Regulatory Landscapes Associated With Changes In Nuclear Architecture

Ioana Olan

Supervisor: Dr. Masashi Narita
Dr. Shamith Samarajiwa

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Girton College September 2018
I dedicate the present body of work to the man who never left my side and to the nurturing community within the University of Cambridge who helped me develop as a scientist and as a human being.

“It turns out that an eerie type of chaos can lurk just behind a facade of order - and yet, deep inside the chaos lurks an even eerier type of order.”

Douglas R. Hofstadter
Declaration

The contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Ioana Olan
September 2018
Acknowledgements

I would like to thank my supervisors, Dr. Masashi Narita and Dr. Shamith Samarajiwa, for continuous guidance and patience and for the large number of hours spent overlooking the progress of this dissertation. I would also like to thank Adelyne Chan, Dr. Masako Narita, Dr. Andy Young, Dr. Guy Slater and all the other members of the Narita lab for being exceptional examples of integrity, optimism and creativity and for all their support and tenacity to read early (and late) drafts. I thank the former members of the Narita lab as well, in particular Dr. Aled Parry and Dr. Kosuke Tomimatsu, for invaluable advice and support. I am also extremely grateful to Prof. Simon Tavaré and to his lab for hosting and advising me for the past three years. In particular, I would like to thank Sam, Meltem and Ed for suggestions and for all the time spent working together. I would also like to thank members of the Samarajiwa lab for their support. Many thanks addressed to Dr. Aaron Lun, Dr. Stefan Schoenfelder and Dr. Andy Lynch for useful discussion and suggestions.

This dissertation would not have been completed without the support and funding from the Wellcome Trust, the Mathematical Genomics and Medicine PhD programme, the University of Cambridge and Cancer Research UK. I also thank Dr. Ann Kaminski for all her advice and support.

I would also like to extend thanks to Dr. David Penman, Prof. Berthold Lausen, Prof. Graham Upton and the University of Essex who supported me during my undergraduate degree in mathematics and without whom I would have never been able to get to this stage in my career.

I am grateful to my partner, Omid Siddiqui, for being there always, rekindling hope whenever I struggled, and learning a little biology and graph theory along the way. I thank my parents and relatives for their good wishes sent from far away. I thank my friends, Ketch, Florin, Lavinia, Tudor, Razvan and Ioana, whom I wish I visited more often. I also thank all the special people I met during my time in Cambridge, especially Constantine, Maria, Jin, Jose, Byron, Adrian and Giulio.
Collaborations

Data used in this dissertation was generously provided as a result of the work of members of the Narita laboratory, in collaboration with Peter Fraser laboratory (Babraham Institute). In particular, Dr. Aled Parry performed Hi-C and capture Hi-C experiments, in collaboration with Peter Fraser laboratory, under the guidance of Dr. Stefan Schoenfelder. Dr. Aled Parry also performed ChIP-seq experiments for HMGA1, H3K4me2 and CTCF DNA binding. Dr. Masako Narita performed ChIP-seq experiments for cohesin binding, using antibodies generously provided by Prof. K Shirahige (Institute of Molecular and Cellular Biosciences, University of Tokyo). Dr. Yoko Ito performed ChIP-seq for C/EBPβ binding.

Preprocessing of several data was aided by Dr. Dora Bihary in the Samarajiwa laboratory (alignments of H3K4me1, H3K4me2, H3K27ac, CTCF, ATAC-seq, H3K9me3 and H3K27me3 histone marks), as well as by Dr. Guy Slater in the Narita laboratory (alignment of Hi-C and capture Hi-C).
Abstract

Gene regulation is essential for establishing a cell’s identity and relation with surrounding cells. A major role in modulating gene regulation is played by the cell’s chromatin architecture. Gene positioning relative to the center of the nucleus or its three-dimensional contacts with other genomic regions can determine whether the gene is expressed or not. Changes in chromatin architecture have been reported in cell differentiation contexts, where the gene expression landscape undergoes striking remodelling. Architectural changes leading to abnormal expression levels have also been reported and linked to diseases, such as dementia or cancer. Chromatin is organised hierarchically, with interactions between regulatory regions such as enhancers and promoters being hosted within megabase-scale domains. Enhancer-promoter alterations have been mostly described in differentiation scenarios, while a number of studies concluded that such interactions are pre-looped within the same lineage. In this study, I characterised chromatin architecture coupled with gene expression changes in RAS-induced senescent (RIS) cells. RIS is a tumour suppressive phenotype associated with cell cycle arrest and inflammation, associated with substantial alterations in both chromatin architecture and expression profile. By combining data from Hi-C, ChIP-Seq and RNA-Seq experiments, I showed that enhancer-promoter interactions were altered during RIS. In particular, pro-inflammatory genes in the IL1 locus showed increased interactions with enhancers. I showed that chromatin alterations also occurred at larger scale, re-arranging the chromatin hierarchy. I used graph theoretical approaches to model the hierarchical organisation of the genome. I linked large scale re-arrangements in RIS to the formation of Senescence Associated Heterochromatic Foci (SAHF), an important phenotypic feature of RIS, consisting of striking re-organisation of heterochromatic regions. In terms of network connectivity, RIS interactions were characterised by increased separation between any two genomic regions. The present study extends the current knowledge regarding the potential for alterations of enhancer-promoter interactions within the same lineage. It also introduces new tools for characterising the chromatin hierarchy and determining multiple layers of organisation and their association. Such tools have the potential of resolving chromosome relative positioning and contextualising the consequences of copy number alterations.
# Table of contents

List of figures xv
List of tables xxvii
Nomenclature xxix

## 1 Introduction
1.1 Chromatin architecture 2
1.2 Chromatin organisation into macro-structures 4
1.3 Graph Theory 5
  1.3.1 Community detection 7
1.4 Determining significant interactions 8
1.5 Senescence 9

## 2 Global features of chromatin interactions in growing and RAS-induced senescence
2.1 Introduction 13
2.2 Hi-C data analysis 14
2.3 Chromatin macro-structures 19
  2.3.1 A/B compartments 19
  2.3.2 Topological associating domains (TADs) 19
2.4 Differential interactions in RAS-induced Senescence 21
  2.4.1 Long-range differential interactions 22
  2.4.2 SAHF H3K9me3 estimation 25
  2.4.3 Interaction changes within and between TADs in RIS 31
  2.4.4 Loop disruption during RIS 34
  2.4.5 Comparison with other types of Oncogene-induced Senescence 38
2.5 Differential Enhancer-Promoter interactions 39
  2.5.1 Annotation with enhancers and promoters 40
2.6 Discussion 63
3 Graph theory approach to characterising chromatin structure 67

3.1 Introduction .................................................. 67
3.2 Non-parametric Bayesian inference adaptation of Stochastic Block Models .... 71
3.3 Applying modified SBM to Hi-C data .......................... 74
  3.3.1 Choosing the number of iterations required .................. 75
3.4 Results ...................................................... 77
  3.4.1 Consistency between resolutions ............................ 79
  3.4.2 Hierarchy depth and community sizes ....................... 80
  3.4.3 Comparison with TADs and A/B compartments ............... 84
  3.4.4 Strategy to improve computational burden .................. 86
  3.4.5 Nested community structure at 40kb (2-step strategy) .... 88
  3.4.6 Comparison with epigenetic marks ........................ 90
  3.4.7 Stability of nested community calling ...................... 91
  3.4.8 Differences between growing and RIS ....................... 92
  3.4.9 SAHF formation ........................................ 94
  3.4.10 Comparison with measured spatial distances .............. 99
  3.4.11 Advantages over other community detection methods .... 101
3.5 Hierarchy visualisation ...................................... 101
3.6 Discussion ................................................ 102

4 Detecting significant interactions from HiC data 105

4.1 Introduction ................................................ 105
4.2 Exploratory Hi-C modelling .................................. 106
  4.2.1 Different chromosomes .................................. 106
  4.2.2 TAD structure effect .................................. 107
  4.2.3 Hi-C biases ............................................ 107
  4.2.4 Comparing distributional assumptions ..................... 109
  4.2.5 Final model ............................................ 113
4.3 Significant interactions at 10kb, 20kb and 40kb ................ 114
  4.3.1 Connectivity .......................................... 114
  4.3.2 Hubs ............................................... 116
  4.3.3 Betweenness centrality ................................ 119
  4.3.4 Cliques .............................................. 120
  4.3.5 Mean distance ....................................... 120
  4.3.6 Enhancer-promoter significant interactions ............... 122
4.4 Discussion ............................................... 128
### 5 Methods

5.1 Experimental Methods ........................................... 129
5.2 Public data ......................................................... 130
5.3 Computational methods .......................................... 130
   5.3.1 Hi-C ......................................................... 130
   5.3.2 Nested community detection from Hi-C data ..................... 132
   5.3.3 Network analysis .............................................. 132
   5.3.4 Determining significant Hi-C interactions ..................... 132
   5.3.5 ChIP-seq ..................................................... 132
   5.3.6 RNA-seq ..................................................... 133

### 6 Conclusions and future study

References ............................................................... 139

Appendix A List of captured Hi-C regions .......................... 151
List of figures

1.1 Chromatin architecture layers: chromosome territories separate into A (active) and B (inactive) compartments consisting of units called topological associating domains (TADs); TADs can harbour enhancer-promoter interactions and loop formation (TAD borders can also form loops) ................................................... 2

1.2 Organisation of the genome into TADs and A/B compartments ........................................ 5

1.3 RAS-induced senescence phenotype in human diploid fibroblasts, the best-established cell model for oncogene-induced senescence (OIS). Oncogenic RAS-induced senescence is a prototype of OIS. Senescence is associated with a number of effector mechanisms, such as autonomous (cell cycle arrest) and non-autonomous (SASP) aspects of senescence. .......................................................... 10

1.4 Formation of SAHF through spatial reorganisation of heterochromatin. Lamin B1 (LMNB1) is down-regulated during senescence. Pericentromeric H3K9me3 heterochromatin is released from the nuclear envelop and spatially re-positioned to form the foci. HMGA proteins (HMGA1 in particular) are required for the process as structural components of SAHFs. .......................................................... 12

2.1 Agreement scores from HiC-spector for Hi-C samples at different resolutions: 20kb - 1Mb; Scores closer to 1 indicate better agreement; Growing: three replicates of growing IMR90 cells, with one stably expressing empty vector, growing with shHMG1: two replicates - growing IMR90 cells stably expressing shHMG1; RIS: two replicates of oncogenic HRAS-G12V-induced senescent IMR90 cells, RIS with shHMG1: two replicates - senescent cells stably expressing shHMG1. .......................................................... 16

2.2 Agreement visualisation with PCA using library size normalised, low counts and diagonal counts removed Hi-C matrices at different resolutions: 20kb - 400kb (bin sizes); G - growing replicates 1; RIS - RIS replicates, G shHMG1 - growing with shHMG1, RIS shHMG1 - RIS with shHMG1; Biological replicates agreed more as the bin size increases. .......................................................... 17
2.3 Schematic view of interactions coming in-and-out of focus between different resolutions; high-resolution interactions may be non-specific neighbourhoods of interactions or “sharp peaks” which show a clear enrichment in a neighbourhood; in the same location at low-resolution, the non-specific interaction may have gained enough reads to become more prominent than the 'sharp peak' one; the reasons for this discrepancy can be technical: under-sequencing of the interaction space or biological: many infrequent all-to-all (non-specific) smaller interactions.

2.4 Conserved A/B compartments (NMI = 1) between growing (G) and RIS in indicated chromosomes (1-22), A (open) compartment is represented by the positive values (green) and B (closed) compartment by negative values (blue).

2.5 Variation in the number of significant differential interactions (a) Number of differential interactions obtained at each resolution in RIS vs growing and in RIS shHMGA1 vs RIS; overall fewer changes in RIS shHMGA1 vs RIS (b) Number of differential interactions at different minimum reads filters showing similar trends against different resolutions (c) Different minimum reads filters, as well as a distance dependency threshold indicating very similar results between different filters at high-resolution values (under 100kb).

2.6 Differential interactions between distant regions at different resolutions; Each dot corresponds to an interaction between two regions of the specified bin size; The x-axis corresponds to the distance between the two interacting regions; Only significantly altered interactions are shown, as calculated with diffHic [73]; A large number of increasing interactions in RIS (red, logFC > 0) are long-range, i.e. > 20Mb apart, as indicated by the dotted vertical line.

2.7 (a) Number of significant interaction changes (increased - ‘green’ and decreased - ‘blue’) at several resolutions between growing and RIS (b) Interaction changes between RIS and RIS shHMGA1 (c) Increase in long-range interactions (> 20Mb between interacting regions) is observed in RIS and (d) decrease in long-range interactions in RIS shHMGA1 compared to RIS at 200kb resolution (e) Proportions of increased long-range interactions in RIS and decreased in RIS shHMGA1 corresponding to interactions between H3K9me3 peaks, as well as between H3K9me3 peaks and other regions.

2.8 Chromosome 4q: ChIP-seq signal in growing and RIS for HMGA1, H3K9me3 and LaminB1 (from Sadaie et al. [110]); HMGA1 accumulations of signal tend to overlap H3K9me3 peaks and regions with reduced LMNB1 binding in RIS.
2.9 chr4:54,640,000-170,111,400: significant H3K9me3-H3K9me3 interactions increased in RIS (green) and decreased (blue) in the presence of shHMGA1; red regions correspond to H3K9me3 peaks and purple signal corresponds to HMGA1 ChIP-seq track in RIS; The interactions pattern is reversed and the H3K9me3 involved in those interactions are bound by HMGA1, constituting potential SAHF.

2.10 (a) A subset of H3K9me3 peaks, termed SAHF-H3K9me3, increase interactions in RIS and decrease in RIS shHMGA1. Chromosome 4 contains the largest number of such H3K9me3 peaks. (b) Frequencies of H3K9me3-H3K9me3 interaction changes in RIS shHMGA1 depending on the direction of change in RIS from growing. Of particular interest are interactions which increase in RIS and decrease in RIS shHMGA1 (n = 746) which potentially form SAHF (c) SAHF-H3K9me3 are larger in size (Mb) than their counterparts and (t-test pairwise difference significant with $p-value \leq 0.001$) (d) HMGA1-rich regions which overlap SAHF-H3K9me3 are also larger in size than non-SAHF HMGA1-rich regions ($p-value \leq 0.001$) (e) SAHF-H3K9me3 are also more gene-poor than their counterparts.

2.11 (a) and (b): H3K9me3-H3K9me3 differential interactions network, where each H3K9me3 is represented as a node and orange/blue lines correspond to increased/decreased interactions in RIS, respectively. H3K9me3 are coloured by: in (a) SAHF H3K9me3 are orange circles, non-SAHF H3K9me3 are blue circles and in (b) purple nodes are HMGA1-rich and grey nodes are non-HMGA1-rich. (c) and (d) represent the SAHF-network only coloured by the size of the H3K9me3 peak in (c) and by the normalised HMGA1 signal in (d).

2.12 Proportion of differential interactions situated intra-TAD (out of all the intra- and inter-TAD interactions) at different resolutions: 20kb, 40kb, 100kb and 300kb.

2.13 Changes in TAD interactions classified into: (a) consecutive TAD-TAD interaction changes (b) distal TAD-TAD interactions, occurring between TADs separated by at least one other TAD (c) altered TAD interiors; and their respective assignments to A/B compartments.

2.14 Changes in interactions within TADs: (a) interaction maps of TAD structure around the NRG1 gene in growing and RIS, associated with loss of intra- and inter-TAD associations, with the black bar corresponding to its TAD, along with RNA-seq, ATAC-seq and H3K27me3 growing (G) and RIS (S) ChIP-seq normalised tracks; (b) interaction maps around the MME gene in growing and RIS, along with H3K27me3 normalised tracks.
2.15 Correlation matrix of 5kb binned THOR-normalised ChIP-seq signals of RAD21, SMC3, CTCF and ATAC-seq in growing and RIS using all replicates; green and blue squares indicate positive correlation and negative correlation, respectively. 35

2.16 (a) CTCF and RAD21 peaks, growing/RIS specific or common between them; (b) Significant increases and decreases of CTCF and RAD21 binding in RIS; (c) CTCF/RAD21 peaks distribution compared to the distribution of binding changes in RIS across regulatory regions such as loops, promoters or enhancers; percentages calculated relative to the total number of peaks or binding changes (negative sign is used for symmetry); (d) potentially disrupted loops - six cases and their overlaps: loss/gain of CTCF or RAD21 binding at loops’ ends or gain of CTCF or RAD21 within a loop’s interior. 36

2.17 Interaction maps at 20kb resolution of two regions which show decreased loop formation in RIS with matched ChIP-seq normalised tracks for histone marks or proteins of interest in growing (G) and RIS (S), as well as marked differentially expressed (DE) genes and IMR90 loop structure from Rao et al. [105] (a) the locus mentioned earlier around the MME gene showing a H3K27me3 TAD enclosed within H3K9me3 regions. Several loops in this region showed RAD21 loss at their ends. (b) another region on chromosome 3 characterised by a complex nested loop structure disrupted in RIS along with decreased RAD21 binding at several loop’s ends. 38

2.18 Analysis summary: differential EP interactions obtained by annotating differential interactions at resolutions between 10kb and 100kb with enhancers and promoters of DE genes. 41

2.19 Chromatin marks profile of enhancers and promoters in growing and RIS; enhancers are represented by every H3K27ac peak and the surrounding 2kb region; promoters are represented by 5kb regions around TSS of every protein coding gene. 43

2.20 (a) An example of a single component of EP network (b) Number of enhancers against gene log-fold change (RIS/growing). Selected genes are highlighted. 44

2.21 Enrichment of super-enhancer targets, from EnrichR [16], using Reactome database; (a) enrichment of up-regulated genes, targets of SE gained in RIS; (b) enrichment of down-regulated genes, targets of SE lost in RIS; all pathways shown are significantly enriched at FDR \( \leq 0.05 \). 45

2.22 (a) Activating (red) and repressing (blue) EP interactions in Hi-C; (b) Combinations of up/down regulated genes with increased and/or decreased enhancer interactions in Hi-C; green and blue arrows correspond to up-regulated and down-regulated genes, respectively (c) and (d) the same as (a) and (b) in capture Hi-C. 47
2.23 Changes in DNA binding or epigenetic marks associated with differential EP interactions, where one end or the other of the interactions displays differential ChIP-seq/ATAC-seq (nucleosome-free regions) signal. ........................................... 48

2.24 Indirect gene regulation by TF via enhancer binding; number of genes whose associated enhancer and/or promoter are bound by a protein such as TP53, E2F7, HMGA1 or BRD4 and the number of genes bound and differentially expressed upon knock-down (KD) of the respective protein. .......................... 49

2.25 Transcription factors (TFs) binding to both promoter and enhancer (left) or enhancer only (right); all the peak sets considered were collected from IMR90 cells (our own data and from REMAP [17]), all rows correspond to growing IMR90 except for the ones marked (S) which correspond to RIS; the x-axis represents the number of differential EP interactions for which either TF binds both enhancer or promoter or just the enhancer. ................................. 50

2.26 Potential negative regulation of TIMELESS regulation by p53 via enhancer binding; ChIP-seq tracks include P53 binding, H3K27ac and H3K4me1 (marking enhancers) and ATAC-seq (nucleosome-free regions) in RIS; top region focuses on the promoter of TIMELESS and the bottom region focuses on the enhancer of TIMELESS, as determined from the differential EP contacts studied with Hi-C in growing and RIS. ................................................................. 51

2.27 Differential EP contacts are infrequently separated by more than two TADs; x-axis represents the number of TADs separating the ends of differential EP interactions; y-axis indicates the frequency of “x” TADs separating enhancers and promoters. ................................................................. 52

2.28 Enhancers released from a loop via loop disruption, free to interact with promoters outside the loop. Yellow rings represent cohesin binding, blue circles represent CTCF binding, purple rectangles are enhancers and orange rectangles are promoters. Green arcs correspond to increased chance of interaction. 53

2.29 Hi-C maps in growing and RIS around the region consisting of E2F7 promoter and its associated enhancers (defined by differential EP interactions in RIS); enhancers: green - increased interactions and blue - decreased interactions with E2F7 promoter; loops - violet segments, derived from high-resolution Hi-C map by Rao et al. [105] in IMR90 cells. ................................................................. 54

2.30 Enhancers interacting preferentially more with SASP genes and less with cell cycle genes; gene set enrichment analysis of up- and down-regulated genes from the competitive pairs of genes which show increased and decreased interactions with the same enhancer, respectively. ................................................................. 55
2.31 (a) Hi-C map at HindIII resolution, corresponding to differential interactions (capture Hi-C) in the IL1 locus, with arcs corresponding to significant interactions between enhancers (grey rectangles) and promoters (coloured and labelled rectangles by their expression: green corresponds to up-regulated and blue to down-regulated genes in RIS); blue arcs correspond to decreased interactions and green arcs to increased interactions; the arcs are drawn using different promoters as anchors (IL1A, IL1B and CKAP2L), in order to extract the relevant interactions; RAD21 binding in growing and RIS is represented by THOR-normalised signal counted in every HindIII fragment; (b) our model for loop disruption in the IL1 locus, leading to two sub-loops forming.  

2.32 Interaction map of differential interactions from capture Hi-C at the MMP locus at 5kb resolution, with arcs anchored on the promoters of MMP10, MMP1, MMP3 and MMP12 corresponding to significantly changing EP contacts; ChIP-seq tracks focus on the region around the promoters of MMP10 and MMP1 showing the new cohesin peak forming in RIS (SMC3 and RAD21), as well as CTCF and H3K27ac ChIP-seq normalised tracks in growing and RIS.  

2.33 Enhancer-driven gene regulation. Current model: in a given cell type, EP interactions are mostly pre-existing and genes are activated upon increased transcription factor binding to the enhancer-promoter (top). For example, IL1A, IL1B and CKAP2L were shown to be co-regulated upon TNFα treatment without any altered interaction of EP, to which NFκB is recruited (middle). In contrast, during RIS (bottom), those shared enhancers preferentially interacted with IL1B contributing to its up-regulation and down-regulation of CKAP2L.  

2.34 Enhancer-enhancer interaction components where each circle represents an enhancer and the colour determines whether it is a typical (grey) or a super-enhancer; blue lines represent decreased interactions and green lines represent increased interactions.  

2.35 The distribution of frequency of enhancer cliques of different sizes; enhancer clique size varied between 1 and 13 enhancers, out of which cliques with 3 and 4 enhancers were the most frequent occurrence.  

3.1 Representation of chromatin interactions as a weighted graph by connecting genomic bins (vertices) using chromatin interactions (edges), weighted by their intensity calculated from chromosome capture data.  

3.2 Nested community structure detection by splitting the genome into equally sized bins whose relative position in the nucleus is reflected in the hierarchy detected, with neighbouring genomic regions being grouped within the same community.
3.3 Different values for the number of iterations and stopping conditions of `mcmc_equilibrate` plotted for chr14 (1Mb resolution); Parameter combinations of interest correspond to multiple states being explored which constitute multiple solutions of the resolved community structure.  
3.4 Nested community structures at several resolutions of chromosome 1 (a) 1Mb bin size (b) 300kb bin size (c) 100kb bin size, with the Hi-C matrix columns rearranged (anagram) to match the order of the bins in the hierarchy (left) side by side with the unordered matrix (right).  
3.5 Agreement of community detection (Normalised mutual information, NMI) between hierarchies of Hi-C matrices at different resolutions: 100kb, 300kb and 1Mb, in growing and RIS, respectively; level 1 corresponds to the basal level of the hierarchy formed of communities of genomic bins, level 2 consists of communities of level 1 communities.  
3.6 Homogeneity Index (HI) between hierarchies of Hi-C matrices at different resolutions: 100kb, 300kb and 1Mb, in growing and RIS, respectively; high values of HI indicate good agreement, e.g. genomic locations which are in the same community at level 1 in the 100kb bins hierarchy tend to also be in the same community at level 1 in the 300kb bins hierarchy.  
3.7 The number of levels (depth) of the hierarchy detected for each chromosome from growing and RIS Hi-C matrices at (a) 1Mb resolution (b) 300kb resolution (c) 100kb resolution.  
3.8 Size distribution of communities in (a) growing level 1 (b) growing level 2 (c) RIS level 1 and (d) RIS level 2, at each of the three resolutions: 1Mb, 300kb and 100kb; size of a community is defined here as the number of bins it consists of; level 1 is the basal level of the hierarchy consisting of communities of genomic bins at the chosen resolution, level 2 consists of level 1 communities.  
3.9 The size distribution of communities of each chromosome in growing (left) and RIS (right) at each of resolutions: (a) 100kb, (b) 300kb and (c) 1Mb.  
3.10 NMI of communities (100kb) compared to TADs and A/B compartments; (a) NMI of communities vs TADs (b) NMI of communities vs A/B compartments; values close to 1 indicate good matches between the sets of regions compared.  
3.11 HI of communities (100kb) compared to TADs and A/B compartments (a) HI of communities vs TADs (b) HI of communities vs A/B compartments; high values correspond to good agreement.  
3.12 NMI and HI values comparing different pairs of levels from the full and the two-step method of community detection in (a) growing and (b) RIS; values are calculated per chromosome.
3.13 Hierarchy levels at 40kb resolution compared to TADs and A/B compartments: (a) growing vs TADs levels 1-4, (b) RIS vs TADs levels 1-4 and (c) both growing and RIS vs A/B compartments levels 1 and 2.

3.14 TADs size distribution compared to 40kb and 100kb community sizes; (a) growing and (b) RIS; level 1 (left), level 2 (middle) and level 3 (right); Level 1 communities at 100kb resolutions and Level 2 40kb communities are the most similar to TADs in terms of their size distributions (red), whereas Level 3 communities are much larger.

3.15 Comparison of the interaction hierarchy with chromatin marks, HMGA1 binding, GC% and gene density in growing and RIS (100kb resolution); each row corresponds to a 100kb bin; the interaction hierarchy is represented on the left of the heatmaps with each colour corresponding to one community and each columns corresponding to one hierarchy level; the heatmaps of epigenetic marks are calculated using THOR normalised ChIP-seq signal which is further quantile normalised so that distributions of different marks can be plotted simultaneously.

3.16 Stability of communities in growing and RIS Levels 1-3, with 0 indicating unstable and 1 indicating stable communities.

3.17 Comparison of the hierarchy levels 1-3 in growing and RIS in terms of NMI (left) and HI (right), at (a) 100kb resolution and (b) 300kb resolution; Level 1 in growing and RIS communities agrees the most, both in terms of NMI and HI, regardless of the resolution used for the comparison; agreement decreases with higher levels but more in terms of NMI than HI values.

3.18 Representation of the hierarchy as a scatter plot that conceptually represents the position of chromosome 10 regions in the two-dimensionally projected cell nucleus; each colour corresponds to a community at each hierarchy level and each point represents a genomic bin; level 1 communities are incorporated in mid level communities which are then further incorporated in top level communities.

