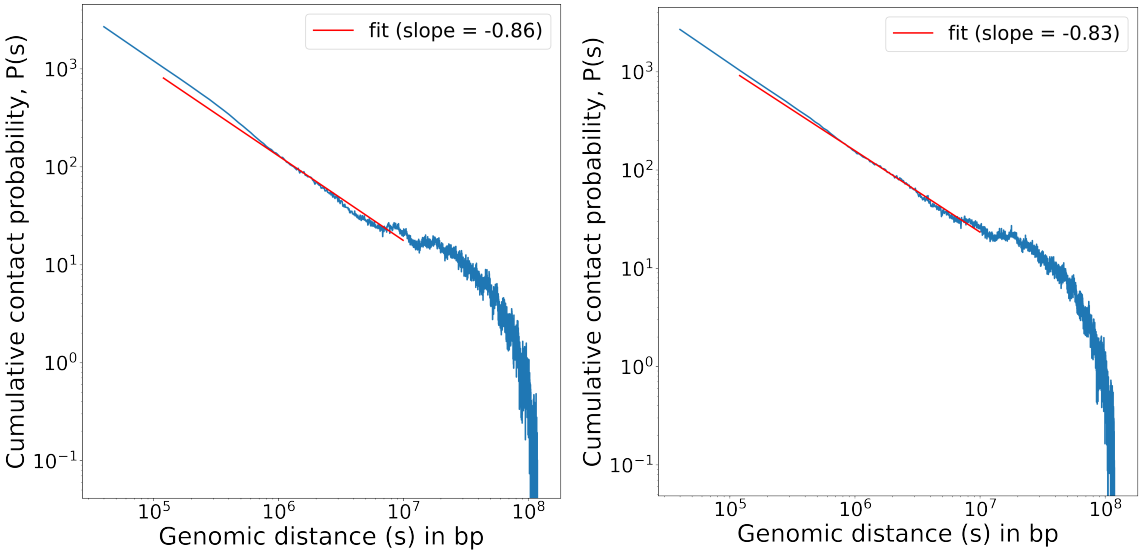
Starting with the highest chromosomal level of organisation, we first investigated the nature of folding of our simulated chromatin structures. It has been shown that the genome is fractal globule in nature unlike the equilibrium globule state as shown in previous studies [7, 84, 85] and also discussed in section 1.2.1. During polymer condensation, topological restrictions that forbid one part of the chain from crossing over another result in the development of a fractal globule, which is essentially a compact polymer state. If the attraction between the monomers is strong enough to overcome the effect of excluded volume repulsion or if the polymer is contained in a sufficiently tiny volume, the polymer instead experiences a coil-globule transition and becomes an equilibrium globule [85]. Based on this background, we carried out the analysis of chromatin folding for our simulated structures.

The standard method is by observing the scaling of contact probabilities $P_{(s)}$ as a function of genomic distance (s) which follow a power-law relationship also represented from its slope [7]. The scaling of s^{-1} means that loci two-fold farther apart having greater genome distance are two-fold less likely to interact with a smaller contact probability. To examine this relationship in our model, we plotted the intra-chromosomal contact probabilities as a function of genomic distance for both the simulated structures that represented the respective cell types (Figure 3.5). The inverse power law scaling with the slope of -0.86 ($s^{-0.86}$) and -0.83 ($s^{-0.83}$) was observed in case of Pre-Pro-B (Figure 3.5a) and Pro-B simulated structures (Figure 3.5b), respectively. These values are very close to the previously reported value of -1 (s^{-1}) for fractal nature of chromosomes as discussed above. These findings allowed us to

validate that the folding and local packing of the polymer structures we used to represent a single chromosome behaved in a way that was compatible with a fractal globule. Due to the fractal globule architecture of our simulated structures, our *in silico* chromatin structures would be able to function similarly to the *in vivo* chromatin structures, showing fast and extensive opening of genomic loci as well as their spatial mobility in the unfolded state. This essentially confirms the reliability of our model. Interestingly, if a chain folds as a fractal globule, each consecutive region of the chain (called subchain) occupies a distinct spatial region termed as genomic territories [85] that is a continuous and spatially compact genomic region with different regions occupying different spatial locations. This spatial segregation due to the fractal globule nature is exhibited at further scales, discussed at the level of chromatin compartments and TAD formations at the sub-chromosomal scale in the upcoming sections. At the nuclear scale, this segregation of subchains was shown to be analogous to the segregation of polymer rings formed due to the topological constraints, and it was proposed as a mechanism for the establishment of the chromosomal territories [35, 84]. Therefore, based on such convincing results on a single chain polymer model at the chromosomal level, we deduce that a genome-wide polymer model at a larger nuclear or genomic scale, generated using the approach implemented in our study would, certainly, demonstrate the formation of genome-wide architecture of chromosome territories observed in a cell nucleus.

3.2.2 3D modularity of chromatin

We were then interested to examine the effect of deletion of regions and comparing those partial chromatin regions with the entire chromosome in order to better understand the spatial modularity in chromatin folding. To do so, we considered different sizes of the chromatin polymer chains of N = 51, 101, 201, 501, 1001 and 2001 beads and simulated these self-avoiding chromatin chains of different lengths. These regions



(a) Pre-Pro-B **(b)** Pro-B

Figure 3.5: Chromatin folding prediction Plot of cumulative contact probabilities ($P_{(s)}$) as a function of genomic distance (s) with a slope (fit shown in red) of -0.86 for (a) Pre-Pro-B and -0.83 for (b) Pro-B which is close to the slope of -1.0 for a fractal globule structure.

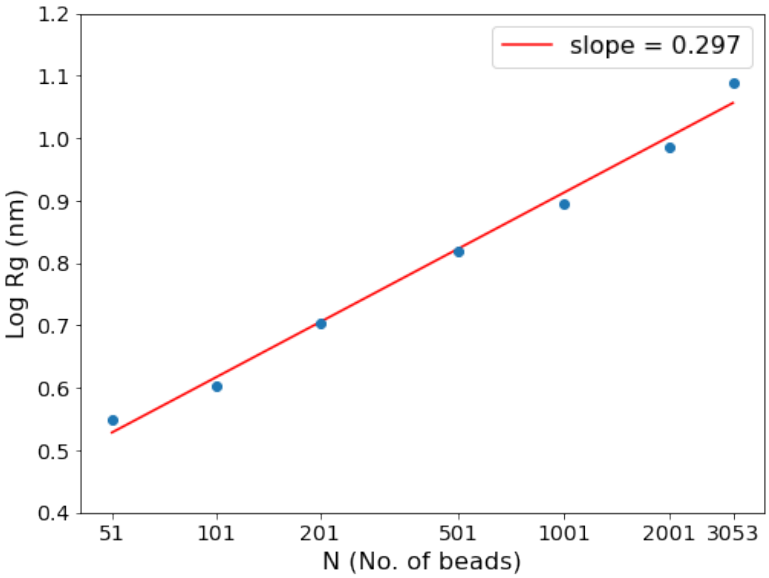
of variable lengths were selected starting from the centre of the chromatin polymer and extending at equal intervals towards the left and right of the chain. We estimated the radius of gyration, R_g, for these various values of N in order to derive the power law scaling from our simulations. The resulting plot is shown in Figure 3.6a. By taking into account how, for large values of N, the polymer's radius of gyration R_g behaves, one may apply the most popular critical exponent, the compactness index, ν, which has been studied previously for the three polymeric phases [112]. The value of ν corresponds to 1/2 for a random-walk polymer, 3/5 for a self-avoiding walk polymers without any restraints and 1/3 if the self-avoiding chain polymer is in collapsed phase inside a confined boundary. From our results in Figure 3.6a, we find that the chromatin behaves in a modular fashion with its R_g correlated as the slope corresponding to 0.297 which is slightly less than 1/3 for the collapsed state of self-avoiding walk polymer in a confinment. This is due to the presence of intra-chromosomal local interactions obtained from Hi-C that have been incorporated as weak harmonic bonds.

Hence, the value (of 0.297) is little lesser than the expected value of 0.33. It is deduced that even though our chromatin polymer faithfully follows the polymeric properties, it is influenced slightly due to the presence of genomic interactions that in turn govern its overall folding. This essentially implies chromatin folding in a bad solvent having more intra-polymeric interactions than polymer-solvent interactions which is certainly the case since the chromatin-chromatin interactions are more prevalent in deciding the chromatin organisation and arrangement inside the nucleus.

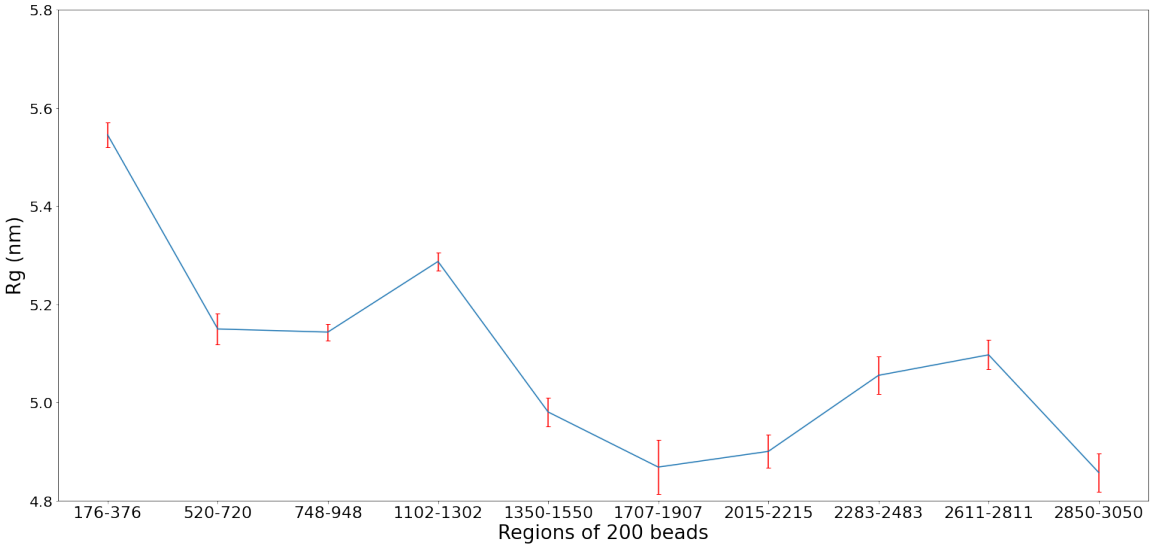
Since, we see that these interactions play a crucial role in governing the folding and dynamics of chromatin polymer, we were interested to investigate their nature of impact. For this, we plotted the R_g of different regions of same size (=200 beads) and examined its behaviour. From Figure 3.6b, we show that short-ranged local interactions are more prevalent than the long-ranged interactions. Also, there is heterogeneity in these interactions as is evident from the different values of R_g for the same size of polymer. It is the presence of these heterogenous local interactions that impacts the chromatin folding and three dimensional architecture of chromatin.

3.2.3 Chromatin state

Further at the next hierarchical chromatin organisation level, we investigated if our model can determine the states of the chromatin regions as transcriptionally active and inactive, i.e. A and B compartments, essentially corresponding to the euchromatin and heterochromatin regions respectively (discussed in section 1.2.2). Owing to the fractal globule nature of our simulated structures seen in the above section 3.2.1, we anticipated chromosomal segregation into compartments. To investigate that, we performed Principal Component Analysis (PCA) to our simulation derived contact probability matrices. It has been discussed earlier in section 2.8 that PCA is the canonical and the most popular method for identifying compartmental status of a given region, where the first principal component (PC1) or eigenvector captur-



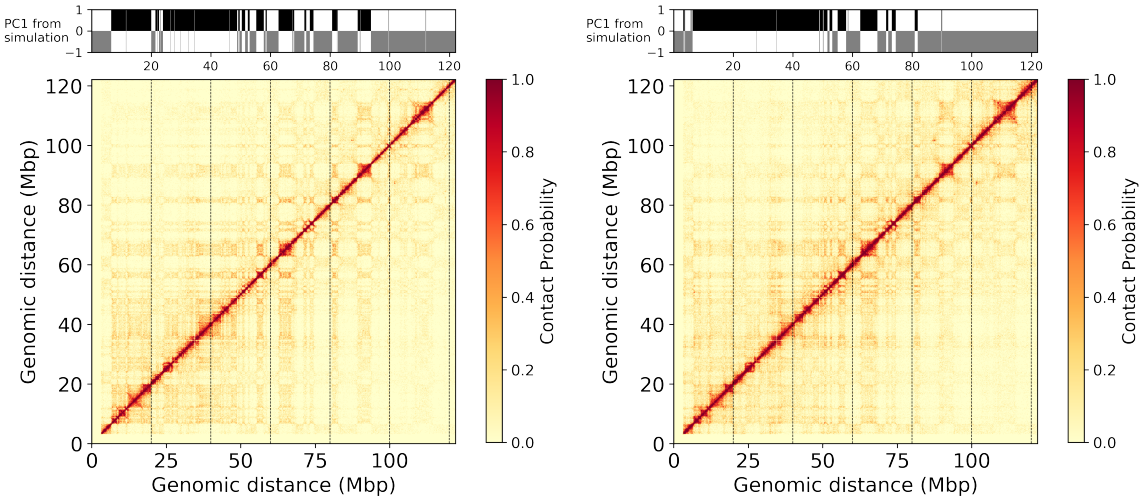
(a) Chromatin Modularity



(b) heterogeneity in local chromatin interactions

Figure 3.6: 3D Chromatin Modularity (a) Behaviour of chromatin regions in terms of their R_g values for different sizes of the polymeric chain. (b) Behaviour in terms of mean R_g values of different chromatin regions of same size. Standard deviation for each region is indicated as red error bars.