3.20 SAHF associated H3K9me3 (orange) and non-SAHF associated H3K9me3 (red) on chromosome 10, in growing, RIS and RIS with shHMGA1; each point represents a 40kb bin and the two-dimensional distances reflect relative distances between pairs of bins in the hierarchy detected using graph-tool; accumulation of SAHF H3K9me3 in RIS is partially reversed in RIS with shHMGA1.
3.19 (a) Hierarchy comparison between growing and RIS chromosome 10, indicating matching genomic regions with coloured lines and each colour corresponds to a different level 1 community; different SAHF-associated H3K9me3 (orange) merge during RIS on chromosome 10 with H3K9me3 (highlighted with dotted orange lines) (b) selected region highlighting SAHF and non-SAHF H3K9me3 and (c) lamin-associated domains (LADs). 98

3.21 Fraction of SAHF regions within level 3 communities in growing and RIS; 0 corresponds to communities with no SAHF and 1 corresponds to communities entirely formed of SAHF regions. 99

3.22 Comparison of physical distances (from Wang et al. [131]) between pairs of TADs from the same and from different communities at (a) levels 2 and (b) 3, on chromosomes 20, 21, and 22; significance values were calculated using a pairwise t-test between the two distributions of distances: within the same community and between different communities (all p-values are below 0.001). 100

4.1 Dependencies of Hi-C read counts (log) on distance, GC%, mappability and nr. RE sites at 40kb resolution (a) y-axis corresponds to the log-variance counts and x-axis corresponds to log-mean counts; each dot corresponds to the mean and variance of an interaction across growing replicates, coloured by the distance separating interacting regions; (b) the x-axis corresponds to the log(distance between interacting regions) and y-axis corresponds to the log-mean interaction counts (c) log(distance) against the log-counts in each growing replicate, as well as the distance fit trends for each replicate; (d) log-counts (y-axis) dependency on scaled GC% (x-axis) and the dependency trends per replicate; (e) log-counts dependency on scaled mappability values; (f) log-counts dependency on the scaled number of restriction enzyme sites (g) log-distance (x-axis) against log-mean counts across growing replicates (y-axis) coloured by position of the interaction relative to TADs; (h) same as (g) but plotted for each replicate instead of the log-mean across replicates (i) log-distance (x-axis) against log-mean count (y-axis) coloured by significant differential interaction status between growing and RIS, as well as the trend of the significant interactions relative to distance. 108

4.2 Chromosome 1 (100kb resolution): epigenetic marks highlighted on the scatter plot of the log-counts (y-axis) against the log-distance between interacting regions (x-axis); the colour gradient corresponds to the intensity of the THOR-normalised ChIP-seq signal of each mark: (a) H3K9me3 and (b) HMGA1 - heterochromatin (c) H3K4me3 - active promoters (d) H3K27ac - active enhancers (e) CTCF - structural protein (f) H3K4me1 - all enhancers. 110
4.3 AIC, BIC, deviance and percentage of zeroes predicted by the models considered; nb, zinb, hunb corresponded to negative binomial with/without zero-inflation and with/without the hurdle variation; hu1, hu2, and hu3 corresponded to variations of hunb (hurdle negative binomial), with hu1 including the intra-TAD term and hu2 and hu3 corresponding to different degrees of B-splines; pois, zipois and hupois correspond to Poisson distribution, zero-inflated and hurdle variations of Poisson distribution. 112

4.4 Degree distributions of the 10kb networks of growing (G) and RIS (S); degree (x-axis) of a node corresponds to the number of neighbouring nodes; y-axis corresponds to the frequency of nodes with certain degree values in the network. 115

4.5 Number of genomic regions corresponding to hubs either in growing or RIS, classified as growing-specific, common between growing and RIS, and RIS-specific. 116

4.6 Epigenetic landscape of hubs common between growing and RIS, growing-specific, and RIS-specific; each row corresponds to a 10kb bin, coloured by the normalised ChIP-seq signal intensities (from low to high) of open chromatin, heterochromatin, structural marks such as CTCF and cohesin, GC% and gene density. 117

4.7 Hubs: (a) Growing and (b) RIS; PHATE dimensionality reduction of the epigenetic features matrix, coloured by various marks corresponding to enhancers, promoters, heterochromatin and HMGA1; different branches correspond to different types of regulatory regions. 118

4.8 Number of genomic regions with high values of betweenness centrality either in growing or RIS, classified as growing-specific, common between growing and RIS, and RIS-specific. 119

4.9 Epigenetic landscape, as well as gene density and GC% of regions with high values of betweenness centrality in growing and RIS; each row corresponds to a 10kb genomic bin from the significant interactions network. 120

4.10 (a) mean distance values for all chromosomes in growing and RIS (b) distribution across chromosomes of all mean distance values (c) diameter values for all chromosomes (d) distribution of diameter values across chromosomes. 121

4.11 EP contacts annotated from significant interactions between 10kb regions, classified into: growing-specific, RIS-specific and common between the two conditions. 122

4.12 Distributions of the number of enhancers to the number of promoters (E to P ratio) and of the number of promoters to the number of enhancers (P to E ratio) in growing and RIS conditions. 123
4.13 Gene enrichment against the Reactome database [26] of the genes with common EP contacts in growing and RIS; only significantly enriched pathways are shown here; the x-axis corresponds to $-\log(\text{adjusted } p \text{-value})$; p-values were FDR-corrected for multiple testing and significance threshold was FDR $\leq 0.05$.

4.14 Capture Hi-C maps of EP interactions common between growing and RIS identified from significant interactions identified at 10kb resolution from genome-wide Hi-C data; the promoters of the genes, shown as green rectangles, interacted with the enhancers shown as black rectangles; the interaction maps are ICE-normalised and distance-corrected, so that warm-coloured interactions correspond to values above expected based on the distance-trend and cold-coloured interactions correspond to values below expected from the distance-trend.

4.15 Pathways which show evidence of co-regulation via shared enhancers in (a) growing and (b) RIS; the x-axis corresponds to the proportion of the pathway consisting of genes which share enhancers with at least one other gene in the same pathway; the pathways correspond to gene lists from the Reactome database [26].
List of tables

2.1 In-nucleus Hi-C and matching capture Hi-C in growing or RIS IMR90 cells with or without shHMGA1. Number of reads corresponding to valid interactions, duplicates and artifacts such as re-ligation, circularised fragments and dangling ends. ................................................................. 15

2.2 Pairs of up-regulated SASP genes - Gene1 and down-regulated cell cycle genes - Gene2, that share enhancers with opposite interaction directions (increased with Gene1, decreased with Gene2). ....................................................... 56
Nomenclature

Acronyms / Abbreviations

BET Bromodomain and extra-terminal domain
DE Differentially expressed
EP Enhancer promoter
HI Homogeneity Index
HMM Hidden Markov Model
LAD Lamina-associated domains
logFC Log-fold change
MCMC Monte Carlo Markov Chains
NMI Normalised Mutual Information
OIS Oncogene-induced senescence
RE Restriction enzyme
RIS RAS-induced senescence
SADS Senescence-associated distension of satellites
SAHF Senescence-associated heterochromatic foci
SASP Senescence-associated secretory phenotype
SBM Stochastic block model
SE Super-enhancer
TAD Topological associating domain
TE Typical-enhancer
Nomenclature

TF  Transcription Factor
TSS  Transcription start site
Chapter 1

Introduction

Gene regulation is a complex process with multiple stages contributing to modulating expression levels. Over the past decade, more and more data has become available depicting several aspects of gene regulation, concomitant with advanced methods to analyse it. Genomic data tends to represent the linear genome, capturing mRNA expression landscapes (RNA-seq) and DNA binding profiles (ChIP-seq) of different regulatory proteins. However, a great deal of insights are not readily available from ChIP-seq and other linear genomic data. Particularly, interactions between different genomic regions require a three-dimensional type of data such as Hi-C, which includes information about the linear coordinates of the interacting regions, as well as the strength of interaction between them. Such data has opened new avenues of research on the different layers of chromatin architecture, such as the presence of large structures which harbour interactions between regulatory regions, such as enhancers and promoters (Figure 1.1).

A multitude of analysis tools exist for Hi-C, with different purposes such as: determining chromatin macro-structures, as well as significant or differential interactions. In this study, I introduced a new tool for determining the hierarchy of chromatin architecture from Hi-C, which models the relative position of genomic regions inside the cell’s nucleus. The analysis exploited recent graph theoretical methodologies of community detection. Determining the chromatin’s hierarchy provides context for interaction changes between biological conditions. Furthermore, I studied significant interactions and integrate them in a network, which highlights genomic regions with interesting properties, such as highly interactive loci.

I used senescence as a model system due to its striking phenotype, associated with both distinct gene regulation and chromatin architecture changes [13], [93], [25], [147]. I found highly dynamic chromatin alterations between growing and senescent cells, consisting of both enhancer-promoter contact differences, as well as changes in associations between large heterochromatic regions. I also determined the hierarchical structure of chromatin in both
growing and senescent cells. I showed that large scale re-structuring occurred in senescence, both at basal and higher-levels of the interaction hierarchy.

Figure 1.1 Chromatin architecture layers: chromosome territories separate into A (active) and B (inactive) compartments consisting of units called topological associating domains (TADs); TADs can harbour enhancer-promoter interactions and loop formation (TAD borders can also form loops)

1.1 Chromatin architecture

Chromatin architecture plays a critical role in determining gene expression levels in a given biological context. The basic unit of chromatin architecture is the nucleosome, where DNA is coiled around a core of histones. Nucleosomes together with other accessory macromolecules, including proteins and RNA, is termed chromatin. Chromatin follows a compaction gradient from tightly compacted regions (heterochromatin) associated with gene repression, to loosely packed “open” regions (euchromatin) associated with permissive gene activation, in part due to the accessibility of transcription factor complexes to these regions. Heterochromatin can be further divided into two classes, constitutive heterochromatin and facultative heterochromatin. Constitutive heterochromatin regions, which include (peri)centromere and telomere regions, are commonly condensed in all cell types, whereas facultative heterochromatin is often determined during development (as exemplified by the inactive X chromosome) and cell lineage specification, contributing to cell type specific gene regulation [90]. Such chromatin states are determined by a number of factors, including DNA modifications, post-translational modifications of histone tails, chromatin remodelling factors [20], chromatin architectural factors, and non-coding RNAs [66], [72]. Nucleosome positioning also affects genome integrity,
regulation and transcription factor (TF) accessibility, and plays an important role in cellular maturation [122].

Interestingly, chromatin states also show some spatial preferences: heterochromatin tends to be situated towards peri-nuclear regions, while euchromatic regions tend to be within the internal regions of the nucleus. The nuclear envelope consists of the nuclear membrane and the nuclear lamina, which is a mesh structure composed of A-type and B-type nuclear lamina. Consistent with the chromosome positioning within nuclei, the nuclear lamina is often considered to be a region associated with gene repression [34], [32]. In addition, genome-wide analyses for Lamin B1 binding sites have shown that lamina-associated domains (LADs) are typically gene poor and enriched for repressive marks [44], [110], [121].

Gene regulatory regions include gene promoters and enhancers. Active promoters, situated proximal to transcriptional start sites (TSSs), are marked by open histone marks, such as H3K4me3. Enhancers are situated at large distances relative to the TSS and they come in contact with promoters via three-dimensional chromatin interactions. In contrast to promoters, enhancers are typically marked by other open histone marks: H3K4me1 and, when active, H3K27ac.

Transcription Factors (TFs) directly bind promoters in a sequence-specific manner and, in conjunction with other regulatory factors, such as co-factors (which do not directly bind to DNA), tightly control gene transcription [45]. In addition, enhancers can also deliver these factors (TFs and co-factors) to the promoter, adding further complexity to gene regulation depending on the cellular context. Although, due to the large distances separating enhancers and promoters, it has been challenging to precisely determine target genes of individual enhancers. Recent advances in genomic technology (e.g. Hi-C, as mentioned below) have allowed researchers to address this question. Interestingly, it has been suggested that the majority of enhancer-promoter pairs are established during lineage-specification and are largely unchanged within the same cell type [57], [61], [79], although studies testing the intra-lineage dynamics of enhancer-promoter binding are still limited. It has also become evident that multiple enhancers can target multiple gene promoters and promoters can be targeted by a large number of enhancers, creating large meshes of co-regulation [11], [10]. However, how the enhancer-promoter network is dynamically regulated within a given cell type is not entirely clear.
1.2 Chromatin organisation into macro-structures

It has been suggested, using imaging techniques, that each chromosome occupies a certain three-dimensional neighbourhood inside the nucleus, called a chromosome territory [24]. A chromosome’s position in relation to the centre of the nucleus is correlated with the gene expression trend mentioned earlier, with chromosomes closer to the centre of the nucleus being associated with higher gene density. Chromosome territory organisation has also been shown to exhibit lineage specificity as shown by Parada et al. [91]. The idea of chromosome territories has now been reinforced by genome-wide 3D interactome techniques.

Chromosome conformation capture (3C) is a technique which involves formaldehyde cross-linking the genome, followed by digestion with a restriction enzyme and the ligation of the proximal ends, to detect specific chromatin interactions by PCR [33]. A number of 3C-related techniques have since been developed in the last decade, such as 4C, 5C, ChIA-PET and Hi-C, providing important new insights into chromatin topology. Of these, Hi-C combines proximity ligation and DNA sequencing to capture chromatin interactions genome-wide. This makes it an invaluable tool for reconstructing the chromatin structure inside a nucleus, identifying the new high-order chromatin entities summarised below. The output corresponds to an interaction matrix reflecting the number of paired-end reads of every pairwise interaction between equally sized genomic regions. Each read in a pair is from the different regions interacting. A high number of reads in the interaction matrix corresponds to frequent interactions between the corresponding regions in a cell population.

Lieberman-Aiden et al. [70] showed, using Hi-C data, that the genome is partitioned into compartments A and B with regions within one compartment preferentially interacting with other regions from the same compartment (Figures 1.1, 1.2). Compartment A consists of open chromatin and is associated with gene activation, while compartment B consists of heterochromatin and is associated with gene repression.

Dixon et al. [36] and Nora et al. [84] showed that the genome is organised into megabase-sized local chromatin interaction domains called topological domains or topologically associating domains (TADs) (Figure 1.2). They are characterised by higher connectivity of genomic regions within compared to between them. A subset of TADs also corresponded to the LADs, mentioned earlier. In contrast to A/B compartments, TADs and TAD boundaries were shown to be highly conserved between cell types and even between different species, suggesting that distinct levels of machinery are engaged in high-order chromatin organisation [36]. It has been shown that TAD boundaries are critical for long-range chromatin interaction and gene regulation. Consistently, TAD boundaries are enriched for the zinc-finger-containing protein CTCF, which can act as an insulator [101]. In addition to CTCF, TAD boundaries are also
known to be enriched for housekeeping genes, tRNAs and short interspersed element (SINE) retrotransposons [36]. In contrast to A/B compartments, which are interspersed along the genome, TADs consist of highly interacting consecutive regions.

Figure 1.2 Organisation of the genome into TADs and A/B compartments

Using very high resolution (∼1kb) Hi-C map, Rao et al. [105] identified contact domains (∼185kb) much smaller than TADs (∼1Mb) and chromatin loops, of which anchors are typically associated with domain boundaries and bound by CTCF and cohesin (Figure 1.1). Loops are associated with gene expression by restricting enhancer-promoter contacts. Similar to TADs and TAD boundaries, contact domains and loops exhibit substantial conservation across cell types and species, although Rao et al. [105] also detected cell-type specific loops, involving gene promoters, which are associated with corresponding gene expression differences.

1.3 Graph Theory

The genome and epigenome are highly interactive and dynamic. A number of analytical methods have been utilised to handle such complex datasets, but they fail to contextualise the findings about chromatin structure and instead focus on individual features. In this study, I applied a graph-theoretical approach in order to intuitively capture the hierarchical structure of chromatin from Hi-C. Graph theoretical concepts, as well as community detection methods are introduced in the present section.

A graph (network) is a collection of vertices (nodes) $V_{1,2...n}$ connected by edges between pairs of vertices. This abstraction is very useful for representing relations such as social links between people, where each person is a vertex and any relation between two people is an edge in a social network, but also for representing biological gene regulation where each gene is a
vertex and the regulatory relation between gene A and gene B (e.g. up/down-regulation) is an edge between gene A and gene B. A subgraph is a subset of vertices of the graph and the edges that connect those vertices. A fully connected subgraph, i.e. every vertex is connected to any other vertex, is called a clique. Vertices and edges in a graph have several properties which are important for describing graphs. The degree of a vertex corresponds to the number of edges which connect that vertex to others. Analysing the degrees of vertices in a graph allows us to define hubs, which correspond to vertices connected to a large number of other vertices. For example, hubs can be master transcription factors in a gene regulatory network or popular people in a social network.

A path is a set of consecutive edges such that all its edges and vertices are distinct. This allows for the definition of the shortest path which is the set of consecutive edges between two nodes obtained by minimising all possible routes between them. The betweenness centrality property of a vertex is built on the concept of the shortest path and it represents the number of shortest paths between any pair of vertices that traverse through the vertex of interest. Such properties are useful for characterising the studied graph and gaining insights about real-world networks. For example, in the network made by all computers connected by the internet in a region, identifying the computers (vertices) with high betweenness centrality is important for assessing the most vulnerable parts of the network. Another property of a graph is its diameter and it corresponds to the length of the longest path (set of edges) in the graph. If there is a path between every vertex in the graph then the graph is connected. Otherwise, the groups of disconnected nodes (with no edge between them) are called the components of the graph.

The concept of a graph can be further transformed to better reflect real-world associations. Directionality can be applied to the edges, so that each edge connects a source vertex to a target vertex. For example, in the gene regulatory network, one gene is the regulator and another is its target. The regulatory edges between them have directionality. This is reflected in the quantification of different graph properties. Adding weights to the edges in the graph can open new avenues of analysis which better reflect the properties of the data. Until now, only binary graphs were described, where each pair of vertices is connected or not by an edge. Sometimes, a probabilistic model is more representative for a real-world graph. This can be introduced in the graph description by connecting every pair of vertices by an edge with a weight, which is the probability that the two vertices are connected. Such probabilistic descriptions are called random graphs.

Generating random graphs which model real-world networks accurately has been a long standing research question. Real-world networks have certain properties which are difficult to
simulate using random graphs, such as a degree distribution with heavy tails or small values for the diameter [65]. A famous example of such properties emerged with the characterisation of scale-free networks. A network is scale-free if its degree distribution follows a power law. The probability that a vertex \( V \) has degree \( k \) is described by equation 1.1 [5]. The World Wide Web and the network of research papers connected by citations are examples of scale-free networks.

\[ P(k) \sim k^{-\gamma} \] (1.1)

The intrachromosomal interaction value derived from Hi-C, written as a function of genomic distance between the two interacting regions, was shown to follow a power-law by Lieberman-Aiden et al. [70]. This indicates that the chromatin interaction network may be similar to a scale-free network.

1.3.1 Community detection

Real-world networks are often structured into groups which show a higher density of edges such as regulatory modules formed by multiple genes, or groups of multiple people formed by friendships, or living in the same neighbourhood. Thus, quantifying such groups and identifying them in a network has been an important graph theory problem with many solutions described in literature. Two different conceptual approaches have developed, one that partitions the graph and one that looks for blocks or a hierarchical structure in the network. The first is represented by applications such as the minimum-cut problem, which finds the minimal number of edges that need to be removed in order to disconnect the graph which gives rise to partitions. The second corresponds to variations of network clustering and its purpose is to detect “communities” in the network. Communities can be defined as groups of vertices which show more connectivity within, than between groups [83]. An example using this approach is represented by variations of stochastic block modelling introduced by Holland et al. [51].

Stochastic block models

Stochastic block models are defined by assigning each vertex in a network to one of \( k \) blocks (groups), \( B_{1,2,...,k} \). Then the probability of every edge between a pair of vertices reflects the block membership of the two vertices. More precisely, the probability of an edge between any vertex in block \( B_i \) and any vertex in \( B_j \) is the same. Assigning small values to this probability when \( i \neq j \) and larger ones when \( i = j \) reflects a community structure in the network. This is because it reflects low probability of connectivity between different blocks and higher probability of connectivity within the same block. Several variations of this method have been studied.
in order to render it more suitable to modelling real-world networks. One such variation was

In the present study, I used a non-parametric Bayesian version of stochastic block models,
introduced by Peixoto et al. [99], as described in chapter 3. Such an approach has not been
previously applied to analyse chromosome conformation data.

1.4 Determining significant interactions

Hi-C data is affected by sequencing biases such as GC and mappability, but also by biases
inherent to the Hi-C experiment. Examples include the lengths of the restriction enzyme (RE)
fragments [142] in the case of chromosomal contact maps at highest achievable resolution. If
the matrix is binned, then the corresponding bias is the number of RE cut-sites present in
each bin [128]. Interactions between equally-sized bins with more RE sites are associated
with higher read counts. Distance between interacting regions was also shown to play an
important role with closer regions being more likely to interact [70] and has been incorporated
in noise models as a covariate [3], [12]. Normalization approaches either explicitly model
those biases and use regression to remove them [142], [52] or use iterative approaches such as
Iterative Correction and Eigenvector decomposition (ICE) to remove them. This is done under
the assumption that all loci in the genome should have similar interaction potential (ICE - [55]).

Several methods have been developed to estimate significant interactions from Hi-C data,
defined as having an interaction value or read count beyond the expected one given the biases
at the loci interacting. Such methods include FitHiC [3], HICCUPS [105], PSYCHIC [108]
and HiC-DC [12]. FitHiC removes biases estimated with ICE and fits the distance bias using
splines assuming a binomial distribution of read counts observed depending on the distance
that separates the corresponding interacting loci. HICCUPS was developed by Rao et al.
[105] and performs peak detection in a Hi-C matrix. Each interaction is compared to its
neighbourhood and if its value is higher than expected, it is considered for further filtering to
improve the confidence of the peaks estimated. PSYCHIC aims to identify promoter-enhancer
significant interactions. It uses mixed models to represent intra- and inter-TAD interactions
using log-normal distributions with mean and variance depending on the distance separating
interacting loci. The intra- and inter-TAD spaces are allocated using dynamic programming.
Over-represented interactions are selected and compared against known information about
enhancers and promoters. Finally, HiC-DC’s authors use hurdle negative binomial regression
to remove biases explicitly, a method which takes into account overdispersion and zero-inflation
in the distribution of Hi-C read counts. They compare their method to FitHiC and show a
1.5 Senescence

Senescence is a state of stable cell cycle arrest that arises from a variety of triggers such as DNA damage, shortening of telomeres, or oncogene stress [47], [30], [93], [112]. In particular, oncogene-induced senescence (OIS), best exemplified by oncogenic RAS-induced-senescence (RIS), is a paradoxical phenotype, where excessive mitotic stimulation triggers senescence, instead of transformation, in normal cells [118]. Senescence is associated with characteristic morphological and biochemical alterations. It is also accompanied by a distinct gene expression profile (Figure 1.3): e.g. (a) some cell cycle genes are stably repressed in a growth factor resistant manner [120]; and (b) senescent cells express diverse soluble factors, including inflammatory cytokines, growth factors, and extracellular matrix proteins [22], [1], [62], [21], collectively called the senescence-associated secretory phenotype (SASP). It has also been shown that senescent cells exhibit unique epigenetic features.
Figure 1.3 RAS-induced senescence phenotype in human diploid fibroblasts, the best-established cell model for oncogene-induced senescence (OIS). Oncogenic RAS-induced senescence is a prototype of OIS. Senescence is associated with a number of effector mechanisms, such as autonomous (cell cycle arrest) and non-autonomous (SASP) aspects of senescence.

A recent study has shown that the inflammatory SASP is correlated with alterations in the enhancer landscape during RIS [125]. DNA methylation (5-Methylcytosine at CpGs) profiles are also altered during replicative senescence, where DNA methylation is globally reduced at heterochromatic regions but focally gained at a subset of CpG islands [27], [141].

Large-scale chromatin rearrangements in senescent cells can also be visualised using fluorescent microscopy. Heterochromatin enriched for H3K9me3 is liberated from the nuclear envelope and forms foci called senescence-associated heterochromatin foci (SAHF), shown in Figure 1.4. This process depends on high mobility group A (HMGA) proteins: non-histone chromatin architectural proteins [82], [81]. The release of heterochromatin from the nuclear envelope during senescence is facilitated by Lamin B1 (LMNB1) down-regulation, which leads to a reduction in LADs, predominantly at H3K9me3-rich regions [110]. In contrast to SAHF, heterochromatin unfolding can also be seen in senescent cells, known as senescence-associated distension of satellites (SADS) [123]. Swanson et al. [123] have shown that α-satellite and satellite II sequences, normally compacted in (peri)centromeric heterochromatin, are distended during senescence. Interestingly, Lamin B1 reduction also appears to be required for SADS formation [123]. Changes in chromatin architecture during senescence are also likely to occur at
a smaller scale, since senescent cells exhibit increased chromatin accessibility [31], [92], as well as changes in enhancer landscape as mentioned above [125]. These studies motivated the development of a 3D model of the dynamic changes in chromatin macro-structure during senescence.

The 3D structure of senescent cells has previously been investigated using Hi-C. OIS was previously characterised by Chandra et al. [13], using oncogenic RAF-induced senescence (RAF is a downstream effector of RAS) in WI-38 fibroblasts. The authors highlight the increased inter-TADs interactions and decreased intra-TADs interactions in OIS cells. The TADs with differential long ($\geq 20$ kb apart) or short-range interactions tended to be H3K9me3-rich. Consistently, Zirkel et al. [147] showed increased long-range interactions during replicative senescence in umbilical vein endothelial cells (HUVECs). This study also showed that A/B compartments are largely conserved, whereas a substantial portion of TADs undergo boundary shifts or fusion: e.g. 17% (boundary shifts) and 26% (fusion) in IMR90 cells. Interestingly, they also showed that HMGB2, which is depleted from nuclei in senescent cells, marks a subset of TAD boundaries and that reductions in HMGB2 are correlated with CTCF spatial clustering, leading to loop reshuffling.

Criscione et al. [25] have shown an alternative view: gains in short-range interactions and loss of long-range interactions (< 2Mb apart) and that a subset of TADs switch between A/B compartments during replicative senescence in human fibroblasts. The source of these discrepancies is not clear. Nevertheless, the data from the OIS system (in which SAHFs are most prominent) are highly consistent with the model that SAHFs are formed through the spatial re-positioning of H3K9me3 regions [14], [15].

Overall, senescence exhibits drastic high-order chromatin structural changes as well as unique gene regulatory mechanisms, potentially beyond the lineage barrier, providing a unique opportunity to study fundamentally distinct gene regulatory machinery.
Figure 1.4 Formation of SAHF through spatial reorganisation of heterochromatin. Lamin B1 (LMNB1) is down-regulated during senescence. Pericentromeric H3K9me3 heterochromatin is released from the nuclear envelop and spatially re-positioned to form the foci. HMGA proteins (HMGA1 in particular) are required for the process as structural components of SAHFs.
Chapter 2

Global features of chromatin interactions in growing and RAS-induced senescence

2.1 Introduction

In the present study, I aimed to characterise regulatory interaction changes (e.g. enhancer-promoter contacts) in RIS, as well as, global re-arrangements which might contribute to SAHF formation. Previous replicative and oncogene-induced senescence Hi-C studies [25], [13] and [147] focused mostly on the global chromatin alterations aspect, investigating the ratio of short- and long-range interaction changes, as well the importance of proteins such as HMGB2 on chromatin structure. Long-range interactions increase reported in OIS is thought to contribute to SAHF formation [13]. HMGA1 reduction was shown to reverse SAHF formation [81] and therefore, information about HMGA1-dependent interactions could help narrow down which regions form SAHF during OIS. For this purpose, I used information from growing, RIS, as well as HMGA1 knock-down Hi-C experiments. Apart from global changes in chromatin architecture, replicative and oncogene-induced senescence was also reported to undergo substantial alterations of the enhancer landscape, marked by the appearance of new super-enhancers ([125], [116]). The targets of the new enhancers were assigned linearly, by investigating the nearest genes. However, the three-dimensional regulatory interaction changes between enhancers and promoters are less studied. I used local changes observed at high-resolution in RIS to annotate which enhancer-promoter contacts might be altered.

As mentioned in chapter 1, I employed a suite of Hi-C analysis methods for comparing growing and RAS-induced senescence (RIS), as well as for extracting individual features of chromatin for each of the conditions. I first compared several existing tools, then I combined
and adapted some of them into the present analysis. DiffHic [73] was used for performing differential interactions analysis between every pair of studied conditions, using the alignments obtained with HiC-Pro [119]. Then, in the following chapters, community detection algorithms, as well as modified negative binomial regression models were used to contextualise the differences in chromatin interactions identified in the present chapter.

I analysed data from Hi-C and matched capture Hi-C experiments for the following biological conditions: growing IMR90 cells (human diploid fibroblasts), RAS-induced senescence (RIS), as well as growing and RIS cells with \textit{HMGA1} knock-down (stable expression of shHMGA1). At least two replicates were available for each condition. The libraries were prepared by Dr. Aled Parry in the Narita laboratory.

### 2.2 Hi-C data analysis

The pairs of reads from Hi-C data aligned to the genome contained technical artefacts which I removed prior to further analysis. Examples of such artefacts included circularised fragments, which self-ligated or dangling ends, which did not ligate. I used HiC-Pro [119] to align, remove artefacts, count read pairs against equally-sized genomic bins and normalise the libraries using iterative correction (ICE). Duplicates were marked with \textit{markdup} from samtools [68]. I removed them with HiC-Pro in the normalised matrices and by diffHic from the processed BAM alignments used in the differential analysis. I decided to remove duplicates due to the increased number of changes and the agreement between replicates observed upon removing them. Each pair of genomic bins corresponded to one unit of interaction and the number of read pairs from two interacting bins represented the intensity of the interaction. Hi-C experiments described in this study were performed using in-nucleus ligation in collaboration with the laboratory of Peter Fraser. Table 2.1 shows the number of valid reads in different conditions, used for analysis after artefacts and duplicates removal.

I also tried the HiCUP aligner [137], as well as the iterative mapping method implemented by the Leonid Mirny laboratory in the python package hiclib [55]. Iterative mapping, employed by HiC-Pro as well, resulted in slightly higher number of aligned reads. I chose HiC-Pro over hiclib due to its efficient implementation of both iterative mapping, as well as ICE normalisation [55] and compatibility of the three-column matrix format with further analysis.

I employed two methods of assessing the reproducibility of the experiments across different replicates. Firstly, I used PCA on binned matrices, normalised for library size, and filtered for low counts using diffHic [73]. Secondly, I used the HiC-spector tool [143] which calculates a comparison score between every two samples based on the first 20 eigenvectors of
Table 2.1 In-nucleus Hi-C and matching capture Hi-C in growing or RIS IMR90 cells with or without shHMGA1. Number of reads corresponding to valid interactions, duplicates and artifacts such as re-ligation, circularised fragments and dangling ends.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Growing 1</th>
<th>Growing 2</th>
<th>Growing 3</th>
<th>RIS 1</th>
<th>RIS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>243,122,461</td>
<td>196,232,011</td>
<td>276,580,244</td>
<td>216,305,356</td>
<td>266,086,016</td>
</tr>
<tr>
<td>Dangling ends</td>
<td>12,997,268</td>
<td>16,955,321</td>
<td>16,399,886</td>
<td>44,864,757</td>
<td>40,098,986</td>
</tr>
<tr>
<td>Re-ligation</td>
<td>22,343,167</td>
<td>14,818,082</td>
<td>28,013,264</td>
<td>16,163,307</td>
<td>22,824,509</td>
</tr>
<tr>
<td>Circularised</td>
<td>3,433,307</td>
<td>2,917,810</td>
<td>4,494,650</td>
<td>2,500,573</td>
<td>4,859,302</td>
</tr>
<tr>
<td>Duplicates</td>
<td>2,097,716</td>
<td>1,874,224</td>
<td>66,097,635</td>
<td>1,619,962</td>
<td>63,442,897</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Growing 1</th>
<th>Growing 2</th>
<th>Growing 3</th>
<th>RIS 1</th>
<th>RIS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>202,250,943</td>
<td>159,666,574</td>
<td>161,574,809</td>
<td>151,156,757</td>
<td>134,860,322</td>
</tr>
<tr>
<td>Cis interactions</td>
<td>185,920,385</td>
<td>146,481,198</td>
<td>148,906,165</td>
<td>134,838,330</td>
<td>120,455,876</td>
</tr>
<tr>
<td>Trans interactions</td>
<td>16,330,558</td>
<td>13,185,376</td>
<td>12,668,644</td>
<td>16,318,427</td>
<td>14,404,446</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Growing 1</th>
<th>Growing 2</th>
<th>Growing 3</th>
<th>RIS 1</th>
<th>RIS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>274,657,857</td>
<td>267,876,194</td>
<td>237,721,346</td>
<td>236,211,479</td>
<td></td>
</tr>
<tr>
<td>Dangling ends</td>
<td>20,596,532</td>
<td>8,507,166</td>
<td>81,012,651</td>
<td>37,289,382</td>
<td></td>
</tr>
<tr>
<td>Re-ligation</td>
<td>35,395,728</td>
<td>18,929,820</td>
<td>24,322,682</td>
<td>19,249,430</td>
<td></td>
</tr>
<tr>
<td>Circularised</td>
<td>4,421,004</td>
<td>3,893,750</td>
<td>3,076,638</td>
<td>3,368,509</td>
<td></td>
</tr>
<tr>
<td>Duplicates</td>
<td>63,107,462</td>
<td>86,585,396</td>
<td>39,039,316</td>
<td>42,448,495</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Growing 1</th>
<th>Growing 2</th>
<th>Growing 3</th>
<th>RIS 1</th>
<th>RIS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>527,485,390</td>
<td>486,852,213</td>
<td>294,445,585</td>
<td>412,354,485</td>
<td>281,964,730</td>
</tr>
<tr>
<td>Dangling ends</td>
<td>9,894,099</td>
<td>8,140,595</td>
<td>6,758,327</td>
<td>35,516,856</td>
<td>17,675,845</td>
</tr>
<tr>
<td>Re-ligation</td>
<td>57,629,539</td>
<td>58,689,581</td>
<td>34,821,449</td>
<td>40,371,093</td>
<td>30,578,971</td>
</tr>
<tr>
<td>Circularised</td>
<td>8,550,848</td>
<td>8,315,588</td>
<td>5,343,646</td>
<td>6,839,221</td>
<td>6,274,538</td>
</tr>
<tr>
<td>Duplicates</td>
<td>95,805,481</td>
<td>93,446,124</td>
<td>77,138,672</td>
<td>63,010,084</td>
<td>73,134,287</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Growing 1</th>
<th>Growing 2</th>
<th>Growing 3</th>
<th>RIS 1</th>
<th>RIS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>355,605,113</td>
<td>318,260,325</td>
<td>170,383,491</td>
<td>154,301,089</td>
<td></td>
</tr>
<tr>
<td>Cis interactions</td>
<td>323,426,945</td>
<td>289,422,374</td>
<td>155,631,783</td>
<td>137,260,273</td>
<td></td>
</tr>
<tr>
<td>Trans interactions</td>
<td>32,178,168</td>
<td>28,837,951</td>
<td>14,751,708</td>
<td>17,040,816</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Growing 1</th>
<th>Growing 2</th>
<th>Growing 3</th>
<th>RIS 1</th>
<th>RIS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>298,670,966</td>
<td>292,810,056</td>
<td>240,454,717</td>
<td>241,556,229</td>
<td></td>
</tr>
<tr>
<td>Dangling ends</td>
<td>9,251,901</td>
<td>4,060,570</td>
<td>41,482,147</td>
<td>19,072,683</td>
<td></td>
</tr>
<tr>
<td>Re-ligation</td>
<td>44,368,249</td>
<td>23,541,551</td>
<td>33,427,629</td>
<td>23,306,220</td>
<td></td>
</tr>
<tr>
<td>Circularised</td>
<td>8,550,848</td>
<td>8,315,588</td>
<td>5,343,646</td>
<td>6,839,221</td>
<td>6,274,538</td>
</tr>
<tr>
<td>Duplicates</td>
<td>95,805,481</td>
<td>93,446,124</td>
<td>77,138,672</td>
<td>63,010,084</td>
<td>73,134,287</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Growing 1</th>
<th>Growing 2</th>
<th>Growing 3</th>
<th>RIS 1</th>
<th>RIS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>151,724,653</td>
<td>168,519,501</td>
<td>98,245,396</td>
<td>117,098,001</td>
<td></td>
</tr>
<tr>
<td>Cis interactions</td>
<td>136,783,914</td>
<td>150,994,252</td>
<td>87,502,985</td>
<td>105,104,031</td>
<td></td>
</tr>
<tr>
<td>Trans interactions</td>
<td>14,940,739</td>
<td>17,525,249</td>
<td>10,742,411</td>
<td>11,993,970</td>
<td></td>
</tr>
</tbody>
</table>
the contact matrices. For the capture Hi-C samples I only used PCA because HiC-spector appears incompatible with capture Hi-C matrices. The HiC-spector reproducibility scores (Figure 2.1) between growing Hi-C samples were all above 0.8 indicating good agreement. Agreement between the first two RIS samples was also above 0.8 across different resolutions, but agreement with the third sample was worse. The cells in the third sample were infected with an “empty” control vector. Therefore, only the first two RIS samples were used for most analyses in order to make them comparable to growing ones, in terms of biological signal. The agreement between replicates with shHMGA1 treatment was overall lower. This might be due to different knockdown efficiencies of HMGA1, which plays a critical role for chromatin structure alteration during RIS.

Figure 2.1 Agreement scores from HiC-spector for Hi-C samples at different resolutions: 20kb - 1Mb; Scores closer to 1 indicate better agreement; Growing: three replicates of growing IMR90 cells, with one stably expressing empty vector, growing with shHMGA1: two replicates - growing IMR90 cells stably expressing shHMGA1; RIS: two replicates of oncogenic HRAS-G12V-induced senescent IMR90 cells, RIS with shHMGA1: two replicates - senescent cells stably expressing shHMGA1.

Similar conclusions were drawn by using the PCA method (Figure 2.2). I observed an increase in agreement between replicates of every condition as resolution increased and plateaued around 200kb. This suggests that 200kb was the resolution where the signal/noise ratio was maximised. With this method, I also observed slightly lower agreement between two RIS replicates and the one with empty vector. Growing samples with and without empty vector did not appear to be very different. Therefore, it was not clear if the third RIS sample was different due to an effect of the empty vector during RIS or due to other technical reasons.
Figure 2.2 Agreement visualisation with PCA using library size normalised, low counts and diagonal counts removed Hi-C matrices at different resolutions: 20kb - 400kb (bin sizes); G - growing replicates 1; RIS - RIS replicates, G shHMGA1 - growing with shHMGA1, RIS shHMGA1 - RIS with shHMGA1; Biological replicates agreed more as the bin size increases.
The agreement between shHMGA1 samples was maximised between 200-500kb, so the analysis of the shHMGA1 effect was performed at those resolution values, focusing on long-range interactions. Agreement values changed between replicates across different resolutions as different interactions came “in-and-out of focus”. For example, a locus which is quite diffuse at high resolution could come into focus at low resolution. Here, diffuse refers to an interaction matrix with a large number of non-zero entries but none high enough to be considered a sharp, significant interaction at high resolution. Symmetrically, a sharp interaction at high resolution surrounded by many zeroes could become diffuse at lower resolution. This argument is presented in Figure 2.3. Such “in-and-out of focus” behaviour suggests that Hi-C interactions should be analysed at multiple resolutions and information gathered from high, as well as, low resolution.

Figure 2.3 Schematic view of interactions coming in-and-out of focus between different resolutions; high-resolution interactions may be non-specific neighbourhoods of interactions or “sharp peaks” which show a clear enrichment in a neighbourhood; in the same location at low-resolution, the non-specific interaction may have gained enough reads to become more prominent than the “sharp peak” one; the reasons for this discrepancy can be technical: under-sequencing of the interaction space or biological: many infrequent all-to-all (non-specific) smaller interactions.
2.3 Chromatin macro-structures

I investigated chromatin megabase-sized structures, such as A/B compartments and TADs, in growing and RIS samples.

2.3.1 A/B compartments

I used PCA analysis to detect A/B compartments (Figure 1.2). The matrices used as input corresponded to the negative binomial mean across replicates of the ICE-normalised contact matrices. I calculated them as described in correctedContact from the R package diffHic [73]. I selected the first principal component that correlated well (correlation value $\geq 0.5$) with the input-normalised ChIP-seq signal of corresponding H3K4me1 (active histone mark) samples as representative for A/B compartments. I chose the sign of the principal component so that positive values correspond to A compartment (open) and negative values to B compartment (closed). The genome-wide correlation between the A/B compartment score and H3K4me1 signal was 0.69. A/B compartments correlated positively with other open chromatin marks as well, such as H3K27ac (0.56 correlation), and negatively with H3K9me3 (-0.55 correlation), as well as HMGA1 binding (-0.42 correlation).

Overall A/B compartments distribution along the genome was remarkably similar between growing and RIS (Figure 2.4). To quantify this comparison, I used Normalised Mutual Information (NMI). NMI is a measure of agreement between two partitions of the same set, scaled so its values range between 0 (no agreement) and 1 (perfect agreement). To calculate NMI, I used the R package NMI [140]. The NMI value comparing A/B compartments in growing and RIS was 1, suggesting that genomic regions largely maintain their open or closed status, respectively. It has been shown that A/B compartments are highly dynamic during development and are cell type-specific [35]. Although A/B compartments were not discussed in the previous OIS study [13], some level of switch between compartments has been reported in replicative senescence models [147], [25]. I re-analysed the Hi-C data from the OIS study from Chandra et al. [13] and called A/B compartments. The A/B distribution was also conserved in their data (NMI = 1), between growing and OIS.

2.3.2 Topological associating domains (TADs)

I computed TADs from 40kb resolution Hi-C matrices, using multiple methods such as TADbit [117], Arrowhead from the Juicer suite [38] and the directionality index calculation in diffHic [73]. TADbit optimises partitioning of the Hi-C matrix into TADs and computes a confidence score between 1 and 10 for each TAD boundary with a score of 10 signifying the boundary call is reliable. It also offers functionality for comparing and combining TADs across biological
Figure 2.4 Conserved A/B compartments (NMI = 1) between growing (G) and RIS in indicated chromosomes (1-22), A (open) compartment is represented by the positive values (green) and B (closed) compartment by negative values (blue).
replicates. I detected a very low number of TADs with Arrowhead and the directionality index based estimates appeared unreliable upon visual inspection of the called TADs. I therefore continued the analysis with the TADbit estimates for domains by calculating a consensus across replicates. This was obtained by selecting the common boundaries across replicates with confidence scores greater than 7. I chose this threshold by visually inspecting borders with different scores and how well they separate regions of frequent interactions.

Few rearrangements were observed between TAD distributions in growing and RIS. I also used NMI to quantify the agreement between TAD partitions. The NMI value between growing and RIS TADs was 0.977, indicating very good agreement. Due to this high level of agreement, a consensus TAD set, consisting of 2,532 TADs, was used across all growing and RIS samples for the rest of this study.

Trends that were reported in earlier studies were also observed in the present data. For example, larger TADs tended to be associated with heterochromatin and smaller ones tended to be associated with open regions [102]. This property was further investigated by comparing the size of TADs overlapping the A and the B compartments, but also by characterising the size of TADs which were rich in H3K27ac, CTCF binding, H3K4me1, ATAC (open chromatin) and H3K9me3 (heterochromatin) peaks. The average lengths of TADs overlapping A compartment and B compartment, respectively was approximately 500kb and 1.2Mb in both growing and RIS. TADs rich in CTCF binding, H3K27ac, H3K4me1 and ATAC peaks were around 4-500kb long, whereas TADs overlapping H3K9me3 peaks were much longer, with an average of 2Mb.

Despite high similarity between the borders of TADs in growing and RIS, the border strength or associations between distant TADs may change in RIS. I continued the investigation by characterising genome-wide interaction changes between the two conditions.

2.4 **Differential interactions in RAS-induced Senescence**

In spite of the similarities between growing and RIS in terms of A/B compartments and TAD borders, chromatin interactions differences may reflect changes in TAD-TAD associations or in regulatory interactions within TADs. Chandra et al. [13] showed that long-range interactions increase in RIS and hypothesised that this corresponds to SAHF formation. SAHF formation is reversed with HMGA1 knock-down [81] and is thought to occur through re-organisation of H3K9me3 associations [15]. Analysing the behaviour of long-range interactions in RIS with HMGA1 knock-down allowed me to narrow down regions which potentially form SAHF and determine their special properties. I used HMGA1 ChIP-seq information (experiment by Dr.
Aled Parry in the Narita Laboratory, described in chapter 5) to determine the direct effect of HMGA1 binding on interaction changes, as well as H3K9me3 and LMNB1 ChIP-seq from the study performed by Chandra et al. [14] and Sadaie et al. [110]. Previous Hi-C senescence studies focused mainly on global changes in chromatin architecture. In the present analysis, I also aimed to estimate potentially dis-regulated local chromatin interactions corresponding to enhancer-promoter contacts and likely residing within TADs.

I performed pairwise differential interaction analysis between the four conditions studied: growing, RIS, growing with shHMGA1 and RIS with shHMGA1, using the R Bioconductor package diffHic [73]. I repeated the analysis at different resolutions, ranging between 10kb to 1Mb, increasing in 5kb steps between 10kb and 100kb to cover the high resolution landscape, and every 100kb between 100kb and 1Mb for low resolution. The number of differential interactions increases steeply at high resolution values from 10kb to 100kb (Figure 2.5 (a)) and decreases at resolutions higher than 400kb. This suggests that the signal/noise ratio of the differences is maximised around 300kb. Different filtering strategies are available for removing interactions with low counts (necessary for the differential analysis). I compared the results of using different values for minimum number of reads (Figure 2.5 (b)), as well as a distance-decay function threshold of the interaction frequency (Figure 2.5 (c)). I applied the latter to interactions at resolutions less than 100kb due to the flattening of the distance decay at large distances imposed by large bin sizes (resolution \(\geq 200kb\)). No significant difference was recorded between different filtering strategies and a lower limit of 10 reads per interaction was chosen (scaled to the smallest library size).

Comparing different resolutions yielded high agreement rates (over 85%) between smaller and lower resolutions, such as 85% of all 40kb differential interactions overlap with 50kb differential interactions. This agreement, combined with the trend of an increase in the number of interactions with bin size (Figure 2.5 (a)), suggests that the interactions obtained at higher resolution were a subset of the ones obtained at lower resolution.

### 2.4.1 Long-range differential interactions

Oncogene-induced senescence (OIS) has been previously associated with an increase in long-range interactions by Chandra et al. [13], defined as interactions between loci separated by at least 20Mb. In general, interactions between distant loci were difficult to detect at high resolution values and only gain enough signal at low resolution. Therefore, the presence of long-range interactions was followed from 10kb to 1Mb resolution. Chandra et al. calculated a score based on 200kb windows in order to characterise those interactions. I also detected gains in long-range interactions in oncogene-induced senescence (RIS) in our dataset, starting from resolution values of 80kb (Figure 2.6). However, at lower resolutions, many long-range...
interactions were also lost in RIS, although the number of the lost interactions was smaller than the number of gained interactions. Long-range interactions discussed here largely corresponded to interacting H3K9me3 regions, with 83% of the long-range interactions showing overlap of at least one of the regions with H3K9me3 peaks. This suggests that, as expected, rearrangement between different H3K9me3 peaks occurs in RIS across large distances.

Next, I investigated the consequences of HMGA1 knock-down in RIS on long-range interactions. At every resolution, RIS showed more increased interactions than decreased compared to growing, while RIS with HMGA1 knock-down showed more decreases than increases (Figure 2.7 (a) and (b)). I observed that the increase in long-range interactions in RIS was reversed with HMGA1 knock-down which showed a large number of decreased long-range interactions (Figure 2.7 (c) and (d)). I calculated the proportion of H3K9me3-H3K9me3 type of interactions among those long-range interactions which are significantly altered in RIS and RIS with HMGA1 knock-down. Most of the long-range increased interactions in RIS and decreased in RIS with HMGA1 knock-down were indeed occurring between H3K9me3 regions (Figure 2.7 (e)). HMGA1 thus plays a role in reversing the pattern of increased long-range interactions observed in RIS.
Figure 2.6 Differential interactions between distant regions at different resolutions; Each dot corresponds to an interaction between two regions of the specified bin size; The x-axis corresponds to the distance between the two interacting regions; Only significantly altered interactions are shown, as calculated with diffHic [73]; A large number of increasing interactions in RIS (red, logFC > 0) are long-range, i.e. > 20Mb apart, as indicated by the dotted vertical line.
2.4 Differential interactions in RAS-induced Senescence

Figure 2.7 (a) Number of significant interaction changes (increased - ‘green’ and decreased - ‘blue’) at several resolutions between growing and RIS (b) Interaction changes between RIS and RIS shHMGA1 (c) Increase in long-range interactions (> 20Mb between interacting regions) is observed in RIS and (d) decrease in long-range interactions in RIS shHMGA1 compared to RIS at 200kb resolution (e) Proportions of increased long-range interactions in RIS and decreased in RIS shHMGA1 corresponding to interactions between H3K9me3 peaks, as well as between H3K9me3 peaks and other regions.

2.4.2 SAHF H3K9me3 estimation

As mentioned in the Chapter 1, senescence-associated heterochromatic foci (SAHFs) form during senescence through spatial rearrangement of heterochromatic regions, primarily H3K9me3 regions forming the SAHF core [15]. This was consistent with the increase in long-range interactions with a large contribution from H3K9me3 interactions (Figure 2.7). In order to overcome the need for multiple resolution-analysis, H3K9me3 peaks were used as interaction units, instead of equally sized genomic bins. I compared interactions between every pair of H3K9me3 peaks in RIS relative to growing and in RIS with shHMGA1 relative to RIS.

I was particularly interested in interactions gained in RIS and lost with shHMGA1 treatment (Figure 2.10 (a)). Since SAHFs are lost upon shHMGA1 treatment [81], without a substantial change of H3K9me3 ChIP-seq pattern [14], the most likely candidates for physical
SAHF components are H3K9me3 peaks which engage in interactions increasing in RIS and decreasing with shHMGA1 treatment. In order to evaluate the effect of HMGA1 binding on interaction changes, I analysed HMGA1 ChIP-seq data produced by Dr. Aled Parry in the Narita Laboratory. HMGA1 binding was characterised by frequent peaks across the genome which also showed particular accumulations over genomic regions I termed “HMGA1-rich regions” (Figure 2.8). I defined them by merging 100kb regions with more HMGA1 normalised signal (details in chapter 5) than the 75% quantile of the genome-wide signal distribution in all samples. 1,809 HMGA1-rich regions were found in this way, averaging 390kb in size.

I used Circos plots to visualise H3K9me3 interaction changes from growing to RIS, to RIS with shHMGA1. In Figure 2.9), chromosome 4 is represented as a circular region marked by H3K9me3 (red) and significant H3K9me3-H3K9me3 interaction changes are represented as arcs. Notably, interactions which increase in RIS decrease with HMGA1 knock-down and vice versa. Moreover, H3K9me3 regions involved in increased interactions in RIS show HMGA1 binding accumulation. Next, I quantitatively investigated those observations genome-wide.

Figure 2.8 Chromosome 4q: ChIP-seq signal in growing and RIS for HMGA1, H3K9me3 and LaminB1 (from Sadaie et al. [110]); HMGA1 accumulations of signal tend to overlap H3K9me3 peaks and regions with reduced LMNB1 binding in RIS.

Figure 2.10 (b) shows the frequency of different trajectories of H3K9me3-H3K9me3 interaction changes with RIS and with RIS with HMGA1 knock-down. Notably, only a small number of same-direction interaction changes were observed, e.g. 3 interactions increasing in RIS and also increasing with HMGA1 knock-down (Figure 2.10 (b)). The 746 interactions increasing in RIS and decreasing with HMGA1 knock-down were used as an estimate of SAHF interactions throughout this study. Those interactions potentially constituted only a subset of all interactions involved in SAHF formation, due to insufficient sequencing depth and differences in signal/noise ratio between the two pairwise comparisons.
2.4 Differential interactions in RAS-induced Senescence

Figure 2.9 chr4:54,640,000-170,111,400: significant H3K9me3-H3K9me3 interactions increased in RIS (green) and decreased (blue) in the presence of shHMGA1; red regions correspond to H3K9me3 peaks and purple signal corresponds to HMGA1 ChIP-seq track in RIS; The interactions pattern is reversed and the H3K9me3 involved in those interactions are bound by HMGA1, constituting potential SAHF.

I observed 294 SAHF H3K9me3 peaks involved in increased interactions in RIS which decrease with shHMGA1 treatment out of a total of 774 H3K9me3 peaks. SAHF formation is marked by release from the nuclear periphery of heterochromatic regions concommittent with loss of lamina-domains in RIS [110]. I checked whether our SAHF-estimated regions indeed overlapped with lamina-associating domains (LADs) lost in RIS, by using the LADs definition from Sadaie et al. [110], lifted over from the hg18 to the hg19 genome coordinates. 81% of the SAHF H3K9me3 and only 34% of the non-SAHF H3K9me3 overlapped LADs lost in RIS.

I also checked whether HMGA1 directly bound SAHF H3K9me3 by analysing HMGA1 ChIP-seq, described in chapter 5. H3K9me3 peaks which potentially formed SAHF were more often covered by HMGA1-rich regions, with 87% of SAHF H3K9me3 and only 36% of the non-SAHF H3K9me3 overlapping HMGA1-rich regions. On the other hand, only 58% of HMGA1-rich regions overlapped H3K9me3 regions.

Interestingly, the H3K9me3 peaks associated with SAHF formation were much larger overall than non-SAHF associated peaks, as seen in Figure 2.10 (c).
Figure 2.10 (a) A subset of H3K9me3 peaks, termed SAHF-H3K9me3, increase interactions in RIS and decrease in RIS shHMGA1. Chromosome 4 contains the largest number of such H3K9me3 peaks. (b) Frequencies of H3K9me3-H3K9me3 interaction changes in RIS shHMGA1 depending on the direction of change in RIS from growing. Of particular interest are interactions which increase in RIS and decrease in RIS shHMGA1 (n = 746) which potentially form SAHF (c) SAHF-H3K9me3 are larger in size (Mb) than their counterparts and (t-test pairwise difference significant with $p - value \leq 0.001$) (d) HMGA1-rich regions which overlap SAHF-H3K9me3 are also larger in size than non-SAHF HMGA1-rich regions ($p - value \leq 0.001$) (e) SAHF-H3K9me3 are also more gene-poor than their counterparts.
Moreover, the HMGA1-rich regions overlapping SAHF-H3K9me3 were also larger than their counterparts (Figure 2.10 (d)). This suggests that size of a heterochromatic region may play an important role in determining its ability to re-organise towards the centre of the nucleus and associate with others. I also checked the gene density of those regions (normalised by their size) and showed that SAHF-H3K9me3 regions are more gene-poor than non-SAHF (Figure 2.10 (e)).

In order to gain further understanding of those associations, I plotted H3K9me3-H3K9me3 interactions as a network and coloured their nodes by various properties (Figure 2.11). Each (disconnected) component of the H3K9me3 network corresponds to one chromosome. Figure 2.11 (a) shows that SAHF-H3K9me3 increased interactions with each other forming highly connected components. On the other hand, non-SAHF H3K9me3 regions showed decreased interactions with SAHF H3K9me3 but showed fewer interaction changes with other non-SAHF H3K9me3 and most of them were decreases. Figure 2.11 (b) show the distribution of HMGA1-rich regions over the same network and confirms the high overlap between HMGA1-rich regions and SAHF H3K9me3. Interestingly, a small number of components is formed entirely by non-SAHF H3K9me3 which are also HMGA1-rich and showed decreased interactions with each other. Then, I plotted the subset of the network formed solely by SAHF H3K9me3 regions in Figure 2.11 (c) and (d). Overlaying the network with information about the size of the nodes (H3K9me3 regions) suggested that most of the larger components, i.e. chromosomes with many connected SAHF H3K9me3, had a central H3K9me3 peak which was much larger and highly connected with other H3K9me3 peaks (Figure 2.11 (c)). This suggests that larger SAHF s might revolve around a central “core” large H3K9me3 region. Further study is required to clarify this possibility. Overlaying HMGA1 "richness" normalised by the size of the H3K9me3 region on the SAHF network revealed various levels of HMGA1 binding, ranking from high HMGA1 binding on the larger components and weaker binding on the smaller components (Figure 2.11 (d)). A few HMGA1-independent H3K9me3 were observed in the SAHF network which also showed a small number of interactions with other SAHF H3K9me3 regions.
**HMGA1** is also expressed in growing cells, although not as strongly as in RIS. I also compared Hi-C libraries of growing cells and growing + shHMGA1, counting reads over H3K9me3 peaks. I found 1,176 significant differential interactions, a smaller number than in RIS compared to RIS + shHMGA1. Interestingly, 88% of the changes involving SAHF H3K9me3 interactions were decreases (488 decreased and 68 increased interactions), involving 220 SAHF-H3K9me3 regions. Some of the potential SAHF interactions appeared to be weaker but pre-existing in growing cells (as seen in Figure 2.10 (a)). The reduced interactions in growing + shHMGA1 suggests that HMGA1 potentially acts as a “molecular glue” for those...
specific SAHF-H3K9me3 regions in multiple contexts.