ing the dimension with the highest variance, is utilised to assess the region's A/B compartmental status. The PC1 has two sets of values assigned, i.e. positive and negative PC1 values. The permissive A compartment regions are represented by the positive PC1 values and the repressive B compartment areas are represented by the negative PC1 values [7]. Further, it was also shown that the areas with positive eigenvalue harboured more genes based on the investigations on the gene expression levels and the histone modifications carried out on both the positive and negative regions. This also resulted in the formation of more genomic interactions in the positive PC1 regions that were also captured in the Hi-C. Therefore, the corresponding gene expression levels were relatively high and these regions correlated well as the transcriptionally active or permissive A compartment regions. On the contrary, the negative eigenvalue is related to gene-poor regions with lesser genomic interactions, thus, correlating well with the transcriptionally inactive or repressive B compartment regions. These interactions were well represented through the checker-board pattern in the Hi-C matrix. We implemented the same concept to our simulation results to identify A/B compartments through PCA and simultaneously overlaid it with the experimental contact probability heatmap for its validation, as shown in Figure 3.7. The results clearly demonstrate the segregation of chromatin into A and B compartments where the positive PC1 values of the simulation data (*black* region in Figure 3.7) correlate to the A compartment in the experimental heatmap while the negative PC1 values (*grey* region in Figure 3.7) correspond to the B compartment in the experimental heatmap in case of both Pre-Pro-B (Figure 3.7a) and Pro-B (Figure 3.7b) simulated structures. This invariably confirms the correctness of our model structures wherein the chromatin status of different regions in the simulated chromatin model is predicted correctly in accordance with the corresponding chromatin status of those regions observed experimentally. Hence, we establish the efficiency of our model even at the megabase level of chromatin organisation.

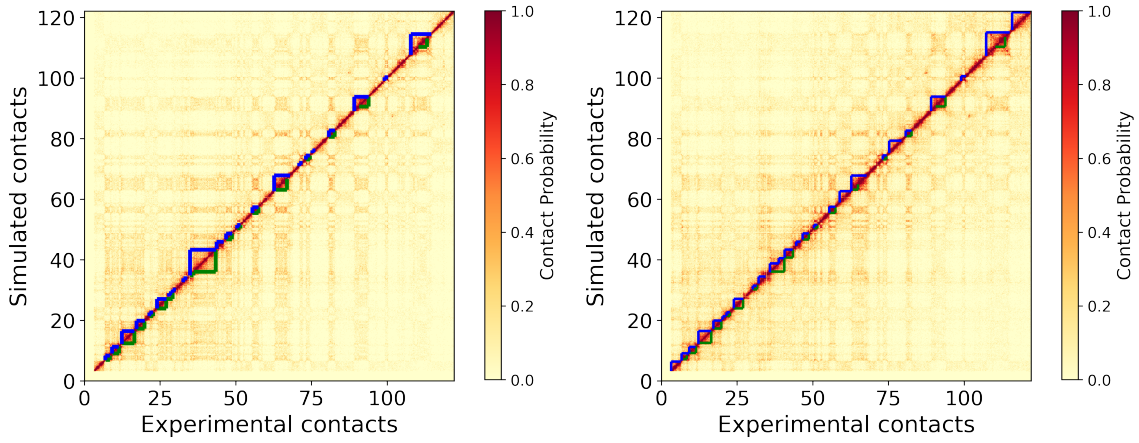


(a) Pre-Pro-B **(b)** Pro-B

Figure 3.7: Prediction of chromatin state Prediction of chromatin states from PCA (PC1 values) of the simulation derived contact probabilities is compared with the heatmap of experimental contact probabilities for (a) Pre-Pro-B and (b) Pro-B. The plot shows that the prediction of A compartments (or permissive regions) from the positive PC1 values (*black* region) in the simulation corresponds to the high contact probabilities in the experimental matrix while the prediction of B compartments (or repressive regions) from the negative PC1 values (*grey* region) in the simulation corresponds to the low contact probabilities in the experimental matrix.

3.2.4 Prediction of TADs

After the encouraging performance of our model at the sub-chromosomal level, we were interested to examine its behaviour at the sub-megabase level also. At this level, the chromatin is compacted and organised into highly self-interacting regions called TADs (discussed in section 1.2.3) where TAD boundaries are important in gene regulation. These can be seen as 'triangles' near the diagonal in the Hi-C contact heatmap. Owing to such a critical role in gene regulation, we were keen if our model could represent TADs and TAD boundaries accurately. There are a number of well-established TAD prediction tools, such as Arrowhead [36], TADbit [113], TADtree [114], TopDom [115] and many others. Based on the evaluation of many TAD callers and eventually choosing the one that produced the most consistent and visually pronounced TADs, we decided to use Armatus TAD caller [10] for our analysis. In Armatus, TADs are defined using algorithms that detect switches in the directionality of interactions. Figure 3.8a shows the results for TAD calling for Pre-Pro-B and Figure 3.8b for Pro-B that compares the results of the simulated structures with their corresponding experimental data. We observe that the number of domains predicted for simulated structure is 480 and 463 for experimental data in case of Pre-Pro-B while for Pro-B, the number of domains predicted for simulated structure is 511 and 410 for the experimental data. The positions of the corresponding TAD in simulated structure versus experimental data for both the cell types is remarkably similar as shown in Figure 3.8. This is a *bonafide* agreement of the simulated chromatin structures with the intrinsic feature of TAD formation in chromosomes, even at such small sub-megabase pair level.

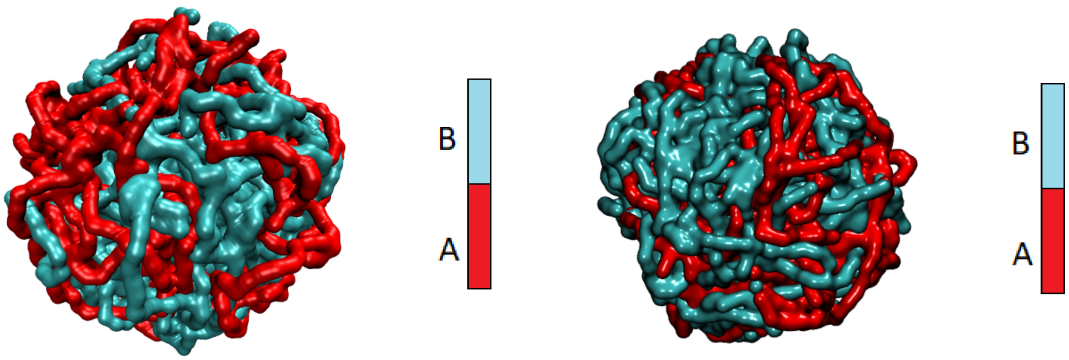


(a) Pre-Pro-B **(b)** Pro-B

Figure 3.8: Prediction of TADs We have used Armatus [10] to predict TADs from the simulated (blue triangles) structures and the experimental (green triangles) contacts obtained from Hi-C. The results are shown for (a) Pre-Pro-B and (b) Pro-B simulated structures representing the two cell stages. In this figure TADs which are at least 1 Mbp long have been shown.

3.2.5 Phase separation

Further, we examined for possible phase separation of A and B compartments in 3D space in our simulated structures. The dynamic phase separation of the genome had been proposed due to the flexible chromatin structure and movements [116]. Phase separation is the consequential effect in 3D arising initially as an outcome of chromatin folding as fractal globule resulting into segregation of genomic regions discussed in earlier sections. Further, due to the respective spatial constraints to allow for differences in interactions of transcriptionally active and inactive regions with other genomic regions, that these similar-state chromatin regions tend to co-localize and become phase separated. Phase separation may also result from a number of non-equilibrium processes occurring inside the cell nucleus, such as transcription, chromatin remodelling, and other processes, in addition to passive interactions caused by the various chromatin regulators and histone markers linked to regions of euchromatin and heterochromatin. The phase-separated multi-molecular assemblies have previously been



(a) Pre-Pro-B

(b) Pro-B

Figure 3.9: Phase Separation Phase separation of A and B compartments in the simulated structures for (a) Pre-Pro-B and (b) Pro-B cell stage. Similar compartment regions (A compartment in red and B compartment in cyan colour) co-localize leading to a phase separation of active (permissive) from inactive (repressive) regions.

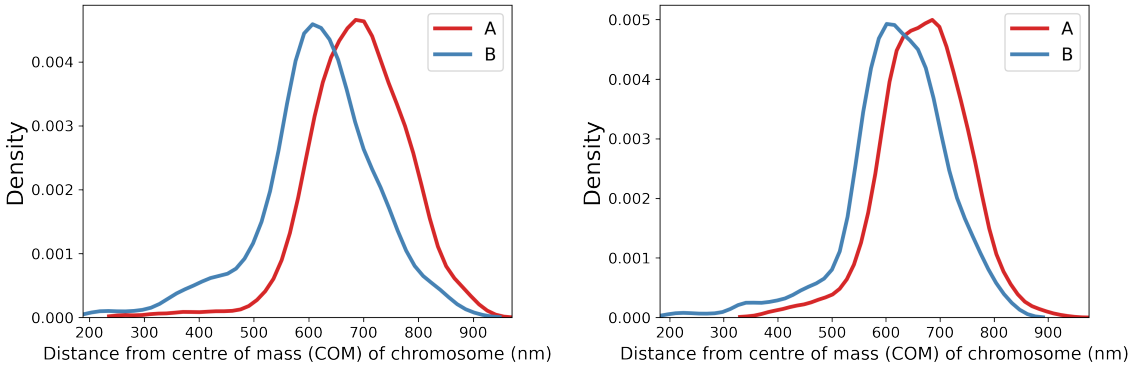
studied to provide a general regulatory mechanism of transcriptional control [116]. We try to investigate the phase separation in our simulated structures from Figure 3.9a for Pre-Pro-B and Figure 3.9b for Pro-B simulated structures and show that the similar compartment regions (A compartment in red and B compartment in cyan colour) co-localize leading to a phase separation of active (permissive) from inactive (repressive) regions. We generated these images by visualising the simulated structures in VMD. Although these results are after qualitative visual inspection only, it will be further interesting to observe differential patterns in these phase-separated compartments which we speculate to largely determine dynamic genome organisation and contribute towards cell fate decisions (discussed further in section 3.3).

3.2.6 Spatial positioning

Further, we were interested to quantitatively assess the preferential spatial positioning of these phase separated compartments in 3D space. To do so, we computed the mean distances of A and B compartments from the centre of mass (COM) of the simulated structures. The resultant plot in Figure 3.10a for Pre-Pro-B simulated

structure and Figure 3.10b for Pro-B, clearly shows that the A compartments have larger mean distance from the COM of the chromatin polymer indicating that these active regions tend to position themselves farther from the centre and towards the periphery of the chromatin. This positioning in 3D space in the exterior surface of the chromatin would allow easy accessibility of the genes harboured by these compartments to the transcriptional machinery of the cell, for their expression. On the other hand, the B compartments have smaller mean distance from the COM of the chromatin polymer indicating that these inactive regions are closer to the COM and are buried in the interior of the chromatin correlating to the inactivation of genes in those compartments. We can, thus, say that the phase separated compartments have preferential positioning in space which is directly related to their gene expression. Therefore, it can be deduced that our model's predictions on the state and positioning of its chromatin regions in 3D space are in-line with the theoretical phenomenon where the spatial arrangement of chromatin has a significant influence on the genome function. In the upcoming sections, we show dynamics of these regions as the cell differentiates during B-cell development.

Taken together, our findings show that our model is able to successfully capture and predict some of the very important characteristic features of chromatin architecture at different levels of chromatin organisation, such as folding of chromatin as a fractal globule, transcriptional state of chromatin resulting into compartmentalization into A/B compartments, formation of TADs and prediction of TAD boundaries, phase separation of similar chromatin state regions and the spatial positioning of the transcriptionally variable regions in context of the chromatin, all of which are independent of any biological inputs other than a small subset of the Hi-C interactions used during model generation and are entirely the resultant properties and behaviour of our generated simulated structures. Hence, we have established the reliability and predictability of our model. We, now, use these characteristics as our model's



(a) Pre-Pro-B **(b)** Pro-B

Figure 3.10: Preferential spatial positioning of compartments Mean distances of A and B compartments from the centre of mass (COM) of the simulated structures of (a) Pre-Pro-B and (b) Pro-B. In both the structures, A compartments have larger mean distance from the COM of the chromatin indicating their preferential positioning farther from the centre, at the chromosomal periphery while smaller mean distances of B compartments indicates that their preferential positioning is in the interior of the chromatin, closer to the centre of the chromatin polymer.