An important insight into SAHF formation from Hi-C was the strength of preferential associations between different H3K9me3 regions. From microscopy studies, the reproducibility of such associations across different cells cannot be fully assessed, due to the large number of cells needed. This means that SAHF formation could occur between different H3K9me3 regions in every cell. However, this study suggested that SAHF-associated interactions were highly preferential, because only interactions with consistent patterns across multiple samples can be detected as significant by the differential interaction analysis. The high log-fold change values of the interaction changes supported this observation. These data reinforced the idea that SAHFs form through three-dimensional repositioning of H3K9me3 regions and that HMGA1 serves as a “molecular glue” to join H3K9me3 regions in SAHF cores. The difference in size between SAHF and non-SAHF associated regions suggests that size plays an important role in the potential for displacement of heterochromatin towards the centre of the nucleus.

2.4.3 Interaction changes within and between TADs in RIS

By studying patterns of differential interactions I was able to characterise changes in TAD-TAD association between growing and RIS. Even if TAD borders were highly conserved, they could differ in strength, which would correspond to many interaction changes between consecutive TADs. Proportions of intra- and inter-TAD differential interactions changed with resolution (Figure 2.12). At high resolution, most of the interactions observed were intra-TAD, while at low resolution most were inter-TAD. Based on those proportions, I used 40kb resolution to analyse changes in interactions between consecutive TADs and within TADs and 200kb resolution for changes between distal TADs.

To define TADs which interact differently in RIS, I looked in particular for areas of either increasing or decreasing interactions. Checking for a large number of changes in the same direction helped select altered TADs more robustly. Only 75 TAD pairs were found to have more than 10 significant unidirectional differential interactions between them at 40kb resolution. The low number of consecutive TADs with significant alterations was consistent with the high NMI value between growing and RIS TAD borders. Interestingly, over 30 of those changes occurred between TADs with different A/B compartment assignments (Figure 2.13 (a)). Consistent with the increase in long-range interactions observed in RIS between heterochromatic regions, 476 distal TADs (separated by at least one other TAD) displayed more than 40 significant unidirectional changes in association at 200kb resolution. The vast majority of those distal TADs belonged to the B compartment (Figure 2.13 (b)). The largest number of changes occurred within TADs, with 767 TADs (out of 2,532) showing more than
10 significant unidirectional changes, distributed evenly across A and B compartments (Figure 2.13 (c)). This suggests a potential for differences in gene regulatory interactions during RIS.

Figure 2.12 Proportion of differential interactions situated intra-TAD (out of all the intra- and inter-TAD interactions) at different resolutions: 20kb, 40kb, 100kb and 300kb.

The largest change I observed in TAD structure occurred at the locus occupied by the Neuregulin 1 (NRG1) gene and consisted of a large area of decreased interactions both within a TAD and between consecutive TADs (Figure 2.14 (a)). The NRG1 gene is formed of multiple isoforms, out of which most of the shorter ones occupied the area of decreased interactions described. This gene was up-regulated in RIS (RNA-seq from Hoare et al. [50] and Tasdemir et al. [125]) and was previously described as a senescence marker [4] that encoded a secretory protein [124]. In the growing condition, the gene body of NRG1 resided within a heterochromatic TAD enriched in H3K27me3, neighbouring another H3K27me3 TAD (Figure 2.14 (a)). However, in RIS, the area of interaction of NRG1 with the rest of the TAD and with the TAD to its left was significantly decreased. This was accompanied with a slight increase over NRG1 in ATAC-seq signal and a decrease in H3K27me3 signal in RIS. This suggests that NRG1 might be up-regulated in RIS by release from a heterochromatic TAD, concomitant with chromatin opening.

I observed a similar pattern, albeit less pronounced, on the gene body of the Membrane metallo-endopeptidase (MME) gene (Figure 2.14 (b)). I observed reduced interactions within its encompassing H3K27me3-enriched TAD which potentially led to the release of MME from heterochromatin and its up-regulation in RIS (RNA-seq from Hoare et al. [50] and Tasdemir et al. [125]). MME is known for mediating senescence and also contributes to the secretory
phenotype [71].

Some of the changes observed in Figure 2.14 potentially correspond to structural loops (as defined by Rao et al. [105]), being marked by localised highly interacting loci. Therefore, I next investigated potentially disrupted loops in RIS due to changes in CTCF and cohesin binding.

Figure 2.13 Changes in TAD interactions classified into: (a) consecutive TAD-TAD interaction changes (b) distal TAD-TAD interactions, occurring between TADs separated by at least one other TAD (c) altered TAD interiors; and their respective assignments to A/B compartments.
Global features of chromatin interactions in growing and RAS-induced senescence

Figure 2.14 Changes in interactions within TADs: (a) interaction maps of TAD structure around the *NRG1* gene in growing and RIS, associated with loss of intra- and inter-TAD associations, with the black bar corresponding to its TAD, along with RNA-seq, ATAC-seq and H3K27me3 growing (G) and RIS (S) ChIP-seq normalised tracks; (b) interaction maps around the *MME* gene in growing and RIS, along with H3K27me3 normalised tracks.

2.4.4 Loop disruption during RIS

Cohesin and CTCF binding is associated with loop formation, as defined by Rao et al. [105] and described in chapter 1. At mRNA level, two out of the three cohesin subunits, *RAD21* and *SMC3*, were significantly down-regulated during RIS, although the *CTCF* level was unchanged (re-analysed RNA-seq data from [50]). Proteomics levels also showed down-regulation of RAD21 and SMC3 (proteomics data produced in the Narita Laboratory not shown). This suggested the possibility of loop disruption during RIS via reshuffling of cohesin binding sites.

In order to analyse this possibility, CTCF and cohesin (RAD21 and SMC3) ChIP-seq data were generated (by Dr. Aled Parry and Dr. Masako Narita, as described in chapter 5). SMC3 ChIP-seq was used mainly to confirm cohesin changes throughout this study. The RAD21 and the SMC3 signals agreed very well, with correlation values between 0.73 and 0.91 in growing samples and between 0.72 and 0.96 in RIS samples. I calculated the correlation values of the THOR-normalised signals counted in 5kb bins. RAD21, SMC3 and CTCF binding also agreed very well and they also correlate positively with nucleosome-free regions represented
by ATAC-seq (Figure 2.15).

I compared CTCF and cohesin binding changes between growing and RIS IMR90 cells. I obtained 44,764 and 53,563 CTCF peaks in growing and RIS, respectively. RAD21 ChIP-seq yielded 26,374 and 24,355 peaks in growing and RIS, respectively (Figure 2.16 (a)). A small number of significant binding changes (estimated with THOR [2], as described in chapter 5) relative to the total number of peaks were obtained for CTCF. Out of the 1,774 CTCF binding changes, 1,709 showed increased binding (Figure 2.16 (b)). RAD21 exhibited 4,553 significant changes in binding, many more relative to the total number of RAD21 peaks in each condition. 3,668 of those changes corresponded to decreased RAD21 binding (Figure 2.16 (b)), consistent with the decrease of gene and protein expression levels. Cohesin peaks largely overlapped CTCF peaks, with 88% and 92% agreement rates in growing and RIS, respectively. This was consistent with reports from other studies such as the one by Hansen et al. [46].

![Figure 2.15 Correlation matrix of 5kb binned THOR-normalised ChIP-seq signals of RAD21, SMC3, CTCF and ATAC-seq in growing and RIS using all replicates; green and blue squares indicate positive correlation and negative correlation, respectively.](image-url)
Figure 2.16 (a) CTCF and RAD21 peaks, growing/RIS specific or common between them; (b) Significant increases and decreases of CTCF and RAD21 binding in RIS; (c) CTCF/RAD21 peaks distribution compared to the distribution of binding changes in RIS across regulatory regions such as loops, promoters or enhancers; percentages calculated relative to the total number of peaks or binding changes (negative sign is used for symmetry); (d) potentially disrupted loops - six cases and their overlaps: loss/gain of CTCF or RAD21 binding at loops’ ends or gain of CTCF or RAD21 within a loop’s interior.
2.4 Differential interactions in RAS-induced Senescence

I next checked the distribution of CTCF and cohesin binding changes compared to that of the peaks across different types of regulatory regions. The definition of structural loops used was obtained from IMR90 loops information from the study by Rao et al. [105]. Some of the loops spanned very short (< 50kb) or very wide (over 5Mb, up to 218Mb) ranges and therefore for the purpose of this analysis I removed loops shorter than 50kb and wider than 1.5Mb in order to remove spurious associations. The final set consisted of 7,647 loops spanning 313kb on average. Figure 2.16 (c) shows that a large percentage of CTCF and cohesin changes reside within loops and at loop ends, especially in the case of RAD21.

The changes in RAD21 were also frequent at enhancers relative to the distribution across all peaks. To determine which loops were potentially disrupted by CTCF and cohesin changes, I checked six scenarios: gains or losses of CTCF or RAD21 binding at loops’ ends and within the interior of the loops. The latter may lead to new sub-loops forming or alternative loop’s ends. Figure 2.16 (d) shows the distribution of each of the six scenarios out of which cohesin loss at loop’s ends was the most frequent. In total, 3,960 out of 7,647 loops showed CTCF or cohesin binding changes and were potentially disrupted in RIS.

Figures 2.17 (a) and (b) show two examples of regions with loops decreasing in RIS associated with loss of cohesin at one or both ends. The same region mentioned earlier which contained the MME gene showed loop disruption (Figure 2.17 (a)), partially explaining the intra-TAD alteration mentioned. Decreased loop interactions in RIS were associated with decreased RAD21 binding at several loops’ ends. However, given the complex nested loop structure observed in both cases, it was difficult to unravel the consequences of such changes beyond the observed interaction alterations. Expression of several genes, such as MME, SLC33A1, CPOX, DCBLD2, COL8A1 or FILIP1L, was likely affected by such alterations.

In order to test whether such loops actually showed disruptions, I investigated differential interactions that overlapped potentially disrupted loops, whose interior area of interactions is likely disrupted. At 40kb resolution, 77.3% of the potentially disrupted loops showed significant interaction changes, while at 100kb resolution, 92.3% showed significant changes. The significant changes resided within the loop territory which was ensured by selecting only interactions where both ends were within the loop. Thus, the loops associated with cohesin or CTCF changes were likely disrupted. I used capture Hi-C to explore in more detail the potential for loop disruption in regions of interest, such as the IL1 locus. Loop disruption is potentially relevant for enhancer-promoter interactions occurring during RIS. I investigated this possibility in the following sections.
Loop reshuffling was reported in the context of senescence by Zirkel et al. [147]. The authors of Zirkel et al. described loop reshuffling as a consequence of CTCF clustering upon HMGB2 depletion during early replicative senescence. This suggests that loop disruption constitutes a possible mechanism of re-structuring chromatin during senescence and can occur due to either CTCF or cohesin changes.

Figure 2.17 Interaction maps at 20kb resolution of two regions which show decreased loop formation in RIS with matched ChIP-seq normalised tracks for histone marks or proteins of interest in growing (G) and RIS (S), as well as marked differentially expressed (DE) genes and IMR90 loop structure from Rao et al. [105] (a) the locus mentioned earlier around the MME gene showing a H3K27me3 TAD enclosed within H3K9me3 regions. Several loops in this region showed RAD21 loss at their ends. (b) another region on chromosome 3 characterised by a complex nested loop structure disrupted in RIS along with decreased RAD21 binding at several loop’s ends.

### 2.4.5 Comparison with other types of Oncogene-induced Senescence

Analysing HiC experiments of other types of senescence using differential interaction analysis allowed me to assess the generality of the present findings in the context of senescence. For this purpose, I re-analysed the data from Chandra et al. [13], who studied a different type of OIS, induced by the oncogene RAF. Despite the difference in the cell type (hTERT-immortalised...
WI38 human fibroblasts) and oncogene (RAF1) used to induce senescence, long-range interactions also increased and TAD boundary were conserved.

Differential interactions identified in our study and in Chandra et al. [13] showed 14% agreement at 10kb resolution and 25% agreement at 100kb resolution. These data suggest that, although the trend of large scale chromatin reorganisation appeared to be similar, particularly in the increase of long-range interaction between H3K9me3 TADs, individual differentially interacting pairs were substantially different depending on the context: different cell lines, oncogenes triggering senescence, and experimental procedures. As mentioned earlier, A/B compartments were also conserved between growing and OIS in the study by Chandra et al. [13].

2.5 Differential Enhancer-Promoter interactions

Previous studies showed that enhancer-promoter pairs are cell-type and lineage specific and, more importantly, that once they have been established they are unlikely to change with few exceptions. The principal mechanism of enhancer-gene regulation within the same cell type appears to be via the delivery of different levels of transcription factors. For example, Jin et al. [57] showed how TNFα induced a strong inflammatory phenotype in IMR90 cells via NFκB binding to enhancers and promoters, which were pre-looped. The enhancer-promoter interactions did not change with TNFα treatment. The only change was the increased NFκB binding to the enhancer and promoter regions, facilitated by the translocation of NFκB into the nucleus from the cytoplasm. RIS is also characterised by an inflammatory phenotype with activation of NFκB signalling [19]. However, as shown in the present study, RIS is governed by substantially different enhancer regulatory mechanisms and it potentially constitutes a breach of the lineage barrier regulation constraint.

I selected enhancer-promoter (EP) interaction changes in order to characterise the functional consequences of the large chromatin architecture changes observed. Promoters with differential EP interactions potentially move from one 3D genomic location into another. Therefore, they would be associated with reduced interactions (negative log-fold change) with the enhancers from their previous genomic location and increased interactions (positive log-fold change) with the enhancers from their new genomic location. Analysing both the provenance and current 3D genomic context of a gene may unravel the complex regulation that underlies the changes from growing to senescent cells.

EP changes were likely to occur based on previously published studies which highlighted extensive increases in chromatin accessibility (Parry et al. [92]) and the formation of new
typical and super-enhancers during oncogene-induced senescence (Tasdemir et al. [125]), coupled with many chromatin interaction changes observed both in this study and others (Chandra et al. [13], [147], [25]). However, enhancer-promoter interactions are thought to be established during lineage specification with small changes occurring within the same lineage (e.g. circadian rhythm context [76]), as mentioned earlier. Therefore, EP associations changing during RIS would constitute an exceptional feature which goes beyond the lineage barrier.

2.5.1 Annotation with enhancers and promoters

I annotated interactions ranging from 10kb to 100kb (increasing in steps of 5kb) with enhancers and promoters (Figure 2.18). GENCODE annotation (https://www.gencodegenes.org/releases/19.html) was queried for transcription start sites (TSS) and a region of 5kb around each TSS to define a promoter. I defined enhancers as regions with H3K4me1 (enhancer mark), H3K27ac (active enhancer mark) and ATAC peaks (chromatin accessibility mark) from ChIP-seq tracks in IMR90 obtained from data produced in the Narita Laboratory by Dr. Aled Parry. In a manner similar to the ROSE algorithm of selecting super-enhancers (SEs)[134] and to the variation used by Tasdemir et al. [125], I collapsed enhancer units closer than 12kb into one enhancer, without overlapping any promoters. I selected SEs by arranging the enhancers in order of their H3K27ac coverage and selecting the point of inflection in this distribution as the cut-off beyond which enhancers are considered super-enhancers. I observed similar numbers of typical and super-enhancers to the ones reported in Tasdemir et al. [125], with 1040 RIS SE and 1170 growing SE in the present dataset and 1250 reported in the study by Tasdemir et al. across multiple conditions.

Bins used for differential interactions were often larger than the promoter or enhancer annotated and therefore, I only used the promoters of differentially expressed genes between growing and RIS for annotation. This was based on the assumption that genes showing expression changes were the most likely targets in a bin involved in differential interactions.

I also calculated a trend of increase or decrease in H3K27ac for each enhancer, by counting the regions which show a differential trend in H3K27ac ChIP-seq data which I calculated using THOR [2]. However, due to the complex structure of enhancers which can be made out of numerous peaks with various changes, all enhancers annotated in growing and RIS were used. I clustered them by their H3K27ac trend into enhancers common between the two conditions, RIS- and growing-specific enhancers (Figure 2.19).
Enhancers are historically defined as having a high to low ratio of H3K4me1 to H3K4me3, compared to promoters which have low H3K4me1 and high H3K4me3. I checked this pattern for the enhancers and promoters used in the present study (Figure 2.19). The heatmaps show the signals of different marks around each peak associated with enhancers and promoters,
I combined interactions between 10kb and 100kb (every 5kb) in order to obtain a more complete network. If an enhancer-promoter (EP) interaction was observed at higher resolution it was not re-annotated in further iterations at lower resolutions. This combined approach was applied due to an effect observed in Hi-C contact matrices with interactions going “in-and-out of focus” (Figure 2.3).

**Differential EP interactions in Hi-C**

I performed differential interaction analysis as described earlier and found 17,050 differential EP interactions, some between the same enhancer and promoter due to multiple interactions spanning those regulatory regions. Such redundant pairs were not filtered out because they provided important information about the extent of the up- or down-regulation of the interactions. The number of unique pairs of enhancers and promoters with differential interactions was 16,862. Differentially interacting EP pairs involved 10,841 enhancers and 2,075 promoters, corresponding to 86% of the significantly differentially expressed genes between growing and RIS. 954 of the genes were up-regulated and 1,121 were down-regulated.

The extensive changes in EP interactions during RIS portrayed a large interwoven system with complex scenarios of regulation for each gene promoter. The differential EP network consisted of 641 (disconnected) components with the largest component consisting of 584 enhancers and promoters. An example of such a component is shown in Figure 2.20 (a), which illustrates the complex interactions linking genes together via enhancer sharing. Moreover, a large number of enhancers interacted with each gene in the differential EP network (Figure 2.20 (b)), with genes such as ACSL5 interacting with up to 27 enhancers. Recent studies have mentioned the existence of “shadow enhancers”, which contribute to buffering the gene expression levels [86]. Thus, our data suggested that the existence of redundant interactions occurring during RIS might contribute to the stability of the gene expression profile in RIS, a fate-determined state.
Figure 2.19 Chromatin marks profile of enhancers and promoters in growing and RIS; enhancers are represented by every H3K27ac peak and the surrounding 2kb region; promoters are represented by 5kb regions around TSS of every protein coding gene.
Differential interactions in capture Hi-C

I also performed differential interaction analysis using the capture Hi-C data in growing and RIS at *HindIII* resolution. The list of captured regions can be found in Appendix A, including loci that potentially undergo structural alterations during RIS based on our preliminary data as well as literature (e.g. the *IL1* cluster, as shown below). The analysis revealed 217 unique pairs of EP interactions. 171 of those were also present in the set obtained using genome-wide information.

Super-enhancer targets

Inflammation has previously been associated with the rapid formation of super-enhancers (SEs) [9]. The increased activity of SEs and their association with BRD4, a bromodomain and extra-terminal domain (BET) protein, in the RIS context have been previously studied by Tasdemir et al. [125]. Tasdemir et al. connect SEs with their nearest promoter as their target. SASP genes were shown to be in close proximity to SE. The targets I identified in this study (141 genes), which differentially interact with RIS-specific SE, were also enriched for SASP (Figure 2.21), suggesting that SASP genes constitute both the short- and long-range targets of RIS SE. This raised the possibility that SASP genes could be co-regulated via shared enhancers. This will be further investigated in chapter 4.
2.5 Differential Enhancer-Promoter interactions

(a) Up-regulated targets of SE gained in RIS

- Toll-Like Receptors Cascades
- Cellular Senescence
- Activation of Matrix Metalloproteinases
- Collagen degradation
- Degradation of the extracellular matrix
- Gastrin-CREB signalling pathway via PKC and MAPK
- Interleukin-1 signaling
- Senescence-Associated Secretory Phenotype (SASP)
- MyD88-Mal cascade initiated on plasma membrane
- Toll Like Receptor 2 (TLR2) Cascade

(b) Down-regulated targets of SE lost in RIS

- Cell Cycle
- Cell Cycle, Mitotic
- Extracellular matrix organization
- Scavenging by Class A Receptors
- Assembly of collagen fibrils and other multimeric structures
- Collagen formation
- Cell Cycle Checkpoints
- Collagen biosynthesis and modifying enzymes
- G2/M Checkpoints
- Leading Strand Synthesis

Figure 2.21 Enrichment of super-enhancer targets, from EnrichR [16], using Reactome database; (a) enrichment of up-regulated genes, targets of SE gained in RIS; (b) enrichment of down-regulated genes, targets of SE lost in RIS; all pathways shown are significantly enriched at $FDR \leq 0.05$.

Frequency of same-direction and repressive enhancer-promoter interactions

Activation events are considered to be increases in interaction of the gene with an enhancer concomitant with increased gene expression. Activation events are frequently mentioned in differentiation contexts [109]. Conversely, inactivation events are defined as decreased gene expression and decreased interaction of the gene with an enhancer. Interestingly, many more up- than down-regulated genes were involved in activation/inactivation enhancer interaction events. I observed a large number of repressive EP pairs (e.g. increased enhancer binding with gene down-regulation), some also accompanied by same-direction events involving the
Global features of chromatin interactions in growing and RAS-induced senescence

same gene promoter (Figure 2.22 (a) and (b)).

In order to assess whether this was due to having many false positives, I implemented several filtering strategies, such as: checking the proportions with minimum 5, 10 and 20 reads per interaction used as initial filtering, using only EP interactions obtained from Hi-C data of resolution higher than 30kb (i.e. 10kb, 15kb, 20kb and 25kb), increasing the FDR interaction threshold, the gene differential expression threshold, increasing the minimum log-fold change for the interactions and for gene expression, respectively.

All the comparisons yielded similar proportions of canonical and non-canonical events, suggesting that this result was not due to false positives. Moreover, the small number of EP interactions obtained from capture Hi-C, which suffer less from resolution biases, also included large numbers of both canonical and non-canonical events (Figure 2.22 (c) and (d)). Notably, a number of genes included in capture Hi-C were down-regulated but were only associated with increased EP interactions.

The functional relevance of such “repressive gene regulation” through EP interaction is not well understood. It is possible that genes associated with repressive events are targets of repressive TF-complexes, although the possibility that same-direction events were not captured in our data due to insufficient sequencing depth cannot be excluded. However, similar repressive regulation was previously observed during epidermal differentiation [109]. Interestingly, their data suggest that enhancers that gain contact with repressed gene promoters tend to be bivalent enhancers (i.e. enriched for both H3K4me1 and H3K27me3).

Repressive enhancer-promoter interactions were also observed by Kolovos et al. [61] in the context of TNFα-driven inflammatory response. How genes with both increased and decreased enhancer interaction are regulated is not clear. A subset of EP pairs might have a dominant role. Changes in insulation at a subTAD level could also influence the accessibility of the promoters to different sets of enhancers creating alternate patterns of increased and decreased interactions.
Figure 2.22 (a) Activating (red) and repressing (blue) EP interactions in Hi-C; (b) Combinations of up/down regulated genes with increased and/or decreased enhancer interactions in Hi-C; green and blue arrows correspond to up-regulated and down-regulated genes, respectively (c) and (d) the same as (a) and (b) in capture Hi-C.
Chromatin landscape of enhancers and promoters with differential interactions

I performed differential ChIP-seq analysis as well as peak calling (methods described in chapter 5) to check the DNA binding signal of several proteins and histone marks at the enhancers and promoters involved in differential interactions. The percentages of EP interactions associated with different epigenetic changes is shown in Figure 2.23. The majority of differential EP interactions were associated with changes in H3K27ac, a mark of active enhancers, and changes in ATAC-seq signal, which marks open chromatin. A number of EP interactions were associated with C/EBPβ and cohesin differential binding, either at the enhancer or at the promoter.

Figure 2.23 Changes in DNA binding or epigenetic marks associated with differential EP interactions, where one end or the other of the interactions displays differential ChIP-seq/ATAC-seq (nucleosome-free regions) signal.

Indirect regulation of genes via enhancer binding of TFs

Regulatory proteins can bind both the enhancer and the promoter of a gene, only the enhancer, or only the promoter. In order to investigate the two scenarios involving enhancer binding, I selected proteins for which both ChIP-seq DNA binding data and RNA-seq (or microarray) from knock-down experiments of the protein are available in growing and RIS. The frequency of genes where the protein selected binds their enhancers or both enhancers and promoters with differential expression upon RNAi for those proteins are shown in Figure 2.24.
A substantial proportion of genes whose enhancer was bound by the proteins of interest, were also differentially expressed. This was true especially in the case of p53, while for E2F7, HMGA1 and BRD4, only a fraction of the genes whose enhancer was bound were differentially expressed.

Figure 2.24 Indirect gene regulation by TF via enhancer binding; number of genes whose associated enhancer and/or promoter are bound by a protein such as TP53, E2F7, HMGA1 or BRD4 and the number of genes bound and differentially expressed upon knock-down (KD) of the respective protein.

I also investigated binding of several proteins to either an enhancer alone or both enhancer and promoter using ChIP-seq data produced in the Narita Laboratory, as well as peak definitions of several other proteins from REMAP [17] (method details in chapter 5). Figure 2.25 summarises the findings, where every row represents the number of EP interactions with
either enhancer binding or both enhancer and promoter binding of the proteins indicated.

C/EBPβ, a key TF driving senescence-associated secretory phenotype (SASP), frequently binds genes via enhancer only as well as via both the enhancer and promoter. C/EBPβ is known to be up-regulated during RIS and regulates both SASP and cell cycle [62], [115], the two main gene regulation axes in RIS. It acts as an activator of pro-inflammatory genes such as \textit{IL6}, \textit{IL1B} or \textit{IL8} and it represses cell cycle genes. This regulator links the activating and repressing EP contacts observed in this study, connecting the regulatory programmes behind SASP and cell cycle.

Consistent with its key role in EP loop structures, cohesin (RAD21) was among the top proteins which bound both enhancer and promoter pairs. I found few genes involved in EP interactions where cohesin bound only the enhancer, indicating a more structural role for cohesin involved in EP contacts. Three of the other top proteins that bound both the enhancer and the promoter in EP differential interactions were MXI, MAZ and MAFK. Those three proteins are related to gene repression and are highly predictive of enhancer-promoter interactions, according to Whalen et al. [133]. NR3C1 and C/EBPβ motifs are enriched in newly-formed superenhancers during adipogenesis in a study by Brown et al. [8]. NR3C1 (glucorticoid receptor) is known to form complexes with NFκB and AP-1 and prevent their
targets from activating [106]. Moreover, NR3C1 regulates the SASP and contributes to the inhibition of IL6, IL1A signalling and NFκB transactivation [63]. NR3C1 also appeared to be down-regulated at the protein level in RIS (proteomics data from the Narita Laboratory not shown). Its repression might contribute to the activation and specification of NFκB and AP-1 targets via enhancer binding in RIS.

An interesting case consists of p53 binding to the enhancer of TIMELESS, but not to its promoter, during RIS (Figure 2.26). The interaction between the TIMELESS enhancer and the promoter was increased, whereas TIMELESS was down-regulated, during RIS. TIMELESS is the mammalian homologue of Tim, a circadian regulator in flies, but has been implicated in cell cycle progression and DNA damage response [74]. It has been proposed that many cell cycle genes, including TIMELESS, are targets of p53-mediated “indirect” repression through the DREAM complex [39]. It would be useful to revisit this possible mechanism for enhancer-bound p53-mediated repression of cell cycle related genes.