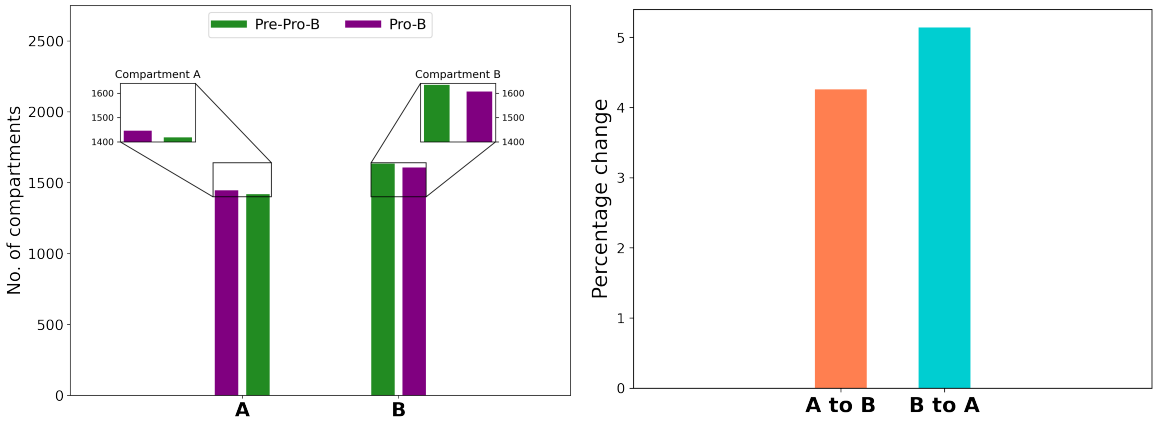
strengths and extend our investigation to further carry out comparative analysis of the two cell types and examine cell type specific differential changes.

3.3 Comparative analysis of simulated structures demonstrates lineage-dependent chromatin architecture

The remarkable agreement of the chromatin interactions, folding behaviour, compartmentalization and TAD formations between the simulated structures and the experimental data of the two cell stages, Pre-Pro-B and Pro-B, led us to extend the model's usage in comparing the chromatin organisation and capturing the structural alterations during cell differentiation that could not be captured in experiments. We proceeded to specifically probe spatial rearrangements of chromatin regions and reorganisation of chromatin architecture signifying functional implications as the cell progresses towards a committed cell stage during B-cell development.

3.3.1 Chromatin undergoes reorganization during B-cell development

We were interested in qualitatively investigating if there were any changes in chromatin organisation by doing the comparative analysis of the two simulated structures representing different stages of B-cell differentiation (Figure 3.11). We first plotted the number of compartments in both the cell stages as identified from their respective simulated structural models (Figure 3.11a). In the figure, it was observed that there is indeed a rearrangement of chromatin architecture in Pro-B cells as evident from the difference in the number of A and B compartments between the two simulated structures. This indicates differential transcriptional states of chromatin regions in the two cell types highlighting their contribution towards maintaining the cell identity and also responsible for governing cellular transitions. To further support it, we quantitatively investigated the number of compartments switching from A to B and B to A compartments and plotted the result in Figure 3.11b. The result confirms that although small, the chromatin undergoes compartmental switching as the cell differentiates from Pre-Pro-B to Pro-B stage during the B-cell development. We anticipate that this developmental change leads to the activation and repression events of lineage specific and multi-lineage genes, respectively, leading to switching of compartments between permissive and repressive states as the cell transitions towards a committed and differentiated cell stage (i.e. Pro-B) from an undifferentiated stage (i.e. Pre-Pro-B). The small-scale difference is justified because, firstly, this transition from an undifferentiated to a differentiated cell stage is a collective outcome of the differential changes contributed by all the chromosomes of the cell which weren't considered in our model. Hence, our results show only the contribution of the chromosome under consideration for this study, which is the sub-set of the concerted dynamics brought by the entire genome. Secondly, the two cell stages under consideration are otherwise



(a) Comparative analysis of compartments

(b) Compartmental rearrangement

Figure 3.11: Comparative analysis of simulated structures. (a) Comparison of differential number of compartments in Pre-Pro-B and Pro-B. (b) Compartmental rearrangement from A to B and vice versa during cell differentiation as the cell progresses from Pre-Pro-B to Pro-B stage.

very similar in their expressions except for the small yet crucial lineage dependent differential expressions. Therefore, the set of differential genes here undergoing the transitions could be very small as compared to the set of other house-keeping genes maintaining the similar state of expression in both the cell stages; but to be able to detect these changes has proved to be a phenomenal achievement by our model. Within the scope of this study, the model's performance is highly remarkable as it succeeds in detecting those crucial consequential changes (with limited initial parameters) that were very difficult to detect otherwise.

3.3.2 A/B compartmental switching promotes cell-type dependent genetic switch for B-cell fate commitment

After identifying chromatin rearrangements upon comparing the two structures representing the different stages during B-cell development, discussed in section 3.3.1 above, we further examined those specific regions which underwent the shift in their chromatin states and resulted in chromatin reorganisation. To do so, we compared the

compartmental status of the entire chromatin of both the simulated structures and identified those regions that showed compartmental switching which consequently, contributed to the differential functional state of the cell. In order to quantitatively identify these switched regions, we compared the PC1 values of both Pre-Pro-B and Pro-B structures (top two panels in Figure 3.12) and identified genomic regions that showed opposite signs in their corresponding PC1 values in the two cell types. The regions shown as green bars in the bottom-most panel in Figure 3.12 are the regions that switched from either permissive to repressive or repressive to permissive compartments in Pro-B cells. In total, >4% of the regions showed compartmental switching from permissive to repressive (A to B) compartments while >5% of the regions showed the reverse trend in Pro-B cell stage (Figure 3.11b). These results substantiate that the compartmental switching between A/B compartments correlates to cell type specific genetic switch. To further verify, it would be interesting to know the expression of genes harboured by these switched regions in order to establish functional relevance associated with the switching observed. To quantitatively carry out this examination, we first annotated these regions and performed expression analysis in the two cell types (discussed later in section 3.4). The goal was to identify if the switched regions possessed any lineage-specific or alternate-lineage genes that underwent compartmental switch from B to A and A to B compartments demonstrating gene activation and gene repression events, respectively, during the developmental transition from Pre-Pro-B to Pro-B cell stage.

3.3.3 Differential spatial positioning of switched compartments reveals dynamic structural rearrangements in chromatin

Within a single chromosomal territory, it has been observed previously that the inner region is comprised of more condensed chromatin domains, while a thin layer of more decondensed chromatin, known as the perichromatin region, can be found around

Figure 3.12: Identification of switched compartments. Comparison of PC1 values of both Pre-Pro-B and Pro-B cells is plotted in the first two panels. The regions shown as green bars in the third panel are the compartments that switched from permissive to repressive and vice versa in Pro-B cells.

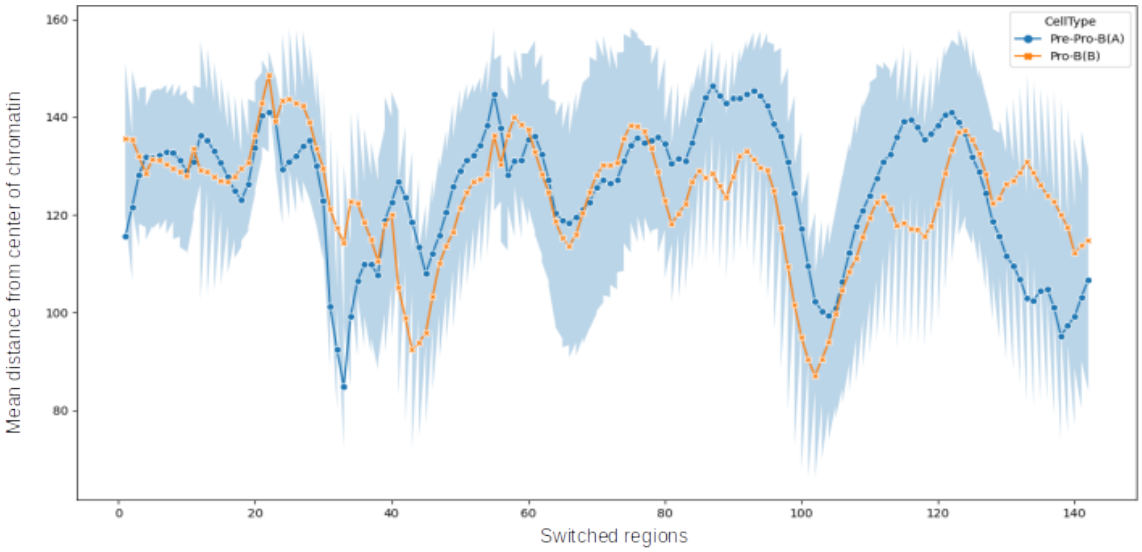
the chromosomal periphery [117]. Also, earlier in the section 3.2.6, we had seen that the spatial positioning of the genome critically impacts its function and that's why genomic regions have preferential positioning in 3D space corresponding to their chromatin state. The most active genomic regions preferentially lie at the surface of the chromosomal territory while the inactive regions are buried inside, which is also demonstrated in our simulated structures for both cells (Figure 3.10). Since we had observed a shift in the chromatin status of some of the genomic regions of the Pre-Pro-B cell, we were intrigued to investigate the corresponding changes in the spatial positioning in 3D of these switched regions in the two cell types from their respective structural models. Therefore, we tried to investigate the distance of these switched regions from the COM of the chromatin. We plotted the mean distance of all the switched bins (from both A to B and B to A) from the COM of the chromatin as shown in Figure 3.13. It was observed that the distance between the spatial positions of most of the regions switching from permissive A compartment in Pre-Pro-B (blue

dots in Figure 3.13a) to repressive B compartment in Pro-B cells (orange dots in Figure 3.13a) and the centre of mass of the chromatin, reduces in Pro-B indicating a shift in their spatial position towards the interior of the chromosomal territory. Since these regions show switching into repressive compartments, they also dynamically rearrange spatially and move from the periphery towards the chromatin interiors which is also indicative of the preferred position of inactive chromatin state.

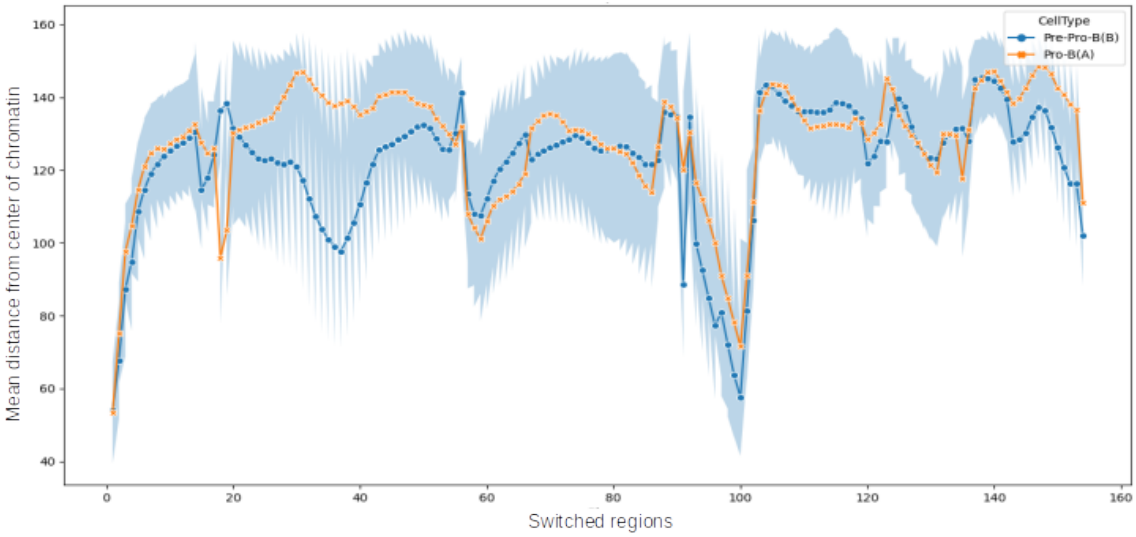
On the other hand, a reverse trend was observed where an increase in the distances between the regions switching from repressive B compartments in Pre-Pro-B (blue dots in Figure 3.13b) to permissive A compartments Pro-B cells (orange dots in Figure 3.13b) and the centre of mass of the chromatin suggested spatial rearrangement of regions in the repressive compartment residing at the interior, to permissive compartment shifting towards the periphery of the chromosomal territory which is indicative of the preferential positioning of activated chromatin state. Together, the overall investigation undoubtedly indicates that the genomic regions spatially rearrange themselves depending upon their acquired active or inactive status and dynamically move towards their preferential positions within the chromosomal territory. This clearly implies that the chromatin undergoes dynamic structural alterations in the Pro-B cell stage, orchestrating functional implications resulting in a committed B-cell stage.