Figure 2.26 Potential negative regulation of TIMELESS regulation by p53 via enhancer binding; ChIP-seq tracks include P53 binding, H3K27ac and H3K4me1 (marking enhancers) and ATAC-seq (nucleosome-free regions) in RIS; top region focuses on the promoter of TIMELESS and the bottom region focuses on the enhancer of TIMELESS, as determined from the differential EP contacts studied with Hi-C in growing and RIS.
Longer-range EP interactions

In our data, the distance separating most EP pairs (90%) was below 2Mb, in accordance with previous studies which only look for EP pairs within this range [57]. Consistently, I found only a few genes separated by their corresponding enhancers by more than 2Mb in our differential EP interactions, including *BUB1* (mitotic gene), *AURKA* and *COL10A1*. In order to determine how many differential EP interactions were within the same TAD, I plotted the frequency of EP interactions against numbers of TADs separating them in Figure 2.27. The majority of EP interactions were within the same TAD (56%), while 26% were in consecutive TADs. This is thought to be due to insulation between different TADs and consistent with the distance effect previously reported, where interactions between closer regions are more frequent [70],[36].

![Figure 2.27](image_url)

Figure 2.27 Differential EP contacts are infrequently separated by more than two TADs; x-axis represents the number of TADs separating the ends of differential EP interactions; y-axis indicates the frequency of “x” TADs separating enhancers and promoters.

Loop disruption as a mechanism of rewiring EP interactions

Conceptually, loop disruption is a possible mechanism behind rewiring EP interactions. For example, enhancers residing inside the same loop could be released to interact with other partners via loop disruption (Figure 2.28). In this context, I refer to loops as defined by Rao et al. [105]. Changes in interaction probability of sets of enhancers with a gene promoter, due to loop disruption, would result in alternating patterns of increased and decreased interactions. Similarly, a newly formed loop could insulate a set of enhancers from a promoter resulting in
decreased interactions with those enhancers. Changes in the relative position of an enhancer and a promoter in the nested loop hierarchy could potentially determine the direction of interaction change. Enhancers clustering dependent on cohesin-bound loops was previously proposed by Ing-Simmons et al. [56].

Loops were reported to be bound by CTCF and cohesin by Rao et al. [105]. I therefore investigated the binding of those two proteins using ChIP-seq (described in chapter 5). Many promoters involved in differential interactions studied overlapped cohesin and CTCF peaks (41% and 54%, respectively). A number of enhancers (differential interaction partners) were also bound by cohesin and CTCF (16% and 22%, respectively). However, only 10% of the differential EP interactions overlap regions with increased or decreased cohesin binding. Therefore, EP interactions might be affected more by changes in the loops encompassing them, rather than changes in cohesin binding over the enhancer or the promoter. To investigate this, I checked enhancers and promoters inside potentially disrupted loops. The possibility of loop disruption in RIS was described earlier in this chapter. Interestingly, 63% of differential EP interactions were linked to potential loop disruptions (defined earlier in this chapter), with either the promoter or enhancer residing at either loop end or inside the disrupted loop.

Moreover, enhancers with positive and negative interactions with a promoter, respectively, appeared clustered along the genome and the clustering seemed to be delimited by loops. The loops mentioned here were defined by Rao et al. [105] in growing IMR90 cells. Figure 2.29 shows the Hi-C interactions at 20kb resolution in the region composed of the promoter of \textit{E2F7} and its associated enhancers. Many of the loops surrounding enhancers, involved in
differential interactions, showed decreased cohesin binding at least at one of their ends. Thus, our findings suggest that changes in EP interactions during RIS were associated with, and possibly partially driven by, loop disruptions.

Figure 2.29 Hi-C maps in growing and RIS around the region consisting of $E2F7$ promoter and its associated enhancers (defined by differential EP interactions in RIS); enhancers: green - increased interactions and blue - decreased interactions with $E2F7$ promoter; loops - violet segments, derived from high-resolution Hi-C map by Rao et al. [105] in IMR90 cells.
Potential decoupling of cell cycle and SASP via competitive enhancer interactions

It has been shown that multiple genes can be co-regulated through binding common sets of enhancers [57]. Extensive enhancer sharing by multiple genes was also found in our study, in the differential EP interactions (see Figure 2.20). Interestingly, 440 gene pairs sharing enhancers were regulated in opposite directions. More precisely, in each pair of genes, one gene was down-regulated and the other up-regulated and the interactions between shared enhancers and one of the genes were increased and decreased with the other gene. This suggests the presence of “competitive enhancer binding” during RIS, leading to increased efficiency of TF delivery to some gene promoters instead of others.

![Diagram of competitive enhancer binding](image)

Figure 2.30 Enhancers interacting preferentially more with SASP genes and less with cell cycle genes; gene set enrichment analysis of up- and down-regulated genes from the competitive pairs of genes which show increased and decreased interactions with the same enhancer, respectively.

I performed gene enrichment analysis to determine which processes are favoured or disfavoured by increased and decreased EP interactions (Figure 2.30). Consistent with the cell cycle arrest phenotype, down-regulated genes were mostly associated with cell cycle genes. Up-regulated genes were enriched for immune-related gene sets, possibly reflecting SASP, a prominent feature of senescence. The actual pairs of cell cycle and SASP genes are listed in Table 2.2. SASP genes were selected from Hoare et al. [50] and cell cycle genes were selected from Reactome cell cycle pathway and manual curation in the case of CKAP2L (mitotic spindle protein, required for cell cycle progression, [54]) and CDCA7L (cell division gene family member 7L).
cycle-associated 7-like protein). Several interesting cases emerged: for example, \textit{IL1A}, \textit{IL1B}, and \textit{IL6}, central components of the SASP shared common enhancers with cell cycle related genes.

Of note, this analysis only included differential interactions between growing and RIS. In chapter 4, I analysed significant EP interactions in growing and RIS, respectively, in order to determine co-regulated cellular processes via enhancer sharing. The \textit{IL1} locus is highlighted below as an example of a mechanistic link between the two senescence-associated phenotypes, stable arrest and the SASP.

<table>
<thead>
<tr>
<th>Gene1 (up)</th>
<th>Gene2 (down)</th>
<th>Gene1 logFC</th>
<th>Gene2 logFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{IL1B}</td>
<td>\textit{CKAP2L}</td>
<td>9.1099858</td>
<td>-2.1540103</td>
</tr>
<tr>
<td>\textit{IL1A}</td>
<td>\textit{CKAP2L}</td>
<td>6.4575638</td>
<td>-2.1540103</td>
</tr>
<tr>
<td>\textit{EMILIN2}</td>
<td>\textit{NDC80}</td>
<td>1.408936</td>
<td>-3.2137304</td>
</tr>
<tr>
<td>\textit{IL1A}</td>
<td>\textit{BUB1}</td>
<td>6.4575638</td>
<td>-3.1380538</td>
</tr>
<tr>
<td>\textit{FGF2}</td>
<td>\textit{CCNA2}</td>
<td>0.9487626</td>
<td>-3.6328749</td>
</tr>
<tr>
<td>\textit{IGFBP5}</td>
<td>\textit{BARD1}</td>
<td>1.2604891</td>
<td>-3.673169</td>
</tr>
<tr>
<td>\textit{PAPPA}</td>
<td>\textit{CNTRL}</td>
<td>1.868747</td>
<td>-3.166054</td>
</tr>
<tr>
<td>\textit{IL6}</td>
<td>\textit{CDCA7L}</td>
<td>3.2862932</td>
<td>-1.7879317</td>
</tr>
<tr>
<td>\textit{ST6GAL1}</td>
<td>\textit{RFC4}</td>
<td>0.906291</td>
<td>-2.0930646</td>
</tr>
<tr>
<td>\textit{VGF}</td>
<td>\textit{MCM7}</td>
<td>2.6412458</td>
<td>-1.2862566</td>
</tr>
<tr>
<td>\textit{WNT9A}</td>
<td>\textit{L1N9}</td>
<td>1.4383297</td>
<td>-1.786814</td>
</tr>
<tr>
<td>\textit{IL1B}</td>
<td>\textit{BUB1}</td>
<td>9.1099858</td>
<td>-3.1380538</td>
</tr>
<tr>
<td>\textit{TMEM156}</td>
<td>\textit{RFC1}</td>
<td>2.2842008</td>
<td>-0.9765483</td>
</tr>
<tr>
<td>\textit{MANBA}</td>
<td>\textit{CENPE}</td>
<td>1.1161488</td>
<td>-3.5013002</td>
</tr>
<tr>
<td>\textit{MANBA}</td>
<td>\textit{H2AFZ}</td>
<td>1.1161488</td>
<td>-0.7959192</td>
</tr>
</tbody>
</table>

Table 2.2 Pairs of up-regulated SASP genes - Gene1 and down-regulated cell cycle genes - Gene2, that share enhancers with opposite interaction directions (increased with Gene1, decreased with Gene2).

\textit{IL1} locus (capture Hi-C)

Interleukin-1 (\textit{IL1}) “ancestral family” members (9 out of 11 “super-family” members) are located around chr.2q14 with several other genes, such as cytoskeleton-associated protein 2 like (\textit{CKAP2L}), encoding a mitotic spindle protein [107]. This cluster includes both \textit{IL1A} and \textit{IL1B}, encoding IL1α and IL1β, respectively, and their strong up-regulation is essential for development of the inflammatory SASP [100]. Thus, the \textit{IL1} cluster was included in the regions for capture HiC. The \textit{IL1} locus has also been associated with super-enhancers, which
are activated during RIS [57], [125].

I observed many alterations in EP interactions involving the promoters of IL1A and IL1B in RIS, mostly increases. In contrast, enhancer interactions were mostly decreased with the promoter of CKAP2L (cell cycle related gene), which was down-regulated during RIS (Figure 2.31). Notably, the IL1B promoter exhibited decreased interactions (blue arcs) with the enhancers to the left and increased interactions (green arcs) to the enhancers on the right side. CKAP2L showed decreased interactions with the same enhancers that showed increased interaction with IL1B. The IL1A promoter showed increased interactions with its immediately surrounding enhancers during RIS. This suggests that there might be a new insulator element, which blocks interactions between the two sides of the locus, represented by the decreased interactions (around IL1B region) in Figure 2.31.

As discussed earlier, the differential EP interaction analysis in our “standard” (not captured) Hi-C data showed the potential regulatory involvement of loop disruption mediated by cohesin alteration (Figure 2.16 (c)). Cohesin binding indeed changed during senescence near IL1B, with a new peak forming around IL1B in RIS cells (Figure 2.31). This supports the hypothesis of a new loop forming, which would only include IL1A, CKAP2L and their enhancers, insulating them from IL1B, which would thereby interact more frequently with the remaining enhancers available on its right side.

The preferential enhancer interaction I observed in the IL1 locus possibly constitutes a mechanism of increasing the efficiency of delivering smaller amounts of NFκB and achieving similar up-regulation of pro-inflammatory genes as in the acute case represented by TNFα treatment. Jin et al. [57] showed that interactions between enhancers and CKAP2L, IL1A and IL1B were pre-existent and that TNFα treatment results in co-regulation of those genes via NFκB binding to the enhancers. Thus, gene regulation is achieved via increased binding of a transcription factor, rather than increased EP interaction. Jin et al. [57] proposed that this trend applies to most tissue-specific EP pairs. Our data suggest that senescence employs a distinct gene regulatory mechanism, which involves the dynamic remodelling of the EP interactive network (2.33).

In addition, CKAP2L is a cell cycle related gene and it is tempting to speculate that direct competition between SASP genes and cell cycle genes for enhancers might contribute to the reinforcement of phenotypic stability. Interestingly, IL6, another central SASP gene, shared enhancers with CDCA7L, a putative cell cycle-related gene, showing a similar trend (preferential EP interaction). CDCA7L (also called transcription factor R1 and JPO2), may act upstream of cyclin D1 and E2F1 [87].
Figure 2.31 (a) Hi-C map at HindIII resolution, corresponding to differential interactions (capture Hi-C) in the IL1 locus, with arcs corresponding to significant interactions between enhancers (grey rectangles) and promoters (coloured and labelled rectangles by their expression: green corresponds to up-regulated and blue to down-regulated genes in RIS); blue arcs correspond to decreased interactions and green arcs to increased interactions; the arcs are drawn using different promoters as anchors (IL1A, IL1B and CKAP2L), in order to extract the relevant interactions; RAD21 binding in growing and RIS is represented by THOR-normalised signal counted in every HindIII fragment; (b) our model for loop disruption in the IL1 locus, leading to two sub-loops forming.

Interestingly, interactions at the MMP locus showed a similar pattern to the IL1 locus. I observed increased insulation represented by a pattern of decreased interactions (Figure 2.32) between MMP10 and MMP1 that coincided with the appearance of a new cohesin peak (RAD21) which lacked CTCF binding.
Figure 2.32 Interaction map of differential interactions from capture Hi-C at the MMP locus at 5kb resolution, with arcs anchored on the promoters of MMP10, MMP1, MMP3 and MMP12 corresponding to significantly changing EP contacts; ChIP-seq tracks focus on the region around the promoters of MMP10 and MMP1 showing the new cohesin peak forming in RIS (SMC3 and RAD21), as well as CTCF and H3K27ac ChIP-seq normalised tracks in growing and RIS.
A large number of differential interactions within the same lineage is senescence specific

In order to assess how common is the observed reorganisation of the EP interaction landscape, I re-analysed the data from Jin et al. [57] with the method for detecting differential interactions described in the present study. As mentioned above, the study by Jin et al. highlights the pre-existence of EP interactions and no change upon TNFα treatment on IMR90 cells. Consistent with their conclusions, no differential interactions (EP or otherwise) were detected at the 5% FDR level, tested at multiple resolutions. I repeated the analysis with all the replicates from their study and with subsets of them to ensure the lack of interactions is not an effect of having multiple variable biological replicates. This reinforces that senescence and TNFα treatment employ distinct gene regulatory mechanism.

Hi-C data from Chandra et al. [13] were also re-analysed for differential interactions, as mentioned earlier: they used RAF-induced senescence in hTERT WI38 cells (WI38 and IMR90 cells are both human embryonic lung fibroblasts). I used the same enhancers and promoters regions for annotation in order to test whether the enhancer-promoter pairs show different interactions patterns in senescence, despite the difference in cell types and oncogene activation.

In contrast to the TNFα treatment, 29,798 differential enhancer-promoter interactions were identified during RAF1-induced senescence, with 7,095 unique EP pairs shared with RAS-induced senescence (analysed in the present study). The caveat of this analysis was the lack of ChIP-seq/expression data for WI38 cells which was replaced with the IMR90 information used in our RIS study in order to test the same interacting regions in both types of senescence. Importantly, the same interaction trend of IL1A and IL1B was present in both OIS studies.
Figure 2.33 Enhancer-driven gene regulation. Current model: in a given cell type, EP interactions are mostly pre-existing and genes are activated upon increased transcription factor binding to the enhancer-promoter (top). For example, \textit{IL1A}, \textit{IL1B} and \textit{CKAP2L} were shown to be co-regulated upon TNF\(\alpha\) treatment without any altered interaction of EP, to which NF\(\kappa\)B is recruited (middle). In contrast, during RIS (bottom), those shared enhancers preferentially interacted with \textit{IL1B} contributing to its up-regulation and down-regulation of \textit{CKAP2L}. 
Enhancers cliques

In this study, I also investigated enhancer-enhancer interactions. Surprisingly, the majority of the enhancers (94%) identified in growing and RIS participated in differential interactions with other enhancers. Moreover, tightly linked structures forming cliques (groups of fully connected nodes – see chapter 1 for introduction to graph theory) or approximate cliques were identified. Those structures consisted of large number of enhancers with increased interactions. Examples of such structures are depicted in Figure 2.34.

![Enhancer-enhancer interaction components](image)

Figure 2.34 Enhancer-enhancer interaction components where each circle represents an enhancer and the colour determines whether it is a typical (grey) or a super-enhancer; blue lines represent decreased interactions and green lines represent increased interactions.

I used community detection via fast-greedy modularity optimisation (implemented in R package igraph [28]) to further investigate enhancer-groups. 200 communities with more than 10 enhancers were determined. The log-fold changes of the interactions in each community were overwhelmingly positive, with 164 communities having over 60% increased interactions (out of all the interactions inside each community). 91 of those communities were almost entirely made out of increased interactions. Exact cliques (groups of enhancers connected all-to-all) were also determined and the distribution of their sizes was depicted in Figure 2.35. Cliques of 3 and 4 enhancers were the most frequently observed, while only 151 cliques of size 12 and 9 cliques consisting of 13 enhancers were observed.
2.6 Discussion

Figure 2.35 The distribution of frequency of enhancer cliques of different sizes; enhancer clique size varied between 1 and 13 enhancers, out of which cliques with 3 and 4 enhancers were the most frequent occurrence.

SEs consist of a large number of H3K27ac peaks which co-localise. SEs also accumulate more H3K27ac than all other enhancers. Thibodeau et al. identified a specific pattern of interaction of SE. Their enhancer sub-units were highly interacting within the SEs but the SEs as a whole exhibited low interaction with other enhancers. The large number of interactions within enhancer clusters, similar to the pattern observed within SE raises the possibility that in three-dimensional space, typical enhancers (TE) clusters and SEs might be indistinguishable, at least in the context of RIS.

2.6 Discussion

The large number of interaction changes detected from Hi-C data in RIS is consistent with the dramatic chromatin alterations described in literature. In particular, I studied HMGA1-dependent long-range interactions and generated an estimate of the genomic regions forming SAHF in RIS. SAHF H3K9me3 regions were larger, more gene-poor, overlapped more with LADs lost in RIS and showed accumulations of HMGA1 binding compared to other H3K9me3 regions. Such distinct properties likely contribute to the potential of heterochromatic regions of being displaced towards the centre of the nucleus and showing higher affinity to similar heterochromatic regions in RIS.
Global features of chromatin interactions in growing and RAS-induced senescence

I also observed local interaction changes during RIS, not just between large regions situated far apart from each other. Such interactions potentially correspond to EP contacts which contribute to gene regulation. The few examples of chromatin conformation studies within the same lineage point towards a TF-driven gene regulation model which utilises pre-looped EP interactions. In contrast, EP changes have been recorded in differentiation and malignancy. RIS shows a large number of changes in EP contacts suggesting it transcends the lineage barrier.

Loop disruption, defined as changes in cohesin (RAD21 and SMC3) binding at loops’ ends, constitutes a potential mechanism for the large number of EP changes observed in RIS in our data. The expression decrease of the genes encoding the RAD21 and SMC3 proteins is consistent with cell cycle arrest and is possibly linked to the changes in binding of cohesin which are mostly decreases. Despite the overall trend of decreased binding, specific loci (IL1 and MMP) show increased binding and the appearance of a larger cohesin peak lacking CTCF binding. The larger size of the cohesin peaks may be due to the lack of CTCF binding which would stabilise the dynamic behaviour of cohesin binding. Cohesin potentially encounters a different binding partner at these locations which is less efficient at stabilising its binding. Further study is required to find the alternative binding partner of de novo cohesin peaks in RIS.

Most studies to date focus on the effects of either cohesin depletion or overexpression. The effect of cohesin alteration on gene expression has been characterised as limited in some studies such as the one by Rao et al. [104]. However, other studies link cohesin disruption to important gene regulatory effects. For example, Cuartero et al. [29] show that cohesin depletion leads to impaired inflammatory phenotype with relevance for acute myeloid leukaemia patients which display cohesin mutations. SMC1A has been reported as elevated in colorectal cancer (Wang et al. [130], Sarogni et al. [113]), as well as androgen-independent prostate cancer cell lines (Pan et al. [88]) and hepatocellular carcinoma (Zhang et al. [146]). Depletion of SMC1A leads to cell cycle arrest and apoptosis in the study by Wang et al. In our study, cohesin binding showed both increases and decreases. This raises the possibility of a more complex mechanism which links cohesin to gene expression and chromatin structure alterations which takes into account multiple directions of binding changes, instead of those observed during overall depletion or overexpression of cohesin, which are artificial systems studying phenotypic extremes.

The network of EP interactions I observed likely constituted a subset of all EP changes between growing and RIS, due to the sub-saturated nature of Hi-C data. Thus, it is likely that our observation is only a subset of the entire events representing the phenotype of interest. In addition, the differences between the bin sizes used (10 - 100kb) and sizes of regulatory regions (promoters ∼ 5kb and enhancers 1 - 134kb) might confer false positives. Although the
Hi-C capture technique used herein alleviates some concern, deeper sequencing is needed to include all genome-wide interactions. The method employed here relies solely on differential chromatin interactions between biological conditions. It highlights differential events but it does not represent common events between the biological conditions – this will be addressed in chapter 4. The list of differential interactions obtained is difficult to contextualise due to the large number of changes between many genomic regions. In the next chapters, graph theory methods will be applied in order to understand the scale of the changes in chromatin architecture that occur in RIS.
Chapter 3

Graph theory approach to characterising chromatin structure

3.1 Introduction

Each graph with $n$ vertices can be represented with an $n \times n$ adjacency matrix $A$, where each row and each column corresponds to a vertex. Each entry of the matrix, $A_{ij}$, is 1 if there is an edge between vertices $i$ and $j$, or 0 if the two vertices are not connected. $A_{ij}$ can also be the weight of the edge between vertex $i$ and vertex $j$. Hi-C interaction matrices can be used as adjacency matrices to represent the chromatin interaction network. In this case, each genomic bin in the Hi-C matrix is a vertex and the interaction value between two bins is the weight of the edge (interaction) between them (Figure 3.1). This allows for interrogation of interesting graph theoretical properties for gaining an understanding of chromatin interactions.

Previous studies performed community detection in order to uncover chromatin structure information from Hi-C data. Two notable attempts include the studies by Norton et al. [85] and Pancaldi et al. [89]. Norton et al. [85] used a modified version of modularity optimisation in order to detect topological associating domains from Hi-C data. Pancaldi et al. [89] combined Hi-C and ChIP-seq data and used chromatin assortativity to detect the community structure of chromatin.

In this study, I applied a recent method of community detection to determine the hierarchy of chromatin in growing and senescent (RIS) cells (Figure 3.2). RIS represents an interesting case-study for detecting large chromatin rearrangements, which would manifest as changes in the hierarchy of chromatin units in the nucleus. The hierarchy of homogeneous cell populations will reflect the actual nuclear distribution of chromatin.
Figure 3.1 Representation of chromatin interactions as a weighted graph by connecting genomic bins (vertices) using chromatin interactions (edges), weighted by their intensity calculated from chromosome capture data.
Heterogeneous populations will highlight the most frequent association of different genomic regions.

The nature of the method chosen allows for estimating the stability of the community calling. Thus, heterogeneous populations would possibly be characterised by unstable communities with groups of regions oscillating between community assignments. For example, assume the cells studied consisted of two sub-populations which had distinct chromatin conformations. This would manifest in the hierarchy as unstable communities with probabilities of community association reflecting the proportions of the cell populations, such as one conformation is observed in 80% of the iterations and the other conformation in the other 20%. Thus, the method has the potential to assess the heterogeneous nature of senescence [136], and other heterogeneous samples. For example, tumour samples suffer from heterogeneity due to the presence of a mix of normal and cancer cells.

SAHF formation is an important phenomenon that could be analysed with community detection for its representation in terms of H3K9me3 associations. In chapter 2, I analysed Hi-C data in the context of SAHF formation, but I only identified individual changes in interactions that possibly participated in SAHF formation. Detecting changes in hierarchical structure would allow for those individual changes to be contextualised and would estimate how do putative SAHF units combine together.

In terms of community structure, popular algorithms impose a rigid range of community sizes and are dependent on input parameters. In this study, I applied a non-parametric Bayesian version of Stochastic Block Models introduced by Peixoto et al. [97]. This method overcomes the dependency of the community structure on input parameters (e.g. number of communities). Moreover, it offers a representation of the stability of the community calling via marginal probabilities calculated over a large number of iterations.

The network is modelled as a nested community structure, allowing for small and large communities to co-exist. The hierarchical nature of chromatin interactions was described previously [132] and is thus, compatible with a nested community structure.
Figure 3.2 Nested community structure detection by splitting the genome into equally sized bins whose relative position in the nucleus is reflected in the hierarchy detected, with neighbouring genomic regions being grouped within the same community.
3.2 Non-parametric Bayesian inference adaptation of Stochastic Block Models

To analyse the Hi-C network, I adapted the methodology introduced and implemented by Peixoto et al. [98], [97], which describes a variation of stochastic block models (SBM) for detecting meaningful blocks in a network. The Hi-C network was obtained by binning the genome and using those regions as nodes and the interactions between different regions as edges.

Traditionally, a stochastic block model is a generative model used to probabilistically generate networks, given their block structure. Here, a block is not restricted to the more popular, yet restricting, definition of communities which refers to groups of nodes more connected within, than with the rest of the network. Blocks can refer to more general structures, defined as groups of nodes with specific probabilities of each pair of blocks having \( n \) edges. Therefore, other types of structures can also be unveiled by using stochastic block models. For example, bipartite graphs are formed by two blocks of nodes such that the only edges in the graph are between pairs of nodes belonging to different groups. This would correspond to maximising the probability of having the maximal number of edges between the two blocks. At the other extreme, minimising the probabilities of having large numbers of edges between blocks can lead to the classic community model, with blocks more connected within than with other blocks.

The probability of observing a network given a set of parameters that describe its structure (e.g. block membership, number of edges between each pair of blocks, degree sequence) can be written as \( P(G|b, \theta) \), where \( G \) is the network, \( b \) is the block structure and \( \theta \) refers to other parameters. This can be re-written using Bayesian probability to express the probability of a block structure given that we observe the network \( G \):

\[
P(b|G) = \frac{P(G|b, \theta)P(b, \theta)}{P(G)}
\]  

(3.1)

Equation 3.1 is the likelihood of the block structure given that network \( G \) is observed. Maximising it results in obtaining meaningful block structures in the data. The problem can be further re-written as an information theoretical one, by using the description length of \( G \): \( \sigma = -\ln P(G|b, \theta) - \ln P(b, \theta) \). Then equation 3.1 becomes equation 3.2.

\[
P(b|G) = \frac{\exp(-\sigma)}{P(G)}
\]  

(3.2)

Minimising the description length of \( G \) becomes equivalent to maximising the likelihood of block structure \( b \), given that \( G \) is observed.
Monte Carlo Markov Chains (MCMC) equilibration

The best partition (block structure of $G$) is obtained by maximising the posterior probability $P(b|G)$. This is the same as minimising its description length. This is achieved by employing Monte Carlo Markov Chains (MCMC) [95]. Starting with an initialised block structure, attempts are made to move nodes (genomic regions) from a block to another, concomitant with observing whether the description length (entropy) decreases.

Problems such as slow mixing and local minima are solved with various strategies: e.g. probabilistically moving a node to another block by taking into account the block membership of its neighbours and accepting/rejecting node moves (Metropolis Hastings) with a probability (dependent on the membership of the neighbouring nodes).

One MCMC sweep consists of attempting to move all the nodes in the network and recording the change in entropy, as well as the number of accepted moves. Several iterations of sweeps can be applied until a large number of iterations undergo no significant changes in entropy. This is parametrised as the waiting time ($\text{wait}$) in which no significant changes are observed ($\text{nbreaks}$). Once this equilibration occurs, model averaging can be performed by sampling from its posterior. This is implemented in graph-tool as $\text{mcmc\_equilibrate}$ [97].

Agglomerative heuristic

Starting from a random configuration leads to significantly longer running times [95]. Therefore, Peixoto developed the agglomerative heuristic as a strategy to obtain a meaningful configuration for the block structure which can be used as a stand-alone result or as the starting point for further refinement using the MCMC procedure described earlier. This consists of starting with the maximal block structure where each node is its own block and then merging them iteratively into larger blocks using the same informed strategy for moving nodes between block as earlier, dependent on the membership of the neighbouring blocks. This is implemented in graph-tool as $\text{minimise\_blockmodel}$ [97].

Edge weights

This methodology was further extended by Peixoto to consider edge weights. Those are treated as covariates and several distributions are incorporated for discrete and continuous edge weights [99], such as the exponential and normal distributions for continuous weights and geometric, binomial and Poisson for discrete weights.
Degree-corrected model

The degree sequence or distribution of a network is the set of degrees of all vertices in the network. This measure is important for determining whether some vertices are much more connected than others or if more uniform connectivity is observed in the network. A degree-corrected formulation is available in graph-tool [97], which is advantageous for describing real-world networks, as discussed in the chapter 1. In the degree-corrected version, the degree distribution in each block of nodes is non-homogeneous, as opposed to the traditional stochastic block models where nodes in each block tend to have very similar degrees. This is achieved by adding the degree sequence as an additional covariate.

Hierarchical block structure

Community detection algorithms tend to suffer from a major limitation characterised by the existence of a lower bound on the size of the blocks identified. Therefore, smaller blocks than the lower limit cannot be detected. For methods such as modularity optimisation, blocks with fewer edges than $\sqrt{E}$, where $E$ is the total number of edges in the network, are unlikely to be detected [43], [96]. Similarly, minimising description length in order to detect block structures is unlikely to identify blocks with fewer nodes than $\sqrt{N}$, where $N$ is the total number of nodes in the network. In 2014, Peixoto introduced a hierarchical variation of the model which solved this problem by estimating a multilevel nested block structure, where the first level consists of groups of nodes and the second level consists of groups of groups of nodes until all groups collapse into a single one [96]. This is achieved by using any higher level block structure as a prior for the block structure at a lower level.

Non-parametric method

The non-parametric nature of this method arises from its ability to infer parameters such as the number of blocks [99] from the data. Most methods which infer block or community structures require a number of input parameters and the results are highly dependent on the values used as input. Peixoto developed a strategy which uses Bayesian inference to determine those parameters from the data, by describing probabilistic generative models for the network but also for its block structure, degree sequence and other parameters.

Computational complexity

One MCMC sweep, implemented in graph-tool as `mcmc_sweep`, which attempts to move every node to another block, has complexity $O(E)$ where $E$ is the number of edges in the network. The computational complexity of the agglomerative heuristic, used to obtain a reasonably good starting block structure to be further refined with MCMC sweeps, is $O(N\ln^2 N)$, where...
N is the number of nodes in the network [95].

Many other extensions to the methodology introduced by Peixoto were published, such as overlapping blocks or predicting missing edges in a network. However, only the ones relevant to its application to Hi-C data were described in this chapter.

### 3.3 Applying modified SBM to Hi-C data

I used the graph-tool implementation by Peixoto [97] of the non-parametric Bayesian inference adaptation of SBM to detect hierarchical structures of chromatin interactions from Hi-C data. I constructed the network with genomic bins as nodes and their pairwise interactions as edges. Edge weights are represented with ICE-normalised counts averaged across replicates (correctedContact from R package diffHiC [73]). I created a series of scripts which built the Hi-C network, removed disconnected nodes and executed the agglomerative heuristic. This was then used as a starting point for the MCMC equilibration. Finally, marginal probabilities corresponding to the stability of the number of blocks and of the block membership of the nodes at every hierarchy level were computed. The output consisted of the hierarchy structure, stability values and a graph with the hierarchy members and their relationship.

The input files required were the interactions file and the BED file with bins coordinates (output by HiC-Pro [119] alignments) and a list of bins to be used for the creation of the network. The network was then created and minimize_nested_blockmodel_dl was run with arguments consisting of degree-correction and the edge weights (interaction ICE-normalised values) following the exponential distribution (implemented in graph-tool as “real-exponential”). The nested block structure obtained was then used as input for mcmc_equilibrate where one iteration consisted of niter sweeps of the network, and was considered equilibrated after nbreaks sequences of wait iterations occurred with no significant change in entropy, defined with \( \sigma \).

After the MCMC was equilibrated, an additional 10,000 iterations of 10 sweeps each were run, in order to estimate the node marginal probabilities. Those values reflect the stability of the communities (blocks) they belong to. Similarly, another 10,000 iterations of 10 sweeps each were run, in order to estimate the marginal probabilities of the number of blocks in each level of the hierarchy. The two final steps were effectuated without any stopping conditions and enforcing the specified number of iterations.
3.3 Applying modified SBM to Hi-C data

3.3.1 Choosing the number of iterations required

Figure 3.3 shows the evolution of entropy values with different values for the variables wait, nbreaks and niter, illustrated for chromosome 14 at 1Mb resolution. Some of the combinations, such as $\text{wait} = 10000$ or $\text{nbreaks} = 100$ significantly increased the running time of the algorithm and were not feasible for analysing higher resolution Hi-C matrices. This was due to the correspondence between high values for those variables and the stringency of the stopping condition. Higher values for niter corresponded to higher thinning since the entropy changes and the number of moves accepted were only recorded after every niter sweeps of the nodes of the network. Multiple sets of iterations of “equilibrated” entropy (corresponding to stable behaviour segments of values) appeared with different values which define the stopping condition and number of sweeps. Those sets became longer and longer as entropy decreased further.

Next, entropy behaviours obtained with different sets of values for wait, nbreaks and niter were compared in order to determine whether significant differences occurred by increasing the stringency of the stopping condition or the number of sweeps performed before recording the changes.

The values chosen were a trade-off between computational burden and the ability to observe large reductions in entropy which were more likely to represent meaningful block structures. All the results presented in this study were obtained with values of niter = 30, wait = 1000 and nbreaks = 30, which correspond to iterations consisting of 30 sweeps and a waiting time of 1,000 iterations. Once 1,000 iterations passed with less than nbreaks = 30 changes in entropy ($\epsilon = 0$), the process was considered to be equilibrated. The values were chosen due to the similar behaviour of the entropy to the one observed for niter = 50, wait = 1000 and nbreaks = 50, but with reduced running time. This set of values corresponded to rather unstable behaviour, consisting of multiple solutions (multiple block models) and in the future, sets of values corresponding to more stable communities will also be investigated.
Graph theory approach to characterising chromatin structure

Figure 3.3 Different values for the number of iterations and stopping conditions of \textit{mcmc\_equilibrate} plotted for chr14 (1Mb resolution); Parameter combinations of interest correspond to multiple states being explored which constitute multiple solutions of the resolved community structure.
3.4 Results

I compared the results of community detection at different resolutions and conditions. I also compared them with TAD structure and with A/B compartments. This was achieved by calculating normalised mutual information between different block assignments (R package NMI [140]).

Mutual information [23] of two variables X and Y is defined in equation 3.3. Normalised mutual information is scaled mutual information between 0 and 1. The main idea behind this measure is to assess the independence of two partitions. Variables X and Y are independent when \( p(X)p(Y) = p(X,Y) \). If the two partitions are independent, the log terms in the definition of mutual information become 0 and the NMI value is 0. High values of NMI indicate similarity and low values indicate independence of the two partitions.

\[
I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.3}
\]

where \( p(x) \) and \( p(y) \) are the marginal probabilities of \( X \) and \( Y \) and \( p(x,y) \) is the joint probability of \( X \) and \( Y \).

However, in the context of comparing chromatin structures, there is one more element of interest which needs to be considered. For example, when comparing TADs and A/B compartments, it is relevant to mention that genomic regions belonging to the same TAD are either all in the A or all in the B compartment. This is a comparison between two partitions of the genome, with different sizes. Specifically, A/B compartments are very large and TADs much smaller. In order to check similar relationships, a measure of agreement named Homogeneity Index (HI) was defined in equation 3.4, which quantifies how mixed is partition P in relation to partition Q of the same set, where the sizes of the groups of partition P are smaller than the groups of partition Q.

\[
HI(P,Q) = \frac{1}{k} \sum_{i=1}^{k} \frac{\text{max}(\text{freq}_i)}{\text{size}_i} \tag{3.4}
\]

where \( \text{max}(\text{freq}_i) \) is the maximum of the frequencies of different labels of partition Q in group \( i \) of partition P, \( \text{size}_i \) is the size of group \( i \) in partition P which has \( k \) groups. For example, in the case of TADs and A/B compartments, TADs correspond to partition P, which has groups of smaller sizes than A/B compartments which is partition Q. The homogeneity between TADs and A/B compartments is calculated by determining whether each TAD is composed of a majority of either A or B compartment bins. This is scaled by the size of the TAD, i.e. the number of bins of the TAD, and summed over all the TADs. Finally, this is
divided by the total number of TADs. The genome-wide HI of TADs and A/B compartments is 0.902 in our data. This is a high value indicating the coherence of TADs in terms of A/B compartments, as expected.

Figure 3.4 Nested community structures at several resolutions of chromosome 1 (a) 1Mb bin size (b) 300kb bin size (c) 100kb bin size, with the Hi-C matrix columns rearranged (anagram) to match the order of the bins in the hierarchy (left) side by side with the unordered matrix (right).
3.4.1 Consistency between resolutions

In order to evaluate consistency of the results across different resolutions, I made a comparison between different hierarchy levels from nested community structures obtained from Hi-C data at 100kb, 300kb and 1Mb resolution. Figure 3.4 shows the evolution of community structure of chromosome 1 across different resolutions. Of note, the Hi-C matrix was reconstructed by re-ordering bin columns according to the hierarchy groups.

![Graphs showing agreement of community detection between different resolutions](image)

Figure 3.5 Agreement of community detection (Normalised mutual information, NMI) between hierarchies of Hi-C matrices at different resolutions: 100kb, 300kb and 1Mb, in growing and RIS, respectively; level 1 corresponds to the basal level of the hierarchy formed of communities of genomic bins, level 2 consists of communities of level 1 communities.

I observed higher interaction values inside communities at different levels. The same split into four or five large blocks was reported at all resolutions with greater detail when using smaller bin sizes.
I quantified the comparison between hierarchies of Hi-C data at different resolutions using the NMI and HI measures described earlier. In particular, level 1, which was the basal level corresponding to the smallest set of communities, and level 2 from data at different resolutions were compared. Despite the large difference in bin sizes, high agreement values were observed, as seen in Figures 3.5 and 3.6. I observed particularly high NMI and HI values for level 1 from different resolutions. This was partly explained by higher stability values of the first level of communities which was further explored in later sections. This high level of agreement between the basal communities decreased at higher hierarchy levels. This suggests that intermediary levels are better resolved at high resolution. High values of the HI index indicated that larger communities, obtained at low resolution, break down further into smaller communities, obtained at high resolution.

### 3.4.2 Hierarchy depth and community sizes

The depth of the hierarchy is equal to its number of levels. In our data, larger chromosomes were associated with deeper hierarchies at every resolution studied, as observed in Figure 3.7. The hierarchies obtained at higher resolutions (100kb and 300kb) were also deeper than the one obtained at 1Mb, in agreement with the behaviour observed earlier, consisting of intermediate levels breaking down into further branches.

To further support the possibility of expansion occurring at intermediate levels, community sizes decreased at every level with increased resolution (Figure 3.8). Growing and RIS exhibited similar behaviours, supporting the possibility that the observations are not an artefact of the data sets used.

Investigating the distribution of the community sizes at level 1 revealed a decrease in the variability across chromosomes concomitant with an increase in resolution (Figure 3.9). 1Mb-resolution block sizes were highly variable across chromosomes with a large range of values, while the distributions observed in different chromosomes at 100kb resolution were relatively even. A few outliers remained in larger chromosomes at 100kb resolution. This indicated a possible invariant related to chromatin folding, such as identification of TAD-like structures. Again, this was supported by the presence of the same trend in both growing and RIS data sets.
3.4 Results

Figure 3.6 Homogeneity Index (HI) between hierarchies of Hi-C matrices at different resolutions: 100kb, 300kb and 1Mb, in growing and RIS, respectively; high values of HI indicate good agreement, e.g. genomic locations which are in the same community at level 1 in the 100kb bins hierarchy tend to also be in the same community at level 1 in the 300kb bins hierarchy.
Figure 3.7 The number of levels (depth) of the hierarchy detected for each chromosome from growing and RIS Hi-C matrices at (a) 1Mb resolution (b) 300kb resolution (c) 100kb resolution.
Figure 3.8 Size distribution of communities in (a) growing level 1 (b) growing level 2 (c) RIS level 1 and (d) RIS level 2, at each of the three resolutions: 1Mb, 300kb and 100kb; size of a community is defined here as the number of bins it consists of; level 1 is the basal level of the hierarchy consisting of communities of genomic bins at the chosen resolution, level 2 consists of level 1 communities.
Graph theory approach to characterising chromatin structure

3.4.3 Comparison with TADs and A/B compartments

I compared several levels of the hierarchy obtained at 100kb with TADs and A/B compartments. I chose this resolution based on the previous observation of smaller community sizes at higher resolutions such as 100kb, as well as of stabilised distributions of basal communities across chromosomes. I used the two measures mentioned before for the comparisons, with NMI measuring the closeness of the match between communities and TADs, while the HI measuring the level of mixing between communities and TADs or A/B compartments.

Figure 3.11 shows high correspondence between the 100kb basal communities and TADs, with consistently high NMI values (over 0.85) across chromosomes, in both growing and senescence. Level 2 communities and TADs showed good agreement as well, but with lower values of NMI, indicating that the basal level is better matched to TADs. The genome-wide overall NMI agreement of level 1 communities and TADs is 0.92. Agreement with A/B compartments was around 0.25 for all chromosomes indicating the existence of dependency.
but a poor match. This is expected due to the much larger sizes of A/B compartments than those of communities. The high HI values presented in Figure 3.11 complemented those observations, showing high agreement between communities and TADs or A/B compartments.

![Figure 3.10 NMI of communities (100kb) compared to TADs and A/B compartments; (a) NMI of communities vs TADs (b) NMI of communities vs A/B compartments; values close to 1 indicate good matches between the sets of regions compared.](image)

Thus, the communities I determined with the method described in this study were coherent in terms of open and closed chromatin compartments. The basal level of 100kb communities in both growing and RIS was very similar to TAD distribution. However, the number of basal communities estimated across all chromosomes was smaller than the number of TADs called genome-wide. Upon closer inspection, some of the communities comprised of a number of TADs, some non-consecutive. This potentially indicates that some groups of TADs form a tight three-dimensional module. This information cannot be retrieved using TAD calling which calls boundaries following the linear genome, making the TAD distribution entirely consecutive.
Figure 3.11 HI of communities (100kb) compared to TADs and A/B compartments (a) HI of communities vs TADs (b) HI of communities vs A/B compartments; high values correspond to good agreement.

Community sizes were shown to decrease with increased resolution and therefore, querying higher resolutions such as 40kb could yield a closer match between communities and TAD distribution. However, the computational running time of the method for large networks rendered the hierarchy calling unfeasible beyond 100kb. At 40kb resolution, chromosome networks had thousands of nodes and millions of edges. A solution was explored in the following section which exploited the high agreement observed between the basal level of different resolutions.

### 3.4.4 Strategy to improve computational burden

Networks with a large number of edges and nodes were difficult to analyse due to the large running times required. I developed a strategy in order to analyse the nested community structure of chromatin at a resolution below 100kb. It consists of two steps: computing communities at lower resolution, such as 1Mb, then using those as an estimate for community detection at higher resolution, such as 40kb. Each community at 1Mb resolution was a set of 1Mb regions, which were re-tiled with 40kb bins. Then community detection was performed on the network formed by those 40kb bins and their interactions (at 40kb
3.4 Results

resolution Hi-C). Finally, 40kb communities computed from every 1Mb community were merged into a set for a whole chromosome (or genome-wide). This was far more efficient, due to allowing the 1Mb communities to be analysed in parallel. Another efficiency improvement consisted of breaking down each 1Mb community into smaller networks of 40kb bins. Those smaller networks had a few hundred nodes at most, which shortened the running time required.

Figure 3.12 NMI and HI values comparing different pairs of levels from the full and the two-step strategy of the two-step strategy. The high HI values for resolution Hi-C). Finally, 40kb communities computed from every 1Mb community were merged into a set for a whole chromosome (or genome-wide). This was far more efficient, due to allowing the 1Mb communities to be analysed in parallel. Another efficiency improvement consisted of breaking down each 1Mb community into smaller networks of 40kb bins. Those smaller networks had a few hundred nodes at most, which shortened the running time required.

Figure 3.12 NMI and HI values comparing different pairs of levels from the full and the two-step strategy of the two-step strategy. The high HI values for resolution Hi-C). Finally, 40kb communities computed from every 1Mb community were merged into a set for a whole chromosome (or genome-wide). This was far more efficient, due to allowing the 1Mb communities to be analysed in parallel. Another efficiency improvement consisted of breaking down each 1Mb community into smaller networks of 40kb bins. Those smaller networks had a few hundred nodes at most, which shortened the running time required.

I chose the high and low values for resolution based on the criterion that the first set of communities can be re-tiled into networks with at most 1000 nodes each. This shortened the running time to a couple of hours compared to days (required for networks with thousands of nodes).

The validity of this two-step method was tested by comparing the resulting hierarchies of the full algorithm for 100kb resolution with the two-step strategy using 1Mb and 100kb resolutions. Figure 3.12 shows the NMI values for every chromosome of different pairs of hierarchy levels compared. Level 1 agreed the best between the two hierarchies and the two-step strategy yielded a deeper hierarchy than the full algorithm. The high HI values indicated low mixing between communities and further supported the validity and consistency of the two-step strategy.
3.4.5 Nested community structure at 40kb (2-step strategy)

The two-step strategy was utilised to estimate the chromatin hierarchy at 40kb resolution. Surprisingly, community structure at 40kb reached beyond TADs, generating smaller groups with better agreement than 100kb groups. The NMI and HI values are shown in Figure 3.13. Levels 1 and 2 of the hierarchy agreed with TADs the most in both growing and RIS. Remarkably, the HI values obtained by comparing the communities in terms of their consistency with A/B compartments are very high.

TAD size distribution was compared with that of communities estimated at 40kb and 100kb resolutions, at levels 1-3 (Figure 3.14). The y-axis of the boxplots, corresponding to the size distribution, was represented on the log-scale due to the large range of TAD sizes. This showed that TAD distribution was larger than those of levels 1 and 2 of 40kb communities, but overall smaller than that of 100kb communities at any level. Interestingly, 40kb communities at levels 2 and 3 were larger in RIS than in growing.
3.4 Results

Figure 3.13 Hierarchy levels at 40kb resolution compared to TADs and A/B compartments: (a) growing vs TADs levels 1-4, (b) RIS vs TADs levels 1-4 and (c) both growing and RIS vs A/B compartments levels 1 and 2.
3.4.6 Comparison with epigenetic marks

Next, I compared the hierarchical structures determined from growing and RIS Hi-C experiments to corresponding ChIP-seq data for epigenetic marks. Figure 3.15 shows the interactions hierarchy determined from Hi-C on the left side and the corresponding chromatin marks of the 100kb bins used for constructing the interaction graphs on the right side, for both growing and RIS. Notably, the chromatin mark distribution clustered well, as determined by the communities called from Hi-C, with genomic regions belonging to the same community showing similar epigenetic features. Each row corresponded to an 100kb bin and the ChIP-seq signal from different marks associated with it. This suggests that groups of regions with similar chromatin features interacted more strongly, forming communities.
3.4 Results

Growing chr8 100kb

ATAC
H3K27ac
H3K4me1
H3K4me2
H3K4me3
CTCF
H3K9me3
H3K27me3
HMGA1
H3K27me2
gene_density

GC Level1 Level2 Level3 Level4 Level5

RIS chr8 100kb

ATAC
H3K27ac
H3K4me1
H3K4me2
H3K4me3
CTCF
H3K9me3
H3K27me3
HMGA1
H3K27me2
gene_density

GC Level1 Level2 Level3 Level4 Level5

Figure 3.15 Comparison of the interaction hierarchy with chromatin marks, HMGA1 binding, GC% and gene density in growing and RIS (100kb resolution); each row corresponds to a 100kb bin; the interaction hierarchy is represented on the left of the heatmaps with each colour corresponding to one community and each columns corresponding to one hierarchy level; the heatmaps of epigenetic marks are calculated using THOR normalised ChIP-seq signal which is further quantile normalised so that distributions of different marks can be plotted simultaneously.

3.4.7 Stability of nested community calling

I represented the stability of the communities in each level by a matrix of frequencies of observing different configurations of the communities. Level 1 communities in both growing and RIS were highly stable. In higher levels, communities became more and more unstable, displaying higher heterogeneity. These results suggest that high-order chromatin interaction is more dynamic and heterogeneous at higher modular levels, although, this might also be partly explained by the under-sequenced nature of Hi-C where interacting regions separated
by large distances are captured less frequently.

Interestingly, both growing and RIS had a small number of level 3 communities which were stable. However, RIS was associated with a higher number of stable level 3 communities which partly overlapped SAHF-associations (Figure 3.16). The stability score was calculated as the maximum probability of observing a specific community in its present configuration.

![Figure 3.16 Stability of communities in growing and RIS Levels 1-3, with 0 indicating unstable and 1 indicating stable communities.](image)

### 3.4.8 Differences between growing and RIS

The first three levels of the nested community structures estimated in growing and RIS were compared for each chromosome (Figure 3.17). Growing and RIS were similar at level 1 ($NMI \geq 0.85$ for all chromosomes). This is consistent with the observation that TAD boundaries were highly similar between growing and RIS. However, the similarity was reduced in higher levels, suggesting that growing and RIS differ in their macro-chromatin arrangement, with different TAD-TAD associations in RIS.

I observed partial (or full) merges, as well as splitting of communities, in spite of the overall good agreement between growing and RIS. As expected, the merges were associated with increased interactions (detected in chapter 2) and the splits with decreased interactions. More precisely, over 80% of the differential interactions between merged communities increased in RIS. In the case of split communities, over 90% of the interactions decreased in RIS.
I noticed more dramatic changes at level 2, where 1193 Mb of the genome rearranged, creating new associations between large genomic regions. Merges and splits at level 2 were also observed, consistent with the direction of differential interactions associated with them. Such differences were partly associated with SAHF formation. I developed a visualisation strategy for those rearrangements, which I described in the next section.

![Figure 3.17 Comparison of the hierarchy levels 1-3 in growing and RIS in terms of NMI (left) and HI (right), at (a) 100kb resolution and (b) 300kb resolution; Level 1 in growing and RIS communities agrees the most, both in terms of NMI and HI, regardless of the resolution used for the comparison; agreement decreases with higher levels but more in terms of NMI than HI values.](image)

Another implication of large chromatin rearrangements identified might be functional. In chapter 2, I showed that the majority of altered EP interactions were within 2Mb. However, this included many interactions between different TADs (44%, only 26% were between consecutive TADs). Switching neighbouring TADs in the same higher level community in senescence might facilitate increases in EP interactions. For this purpose, communities which merged with others in RIS level 1 were selected. 227 such communities were found. They were associated with 944 differential EP interactions (76% increasing). Around 50% of such EP interactions were separated by large distances, over 500kb.

Even merges of growing level 2 communities potentially facilitated EP interactions. I found 26 EP interactions, involving 16 genes between merging level 2 communities. Such merges occurred over large distances, over 1Mb, and constitute major re-arrangements of chromatin.
Percentages of differential EP interactions within and between communities were calculated. Similar percentages to those observed between and within TADs were observed for level 1 communities. 52% occurred within the same community both in growing and senescence, 29% occurred between different communities and around 19% were associated to the same community in one condition, but not the other. This corresponded to merges and splits of level 1 communities.

Interestingly, at level 2 in the hierarchy, I observed similar percentages of within and between communities, with only a minor increase in EP interactions occurring within the same community - 55%. Even at level 3 only 64% of EP occur within the same community. A substantial percentage (34%) of the differential EP interactions occurred between different level 3 communities in either growing or RIS. This observation was reminiscent of the transcription factories model where enhancers and promoters from different parts of the chromosome would interact in a specific environment rich in transcription factors, which could be at the borders of large genomic regions, such as level 3 communities.

3.4.9 SAHF formation

In chapter 2, I determined SAHF-associated H3K9me3 peaks (SAHF K9me3), using the information about HMGA1-dependent long-range interactions. To identify large genomic restructuring, which represent potential SAHFs formation, SAHF H3K9me3 were highlighted in the hierarchy models fitted for growing and RIS.

I observed two main behaviours regarding SAHF H3K9me3 reorganisation. Smaller communities consisting of SAHF H3K9me3 separated and associated with other regions. Other small communities, from different branches of the chromosome, merged into large communities representing potential SAHF.

Figure 3.19 highlights an example of potential SAHF formation. In growing, chromosome 10 was organised into four large communities at level 3, followed by a fragmented community corresponding to pericentromeric regions. In RIS, a larger number of smaller communities was observed within the same level, due to multiple rearrangements occurring. SAHF H3K9me3 are highlighted in orange and non-SAHF H3K9me3 in red (Figure 3.19). Merging of several SAHF H3K9me3 communities was observed between growing and RIS, with the SAHF H3K9me3 from all the five different large communities in growing rearranging into two communities (dotted orange ellipses), largely consisting of H3K9me3. In Figures 3.19 (b) and (c), the selected region from (a) (represented with a black dotted rectangle) was further characterised by highlighting not only H3K9me3 regions, but also Lamin B1-associated domains (LADs). As mentioned in Chapter 1, LADs are reduced during senescence, releasing H3K9me3 to
3.4 Results

rearrange in SAHFs, from the nuclear periphery. The communities rich in SAHF H3K9me3 in RIS overlapped LADs (definition of LADs from [110]), as seen in Figure 3.19 (c).

In order to visualise SAHF formation in a more intuitive way, I designed another plotting strategy by simulating the relative position of genomic regions inside a chromosome. It involved plotting genomic regions as points in a scatter plot, whose position would respect the hierarchy determined earlier. The regions which were in the same community at level 1 (smallest groups) would be the closest, followed by regions which were in the same community at level 2, but not at level 1, and so on. The full details of this visualisation were described in later sections.

Figure 3.18 illustrates the way in which the scatter plot corresponded to the hierarchy identified. Each cluster of genomic regions (each point is a genomic bin) corresponds to neighbouring regions in the cell’s nucleus. In Figure 3.20, SAHF H3K9me3 and non-SAHF H3K9me3 are coloured in orange and red, respectively. The same effect that was observed in the hierarchy plot in chromosome 10, at 100kb resolution was reflected in this figure. SAHF H3K9me3 tended to cluster in the same locations or communities in RIS, whereas in growing they were part of all the larger communities. This effect was partially reversed with shHMGA1 treatment in RIS.

I observed similar patterns of SAHF-associated genomic regions aggregating into large communities for other chromosomes. In order to quantify this aggregation trend, I plotted the distribution of SAHF fractions per communities in level 3 in Figure 3.21. The communities with no overlap with SAHF H3K9me3 were removed. At level 3, communities incorporated many more SAHF regions, up to 100%. This suggests the presence of less mixing in terms of SAHF H3K9me3 and other regions within the same community during RIS. This was in agreement with trends observed visually of SAHF regions from different communities re-arranging in the same level 2 and level 3 communities. A small number of communities which consist mostly of SAHF H3K9me3 were observed in level 2 growing. However, at level 3, communities consisting mostly of SAHF H3K9me3 were only observed in RIS. These results suggest that the method for hierarchical chromatin community detection developed here captures the global high-order chromatin reorganisation.
Figure 3.18 Representation of the hierarchy as a scatter plot that conceptually represents the position of chromosome 10 regions in the two-dimensionally projected cell nucleus; each colour corresponds to a community at each hierarchy level and each point represents a genomic bin; level 1 communities are incorporated in mid level communities which are then further incorporated in top level communities.
Figure 3.20 SAHF associated H3K9me3 (orange) and non-SAHF associated H3K9me3 (red) on chromosome 10, in growing, RIS and RIS with shHMGA1; each point represents a 40kb bin and the two-dimensional distances reflect relative distances between pairs of bins in the hierarchy detected using graph-tool; accumulation of SAHF H3K9me3 in RIS is partially reversed in RIS with shHMGA1.
Figure 3.19 (a) Hierarchy comparison between growing and RIS chromosome 10, indicating matching genomic regions with coloured lines and each colour corresponds to a different level 1 community; different SAHF-associated H3K9me3 (orange) merge during RIS on chromosome 10 with H3K9me3 (highlighted with dotted orange lines) (b) selected region highlighting SAHF and non-SAHF H3K9me3 and (c) lamin-associated domains (LADs).
3.4 Results

3.4.10 Comparison with measured spatial distances

In 2016, Wang et al. [131] used 3D FISH imaging to study actual 3D coordinates of regions situated at the centre of TADs on chromosomes 20, 21, and 22 in IMR90 cells. I used the coordinates published by Wang et al. [131] to determine whether regions from the same community were closer to each other. This was repeated at multiple levels of the nested community structure, determined at 100kb resolution in growing IMR90 cells.