3.3.4 Degree of compactness of switched regions corresponds to lineage-dependent alterations in chromatin structural framework

Further, in support of our previous findings, we moved forward to investigate if there exists any change in the compactness and folding of the switched regions. The compactness of a region measures the degree of openness or closeness in 3D space which is also associated to the chromatin state and function. In order to examine these



(a) Compartment switching from A to B

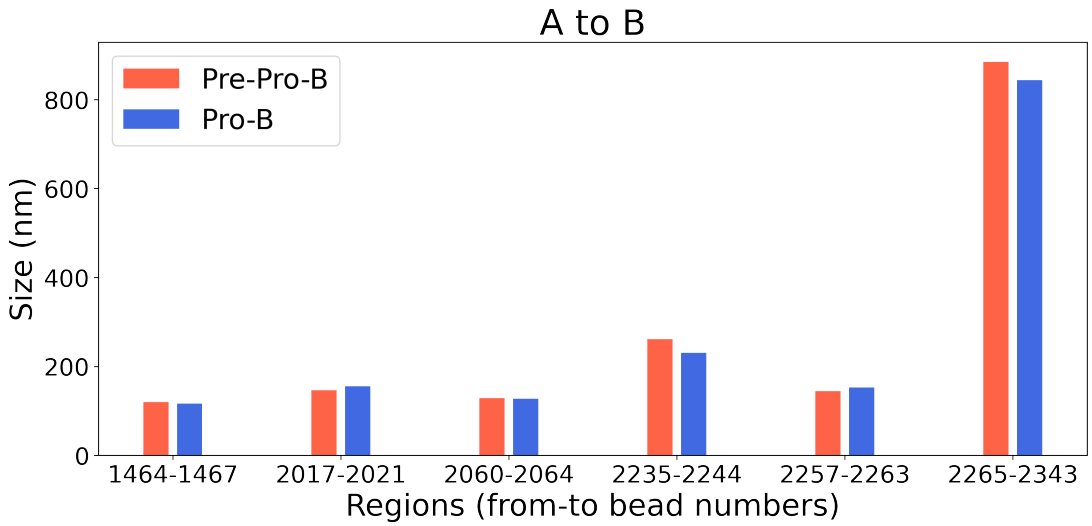


(b) Compartment switching from B to A

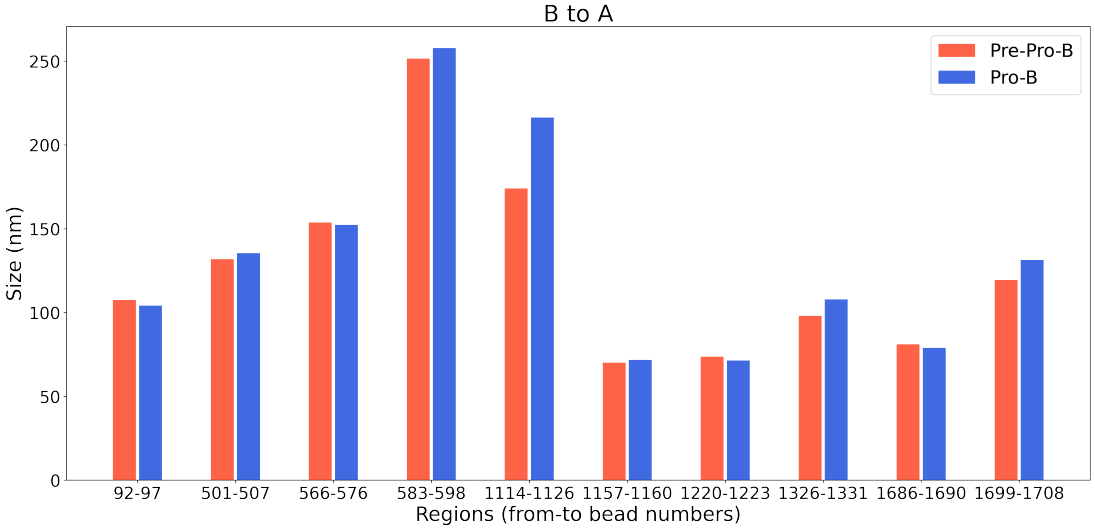
Figure 3.13: Spatial positioning of regions switching compartments. (a) Compartments switching from A compartment in Pre-Pro-B to B compartment in Pro-B. Blue dots indicate the distance from the COM of regions in the permissive compartment in Pre-Pro-B while orange dots indicate their distance from COM in the switched repressive compartments in Pro-B. (b) Compartments switching from B compartment in Pre-Pro-B to A compartment in Pro-B. Blue dots indicate the distance from the COM of regions in the repressive compartment in Pre-Pro-B while orange dots indicate their distance from COM in the switched permissive compartments in Pro-B.

differential structural changes in the switched regions of chromatin in Pro-B cell, we computed the radius of gyration, R_g (a popular metric in polymer-physics) which measures the compactness of a region that also correlates to the accessibility of that region; lower R_g value represent a more condensed or compacted state indicating an inactive repressed region while active permissive regions are less compacted and decondensed having a larger R_g value providing an easy access to the transcriptional machinery. As a first step to calculate R_g, we identified and selected regions showing compartmental switching consisting of a continuous stretch of atleast four beads. Then we computed the R_g of these regions in both Pre-Pro-B and Pro-B simulated structures. From Figure 3.14a, we find that the R_g value of most regions in the permissive compartment in Pre-Pro-B structure show a slight reduction when they switch to repressive compartment in Pro-B cell. It is to be noted that the effect is more pronounced and easily visible in regions that are longer in length (regions from 2235 to 2244 and from 2265 to 2343 bead in Figure 3.14a) than the regions of smaller length comprising of 4 beads (regions from 1464 to 146 and from 2060 to 2064), as the measurement of compactness makes more sense as the length of the region increases. The same holds true in case of figure 3.14b where an increase in the R_g value of most regions switching from repressive compartment in Pre-Pro-B cell to permissive compartment in Pro-B cell was observed. These results clearly show that the compartmental switching is favoured by relative change in the compactness of those regions where active regions in the Pre-Pro-B stage acquire a more compacted structure when they switch into inactive compartments in the Pro-B stage while the inactive regions in Pre-Pro-B open up and attain a comparatively less compacted decondensed structure when they switch to active compartments in Pro-B cells. Hence, we demonstrate a shift in the chromatin structural framework that governs the functional state as the cell differentiates into lineage-specific developmental stages.

We specifically show the above changes occurring in Ebf1, the master regulator



(a) R_g of regions switching from A to B

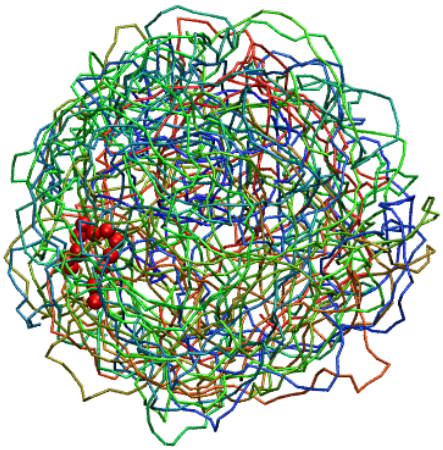


(b) R_g of regions switching from B to A

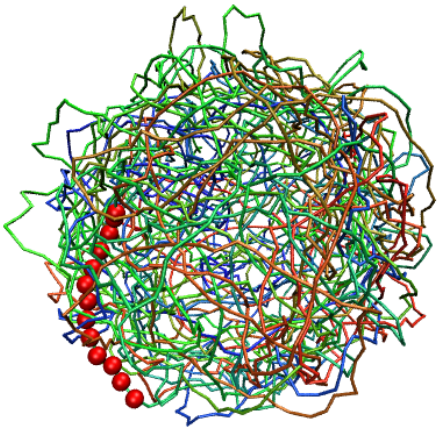
Figure 3.14: Analysis of compactness of regions that show compartmental switching. (a) Radius of gyration (R_g) of regions switching from permissive A compartment in Pre-Pro-B to repressive B compartment in Pro-B. (b) Radius of gyration (R_g) of regions switching from repressive B compartment in Pre-Pro-B to permissive A compartment in Pro-B.

and crucial factor for B-cell commitment [107,118], through our simulated structures in Figure 3.15 *(top)*. It is evident that the Ebf1 region in the repressive compartment of Pre-Pro-B structure (red beads in Figure 3.15a) is a compact region buried in the interior of the chromatin that rearranges itself towards the chromosomal surface and acquires an open chromatin state as it switches to permissive compartment in the Pro-B structure shown in Figure 3.15b. This provides clear evidence of the structural change in the chromatin organisation framework during differentiation having consequential functional implication of activation of lineage dependent gene, Ebf1, in Pro-B cells confirming B-cell fate commitment. Thus, through our model, we were able to show activation of lineage-dependent genes is related to 3D changes in the structure and architecture of chromatin.

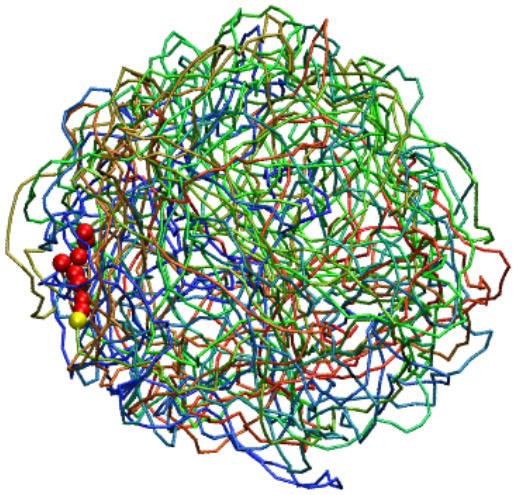
Similarly, Ccl11 is a chemokine gene from the CC subfamily that displays chemotactic activity for eosinophils only. It is an eosinophil-specific chemokine that has no significant functional relevance in the B-cell development and hence, is an alternate lineage-dependent gene. From the 3D positioning of Ccl11 (in read beads) in Figure 3.15c and 3.15d, it is distinctly evident that 3D spatial positioning of Ccl11 gene shifts from the exterior of the chromatin structure (in the undifferentiated Pre-Pro-B cell stage) and buries towards the interior in the committed Pro-B cell stage. Although small, a relative increase in the compactness of the gene is observed in the simulated structure of Pro-B stage indicating an inactivation event of the function of the gene. Together from these results of both the genes, we confirm the lineage-dependent dynamic structural alterations in the chromatin architecture during B-cell development.



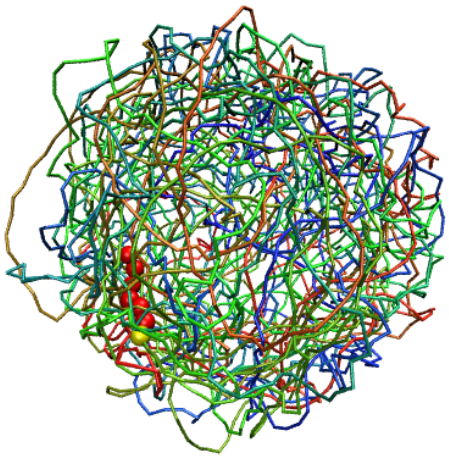
(a) Ebf1 in Pre-Pro-B



(b) Ebf1 in Pro-B



(c) Ccl11 in Pre-Pro-B



(d) Ccl11 in Pro-B

Figure 3.15: Differential 3D positioning of lineage dependent and alternate-lineage dependent gene examples. *(top)* 3D position of lineage-dependent gene Ebf1 (in *red* beads) in (a) Pre-Pro-B and (b) Pro-B simulated structures. *(bottom)* 3D position of alternate-dependent gene Ccl11 (in *red* beads) in (a) Pre-Pro-B and (b) Pro-B simulated structures.

3.4 Prediction of novel differential regions and their role in maintaining cell identity during differentiation

So far through our model, we were able to identify and compare the chromatin architectural changes between two cell types during differentiation. From the convincing performance of our model to accurately capture these intrinsic and differential features of the chromatin organisation, we extended its capabilities to predict novel differential genes that weren't captured in the experiments but showed evident changes in our simulated structures of the two cell types. We further supported our model's prediction through experimental validations in order to establish this predictive behaviour to our model's existing features.

We first annotated the genomic regions which showed compartmental switching, with genes from the publicly available data in UCSC (`http://genome.ucsc.edu`) and other published resources [119] in order to cross-examine their functional roles that can be associated to the observed compartmental switching. All the genes switching compartments from Pre-Pro-B to Pro-B, identified through our simulated structural model of Pre-Pro-B and Pro-B cells are listed in Table 3.1. Next, we compared this list with the publicly available RNA-seq expression data of both cell types [120] and identified genes in our results which showed differential patterns in the two cell types that were not captured in the RNA-seq expression data. We call the novel list of these genes as predicted exclusively from the simulated structures (Table 3.2).