The TAD definition used in Wang et al. corresponds to hg18 genome assembly coordinates. I used the liftOver utility from UCSC [49] to convert the hg18 to hg19 coordinates. Four TADs were removed from further analysis, due to deletion (partial or full) between hg18 and hg19. The distance between pairs of TADs was calculated as the Euclidean distance between the three-dimensional coordinates available from Wang et al. [131]. The authors repeated the measurements for each TAD position in multiple cells. Therefore, I calculated multiple measurements of each pair-wise distance.

Figure 3.22 shows that TADs belonging to the same community at either level are separated by shorter distances, than TADs belonging to different communities. A pair of TADs was considered to be in the same community if it had an overlap of more than 80% with the same community. A one-sided t-test was used to compare the distance distribution in each case. The alternative hypothesis corresponded to the difference of distances between TADs from the same community and TADs from different communities being negative. All p-values
calculated were below $10^{-10}$ in favour of the alternative hypothesis.

Thus, the results suggest that the nested community structure detected reflects actual three-dimensional organisation of genomic regions.

Figure 3.22 Comparison of physical distances (from Wang et al. [131]) between pairs of TADs from the same and from different communities at (a) levels 2 and (b) 3, on chromosomes 20, 21, and 22; significance values were calculated using a pairwise t-test between the two distributions of distances: within the same community and between different communities (all p-values are below 0.001).
3.4.11 Advantages over other community detection methods

The community detection method employed in the present study is advantageous over the modified Louvain optimisation by Norton et al. [85], due to its non-parametric nature and lack of post-processing required. The nested community structure estimated naturally fits the hierarchical nature of chromatin interactions and can be called from a single chosen resolution value, whereas Norton et al. re-run the community calling for different resolution values in order to determine a hierarchy of communities. The re-runs are followed by post-processing needed for deciding on an agreement set of communities for each resolution. Non-parametric Bayesian SBM uses Monte Carlo Markov Chains to stabilise and sample the nested community structure called and results in a consensus hierarchy with posterior probabilities reflecting the stability of the call without further post-processing.

To call communities, Pancaldi et al. [89] combine Hi-C data with numerous ChIP-seq datasets. Such large amounts of data are not available for most biological experiments. Community detection was performed in the present study using Hi-C data exclusively as input. The output hierarchy was consistent with the distribution of chromatin marks, corresponding to several features such as open and closed chromatin, (active) enhancers and promoters.

3.5 Hierarchy visualisation

In order to visualise epigenetic features and compare the hierarchical structures derived, I developed a visualisation tool, which draws hierarchies and Hi-C matrices. It also aids in the comparison of different conditions and colours matching epigenetic features. I added functionality that allows drawing individual interactions as arcs between selected regions. The regions are represented with rectangles coloured by properties such as gene up/down-regulation.

Drawing the matrix

In order to draw the matrix, the three column Hi-C matrix (start bin, end bin and value) is used as input. Such matrices were obtained from HiC-Pro [119] alignment. The `geom_polygon()` functionality from ggplot2 [135] is used to draw each interaction between two bins as a rhombus, defining the coordinates of each corner. This is performed in order to plot the Hi-C matrix as a triangular matrix. Colouring reflects either the strength of interaction of logFC between conditions, as determined by the colour scheme choices. Missing data is represented with white spaces.
Drawing the hierarchy

The design of the hierarchy plot starts with the spacing of the first level bins with two values: one between nodes in the same community and one between different communities. Next, we define a space value between different levels (height) and the next level nodes are drawn by calculating the middle point of the community they represent as an x-coordinate and the y-coordinate is calculated by adding the height value to the previous y-coordinate determined. Finally, a space value is determined which shifts the entire hierarchy upwards on the y-axis in order to allow for either a mirrored hierarchy plot below or a mirrored Hi-C matrix.

Comparing hierarchies

Comparing hierarchies was represented with a tanglegram, by mirroring of two hierarchies and drawing a line between each matching bin of the first level of the hierarchies. Strategies to untangle the structure were applied which employ calculating possible re-ordering of communities which results in minimising the sizes of the lines drawn to match identical bins. This reflects the re-ordering of chromatin interactions between different biological conditions.

Drawing the hierarchy as a simulation of the cell nucleus

The hierarchical structure of chromatin regions was re-drawn as a scatter plot. Each point in the scatter plot corresponds to a genomic region. The distance between every two points depends on their distance in the hierarchy. If they are in the same community at level 3, then their distance is higher than points which are in the same community at levels 1 and 2. Then, edges between those points are drawn by sampling with a probability from the Bernoulli distribution (choosing between 0: no edge and 1: edge). The probability of drawing an edge decreases with the distance between the regions. Edges are also weighted such that closer regions have higher edge weights. Then a layout of the network is drawn using the igraph R package [28]. The layout chosen was Fruchterman-Reingold - a force-directed layout which determines coordinates for each region. Groups of regions which interact more are closer together and regions with higher edge weights are attracted to each other. This results in a set of coordinates used to draw the scatter plot of chromatin regions. The nested community structure is represented such that closer regions in the hierarchy are also closer in the scatter plot. This simulates the hierarchy of the chromosome structure more intuitively, by being a conceptual representation of a section of the chromosome in 3D space.

3.6 Discussion

The present study highlights the advantages of using community detection algorithms which are capable of detecting multiple possible conformations and their probabilities, as well as
3.6 Discussion

unveiling the hierarchical structure of chromatin interactions. Detecting such hierarchies is very useful for defining areas of interactions which greatly change between conditions, as well as global behaviours of chromatin architecture. In our case, SAHF formation corresponds to specific changes at higher hierarchy levels, whereas regulatory interactions such as enhancer-promoter contacts correspond to lower hierarchy levels changes.

The communities identified in growing and RIS showed good agreement with TADs at the basal level, but also uncovered strongly connected sub-TAD regions, as well as groups of TADs which might form a single unit in three-dimensional space. Moreover, the communities were consistent in terms of their epigenetic features, representing groups of epigenetically similar regions. Interestingly, heterochromatic H3K27me3 regions co-localised more often with open chromatin in the same genomic large bins (100kb), not with H3K9me3 heterochromatin.

Performing nested community detection on chromatin networks with more than 1,000 regions proved to be very computationally intensive. One solution involved combining results from different resolutions. Another solution was described in chapter 4. It involved using a non-weighted interactions network, selecting only significant interactions detected from Hi-C. The number of nodes and edges considered was greatly reduced. This relieved some of the computational burden of a large network and allowed for extraction of topological information. Such properties were not obvious from the weighted network, due to the overwhelming presence of non-specific interactions.

Using distance corrected interaction values is another approach for future development which would greatly reduce the number of non-specific interactions considered in detecting the chromatin hierarchy, as well as reducing the computational burden if interactions are filtered for ones which pass the distance correction threshold.
Chapter 4

Detecting significant interactions from HiC data

4.1 Introduction

In this chapter, I explored Hi-C modelling in order to select an appropriate method of detecting significant interactions. I then used connected genomic regions as edges in a graph. This offered new insights into interaction trends which could not be analysed using the weighted graph model in chapter 3. This was because the latter contained an overwhelming number of interactions, some of which may not be of interest.

Several modelling approaches of significant Hi-C interactions exist, such as FitHiC [3], HICCUPS [105], [38], PHYCHIC [108] and HiC-DC [12], which use different distributional assumptions. They were assessed for suitability of analysing data in this study. HICCUPS focused on loops present in Hi-C data and was very computationally intensive. PSYCHIC was also computationally intensive involving several processing steps in different computational languages. It used a continuous distribution for discrete counts and relied heavily on definition of TADs. In the HiC-DC study, the authors showed HiC-DC is more accurate than FitHiC in terms of significance values assigned to each interaction. I tested FitHiC on matrices of different resolutions in our data and the analysis yielded limited reproducibility across different biological replicates in terms of the significant interactions obtained. Therefore, HiC-DC was chosen for further study.

However, pre-processing involved in HiC-DC proved to be computationally intensive due to steps which loaded a large amount of data into virtual memory, such as individual reads. The authors also focused only on regions separated by at most 2Mb. They performed 1% subsampling of the data used for fitting the Hi-C model. Therefore, I effectuated an
exploratory study to describe different distributional assumptions. I also investigated the potential of using all Hi-C data available by comparing results from different biological replicates. I chose the statistical R package glmmTMB [7] due to its capabilities to model several different distributions, including Poisson and negative binomial, with or without zero-inflation.

Finally, I used a modified method and implementation of HiC-DC to call significant interactions in all replicates available for every condition. This assumed that most of the interactions observed occurred due to random ligation and that read counts varied according to sequencing biases. I modelled non-specific interactions by fitting the distribution of Hi-C data. Based on the assumption stated, the distribution of Hi-C data should reflect that of non-specific interactions. The extreme values (higher than expected from the model fitted) represent significant interactions. The sequencing biases considered were GC%, mappability and number of restriction enzyme (RE) sites. The majority of the interactions appeared to occur due to the distance-decay effect, as mentioned in earlier studied which state that two regions close to each other are more likely to interact [70]. I explored accounting for different biological replicates, different chromosomes, intra-TAD effect, as well as sub-sampling the data.

4.2 Exploratory Hi-C modelling

I used the three Hi-C replicates of the growing condition to test the dependency of read counts on distance (between interacting regions), GC %, mappability and number of RE sites. I repeated this for multiple resolutions. I then used those variables in the regression model for detecting significant Hi-C interactions. I also characterised variation intra- and inter-TAD, as well as differences between replicates.

4.2.1 Different chromosomes

Previous studies such as the one by Wu et al. [139] showed that each chromosome exhibits a different distance trend. This difference can be taken into account by adding chromosome as a fixed effect in the regression model. Another option is fitting the model for each chromosome separately. In order to relieve some of the computational burden due, I chose to fit the model separately for each chromosome. Previous models such as FitHiC [3] or HiC-DC [12] do not specifically mention modelling chromosomes individually, but their implementations and examples focus on single chromosomes.
4.2 Exploratory Hi-C modelling

4.2.2 TAD structure effect

The interaction frequencies inside TADs are another possible source of variation. Ron et al. [108] showed that interaction values inside a TAD are consistently higher than between TADs. Models such as HiC-DC or FitHiC do not discriminate between intra- and inter-TAD interactions which potentially results in less significant inter-TAD interactions. This is likely especially between consecutive TADs, between regions separated by shorter distances on different sides of TAD borders. A model which does not discriminate between intra- and inter-TAD positioning shifts the fit in-between, underestimating intra-TAD and overestimating inter-TAD interactions. HiC-DC authors considered only interactions between regions separated by at most 2Mb, which were mostly intra-TAD interactions.

4.2.3 Hi-C biases

Investigating the mean-variance relationship of the Hi-C counts distribution is essential for choosing between the Poisson and negative binomial distributions. By definition, under the Poisson distributional assumption, the variance is equal to the mean ($\mu = \sigma^2$, where $\mu$ is the mean and $\sigma^2$ is the variance). Under negative binomial, the variance varies quadratically with the mean ($\sigma^2 = \mu + \mu^2/\theta$, where $\theta$ is called the shape parameter).

The log-log plot of the variance against the mean of the three Hi-C replicates of growing was shown in Figure 4.1 (a). The plot shows a departure from linear dependency, where the dotted line is the $x = y$ dependency which corresponds to the Poisson distribution. The variance depended quadratically on the mean and this supported using the negative binomial distribution instead of Poisson, to model Hi-C counts. Figure 4.1 represents the counts at 40kb resolution of chromosome 1. I plotted similar exploratory figures for all chromosomes with Hi-C data at 40kb and 100kb resolution.

The log-mean count was shown to depend almost linearly on the log-distance between interacting regions in Figure 4.1 (b). This trend was also very similar between different replicates (Figure 4.1 (c)). I estimated the trends fitted in Figure 4.1 with ggplot2 functionality in `geom_smooth()` [135]. The dependencies of the log-counts on GC%, mappability and number of RE sites were also very similar across replicates (Figure 4.1 (d),(e),(f)).

The dependency of the log-mean count on distance varied greatly between intra- and inter-TAD, respectively, as shown in Figure 4.1 (g). I observed this effect in each of the replicates (Figure 4.1 (h)). A possible solution is to fit a different model for inter- and intra-TAD interactions, like in PSYCHIC [108]. Another option is to add intra-TAD as a fixed effect. The TAD definition used here was the same as the one used in chapter 2, estimated
Figure 4.1 Dependencies of Hi-C read counts (log) on distance, GC%, mappability and nr. RE sites at 40kb resolution (a) y-axis corresponds to the log-variance counts and x-axis corresponds to log-mean counts; each dot corresponds to the mean and variance of an interaction across growing replicates, coloured by the distance separating interacting regions; (b) the x-axis corresponds to the log(distance between interacting regions) and y-axis corresponds to the log-mean interaction counts (c) log(distance) against the log-counts in each growing replicate, as well as the distance fit trends for each replicate; (d) log-counts (y-axis) dependency on scaled GC% (x-axis) and the dependency trends per replicate; (e) log-counts dependency on scaled mappability values; (f) log-counts dependency on the scaled number of restriction enzyme sites (g) log-distance (x-axis) against log-mean counts across growing replicates (y-axis) coloured by position of the interaction relative to TADs; (h) same as (g) but plotted for each replicate instead of the log-mean across replicates (i) log-distance (x-axis) against log-mean count (y-axis) coloured by significant differential interaction status between growing and RIS, as well as the trend of the significant interactions relative to distance.
with TADbit [117] at 40kb resolution. This estimation suffered from the inaccuracy of every TAD caller, which meant that the true TAD borders could be slightly different. In order to overcome this, I investigated CTCF insulation as an alternative delimitation of intra- and inter-TAD spaces. Assuming that CTCF insulation can be estimated from CTCF ChIP-seq signal, the regions which are not separated by an insulator should have higher interaction values. I defined CTCF insulation as the presence of two nearby CTCF peaks with motifs with the same direction (forward-forward or reverse-reverse). Such occurrences were reported at TAD boundaries in many studies, such as Vietri Rudan et al. [129]. However, CTCF insulation explained less of the variation in the log-mean count than intra-TAD in our data.

The significant differential interactions I calculated in the analysis in chapter 2 were also highlighted on the distance-dependency plot in Figure 4.1 (i). The majority were situated above the distance trend estimated. This indicates that their values are higher than expected just from the distance trend and are more likely to correspond to true interactions.

I also highlighted several epigenetic marks on the distance-dependency plot of the Hi-C counts. Interestingly, interactions involving H3K9me3 regions had high values despite occurring across large distances (3-40Mb). Interactions between HMGA1 bound regions also appeared to have higher counts than expected, especially in the distance range 500kb-3Mb. Open chromatin marks of promoters and enhancers such as H3K4me1-3 and H3K27ac showed lower values of interaction counts over large distances, with few exceptions, which appeared as traces of high counts separated by long distances. One of those exceptions appeared as a cluster of points separated by the same distance. It included H3K4me1 and H3K4me3 regions and corresponded to the interactions of every 40kb bin pair between a 1.2Mb region on chromosome 1:120,200,000-121,400,000 and a 650kb region on chromosome 1:144,620,000-145,282,500. Active enhancers and promoters of genes such as NOTCH2, NOTCH2NL and PDE4DIP were situated in those regions. This strong, atypical interaction between open regions 23Mb apart was formed possibly due to an old duplication event, observed in the evolution of the human genome in the 1q21 region [42].

4.2.4 Comparing distributional assumptions

I systematically tested different distributions for modelling count data in order to choose the optimal one. I chose the R package glmmTMB [7] to fit models assuming counts distributions such as Poisson, zero-inflated Poisson, hurdle Poisson, negative binomial, zero-inflated negative binomial or hurdle binomial. In order to compare between models, I used several measures, such as: Akaike-information criterion (AIC), Bayesian information criterion (BIC), as well as deviance and the number of zeroes predicted by each model. Models which minimised AIC, BIC and deviance and did not under-predict the number of zeroes in the data were preferred.
Figure 4.2 Chromosome 1 (100kb resolution): epigenetic marks highlighted on the scatter plot of the log-counts (y-axis) against the log-distance between interacting regions (x-axis); the colour gradient corresponds to the intensity of the THOR-normalised ChIP-seq signal of each mark: (a) H3K9me3 and (b) HMGA1 - heterochromatin (c) H3K4me3 - active promoters (d) H3K27ac - active enhancers (e) CTCF - structural protein (f) H3K4me1 - all enhancers.
The models I tested (Equation 4.1) were variations of the one used in HiC-DC [12]. Those included: using different distributional assumptions and with or without considering TAD structure (Equation 4.2). The distributions I tested were: negative binomial and Poisson distribution with and without zero-inflation or hurdle (named nb, zinb, hunb for negative binomial models and poisson, zipoisson and hupoisson for Poisson variation). I modelled the distance variable using a B-spline of order 3 and 6 degrees of freedom, as described in Carty et al. [12]. I also tested splines of different orders and degrees in models named hu2 and hu3.

\[
\log(\mu_{ij}) = \beta_0 GC_{ij} + \beta_1 \text{Mapp}_{ij} + \beta_2 \text{REsites}_{ij} + B_{\text{spline}}(\text{distance}_{ij}) + \epsilon
\]  

where \( \mu_{ij} \) is the mean read count of the interaction between bins i and j

\[
\log(\mu_{ij}) = \beta_0 GC_{ij} + \beta_1 \text{Mapp}_{ij} + \beta_2 \text{REsites}_{ij} + \beta_3 \text{intraTAD}_{ij} + B_{\text{spline}}(\text{distance}_{ij}) + \epsilon
\]  

where \( \text{intraTAD}_{ij} = 1 \) if loci i and j are in the same TAD and 0 otherwise

Figure 4.3 shows the results of the fitting of the models described. Hurdle negative binomial, followed by zero-inflated and simple negative binomial were better suited for the data from all three perspectives, with small differences between them in terms of AIC and deviance. All Poisson models had higher AIC and deviance. I scaled all variables (GC, mappability and number of RE sites) in the same way as in HiC-DC. However, as pointed out in [128], the interaction counts have a highly non-linear dependency on variables such as number of restriction enzyme sites. Therefore, using smoothing terms might be a better choice. I also tested models such as generalised additive model (GAM) fitted using the R package mgcv [138], with and without using regression thin-plate splines for the terms GC, mappability, number of RE sites and distance, assuming negative binomial distribution of the counts. However, models fitted with mgcv ranked similarly to the other negative binomial models in terms of AIC and deviance. In terms of the number of zeroes predicted, all the models except the hurdle negative binomial models under-predicted the amount of zeroes in the data. Regression models which allowed the negative binomial dispersion parameter to vary with distance or with intra-/inter-TAD were very similar in terms of the fit, but much more computationally intensive. Interestingly, all models considered under-predicted the number of zeroes in the data on chromosomes 18 and 21, suggesting there might be other factors blocking interactions or read allocation to regions on those chromosomes.
Figure 4.3 AIC, BIC, deviance and percentage of zeroes predicted by the models considered; nb, zinb, hunb corresponded to negative binomial with/without zero-inflation and with/without the hurdle variation; hu1, hu2, and hu3 corresponded to variations of hunb (hurdle negative binomial), with hu1 including the intra-TAD term and hu2 and hu3 corresponding to different degrees of B-splines; pois, zipois and hupois correspond to Poisson distribution, zero-inflated and hurdle variations of Poisson distribution.
4.2 Exploratory Hi-C modelling

Adding a replicate term as a random effect in the model significantly increased the computational burden without improving the results in terms of agreement between replicates and therefore was opted against. I encountered convergence issues when fitting the model at high resolution values between 10-40kb due to the large amount of zeroes between loci in different TADs. I also encountered convergence issues when employing sub-sampling, due to the shape of the sub-sampled data not always resembling the negative binomial distribution.

Carty et al. [12] use only 1% of the data to fit the model, to reduce computational burden. However, this can introduce bias in the analysis if sub-sampling is not repeated to ensure the subset of the data is representative. Therefore, I applied repeated sub-sampling of 1% and 5% of the data and I selected a consensus set of significant interactions as output. However, convergence issues occurred and predicting values even when repeating the sub-sampling and fitting only 10 times were computationally intensive. This rendered fitting the full data more desirable. Sub-sampling the data also appeared to be under-fitting the data substantially. The residuals of the models fitted with sub-sampling were overwhelmingly corresponding to positive values. This suggests that the real interaction counts are mostly greater than fitted data. This was in contradiction with the main assumption of modelling Hi-C: most interactions occur at random (due to the distance effect) and the counts are largely due to the biases considered. Instead, an acceptable model would show a relatively symmetric distribution of residuals around 0 with a slight skew towards positive values representing a small number of significant interactions.

4.2.5 Final model

Based on the considerations explored, I modelled the 100kb Hi-C data using Equation 4.2, with the full data as input for fitting. I calculated significance scores as p-values, similarly to FitHiC [3] and HiC-DC [12]. The probability of observing an interaction count greater than or equal to the one observed given the biases at the location considered, \( P(X \geq x) \), was calculated. \( P(X \geq x) \) was rewritten as \( 1 - P(X \leq x) + P(X = x) \). I estimated this probability using the R functions \( \text{dnbinom} \) and \( \text{pnbinom} \). I executed the fitting in two rounds, like in FitHiC and HiC-DC, with the first round fitting the full data and selecting the extreme interaction values. The second run fitted the full data with the extreme values from the first run removed. The second fit was more likely to accurately represent the majority of the interactions that occur as a result of random ligation. Then, I calculated final significance values and applied FDR correction, to account for multiple testing. I selected the final 1% significant interactions as significant interactions for each replicate and for each chromosome. I then considered the union of the interactions for each condition, from every replicate available.
I performed a similar calculation for modelling data at high resolution, such as 10kb, 20kb and 40kb bins. However, only the interactions between regions separated by at most 2Mb were used, like in HiC-DC [12]. This was due to overwhelming computational burden and issues with convergence due to long-range interactions being mostly absent at high-resolution.

I used significant interactions called at several resolutions (10kb, 20kb and 40kb) to determine networks of interactions in growing and RIS, respectively. This complemented the differential interaction analysis, described in chapter 2. Interactions specific to each condition, as well as common interactions between growing and RIS were investigated.

4.3 Significant interactions at 10kb, 20kb and 40kb

784,795 and 755,646 significant interactions were called at 10kb resolution in growing and RIS, respectively. Around 145,000 and 30,000 interactions were called at 20kb and 40kb resolutions, respectively, in both conditions. The overlap between the interaction sets showed that 91% of the interactions called at 20kb overlapped the ones called at 10kb resolutions and that 95% of the interactions called at 40kb overlapped the ones called at 20kb. In the other direction, only 30% of the interactions called at 10kb overlapped the ones called at 20kb resolution and only 42% of the interactions called at 20kb overlapped the ones called at 40kb resolution. Therefore significant interactions determined at 10kb resolution were used for determining graph-theoretical properties of the Hi-C network, due to higher accuracy for annotation with epigenetic marks and gene promoters and enhancers and also due to more information about the interactions occurring in growing and RIS, respectively.

In terms of the overlap between growing and RIS, the two conditions showed 50% agreement at 40kb resolution in terms of their significant interactions, 35% at 20kb and only 21% at 10kb resolution. This was consistent with the results presented in chapter 2 which showed major rewiring of interactions in RIS.

4.3.1 Connectivity

The network in growing cells (10kb resolution) was organised across 35 (disconnected) components, while the RIS network had 44 components. Interestingly, the components involved both open and closed regions suggesting the presence of bridging regions between the A and B compartments. Highly similar power-laws governed the degree distributions in the growing and RIS networks, as observed in Figure 4.4. This indicated the presence of hubs in the networks. Some of the regions interact with more than 50 other regions in both growing and RIS.
The components with fewer than 10 nodes were removed from further analysis due to the high interaction sparsity they were associated with. The final networks consisted of 228,791 nodes (genomic regions) and 783,363 edges (interactions) in growing and 228,120 nodes and 753,851 interactions in RIS.

Figure 4.4 Degree distributions of the 10kb networks of growing (G) and RIS (S); degree (x-axis) of a node corresponds to the number of neighbouring nodes; y-axis corresponds to the frequency of nodes with certain degree values in the network.

Around 20% of the significant interactions in each of the networks, growing and RIS, occurred between the A and the B compartment. The overwhelming majority of interactions occurred within the same compartment. In chapter 2, most of the differential interactions between consecutive TADs belonged to different A/B compartments. This suggests that interactions tend to occur within the same compartment, but not exclusively. Interestingly, 50% of the mixed interactions were associated with HMGA1 binding on both ends and 86% exhibited HMGA1 binding at least on one end. HMGA1 binding was negatively correlated with A/B compartments score (chapter 2), with 60% of the HMGA1 rich regions belonging to the B compartment and 40% to the A compartment. This suggested that HMGA1-mediated interactions may be responsible for many of the mixed interactions observed.
4.3.2 Hubs

I determined hubs in the growing and RIS networks using the HITS algorithm implemented in the R package igraph [28]. 12,747 hubs were common between the two conditions, while 10,645 were growing-specific and a much larger number, 22,077, were RIS-specific (4.5).

![Figure 4.5 Number of genomic regions corresponding to hubs either in growing or RIS, classified as growing-specific, common between growing and RIS, and RIS-specific.]

The epigenetic landscape of hubs (Figure 4.6) showed a variety of regions which correspond to promoters (H3K4me3), enhancers (H3K4me1, H3K27ac), gene-rich regions, HMGA1- and H3K9me3-rich regions. In order to make this separation clearer, PHATE dimensionality reduction [78] was used on the histone marks data associated with hubs. In growing (Figure 4.7 (a)), the right most branch corresponded to enhancers, the next to promoters, followed by H3K9me3 and HMGA1. Interestingly, a subset of the H3K9me3 peaks overlapped regions also represented by ATAC-seq signal. Such regions were potentially observed due to heterogeneity of cell populations or corresponded to the small subset of H3K9me3 regions which are permissive, as described by Becker et al. [6]. RIS hubs revealed a similar split into promoters, enhancers, H3K9me3 and HMGA1 regions (Figure 4.7 (b)).

198 growing-specific promoter hubs were identified. They included promoters of HISTH2 and Wnt pathway genes. The 553 RIS-specific promoter hubs were enriched for DNA-damage response. All categories of promoter hubs interacted with other enhancers and promoters, but most of all with regions corresponding to gene bodies. The limited gene enrichment observed in both conditions suggests that the hub-property is more likely associated with structural properties, such as high H3K4me3, rather than functional. Regions such as histone gene clusters appeared to be consistently representing hubs both in growing and in RIS.

In terms of H3K9me3 peaks, the number of highly interacting H3K9me3 nodes was higher in RIS than in growing, as expected, consistent with SAHF formation. Interestingly, a number of regions bound by HMGA1, but not associated with H3K9me3, also displayed the hub-property. Many more such regions were recorded in RIS (2,508 10kb bins), than in growing (1,382 10kb bins).
Figure 4.6 Epigenetic landscape of hubs common between growing and RIS, growing-specific, and RIS-specific; each row corresponds to a 10kb bin, coloured by the normalised ChIP-seq signal intensities (from low to high) of open chromatin, heterochromatin, structural marks such as CTCF and cohesin, GC% and gene density.
Figure 4.7 Hubs: (a) Growing and (b) RIS; PHATE dimensionality reduction of the epigenetic features matrix, coloured by various marks corresponding to enhancers, promoters, heterochromatin and HMGA1; different branches correspond to different types of regulatory regions.
The relationship between HMGA1 with no H3K9me3 interaction and the SAHF structure is currently unclear. It would be interesting to test a hypothesis that such HMGA1 hubs with no H3K9me3 might be in between euchromatin and SAHF-core, considering a dual role of HMGA1 [81].