Table 3.1: List of genes switching compartments from Pre-Pro-B to Pro-B, identified through the simulated structures of the two cell types

Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B
2210407C18Rik	0610010F05Rik, Gm12167, Myo1g
4930405D11Rik	1700030C12Rik, Gm12184, Nacad
4930507D10Rik	1700061J23Rik, Gm12185, Ntn1
4930527B05Rik	1700093K21Rik, Gm12188, Nudcd3
5530401A14Rik	2610024D14Rik, Gm12192, Olfr1393
Ankfn1	4921536K21Rik, Gm12193, Olfr1396
Asic2	4930512M02Rik, Gm12194, Olfr56
Car10	8430429K09Rik, Gm12195, Osbp2
Ccl11	9130017K11Rik, Gm12196, Papolg
Ccl12	9130230N09Rik, Gm12208, Peli1
Ccl2	9230020A06Rik, Gm12209, Pex13
Ccl7	9530068E07Rik Gm12210, Phykpl
Ccl8	9930111J21Rik1, Gm12235, Pik3ip1
Cox11	Actr2, Gm12301, Pla2g3
Fam183b	Aftph, Gm12303, Psme2b
Gm11207	Ahsa2, Gm12304, Pus10
Gm11416	Atox1, Gm12305, Rab1a
Gm11417	B3gnt2, Gm12592, Rack1
Gm11419	C78197, Gm16170, Rasgef1c

Continued on next page

Table 3.1: List of genes switching compartments from Pre-Pro-B to Pro-B, identified through the simulated structures of the two cell types (Continued)

Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B
Gm11494	Canx, Gm16518, Rel
Gm11498	Cby3, Gm20169, Rnf130
Gm11500	Ccm2, Gm20456, Rnf185
Gm11501	Cct4, Gm22600, Rpl12-ps2
Gm11502	Cct4, Gm22753, Rufy1
Gm11506	Cep68, Gm22807, Selenok-ps1
Gm11511	Cfap52, Gm22990, Selenom
Gm11512	Cnot6, Gm23114, Sertad2
Gm11516	Col23a1, Gm23492, Slc1a4
Gm12251	Commd1, Gm23582, Slc35e4
Gm12252	Cyfip2, Gm23681, Slc36a1
Gm12253	Dusp18, Gm23772, Slc36a1os
Gm12254	Ebf1, Gm23813, Slc36a2
Gm12255	Efcab9, Gm23827, Slc36a3
Gm12570	Ehbp1, Gm24013, Slc36a3os
Gm17268	Eif4enif1, Gm24313, Smtn
Gm22599	Eml6, Gm24398, Snora5c
Gm22702	Fam161a, Gm24439, Snord95
Gm22762	Fam71b, Gm24917, Snord96a

Continued on next page

Table 3.1: List of genes switching compartments from Pre-Pro-B to Pro-B, identified through the simulated structures of the two cell types (Continued)

Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B
Gm24612	Fbxw11, Gm25296, Stk10
Gm24856	Fstl4, Gm26157, Stx8
Gm25113	G3bp1, Gm26253, Tbc1d9b
Gm31522	Gas7, Gm26393, Tbrg4
Hlf	Gfpt2, Gm27194, Tcn2
Kif2b	Gm10428, Gm27517, Trim41
Lypd8	Gm11186, Gm27624, Trim7
Lypd8l	Gm11189, Gm27640, Tug1
Lypd9	Gm11944, Gm27937, Ugp2
Myo1d	Gm11945, Gm28048, Usp34
Olfr224	Gm11948, Gm30942, Usp43
Olfr30	Gm11949, Gm33351, Vps54
Olfr311	Gm11950, Gm3718, Wap
Olfr312	Gm11951, Gm40824, Wdpcp
Olfr313	Gm11952, Gm47279, Wsb2-ps
Olfr314	Gm11973, Gm51877, Xpo1
Olfr315	Gm11998, Gm5431, Zfp287
Olfr318	Gm12030, Hnrnph1, Zrsr1
Olfr319	Gm12031, Hspa4

Continued on next page

Table 3.1: List of genes switching compartments from Pre-Pro-B to Pro-B, identified through the simulated structures of the two cell types (Continued)

Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B
Olfr320	Gm12034, Ifi47
Olfr322	Gm12035, Inpp5j
Olfr323	Gm12036, Irgm1
Olfr324	Gm12037, Itk
Olfr325	Gm12038, Lcp2
Olfr326-ps1	Gm12039, Lgalsl
Olfr328	Gm12040, Limk2
Olfr329	Gm12041, Mapk9
Olfr329-ps	Gm12042, Mdh1
Olfr330	Gm12043, Med7
Olfr331	Gm12044, Mgat1
Olfr332	Gm12055, Mir1933
Olfr333-ps1	Gm12056, Mir340
Spaca3	Gm12057, Mir3470a
Stxbp4	Gm12058, Mir6406
Tmem132e	Gm12061, Mir804
Tmem98	Gm12062, Morc2a
Tom1l1	Gm12158, Mup-ps22
Trim58	

Table 3.2: List of novel predictions of genes exhibiting compartmental switching in Pro-B, exclusively identified in the simulated structural models

Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B
2210407C18Rik, Lypd8l	4921536K21Rik, Gm27194
4930405D11Rik, Lypd9	8430429K09Rik, Gm27517
4930507D10Rik, Myo1d	9130017K11Rik, Gm27624
Ankfn1, Olfr224	9230020A06Rik, Gm27640
Car10, Olfr30	9530068E07Rik, Gm27937
Ccl11, Olfr311	9930111J21Rik1, Gm33351
Ccl12, Olfr312	Atox1, Gm51877
Ccl2, Olfr313	C78197, Hnrnph1
Ccl7, Olfr314	Canx, Hspa4
Ccl8, Olfr315	Cby3, Inpp5j
Cox11, Olfr318	Cfap52, Limk2
Fam183b, Olfr319	Dusp18, Mir3470a
Gm11207, Olfr320	Eif4enif1, Mir6406
Gm11419, Olfr322	G3bp1, Mir804
Gm11494, Olfr323	Gm10428, Morc2a
Gm11498, Olfr324	Gm11189, Mup-ps22
Gm11500, Olfr325	Gm11944, Myo1g

Continued on next page

Table 3.2: List of novel predictions of genes exhibiting compartmental switching in Pro-B, exclusively identified in the simulated structural models (Continued)

Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B
Gm11501, Olfr326-ps1	Gm11945, Ntn1
Gm11502, Olfr328	Gm11948, Nudcd3
Gm11506, Olfr329	Gm11949, Osbp2
Gm11511, Olfr329-ps	Gm11950, Phykpl
Gm11512, Olfr330	Gm11951, Pik3ip1
Gm11516, Olfr331	Gm11952, Pla2g3
Gm12251, Olfr332	Gm11973, Psme2b
Gm12252, Olfr333-ps1	Gm12062, Rnf185
Gm12253, Spaca3	Gm12194, Selenom
Gm12254, Stxbp4	Gm12195, Slc35e4
Gm12255, Tmem132e	Gm12196, Slc36a1
Gm12570, Tmem98	Gm12208, Slc36a1os
Gm17268, Tom1l1	Gm12235, Slc36a2
Gm22599, Trim58	Gm12301, Slc36a3
Gm22702,	Gm12303, Slc36a3os
Gm22762,	Gm12304, Smtn

Continued on next page

Table 3.2: List of novel predictions of genes exhibiting compartmental switching in Pro-B, exclusively identified in the simulated structural models (Continued)

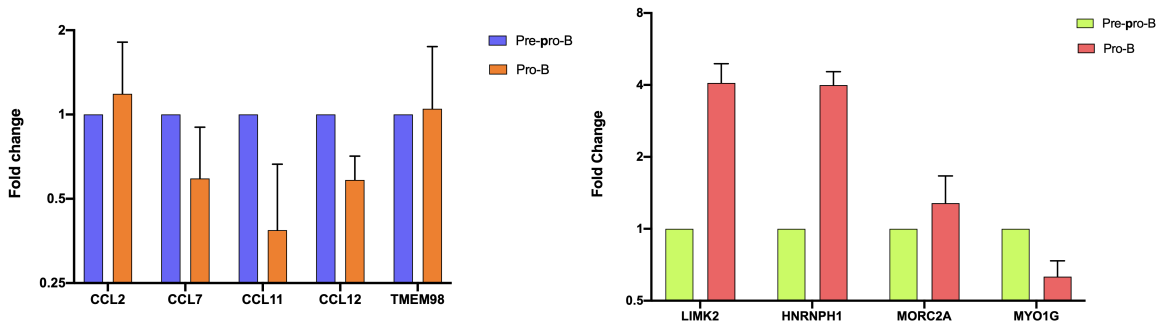
Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B
Gm24612,	Gm12305, Stx8
Gm24856,	Gm12592, Tbc1d9b
Gm25113,	Gm16518, Tcn2
Gm31522,	Gm20169, Trim41
Hlf,	Gm24013, Tug1
Kif2b,	Gm24439, Usp43
Lypd8,	Gm26157, Wsb2-ps
	Gm26393, Zfp287

Permissive (A) in Pre-Pro-B to Repressive (B) in Pro-B	*Repressive (B) in Pre-Pro-B to Permissive (A) in Pro-B*
Ccl2	Limk2
Ccl7	Hnrnph1
Ccl11	Morc2a
Ccl12	Myo1g
Tmem98	

Table 3.3: Selected list of genes for experimental validation through RT-PCR

3.4.1 Quantitative validation of the exclusively predicted genes

Next, we functionally annotated these genes and selected a few genes (Table 3.3) to be analysed quantitatively through RT-PCR for an experimental validation. We



(a) Permissive to Repressive

(b) Repressive to Permissive

Figure 3.16: Quantitative analysis of annotated genes by RT-PCR. (a) Genes predicted to switch from permissive to repressive (A to B) compartments show downregulation in Pro-B cells due to their involvement in the development of alternate-lineages. (b) Genes predicted to switch from repressive to permissive (B to A) compartments show upregulation in Pro-B cell which is a B-cell committed cell stage and marks the expression of B-cell related genes.

found that the majority of the predicted genes in regions switching from permissive to repressive compartments were related in the developmental expression in alternate lineage immune cells. For example, genes such as Ccl7, Ccl11 and Ccl12 are not expressed in Pro-B cells but in neutrophils, mast cells, macrophages and other cells of alternate lineages. On the other hand, genes in regions switching from repressive to permissive compartments had characteristic roles in the development and maintaining the identity of B cell and begin to express in the Pro-B cell stage. The same trend has been observed through the results of RT-PCR analysis showing the reduced expression of genes involved in the development of alternate lineages as they switch from permissive to repressive compartments in Pro-B cells (Figure 3.16a) while B-cell related genes harboured by the regions switching from repressive to permissive compartments are upregulated in Pro-B cell stage (Figure 3.16b).

Through these results, the prediction of gene switching has been duly validated and we were able to confirm the predictive potential of our model.

Chapter 4

Discussion

The complex yet indispensable relationship between chromatin architecture and its impact on the functional state of the cell has been an interest of research for many years in gaining insightful learnings on the underlying mechanisms associated with different cellular functions ranging from cell development, differentiation, maintenance, cell repair etc. This can be better understood by appreciating the three-dimensional organisation of the genome that mediates genomic interactions in 3D nuclear space to bring out the desired functional implications. Hence, a lot of work including experimental and computational studies has been carried out and continues to improve our knowledge on this aspect.

It is understood that despite each cell having an identical genetic makeup, there exists cell type-specific gene expression patterns that play an immensely crucial role in the variety of cellular events. The investigation of these cell-type specific patterns in the context of 3D during the process of cell differentiation was the overall aim of this study. In particular, we looked into the three-dimensional structural architecture and dynamics of chromatin during the formation of B-cells through mechanistic modelling using a combinatorial approach of polymer physics and high-throughput chromosomal conformation capture Hi-C data. We presented a computational model that not only captured the hierarchical structural organisation but also provided mechanistic insights into the spatial rearrangements of chromatin during developing lymphoid lineage cells.

Through this study, we were able to show the *spatial dynamics* and 3D transitional rearrangements in chromatin organisation upon cell differentiation, which were not possible through data-driven reconstruction-based modelling approaches that are limited to reconstruction of only static chromatin structures based on the input provided. Also, benefitting from our polymer-based predictive approach, we went ahead to make significant differential structural predictions through our simulated structures, which could not be captured in other high-throughput experiments but were

detected by our simulated structures.

To understand the chromatin's 3D alterations during cell differentiation, we generated a homopolymeric bead-on-a-string model that could represent the chromatin of respective cell stages and help in understanding how it evolves in 3D through simulations. The model structures represent chromosome 11 of mouse that essentially consisted of spherical non-overlapping beads with each bead mapping a genomic region of 40kb defining the resolution of our model. We show that our simulated structures succeed in independently recapitulating the salient features of different levels of chromatin architecture and additionally, help in identifying the cell-type specific 3D organisation of the chromosome between the two different cell stages considered in our study. In particular, it faithfully reproduced all the considered Hi-C interactions while showing remarkable agreement to chromatin interactions that were not included while generating the initial model structures. Further, we were able to show the intrinsic features of chromatin organisation including folding and local packing as a fractal globule, compartmentalization into permissive A and repressive B compartments and formation of TADs even at the sub-chromosomal scale. The model's predictions were in agreement for long-range interactions with some amount of noise observed for short-ranged chromatin interactions. These results established the integrity of our model with minimalistic inputs without relying on the proximity-based experimental data. Additionally, in 3D space, we demonstrated through our simulated structures the spatial dynamics and positioning of chromatin into phase separated regions based on their similar chromatin states. Through the mean distances of different regions from the centre of the chromosome, we confirmed that the preferential position in 3D of permissive regions is at the periphery of the chromosomal territory, while repressive regions tend to reside at the chromosomal interiors.

After the successful predictions of the 3D chromatin structures & organisation, we further extended its use in investigating the cell type specific differential changes by

performing a comprehensive comparative analysis of the two cell types of the B-cell developmental stages. Our model revealed that chromatin undergoes compartmental switching and dynamic 3D spatial rearrangements during cell differentiation towards B-cell commitment. Although the transitions of lineage specific genes were observed to be small as compared to other genes maintaining the similar state of expression in both the cell stages, yet being able to detect these changes with the help of our model, has proved to be a phenomenal achievement. Within the scope of this study, the model's performance is highly remarkable as it succeeded in detecting those crucial consequential changes (with limited initial parameters) that were otherwise very difficult.

From the investigations of compactness of switched regions, we showed that the genomic regions acquire an open or closed state depending on their switched active or inactive status and dynamically move in 3D space towards their preferential positions within the chromosomal territory. Our results clearly implied that chromatin undergoes dynamic structural alterations in the Pro-B cell stage, orchestrating functional implications resulting in a committed B-cell stage.

A major advantage of our model is that it further allowed us to make important structural and functional predictions about chromatin rearrangements & folding and its relationship with gene regulation which would not have been detected by simple qualitative examination of the Hi-C data. We were able to predict switching of novel regions from permissive to repressive and vice versa during cell differentiation through our simulated structures. These cell type specific chromatin organisation predictions were further quantitatively validated *in vitro*. The role of the genes in the predicted regions showing downregulation in Pro-B cells is largely associated with alternate lineage development related events, confirming the cell's commitment towards B-cell fate, thereby also confirming the reliability and predictivity of our model. This predictive model, thus, presents a significant leap forward in understanding the 3D chro-

matin architecture and *in silico* study of the 3D chromatin architecture and dynamics of differentiated versus undifferentiated cells during development of lymphoid-lineage cells.

In spite of the successful performance of our model, there are few constraints to which our model is limited to. First, is the resolution of coarse-graining, i.e. 40kb, in the current study that depends on the resolution of the Hi-C data considered. Although, the model shows phenomenal achievement in capturing the 3D organisation and detecting significant spatial changes, a higher resolution can help in throwing more light on the structural changes of genes and *cis*-regulatory elements that are smaller than 40kb in size, along with the associated gene-gene or enhancer-promoter interactions. Presently, a bead represents a genomic size of 40kb that may harbour many genes. Therefore, changes in one bead may account for the cumulative effect of all the genes in that bead which might be unnecessary. However, a higher resolution would provide granularity and upgrade the scope of the model to single-gene level and its regulation through epigenetic mechanism by treating each gene as a single bead. Certainly, this would result in heterogeneity in size of each bead which might require further assumptions. Additionally, since the resolution considered is much lower than the magnitude of persistence length of a chromatin polymer, analysis of the bending and stiffness of the chromatin is beyond the scope of this study and cannot be studied in the context of interpreting the structural aspects of gene regulation through loop formations. Increasing the resolution close to the nucleosomal scale might prove useful to understand it better. Nevertheless, it can not be neglected that a balance between computational cost and further enhancement of this model would be a major challenge. Second, is the number of chromosomes considered in our investigations. It is not denied that cellular changes are a collective outcome brought by the entire genome. However, in our study, we were restricted to perform our analysis on a single chromosome due to the high computational costs and increased

complexities upon incorporation of more chromosomes. Modelling of an entire genome would definitely provide multiscale inter-chromosomal information and understanding of genome-wide structural dynamics while simultaneously facing difficulties arising from it, such as the heterogeneity in polymer sizes, adjustment of corresponding model parameters, complexities arising due to increased interactions and associated noise, computationally expensive simulations etc, to mention a few. Hence, we limited our present work to a single chromosomal analysis for a start since the genome-wide 3D study of chromatin organisation is indeed challenging requiring various compromises at different levels.

Chapter 5

Conclusion

We have introduced a novel computational model for studying the 3D chromatin architecture by integrating high-throughput chromosome conformation capture data with polymer modelling. Our combinatorial approach of this hybrid-model has helped in overcoming the limitations of previously reported direct and indirect modelling techniques, while retaining the best of both for amalgamation into a powerful predictive tool.

This computational model aided in quantitatively understanding the chromatin organisation & dynamics in 3D unlike the chromatin conformation capture and other genomic and biochemical assays that only give a 2D depiction of it. The results gave broad insights into the structural dynamics of architectural changes via the 3D spatial rearrangements of genomic regions during the quantitative study of differentiation towards B-cell fate, as we carried out a comprehensive comparative analysis of two different cell types representing undifferentiated and differentiated cell stages.

The findings from our study suggest that:

a. Chromatin undergoes compartmental switching and dynamic 3D spatial rearrangements during cell differentiation for B-cell commitment. Upon directly comparing the regions of differential compartmental status in the two cells, we found that the chromatin in Pro-B cell stage undergoes spatial repositioning and changes in its compactness, corresponding to the switched compartmental status and consequential 3D regulation of gene expression in those regions. This suggested a coordinated chromosomal dynamics towards B-cell fate commitment.

b. Importantly, we also established the predictive nature of our model by identifying regions associated with these alterations that were not detected in the experiments and were further quantitatively verified *in vitro*.

c. Collectively, our studies demonstrate that during B-cell fate commitment, dynamic three-dimensional re-organization of chromatin induces lineage-specific gene expression patterns. Thus, our prediction model represents both an *in silico* study

of the 3D chromatin architecture & dynamics and a substantial advancement in our knowledge of differential 3D chromatin organisation.

d. There exists a scope for potentially extending this study for an entire genome-wide structural analysis at a higher resolution which should address the principles of 3D organisation at a multi-scale level while parallelly maintaining the integrity, reliability and predictability of our model. These future studies should also address chromatin interactions and interpretation at the single-cell level.

e. Besides, the study opens up a plethora of other horizons to expand the dimensions of this research. To start with, a few factors can be enhanced to build upon the approach considered in our study, for example, improving the resolution of Hi-C data and incorporation of other chromosomes for a genome-wide multi-scale analysis. Additionally, inclusion of *trans*-regulatory elements such as TFs and other cellular proteins as separate entities during model generation along with their epigenetic landscape can also be helpful in providing improved mechanistic insights in studying the epigenetics of gene regulation in the context of three-dimensional understanding of the events involved. Recent advances in high throughput experiments has evolved into single-cell Hi-C (scHi-C) data that can also be utilised in future for enhancing the existing model rather than the ensemble averaged-based study. scHi-C can help in identifying cell-to-cell variability of 3D chromatin organisation, however the sparseness of measured interactions can pose an analysis challenge.

Clearly, this fascinating voyage is only getting started. In the future, with the improvement in the data that is more extensive and sophisticated, we will be able to better understand the underlying mechanism of many cellular functions in the context of 3D. It would be exciting to see the ever-evolving modelling strategies and the enhancement of those that prove to be most promising. The close relationship between computational modelling and experimentation will aid in elucidating the mechanisms governing genomic architecture and its relationship to biological processes. The up-

coming boom of Artificial Intelligence and Machine Learning (AI/ML) might soon overcome the present hurdles and is expected to introduce smarter ways to handle and include the multi-omic data and improve model interpretability to deepen our understanding of the mechanisms governing gene regulation via the prism of 3D genome organisation.

Highlights of the study:

- Hybrid polymer model for chromatin defines the chromatin structure and properties using minimalistic experimental inputs.
- The predictive nature of the model demonstrated intrinsic features of chromatin folding, hierarchical organisation and co-localization of similar regions in 3D nuclear space which were independent of proximity based experimental inputs.
- Comparative analysis of simulated structures demonstrated that chromatin undergoes lineage-dependent chromatin reorganisation during B-cell fate commitment.
- Cell type specific spatial rearrangement showed transition of repressive compartments towards the periphery and permissive compartments towards the chromatin interiors, as they switch into permissive and repressive compartments respectively, in differentiated & committed B-cell stage.
- Chromatin dynamics showed changes in the compactness of the switched regions that is associated with their acquired functional state.
- Identification of novel regions through comparative study of the two simulated structures and its quantitative experimental validation revealed that the role of genes, associated to the switching into repressive compartments, shown to be

downregulated in Pro-B, is largely associated with alternate lineage development, confirming the cell's commitment towards B-cell fate.

- Compartmental switching promotes cell-type dependent genetic switch via 3D spatial rearrangement for B-cell fate commitment.

Bibliography

[1] J. M. R. Pongubala and C. Murre, "Spatial organization of chromatin: Transcriptional control of adaptive immune cell development," *Frontiers in Immunology*, vol. 12, 3 2021.

[2] R. P. McCord, "McCord Lab Deciphering 3D Genome Structure."

[3] G. Bonora, K. Plath, and M. Denholtz, "A mechanistic link between gene regulation and genome architecture in mammalian development," *Current Opinion in Genetics and Development*, vol. 27, pp. 92–101, 8 2014.

[4] F. Li, Z. An, and Z. Zhang, "The dynamic 3d genome in gametogenesis and early embryonic development," *Cells*, vol. 8, 8 2019.

[5] B. Bonev and G. Cavalli, "Organization and function of the 3d genome," *Nature Reviews Genetics*, vol. 17, pp. 661–678, 11 2016.

[6] J. J. Parmar, M. Woringer, and C. Zimmer, "How the genome folds: The biophysics of four-dimensional chromatin organization," *Annual Review of Biophysics*, vol. 48, pp. 231–253, 5 2019.

[7] E. Lieberman-Aiden, N. L. V. Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos,

L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, pp. 289–293, 10 2009.

[8] S. V. Ulianov, A. A. Gavrilov, and S. V. Razin, "Nuclear compartments, genome folding, and enhancer-promoter communication," *International Review of Cell and Molecular Biology*, vol. 315, pp. 183–244, 2015.

[9] S. Takagaki, R. Yamashita, N. Hashimoto, K. Sugihara, K. Kanari, K. Tabata, T. Nishie, S. Oka, M. Miyanishi, C. Naruse, and M. Asano, "Galactosyl carbohydrate residues on hematopoietic stem/progenitor cells are essential for homing and engraftment to the bone marrow," *Scientific Reports 2019 9:1*, vol. 9, pp. 1–11, 5 2019.

[10] D. Filippova, R. Patro, G. Duggal, and C. Kingsford, "Identification of alternative topological domains in chromatin," *Algorithms for molecular biology : AMB*, vol. 9, 5 2014.

[11] J. Dekker and L. Mirny, "The 3d genome as moderator of chromosomal communication," *Cell*, vol. 164, pp. 1110–1121, 2016.

[12] H. K. Long, S. L. Prescott, and J. Wysocka, "Ever-changing landscapes: Transcriptional enhancers in development and evolution," *Cell*, vol. 167, pp. 1170–1187, 11 2016.

[13] T. Misteli, "Higher-order genome organization in human disease.," *Cold Spring Harbor perspectives in biology*, vol. 2, pp. a000794–a000794, 8 2010.

[14] M. Spielmann, D. G. Lupiáñez, and S. Mundlos, "Structural variation in the 3d genome," *Nature Reviews Genetics*, vol. 19, pp. 453–467, 2018.

[15] D. G. Lupiáñez, M. Spielmann, and S. Mundlos, "Breaking tads: How alterations of chromatin domains result in disease," *Trends in Genetics*, vol. 32, pp. 225–237, 4 2016.

[16] C. Anania and D. G. Lupiáñez, "Order and disorder: Abnormal 3d chromatin organization in human disease," *Briefings in Functional Genomics*, vol. 19, pp. 128–138, 3 2020.

[17] W. A. Bickmore and B. V. Steensel, "Genome architecture: Domain organization of interphase chromosomes," *Cell*, vol. 152, pp. 1270–1284, 3 2013.

[18] G. Cavalli and T. Misteli, "Functional implications of genome topology," *Nature Structural and Molecular Biology*, vol. 20, pp. 290–299, 3 2013.

[19] L. B. Edelman and P. Fraser, "Transcription factories: Genetic programming in three dimensions," *Current Opinion in Genetics and Development*, vol. 22, pp. 110–114, 4 2012.

[20] M. Ruiz-Velasco and J. B. Zaugg, "Structure meets function: How chromatin organisation conveys functionality," *Current Opinion in Systems Biology*, vol. 1, pp. 129–136, 2 2017.