4.3.3 Betweenness centrality

In a network, vertices with high values of betweenness centrality are most often situated on the shortest paths between other vertices. Their removal would be the most disruptive in terms of network connectivity.

I defined regions with high values for betweenness centrality in the significant interactions graphs as the ones with values higher than the mean betweenness centrality in each graph. Interestingly, the regions with high betweenness centrality vary greatly between growing and RIS, consistent with the chromatin interactions rewiring observed in previous chapters. 13,664 and 12,456 regions had high betweenness centrality specific to growing and RIS, respectively. However, only 3,825 regions had high betweenness centrality in both conditions (Figure 4.8).

Interestingly, a subset of bins near TAD boundaries exhibited high values of betweenness centrality in either condition. The bins representing TAD boundaries were chosen by intersection with every TAD boundary ±10kb. This was due to TADs being called at 40kb resolution leading to inaccuracies regarding the precise position of the TAD boundary. Thus, although only a subset of the bins selected truly represented the TAD boundaries, 24% exhibited high betweenness centrality values.

Similarly to the epigenetic identity of hubs, a variety of regions showed high betweenness centrality, such as enhancers, promoters, regions rich in CTCF and cohesin, H3K9me3 and HMGA1 peaks (Figure 4.9).

![Figure 4.8 Number of genomic regions with high values of betweenness centrality either in growing or RIS, classified as growing-specific, common between growing and RIS, and RIS-specific.](image-url)
Figure 4.9 Epigenetic landscape, as well as gene density and GC% of regions with high values of betweenness centrality in growing and RIS; each row corresponds to a 10kb genomic bin from the significant interactions network.

4.3.4 Cliques

Around 16,000 cliques (fully connected sub-graphs) consisting of at least 5 regions were detected in growing, while only 1,312 were found in RIS. 40% of the regions involved in cliques in each of the conditions were associated with cohesin, which exhibited reduced binding during RIS (as described in chapter 2). This possibly contributed to the reduced number of cliques observed in RIS.

4.3.5 Mean distance

The mean distance between two regions in the network is the length of the shortest path between them, i.e. the smallest number of interactions connecting them. The diameter of a network is equal to the maximum length of all shortest paths in the network. The diameter or the mean distance in the Hi-C network reflect the number of interactions between every two regions. Small values for the diameter and the mean distance would suggest that every two regions are relatively close to each other, whereas high values for diameter and mean
distance would suggest increased separation.

In order to investigate the separation of genomic regions in growing and RIS, the mean distance and the diameter of the networks were investigated. Both measures were used for completeness. For example, a larger value in diameter in growing compared to RIS, concomitant with a small value in mean distance, indicates that regions separated by a large number of interactions are outliers and that most regions are relatively close to each other.

![Graphs of mean distance and diameter](image.png)

Figure 4.10 (a) mean distance values for all chromosomes in growing and RIS (b) distribution across chromosomes of all mean distance values (c) diameter values for all chromosomes (d) distribution of diameter values across chromosomes

Both the global (whole-network) mean distance and diameter were larger in RIS, than in growing, with the mean distance in growing and RIS being equal to 30 and 35, respectively.
and with the diameter being equal to 160 and 173, respectively. This meant that, on average, every two regions were separated by 29 other regions in growing and 34 other regions in RIS. This suggested increased separation between genomic regions in RIS, relative to growing.

Each chromosome’s mean distance and diameter were also investigated. The values are shown in Figure 4.10. The values for chromosomes 3, 5, 10, 18 and 19 changed dramatically between growing and RIS. Chromosomes 3, 5 and 10 exhibited increases, whereas 18 and 19 exhibited decreases in both mean distance and diameter. The distribution across all chromosomes reflected those changes, with RIS being associated with larger variation. Regions such as the ones on chromosome 18 and 19 were closer to each other, whereas chromosomes 3, 5 and 10 showed increased separation.

4.3.6 Enhancer-promoter significant interactions

Genomic regions involved in significant interactions as detected at 10kb resolutions corresponded to regions marked by H3K4me3 or H3K4me1 and H3K27ac, indicating they might play a regulatory role. Therefore, I annotated significant interactions from the 10kb resolution growing and RIS networks with enhancers and promoters, similarly to the annotation performed in chapter 2. I only used uniquely annotated enhancers and promoters in this case. 17,420 EP interactions were detected in growing and RIS, with 3,342 contacts common between the two conditions (Figure 4.11).

On average, each enhancer interacted with 1.6 and 1.4 promoters in growing and RIS, respectively. Enhancers interacted with at most 8 promoters in both conditions (Figure 4.12). Promoters interacted with more enhancers, with 2.6 and 2.5 enhancers on average in growing and RIS, respectively. The maximum number of enhancers a promoter interacted with was 22 in growing and 24 in RIS. This was consistent with previous reports [108], [48] stating that enhancer to promoter ratio is smaller than the promoter to enhancer one.
I was particularly interested in the 3,342 EP contacts common between growing and RIS. Such contacts were pre-looped between the two conditions and potentially corresponded to key regulatory programmes. Note that differences in contact frequency could still occur in RIS but each such EP contact was detected significantly in both growing and RIS. Gene enrichment analysis of the genes involved in common EP contacts revealed important processes being regulated in both growing and RIS via chromatin interactions. Gene enrichment was performed using EnrichR [16], against the Reactome pathway database [26]. Pathways such as response to stress, DNA damage repair, EGFR and TGFβ signalling, as well as cell cycle were among the top enriched categories (Figure 4.13).
Figure 4.13 Gene enrichment against the Reactome database [26] of the genes with common EP contacts in growing and RIS; only significantly enriched pathways are shown here; the x-axis corresponds to \(-\log(adjusted.p - value)\); p-values were FDR-corrected for multiple testing and significance threshold was FDR \(\leq 0.05\).
Figure 4.14 Capture Hi-C maps of EP interactions common between growing and RIS identified from significant interactions identified at 10kb resolution from genome-wide Hi-C data; the promoters of the genes, shown as green rectangles, interacted with the enhancers shown as black rectangles; the interaction maps are ICE-normalised and distance-corrected, so that warm-coloured interactions correspond to values above expected based on the distance-trend and cold-coloured interactions correspond to values below expected from the distance-trend.
I also validated some of the significant EP interactions mentioned here, which were common between growing and RIS. This was achieved by overlapping capture Hi-C data with the significant contacts studied. Figure 4.14 shows three example regions, involving the genes APP, CDKN2B and GNG11. Note that there were contact frequency differences between growing and RIS but the interactions occurred in both conditions. Interestingly, the promoter of the APP gene interacted with enhancers situated at the end of its gene body. This interaction was reduced in RIS and coincided with the up-regulation of the gene. APP (Amyloid Precursor-Protein) is an important protein in the development of Alzheimer Disease [126]. The interactions of enhancers with the promoter of the APP gene are potentially conserved in other cell types and constitute avenues of research into the regulation the APP gene.

Another important aspect observed using the significant EP contacts determined was co-regulation of genes belonging to the same pathway via shared enhancers. 5,023 and 3,458 pairs of genes were co-regulated via shared enhancers in growing and RIS, respectively. I performed gene set enrichment on the co-regulated gene lists and selected pathways enriched based on at least 5 genes. The gene set enrichment was performed using EnrichR [16] against the Reactome pathway database [26]. Significantly enriched pathways were selected based on an FDR threshold of 0.05. I then used this list of pathways to check how much co-regulation existed between each pair of pathways. This was quantified as a tuple of the number of genes in each pathway, which were co-regulated via shared enhancers. For example, 295 genes from the Immune System pathway and 37 genes from the Cell Cycle pathway showed co-regulation. Interesting pairs include IRAK3 and MDM2, as well as IL6 and NUPL2.

Interestingly, 21 and 18 pathways showed large co-regulation within the same pathway in growing and RIS, respectively. This suggests that regulation via enhancers affects the gene expression of multiple members of the same pathway. Such pathways with high levels of co-regulation via shared enhancers are shown in Figure 4.15. They were ordered decreasingly by a score corresponding to the proportion of all the genes in the pathway which share enhancers with at least one other gene in the same pathway. In both growing and RIS, around 25-30% of each pathway showed evidence of co-regulation and EGFR and FGFR signalling were amongst the most represented pathways in both conditions. Genes belonging to the Metabolism and Wnt pathways also showed evidence of co-regulation in both growing and RIS.
Figure 4.15 Pathways which show evidence of co-regulation via shared enhancers in (a) growing and (b) RIS; the x-axis corresponds to the proportion of the pathway consisting of genes which share enhancers with at least one other gene in the same pathway; the pathways correspond to gene lists from the Reactome database [26].
4.4 Discussion

Many methods exist currently for detecting significant interactions from Hi-C data, each with specific inaccuracies. I investigated several statistical models and adjustments to current detection methods in the present body of work. In particular, taking into consideration the interactions intra- and inter-TAD appeared to improve the detection of significant interactions, along with using zero-inflated negative binomial models. A number of chromatin marks appeared to correlate differently with Hi-C read counts against the distance between interacting regions, with open chromatin and heterochromatic regions being associated with high interaction counts across short and large distances, respectively. A small number of regions on chromosome 1 were found to be interacting across large distances in spite of being associated with open chromatin and were potentially linked to a translocation event in the evolution of the human genome.

Using the optimised model for detecting significant interactions, I built interaction networks at several resolutions in growing and RIS. I have analysed several network properties reflecting the nature of individual nodes’ connectivity in each condition, reinforcing the chromatin network remodelling during RIS. In particular, I showed that regions associated with various chromatin marks had the potential to be highly connected and no single chromatin feature was associated with high values of betweenness centrality or hub potential. In terms of condition specificity, the RIS network was associated with higher values for properties such as diameter and mean distance, than the growing network. This was consistent with increased physical separation between interacting regions due to the larger size of the nucleus of a senescent cell. Physical and functional relationship between each network properties will be investigated in the future. Phenotype specific betweenness centrality values are potentially useful to prioritise the nodes, which are likely to be essential for network integrity. Those regions might be somehow altered in related pathologies, such as cancer and cancer-related disorders.
Chapter 5

Methods

5.1 Experimental Methods

Cell culture and vectors

IMR90 cells (normal human diploid fibroblasts) (ATTC) were cultured as previously described under the 5% O2 condition [145]. The following retroviral vectors were used in this study: pLNCX2-Neo (ER:HRAS-G12V, encoding a fusion protein of the estrogen receptor ligand-binding domain and HRAS-G12V) [145], pMSCV -puro-miR30 RNAi vector for control (empty) and shHMGA1 [81]. RAS-induced senescence (RIS) was established by treating ER:RAS-G12V IMR90 cells with 4- hydroxytamoxifen for 6 days [145].

Hi-C

Hi-C and capture Hi-C data was performed by Dr. Aled Parry in the Narita laboratory, in collaboration with Peter Fraser (Babraham Institute). The Hi-C was performed in-nucleus and the protocol was followed as previously published in Nagano et al. [80] and Lieberman-Aiden et al. [70]. Capture Hi-C was performed as published in Shoenfelder et al. [114]. Captured regions were listed below, in Appendix A.

ChIP-seq

HMGA1, CTCF and H3K4me2 ChIP-seq data was obtained by Dr. Aled Parry in the Narita laboratory. The ChIP-seq was performed as described in Parry et al. [92], with the exception of HMGA1 ChIP-seq. This was performed using 50µg of sonicated chromatin except that cell fixation was performed for 1 hour at 4°C (rather than 10 minutes at room temperature). CEBPβ ChIP-seq was obtained by Dr. Yoko Ito in the Narita laboratory. It was performed as described in Hoare et al. [50]. Cohesin ChIP-seq (RAD21 and SMC3) was obtained by
ChIP-seq, effectuated by Dr. Masako Narita in the Narita laboratory. This was performed using additional cell lysis steps using buffer with 0.1% and 0.2% SDS, with optimised sonication cycles. Dynabeads\textsuperscript{TM} Protein A for Immunoprecipitation (Thermo Fisher 10002D) was used for IP (rather than Secondary Antibody-Coupled Magnetic Beads) and high salt 500mM NaCl buffer was utilised during washing steps [111].

\textbf{Antibodies used:}

- CTCF - Cell Signalling, 3418S, 10µg
- HMG1A1 - Abcam, Ab129153, 10µg
- H3K4me2 - CMA303 (gift from Dr. Hiroshi Kimura, [59])
- CEBP\textsubscript{β} - C-19 Santa-Cruz, sc-150, 1:500 (used in [50])
- Cohesin - Rabbit polyclonal anti-Rad21 antibody (generously provided by Prof. K Shirahige, Institute of Molecular and Cellular Biosciences, University of Tokyo), generated using the peptides DEPIIEEPSR and ATPGPRFHII as immunogens [77]
- Cohesin - Abcam, Rabbit polyclonal Anti-SMC3 Ab9263

\textbf{5.2 Public data}

RNA-seq and ChIP-seq data previously developed and published by members of the Narita laboratory, as well as other publicly available data, was used for analysis in this project. RNA-seq data was re-analysed from Hoare et al. [50] for growing and senescence expression values (GSE72407). Public ChIP-seq data from Parry et al. [92] was re-analysed for H3K4me1, H3K27ac and ATAC-seq information about growing and senescence (GSE103590). Data from Chandra et al. [14] for H3K9me3, H3K9me3, H3K27me3 and H3K4me3 marks was also used (GSE38448). TP53 binding information in senescence was obtained from ChIP-seq data from Kirschner et al. [60] (GSE53491).

Additional H3K4me3 information was obtained from ChIP-seq data from Chicas et al. [18] (GSE43922). BRD4 ChIP-seq data was re-analysed from Tasdemir et al. [125] (GSE74238).

\textbf{5.3 Computational methods}

\textbf{5.3.1 Hi-C}

Ten Hi-C and ten matched capture Hi-C libraries were produced by Dr. Aled Parry in the Narita laboratory, as mentioned earlier. Two samples (Hi-C and matched capture Hi-C) were
removed due to low library size after artifact and duplicates removal, leaving three replicates for growing, two for senescence, two for growing shHMGA1 and two for RIS shHMGA1, both in Hi-C and capture Hi-C data sets (matched). Reproducibility of the libraries across conditions was assessed using PCA with low counts removed and library normalisation, as well as with HiC-spector [144].

Alignment of Hi-C and capture Hi-C samples was performed using HiC-Pro [119] to the hg19 genome, duplicates marking and removal was performed using samtools [68], while artifact removal, counting into bins and iterative correction normalisation (ICE) was performed both with HiC-Pro [119] and with R Bioconductor package diffHic [73], [53]. HiC-Pro counts were calculated against equally sized bins and ICE normalisation was performed for each sample, while diffHic counts were calculated more robustly against bins adjusted for the presence of HindIII cutting sites. ICE normalisation of summarised replicates was also performed using diffHic, where the negative binomial estimate of the mean count is used instead of the counts in each library for iterative correction (correctedContact).

Differential interaction analysis was performed with R Bioconductor package diffHic using 10 as the threshold for minimum reads per bin considered in order to eliminate the large number of bins with low counts, necessary for the analysis. Significant differential interactions were selected at 5% false discovery rate (FDR).

TADbit [117] was used to call topological associating domains in every Hi-C sample. Different TAD calls for biological replicates were coalesced into consensus TADs using TADbit function align_experiment.

A/B compartments were called by performing PCA for each chromosome on the normalised (ICE) Hi-C matrix corresponding to the negative binomial mean across different replicates for each condition, calculated with correctedContact from diffHic [73]. The PCA with the best correlation (absolute values over 0.5) with H3K4me1 signal was selected as representative for A/B compartments and the A compartment was assigned to the positive values if the correlation with H3K4me1 was positive and to the negative values if the correlation was negative.

The visualisation of Hi-C matrices and enhancer-promoter interactions were performed with custom scripts made in R, using ggplot2 [135]. The design of the interaction matrix plots was described in chapter 3. The matrices used for individual condition plots were averaged across replicates, ICE-normalised and distance corrected using diffHic [73]. The colour scheme
used is centred around 0, reflecting interaction values above and below the expected values due to the distance trend.

5.3.2 Nested community detection from Hi-C data

Community detection was performed using graph-tool [97], developed and implemented by Dr. Tiago Peixoto. A script was made for reading in Hi-C data, creating a network compatible with graph-tool functionality and wrapping functions needed to perform nested community detection initialisation, equilibration and calculation of marginal probabilities. Visualisation of the hierarchies obtained was performed using custom R functions which use the `geom_polygon` functionality from R package ggplot2 [135] and a set of coordinates calculated for the hierarchy and for the matrix, as described in chapter 3. The scatter plot visualisation method was obtained by calculating coordinates of the Fruchterman-Reingold layout for the network described in chapter 3, for each chromosome, using the igraph R package [28].

Relevant comparisons between different labellings of genomic regions were computed using NMI from the R package NMI [140], as well as a custom measure implemented in R, called homogeneity index.

5.3.3 Network analysis

In chapter 2, a network of differential H3K9me3-H3K9me3 interactions was described and plotted, as well as networks of enhancer-promoter and enhancer-enhancer interactions. Such analyses and plots using custom network layouts were created using the functions in the igraph [28] R package.

5.3.4 Determining significant Hi-C interactions

Significant Hi-C interactions were determined using a modified version of HiC-DC [12], as described in chapter 4. The union of significant interactions across biological replicates was used for further analysis. The igraph [28] R package was used to create a network of significant interactions for growing and senescence, respectively and to calculate relevant network properties.

5.3.5 ChIP-seq

Alignment of ChIP-seq data was performed using bowtie2 [64] using the hg19 genome assembly and post-processing of alignment files was performed using samtools [68] and bedtools https://bedtools.readthedocs.io/en/latest/. Alignment of the H3K4me1, H3K4me2, H3K27ac, CTCF, ATAC-seq, H3K9me3 and H3K27me3 was done by Dr. Dora Bihary in the Samarajiwa
5.3 Computational methods

Differential ChIP-seq analysis was performed using THOR [2] with the TMM normalization and input subtraction, merge and remove blacklisted regions options. This yielded regions with differential binding, as well input and library normalised profiles. Peak calling was performed using MACS2 [41]. Heatmaps of ChIP-seq signal were drawn using deepTools [103].

ChIP-seq binned signals used to characterise regions of interest were obtained using the utility bigWigAverageOverBed (https://www.encodeproject.org/software/bigwigaverageoverbed/) and the bigWig files which were THOR-normalised after input subtraction. All the ChIP-seq signal tracks presented were THOR-normalised after input subtraction and scaled to the same data ranges between conditions.

Growing and RIS samples with two to four replicates per condition was available for every ChIP-seq data set.

5.3.6 RNA-seq

Alignment of the RNA-seq data from Hoare et al. [50] was performed using the STAR aligner [37] and the hg19 genome assembly (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/) and the reads were counted over genes using GENCODE19 (https://www.gencodegenes.org/releases/19.html) annotation and the featureCounts functionality from the subread package [69]. Differential expression analysis was performed using the R Bioconductor package edgeR [75]. For significance testing, I used the glmTREAT functionality which combined an indication for a minimum log-fold change of interest with multiple testing correction. In this way, I obtained a reasonable number of significantly differentially expressed genes in RIS relative to growing. Other methods of testing such as glmLRT yielded a high number of differentially expressed genes that contradicted the underlying assumption that the majority of genes did not change their expression level. Gene set enrichment was performed using EnrichR [16].
Chapter 6

Conclusions and future study

One of the most obvious morphological changes that occurs upon the onset of senescence in human diploid fibroblasts is the formation of heterochromatic foci (termed SAHF [82]). Previous studies in our lab have been based on fluorescent light microscopy observations, while this study is the first one to focus on the chromatin alterations that occur upon senescence via a chromosome conformation capture technique, in this case Hi-C. RAS-induced senescence (RIS) was found to be associated with multiple chromatin architectural alterations, compared with growing cells.

In terms of gene regulation, numerous EP interactions were altered during RIS. Promoters involved in differential EP increased or decreased interactions with large number of enhancers, possibly corresponding to a phenotypical anchoring effect. Moreover, a number of pairs of genes were shown to compete for enhancer interactions, with one gene displaying increased interactions, and the other decreased interactions with the same enhancers. The genes juxtaposed in this way were enriched for inflammatory-related processes, cell cycle and metabolism. In particular, SASP and cell cycle are potentially decoupled during RIS, with increased enhancer interactions with SASP genes and concomitant decreased interaction of the same enhancers with cell cycle genes. SASP factors are some of the most up-regulated genes, in terms of gene expression, in senescence. Consistent with the phenotype, cell cycle genes are some of the most down-regulated. The changes in EP interactions would therefore be very interesting to follow up on in future studies.

A gene related to cell cycle, CKAP2L, was shown to be competing for enhancers with \textit{IL1A} and \textit{IL1B}, key SASP genes. The changes in interaction landscape at this locus were associated with possible loop disruption due to a new cohesin peak observed in RIS near \textit{IL1B}. Future study involves experimentally validating the competitive enhancer regulation in the \textit{IL1} locus. The cohesin peak forming near \textit{IL1B} will also be investigated in other biological contexts such as ageing or immune cells. Cohesin binding in IMR90 cells with TNF\textalpha{} treatment will
also be studied to establish whether cohesin binding changes are RIS specific. Interestingly, cohesin binding near *IL1B* in RIS was not accompanied by CTCF. Future plans include finding cohesin’s binding partner at this locus, which potentially contributes to increased insulation. EP interaction differences associated with cohesin changes in other captured loci remain to be more deeply characterised.

Future directions include analysing the interaction changes in quiescence, induced by serum starvation, relative to growing cells. The extent of the changes discover will allow us to elucidate the extent to which EP interaction changes are RIS-specific. We are particularly interested in finding out whether cell cycle and SASP are decoupled in other conditions as well, or if this behaviour is a marker of RIS. Such decoupling potentially contributes to the irreversible nature of cell cycle arrest that defines senescence.

Graph-theoretical approaches were used to model the chromatin interactions network. Hi-C matrices of growing and RIS were used to determine nested community structures, using a non-parametric Bayesian version of stochastic block models. The community detection method was introduced by Tiago Peixoto [97] and was previously used in studying earthquakes, disease localisation, company ownership networks, but not for chromatin interaction networks. Chromatin interactions are highly dynamic and can exhibit heterogeneity. The method chosen for detecting community structure is particularly suitable for measuring such heterogeneity and assess the stability of the communities found.

The layered structure of the hierarchy estimated in growing and RIS, respectively, was particularly insightful regarding the scale of changes observed in RIS. This was reminiscent of what had been seen previously [14], where SAHFs formed through the spatial rearrangement of existing blocks of H3K9me3/H3K27me3. Large genomic regions were shown to merge or split during senescence, re-organising at different levels of the hierarchy. Potential SAHF formation was represented by the formation of large communities at level 3, consisting mostly of SAHF-associated H3K9me3 regions. The stability of community detection across many iterations showed that the smallest communities were the most stable. Future investigation includes determining the number of iterations that maximises stability computationally. This will then lead to more accurate information about stability changes between conditions. It will also allow investigation of possible functional roles of genomic regions which appear unstable in terms of community structure. Those regions potentially participate in dynamic interactions or bridge different communities. The method for community detection allows for detecting nodes which belong to multiple communities but so far, has been proven to be too computationally intensive to be applied in the case of chromatin interactions.
Another important future direction consists of detecting the nested community structure of RIS with shHMGA1 treatment, to complement the SAHF formation analysis. The large communities consisting of SAHF-associated H3K9me3 are expected to re-arrange with shHMGA1 treatment, based on the analysis in chapter 2, which shows a group of H3K9me3- and HMGA1-associated interactions increasing in RIS and decreasing with shHMGA1 treatment.

All the positions of genomic regions in the hierarchy determined are relative to each other. This can be complemented with information about regions situated at the edge and at the centre of the nucleus, respectively. A similar approach was employed in a study by Paulsen et al. [94], where the relative positions of TADs to each other and to the edge of the nucleus is inferred, by modelling TADs as globules. Paulsen et al. use Lamin B1, Lamins A/C information to infer whether the regions modelled are closer to the edge or the centre of the cell’s nucleus, respectively. Lamin B1 data is also available for IMR90 cells which would be suitable for our study as a point of reference for regions situated closer to the edge of the nucleus.

In the present study, community structure was studied chromosome by chromosome, in order to investigate high resolution structures. The method will be further utilised to infer relative positions of chromosomes in the cell nucleus, for which a Hi-C matrix of low resolution of 1-10 Mb would suffice. The relative chromosome positions can then be validated either experimentally or using publicly available datasets.

We are also interested in applying the community detection method to Hi-C datasets of other types of senescence, especially replicative senescence and in vivo senescence contexts. Another interesting comparison would be between the chromatin structure of tumourigenic and normal cells. Genomic alterations in cancer would result in dramatic chromatin rearrangement and possible new associations of affected regions in terms of the community they belong to.

The community detection approach used all the information about interactions from Hi-C data. It was complemented by graph theoretical insights about the connectivity of significant interactions. A study on different methods to detect significant interactions was described in chapter 4. Significant interactions in growing and RIS were determined using a modified version of HiC-DC [12]. Selecting only significant interactions allows for modelling a much smaller, unweighted network of interactions. Hubs and regions with high betweenness centrality were shown to be different between growing and RIS. Their biological relevance remains to be further explored in future studies, particularly in the case of the HMGA1-bound regions that do not overlap H3K9me3 peaks. HMGA1’s role in mixed chromatin interactions (between active and repressive regions) also constitutes a possible future direction of study.
along with the behaviour of HMGA1 hubs during shHMGA1 treatment.

Studying chromatin interaction networks using graph theory opened up many avenues of research for a better understanding of chromatin architecture. Insights were obtained into highly connected regions, as well as the relevance of genomic regions which undergo changes in different biological contexts. RAS-induced senescence was shown to undergo substantial chromatin alterations, both linked to gene regulation (EP interactions) and with larger re-arrangements that potentially contribute to SAHF formation.
References


Appendix A

List of captured Hi-C regions

The regions that were captured with capture Hi-C (genome assembly hg19):

- chr1:760001-1840000
- chr1:6120001-6440000
- chr1:6440001-8000000
- chr1:4000001-40280000
- chr1:149636670-150281559
- chr1:151999925-153528470
- chr1:195628197-197853469
- chr1:204690825-205805078
- chr1:218200001-219000000
- chr2:51658706-52158704
- chr2:112264738-113877018
- chr2:237974248-238490045
- chr3:69651333-70172356
- chr3:98454861-99661559
- chr3:193720001-194800000
- chr4:118246657-118746655
- chr4:121192214-122799743
List of captured Hi-C regions

- chr4:146280001-147080000
- chr5:13855890-14375686
- chr5:125988367-126892177
- chr5:131692313-132462153
- chr6:25309245-26688096
- chr6:33343435-34252792
- chr6:43786243-44289819
- chr7:26957522-27457520
- chr7:93547306-94517056
- chr8:79440001-81120000
- chr8:144440001-145680000
- chr8:145720001-146274826
- chr9:21794781-22282915
- chr9:138623251-139694720
- chr10:62178292-63700032
- chr11:1743417-2349463
- chr11:4774421-5776011
- chr11:17486246-18907128
- chr11:34446087-36532530
- chr11:36747649-38838873
- chr11:102362505-102987813
- chr12:52299443-53422841
- chr13:48754687-50555579
- chr14:76360001-77320000
- chr15:66320001-67040000
• chr15:67080001-67560000
• chr15:67600001-68720000
• chr15:70280001-72440000
• chr16:60002-559999
• chr16:62276450-62776448
• chr17:7509905-8253234
• chr17:31898178-32742752
• chr17:37958004-38812180
• chr17:40586941-41012853
• chr17:66103200-66617013
• chr18:45280001-46200000
• chr18:46400001-49920000
• chr19:41200001-43080000
• chr19:50991287-51922583
• chr20:39181020-39775400
• chr20:48000001-49000000
• chr21:27076705-27583249
• chr22:30372658-30912614
• chr22:36531666-37038352