[21] A. Bianchi and C. Lanzuolo, "Into the chromatin world: Role of nuclear architecture in epigenome regulation," *AIMS Biophysics*, vol. 2, pp. 585–612, 2015.

[22] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, "Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data," *Nature Reviews Genetics*, vol. 14, pp. 390–403, 6 2013.

[23] T. Misteli, "Chromosome territories: The arrangement of chromosomes in the nucleus," *Nature Education*, 2008.

[24] M. R. Branco and A. Pombo, "Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations," *PLoS Biology*, vol. 4, pp. 780–788, 5 2006.

[25] A. Pombo and N. Dillon, "Three-dimensional genome architecture: Players and mechanisms," *Nature Reviews Molecular Cell Biology*, vol. 16, pp. 245–257, 4 2015.

[26] L. A. Parada, P. G. McQueen, and T. Misteli, "Tissue-specific spatial organization of genomes.," *Genome biology*, vol. 5, p. R44, 2004.

[27] H. Tanabe, S. Müller, M. Neusser, J. V. Hase, E. Calcagno, M. Cremer, I. Solovei, C. Cremer, and T. Cremer, "Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 4424–4429, 4 2002.

[28] A. Bancaud, C. Lavelle, S. Huet, and J. Ellenberg, "A fractal model for nuclear organization: Current evidence and biological implications," *Nucleic Acids Research*, vol. 40, pp. 8783–8792, 10 2012.

[29] G. Bajpai, D. A. Pavlov, D. Lorber, T. Volk, and S. Safran, "Mesoscale phase separation of chromatin in the nucleus," *eLife*, vol. 10, 5 2021.

[30] B. Chen, M. Yusuf, T. Hashimoto, A. K. Estandarte, G. Thompson, and I. Robinson, "Three-dimensional positioning and structure of chromosomes in a human prophase nucleus," *Science Advances*, vol. 3, 7 2017.

[31] K. J. Meaburn and T. Misteli, "Cell biology: Chromosome territories," *Nature*, vol. 445, pp. 379–781, 1 2007.

[32] R. Schneider and R. Grosschedl, "Dynamics and interplay of nuclear architecture, genome organization, and gene expression," *Genes and Development*, vol. 21, pp. 3027–3043, 12 2007.

[33] M. Cremer, J. V. Hase, T. Volm, A. Brero, G. Kreth, J. Walter, C. Fischer, I. Solovei, C. Cremer, and T. Cremer, "Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells," *Chromosome Research*, vol. 9, pp. 541–567, 2001.

[34] M. N. Nikiforova, J. R. Stringer, R. Blough, M. Medvedovic, J. A. Fagin, and Y. E. Nikiforov, "Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells," *Science*, vol. 290, pp. 138–141, 10 2000.

[35] A. Rosa and R. Everaers, "Structure and dynamics of interphase chromosomes," *PLoS Computational Biology*, vol. 4, p. e1000153, 8 2008.

[36] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, pp. 1665–1680, 12 2014.

[37] I. Solovei, K. Thanisch, and Y. Feodorova, "How to rule the nucleus: divide et impera," *Current Opinion in Cell Biology*, vol. 40, pp. 47–59, 6 2016.

[38] B. van Steensel and A. S. Belmont, "Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression," *Cell*, vol. 169, pp. 780–791, 5 2017.

[39] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. D. Klein, L. Wessels, W. D. Laat, and B. V. Steensel, "Domain

organization of human chromosomes revealed by mapping of nuclear lamina interactions," *Nature*, vol. 453, pp. 948–951, 6 2008.

[40] J. P. Fortin and K. D. Hansen, "Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data," *Genome Biology*, vol. 16, p. 180, 8 2015.

[41] A. L. Sanborn, S. S. Rao, S. C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden, "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. E6456–E6465, 11 2015.

[42] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, "Formation of chromosomal domains by loop extrusion," *Cell Reports*, vol. 15, pp. 2038–2049, 5 2016.

[43] J. Nuebler, G. Fudenberg, M. Imakaev, N. Abdennur, and L. A. Mirny, "Chromatin organization by an interplay of loop extrusion and compartmental segregation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, pp. E6697–E6706, 7 2018.

[44] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, 5 2012.

[45] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. V. Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot,

N. Blüthgen, J. Dekker, and E. Heard, "Spatial partitioning of the regulatory landscape of the x-inactivation centre," *Nature*, vol. 485, pp. 381–385, 5 2012.

[46] B. A. Bouwman and W. de Laat, "Getting the genome in shape: The formation of loops, domains and compartments," *Genome Biology*, vol. 16, p. 154, 12 2015.

[47] J. H. Gibcus and J. Dekker, "The hierarchy of the 3d genome," *Molecular Cell*, vol. 49, pp. 773–782, 3 2013.

[48] Q. Szabo, F. Bantignies, and G. Cavalli, "Principles of genome folding into topologically associating domains," *Science Advances*, vol. 5, 4 2019.

[49] A. V. Luzhin, I. M. Flyamer, E. E. Khrameeva, S. V. Ulianov, S. V. Razin, and A. A. Gavrilov, "Quantitative differences in tad border strength underly the tad hierarchy in drosophila chromosomes," *Journal of Cellular Biochemistry*, vol. 120, pp. 4494–4503, 3 2019.

[50] C. T. Ong and V. G. Corces, "Ctcf: An architectural protein bridging genome topology and function," *Nature Reviews Genetics*, vol. 15, pp. 234–246, 4 2014.

[51] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren, "Chromatin architecture reorganization during stem cell differentiation," *Nature*, vol. 518, pp. 331–336, 2 2015.

[52] A. S. Hansen, C. Cattoglio, X. Darzacq, and R. Tjian, "Recent evidence that tads and chromatin loops are dynamic structures," *Nucleus*, vol. 9, pp. 20–32, 12 2018.

[53] R. Boya, A. D. Yadavalli, S. Nikhat, S. Kurukuti, D. Palakodeti, and J. M. Pongubala, "Developmentally regulated higher-order chromatin interactions or-

chestrate b cell fate commitment," *Nucleic Acids Research*, vol. 45, pp. 11070–11087, 2017.

[54] O. Symmons, V. V. Uslu, T. Tsujimura, S. Ruf, S. Nassari, W. Schwarzer, L. Ettwiller, and F. Spitz, "Functional and topological characteristics of mammalian regulatory domains," *Genome Research*, vol. 24, pp. 390–400, 3 2014.

[55] J. E. Phillips-Cremins, M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C. T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, and V. G. Corces, "Architectural protein subclasses shape 3d organization of genomes during lineage commitment," *Cell*, vol. 153, pp. 1281–1295, 6 2013.

[56] S. Berlivet, D. Paquette, A. Dumouchel, D. Langlais, J. Dostie, and M. Kmita, "Clustering of tissue-specific sub-tads accompanies the regulation of hoxa genes in developing limbs," *PLoS Genetics*, vol. 9, p. e1004018, 12 2013.

[57] B. Bonev, N. M. Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J. P. Hugnot, A. Tanay, and G. Cavalli, "Multiscale 3d genome rewiring during mouse neural development," *Cell*, vol. 171, pp. 557–572.e24, 10 2017.

[58] J. M. Dowen, Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schuijers, T. I. Lee, K. Zhao, and R. A. Young, "Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes," *Cell*, vol. 159, pp. 374–387, 10 2014.

[59] Z. Tang, O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L. M. Sachs, X. Ruan, C. L. Wei,

E. T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li, and Y. Ruan, "Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription," *Cell*, vol. 163, pp. 1611–1627, 12 2015.

[60] X. Ji, D. B. Dadon, B. E. Powell, Z. P. Fan, D. Borges-Rivera, S. Shachar, A. S. Weintraub, D. Hnisz, G. Pegoraro, T. I. Lee, T. Misteli, R. Jaenisch, and R. A. Young, "3d chromosome regulatory landscape of human pluripotent cells," *Cell Stem Cell*, vol. 18, pp. 262–275, 2 2016.

[61] B. J. Beliveau, A. N. Boettiger, M. S. Avendaño, R. Jungmann, R. B. McCole, E. F. Joyce, C. Kim-Kiselak, F. Bantignies, C. Y. Fonseka, J. Erceg, M. A. Hannan, H. G. Hoang, D. Colognori, J. T. Lee, W. M. Shih, P. Yin, X. Zhuang, and C. T. Wu, "Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using oligopaint fish probes," *Nature Communications*, vol. 6, p. 7147, 11 2015.

[62] S. Wang, J. H. Su, B. J. Beliveau, B. Bintu, J. R. Moffitt, C. T. Wu, and X. Zhuang, "Spatial organization of chromatin domains and compartments in single chromosomes," *Science*, vol. 353, pp. 598–602, 8 2016.

[63] A. S. Belmont, "Visualizing chromosome dynamics with gfp," *Trends in Cell Biology*, vol. 11, pp. 250–257, 6 2001.

[64] S. M. Gasser, "Nuclear architecture: Visualizing chromatin dynamics in interphase nuclei," *Science*, vol. 296, pp. 1412–1416, 5 2002.

[65] M. Eltsov, K. M. MacLellan, K. Maeshima, A. S. Frangakis, and J. Dubochet, "Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 19732–19737, 12 2008.

[66] H. D. Ou, S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, and C. C. O'Shea, "Chromemt: Visualizing 3d chromatin structure and compaction in interphase and mitotic cells," *Science*, vol. 357, p. eaag0025, 7 2017.

[67] M. A. L. Gros, E. J. Clowney, A. Magklara, A. Yen, E. Markenscoff-Papadimitriou, B. Colquitt, M. Myllys, M. Kellis, S. Lomvardas, and C. A. Larabell, "Soft x-ray tomography reveals gradual chromatin compaction and reorganization during neurogenesis in vivo," *Cell Reports*, vol. 17, pp. 2125–2136, 11 2016.

[68] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *Science*, vol. 295, pp. 1306–1311, 2 2002.

[69] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. D. Santiago, L. M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. Edwards, M. Nicodemi, and A. Pombo, "Complex multi-enhancer contacts captured by genome architecture mapping," *Nature*, vol. 543, pp. 519–524, 3 2017.

[70] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M. Guttman, "Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus," *Cell*, vol. 174, pp. 744–757.e24, 7 2018.

[71] A. Rosa and C. Zimmer, "Computational models of large-scale genome architecture," 2014.

[72] M. V. Imakaev, G. Fudenberg, and L. A. Mirny, "Modeling chromosomes: Beyond pretty pictures," *FEBS Letters*, vol. 589, pp. 3031–3036, 10 2015.

[73] A. Hofmann and D. W. Heermann, "Deciphering 3d organization of chromosomes using hi-c data," *Methods in Molecular Biology*, vol. 1837, pp. 389–401, 2018.

[74] M. Nicodemi and A. Pombo, "Models of chromosome structure," *Current Opinion in Cell Biology*, vol. 28, pp. 90–95, 6 2014.

[75] R. K. Sachs, G. V. D. Engh, B. Trask, H. Yokota, and J. E. Hearst, "A random-walk/giant-loop model for interphase chromosomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, pp. 2710–2714, 3 1995.

[76] N. M. Toan, D. Marenduzzo, P. R. Cook, and C. Micheletti, "Depletion effects and loop formation in self-avoiding polymers," *Physical Review Letters*, vol. 97, p. 178302, 10 2006.

[77] M. Bohn, D. W. Heermann, and R. V. Driel, "Random loop model for long polymers," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 76, p. 051805, 11 2007.

[78] M. Bohn and D. W. Heermann, "Diffusion-driven looping provides a consistent provides a consistent framework for chromatin organization," *PLoS ONE*, vol. 5, p. e12218, 8 2010.

[79] Y. Zhang and D. W. Heermann, "Loops determine the mechanical properties of mitotic chromosomes," *PLoS ONE*, vol. 6, p. e29225, 12 2011.

[80] M. Barbieri, M. Chotalia, J. Fraser, L. M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, "Complexity of chromatin folding is captured by the strings and binders switch model," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, pp. 16173–16178, 10 2012.

[81] K. Rippe, "Making contacts on a nucleic acid polymer," *Trends in Biochemical Sciences*, vol. 26, pp. 733–740, 12 2001.

[82] S. Jun and B. Mulder, "Entropy-driven spatial organization of highly confined polymers: Lessons for the bacterial chromosome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 12388–12393, 8 2006.

[83] A. Y. Grosberg, A. R. Khokhlov, H. E. Stanley, A. J. Mallinckrodt, and S. McKay, "Statistical physics of macromolecules," *Computers in Physics*, vol. 9, p. 171, 1995.

[84] A. Y. Grosberg, S. Nechaev, and E. Shakhnovich, "The role of topological constraints in the kinetics of collapse of macromolecules," *Journal de Physique*, vol. 49, pp. 2095–2100, 1988.

[85] L. A. Mirny, "The fractal globule as a model of chromatin architecture in the cell," *Chromosome Research*, vol. 19, pp. 37–51, 1 2011.

[86] G. Kreth, J. Finsterle, J. V. Hase, M. Cremer, and C. Cremer, "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*, vol. 86, pp. 2803–2812, 5 2004.

[87] M. Tark-Dame, R. V. Driel, and D. W. Heermann, "Chromatin folding - from biology to polymer models and back," *Journal of Cell Science*, vol. 124, pp. 839–845, 3 2011.

[88] D. Marenduzzo, C. Micheletti, and P. R. Cook, "Entropy-driven genome organization," *Biophysical Journal*, vol. 90, pp. 3712–3721, 5 2006.

[89] S. Jhunjhunwala, M. C. van Zelm, M. M. Peak, S. Cutchin, R. Riblet, J. J. van Dongen, F. G. Grosveld, T. A. Knoch, and C. Murre, "The 3d structure

of the immunoglobulin heavy-chain locus: Implications for long-range genomic interactions," *Cell*, vol. 133, pp. 265–279, 4 2008.

[90] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, pp. 363–367, 5 2010.

[91] H. Tanizawa, O. Iwasaki, A. Tanaka, J. R. Capizzi, P. Wickramasinghe, M. Lee, Z. Fu, and K. I. Noma, "Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation," *Nucleic Acids Research*, vol. 38, pp. 8164–8177, 12 2010.

[92] D. Baú, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom, "The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules," *Nature Structural and Molecular Biology*, vol. 18, pp. 107–115, 1 2011.

[93] M. A. Umbarger, E. Toro, M. A. Wright, G. J. Porreca, D. Baù, S. H. Hong, M. J. Fero, L. J. Zhu, M. A. Marti-Renom, H. H. McAdams, L. Shapiro, J. Dekker, and G. M. Church, "The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation," *Molecular Cell*, vol. 44, pp. 252–264, 10 2011.

[94] M. Rousseau, J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette, "Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling," *BMC Bioinformatics*, vol. 12, 2011.

[95] J. Fraser, M. Rousseau, S. Shenker, M. A. Ferraiuolo, Y. Hayashizaki, M. Blanchette, and J. Dostie, "Chromatin conformation signatures of cellular differentiation," *Genome Biology*, vol. 10, 2009.

[96] M. Hu, K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, B. Ren, and J. S. Liu, "Bayesian inference of spatial organizations of chromosomes," *PLoS Computational Biology*, vol. 9, p. e1002893, 1 2013.

[97] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen, "Genome architectures revealed by tethered chromosome conformation capture and population-based modeling," *Nature Biotechnology*, vol. 30, pp. 90–98, 1 2012.

[98] A. Lesne, J. Riposo, P. Roger, A. Cournac, and J. Mozziconacci, "3d genome reconstruction from chromosomal contacts," *Nature Methods*, vol. 11, pp. 1141–1143, 11 2014.

[99] Z. Zhang, G. Li, K. C. Toh, and W. K. Sung, "3d chromosome modeling with semi-definite programming and hi-c data," *Journal of Computational Biology*, vol. 20, pp. 831–846, 11 2013.

[100] M. D. Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, "Transferable model for chromosome architecture," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 12168–12173, 10 2016.

[101] M. D. Pierro, R. R. Cheng, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, "De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, pp. 12126–12131, 11 2017.

[102] L. R. Gehlen, G. Gruenert, M. B. Jones, C. D. Rodley, J. Langowski, and J. M. O'Sullivan, "Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions," *Nucleus (United States)*, vol. 3, pp. 370–383, 7 2012.

[103] N. Tokuda, T. P. Terada, and M. Sasai, "Dynamical modeling of three-dimensional genome organization in interphase budding yeast," *Biophysical Journal*, vol. 102, pp. 296–304, 1 2012.

[104] B. Zhang and P. G. Wolynes, "Topology, structures, and energy landscapes of human chromosomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 6062–6067, 5 2015.

[105] N. Varoquaux, F. Ay, W. S. Noble, and J. P. Vert, "A statistical approach for inferring the 3d structure of the genome," *Bioinformatics*, vol. 30, pp. i26–i33, 6 2014.

[106] H. Iwasaki and K. Akashi, "Hematopoietic developmental pathways: On cellular basis," *Oncogene*, vol. 26, pp. 6687–6696, 10 2007.

[107] K. L. Medina, J. M. Pongubala, K. L. Reddy, D. W. Lancki, R. DeKoter, M. Kieslinger, R. Grosschedl, and H. Singh, "Assembling a gene regulatory network for specification of the b cell fate," *Developmental cell*, vol. 7, pp. 607–617, 10 2004.

[108] L. Li, M. Leid, and E. V. Rothenberg, "An early t cell lineage commitment checkpoint dependent on the transcription factor bcl11b," *Science*, vol. 329, pp. 89–93, 7 2010.

[109] E. J. Wood, "Cellular and molecular immunology (5th ed.): Abbas a. k., and lichtman, a. h.," *Biochemistry and Molecular Biology Education*, vol. 32, pp. 65–66, 1 2004.

[110] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindah, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, 9 2015.

[111] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 2 1996.

[112] M. N. Chernodub, M. Lundgren, and A. J. Niemi, "Elastic energy and phase structure in a continuous spin ising chain with applications to chiral homopolymers," *Physical Review E*, vol. 83, p. 011126, 1 2011.

[113] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G. Filion, and M. A. Marti-Renom, "Automatic analysis and 3d-modelling of hi-c data using tadbit reveals structural features of the fly chromatin colors," *PLOS Computational Biology*, vol. 13, p. e1005665, 7 2017.

[114] C. Weinreb and B. J. Raphael, "Identification of hierarchical chromatin domains," *Bioinformatics (Oxford, England)*, vol. 32, pp. 1601–1609, 6 2016.

[115] H. Shin, Y. Shi, C. Dai, H. Tjong, K. Gong, F. Alber, and X. J. Zhou, "Topdom: an efficient and deterministic method for identifying topological domains in genomes," *Nucleic Acids Research*, vol. 44, pp. e70–e70, 4 2016.

[116] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp, "A phase separation model for transcriptional control," *Cell*, vol. 169, pp. 13–23, 3 2017.

[117] S. Fakan and R. van Driel, "The perichromatin region: A functional compartment in the nucleus that determines large-scale chromatin folding," *Seminars in Cell and Developmental Biology*, vol. 18, pp. 676–681, 10 2007.

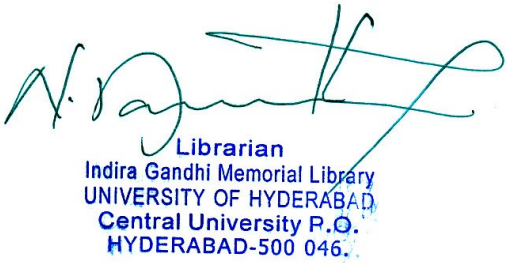
[118] J. M. R. Pongubala, D. L. Northrup, D. W. Lancki, K. L. Medina, T. Treiber, E. Bertolino, M. Thomas, R. Grosschedl, D. Allman, and H. Singh, "Transcription factor ebf restricts alternative lineage options and promotes b cell fate commitment independently of pax5," *Nature Immunology*, vol. 9, pp. 203–215, 2 2008.

[119] J. R. Smith, G. T. Hayman, S.-J. Wang, S. J. F. Laulederkind, M. J. Hoffman, M. L. Kaldunski, M. Tutaj, J. Thota, H. S. Nalabolu, S. L. R. Ellanki, M. A. Tutaj, J. L. D. Pons, A. E. Kwitek, M. R. Dwinell, and M. E. Shimoyama, "The year of the rat: The rat genome database at 20: a multi-species knowledgebase and analysis platform," *Nucleic Acids Research*, 11 2019.

[120] M. Heydarian, "Prediction of gene activity in early b cell development based on an integrative multi-omics analysis," *Journal of Proteomics and Bioinformatics*, vol. 07, 2014.

Probing structural organisation of chromatin during B cell development through mechanistic modelling

by Anubhooti .



Submission date: 27-Dec-2022 12:28PM (UTC+0530)
Submission ID: 1986849759
File name: Thesis_Anubhooti_13LAPH19_Final.pdf (26.87M)
Word count: 21571
Character count: 118580

Probing structural organisation of chromatin during B cell development through mechanistic modelling

ORIGINALITY REPORT

9%	4%	8%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Angelo Rosa, Christophe Zimmer. "Computational Models of Large-Scale Genome Architecture", Elsevier BV, 2014 Publication	2%
2	"Bacterial Chromatin", Springer Science and Business Media LLC, 2018 Publication	1%
3	escholarship.org Internet Source	1%
4	www.nature.com Internet Source	<1%
5	Robin D. Lee, Todd P. Knutson, Sarah A. Munro, Jeffrey T. Miller et al. "Nuclear corepressors NCOR1/NCOR2 regulate B cell development, maintain genomic integrity and prevent transformation", Nature Immunology, 2022 Publication	<1%
6	Jinwook Choi, Myunggon Ko, Shin Jeon, Yoon Jeon, Kyungsoo Park, Changjin Lee, Ho Lee,	<1%

	Rho H. Seong. "The SWI/SNF-like BAF Complex Is Essential for Early B Cell Development", The Journal of Immunology, 2012 Publication	
7	repub.eur.nl Internet Source	<1 %
8	Leonid A. Mirny. "The fractal globule as a model of chromatin architecture in the cell", Chromosome Research, 2011 Publication	<1 %
9	Nicodemi, Mario, and Ana Pombo. "Models of chromosome structure", Current Opinion in Cell Biology, 2014. Publication	<1 %
10	Lin, Yin C, Christopher Benner, Robert Mansson, Sven Heinz, Kazuko Miyazaki, Masaki Miyazaki, Vivek Chandra, Claudia Bossen, Christopher K Glass, and Cornelis Murre. "Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate", Nature Immunology, 2012. Publication	<1 %
11	Submitted to Cardiff University Student Paper	<1 %
12	www.esp.org Internet Source	

		<1%
13	Submitted to University of Hyderabad, Hyderabad Student Paper	<1%
14	Ravi Boya, Anurupa Devi Yadavalli, Sameena Nikhat, Sreenivasulu Kurukuti, Dasaradhi Palakodeti, JaganM.R. Pongubala. "Developmentally regulated higher-order chromatin interactions orchestrate B cell fate commitment", Nucleic Acids Research, 2017 Publication	<1%
15	Tianming Zhou, Ruochi Zhang, Jian Ma. "The 3D Genome Structure of Single Cells", Annual Review of Biomedical Data Science, 2021 Publication	<1%
16	K Akashi. "Hematopoietic developmental pathways: on cellular basis", Oncogene, 10/15/2007 Publication	<1%
17	Luca Giorgetti, Rafael Galupa, ElphègeP. Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, Edith Heard. "Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription", Cell, 2014 Publication	<1%

18	www.biorxiv.org Internet Source	<1 %
19	www.swinburne.edu.au Internet Source	<1 %
20	Jyotsana J. Parmar, Maxime Woringer, Christophe Zimmer. "How the Genome Folds: The Biophysics of Four-Dimensional Chromatin Organization", Annual Review of Biophysics, 2019 Publication	<1 %
21	opendsi.cc Internet Source	<1 %
22	s3.amazonaws.com Internet Source	<1 %
23	Atsuko Masumi, Shoichiro Miyatake, Tomoko Kohno, Toshifumi Matsuyam. "Chapter 16 Interferon Regulatory Factor-2 Regulates Hematopoietic Stem Cells in Mouse Bone Marrow", IntechOpen, 2012 Publication	<1 %
24	ir.uiowa.edu Internet Source	<1 %
25	Gui-Xin Ruan, Yuxing Li, Wenjing Chen, Hengjun Huang, Rui Zhang, Changxu Chen, Kong-Peng Lam, Shengli Xu, Xijun Ou. "The spliceosome component Usp39 controls B cell	<1 %

development by regulating immunoglobulin gene rearrangement", Cell Reports, 2022
Publication

26	Priscillia Lhoumaud, Sana Badri, Javier Rodriguez Hernaez, Theodore Sakellaropoulos et al. "NSD2 overexpression drives clustered chromatin and transcriptional changes in a subset of insulated domains", Cold Spring Harbor Laboratory, 2019 Publication	<1 %
27	K F Yoong. "Tumour infiltrating lymphocytes: insights into tumour immunology and potential therapeutic implications", Molecular Pathology, 10/1/1996 Publication	<1 %
28	Xiaobin Zheng, Yixian Zheng. "CscoreTool: Fast Hi-C Compartment Analysis at High Resolution", Cold Spring Harbor Laboratory, 2017 Publication	<1 %
29	Abdul Wasim, Ankit Gupta, Jagannath Mondal. " A Hi–C data-integrated model elucidates chromosome's multiscale organization at various replication stages ", Nucleic Acids Research, 2021 Publication	<1 %
30	ir.vanderbilt.edu Internet Source	<1 %

31	ipfs.io Internet Source	<1%
32	Antonio B. Oliveira Junior, Cynthia Perez Estrada, Erez Lieberman Aiden, Vinícius G. Contessoto, José N. Onuchic. "Chromosome Modeling on Downsampled Hi-C Maps Enhances the Compartmentalization Signal", The Journal of Physical Chemistry B, 2021 Publication	<1%
33	pure.tue.nl Internet Source	<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		