Chromatin topology defines cell identity and phenotypic transition in human cancer and fungal pathogen

Chen, Yao

2020

Chen, Y. (2020). Chromatin topology defines cell identity and phenotypic transition in human cancer and fungal pathogen. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/140893>

<https://doi.org/10.32657/10356/140893>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 19:56:13 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**CHROMATIN TOPOLOGY DEFINES CELL IDENTITY
AND PHENOTYPIC TRANSITION
IN HUMAN CANCER AND FUNGAL PATHOGEN**

CHEN YAO

SCHOOL OF BIOLOGICAL SCIENCES

2020

**CHROMATIN TOPOLOGY DEFINES CELL IDENTITY
AND PHENOTYPIC TRANSITION
IN HUMAN CANCER AND FUNGAL PATHOGEN**

CHEN YAO

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological
University in partial fulfilment of the requirement for the
degree of Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

23rd August 2019

.....
Date

Yao Chen

.....
CHEN YAO

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

23rd August 2019

.....

Date

Amartya Sanyal.

.....

ASST. PROF. AMARTYA SANYAL

Authorship Attribution Statement

This thesis contains material from 2 papers published in the following peer-reviewed journal in which I am listed as an author.

The work presented in Chapter 4 is included in Krishnendu Guin, Yao Chen, Radha Mishra, Siti Rawaidah Binte Mohammad Muzaki, Bhagya C. Thimmappa, Caoimhe O'Brien, Geraldine Butler, Amartya Sanyal, and Kaustuv Sanyal. Spatial inter-centromeric interactions facilitated the emergence of evolutionary new centromeres. *Elife*, 2020 May 29;9:e58556. DOI: 10.7554/eLife.58556. PMID: 32469306.

The contributions of the co-authors are as follows:

- Professor Kaustuv Sanyal and Assistant Professor Amartya Sanyal supervised the study, guided the data analysis and interpretation, revised and edited the manuscript drafts.
- Professor Geraldine Butler supervised the initial step in contig assembly of *Candida tropicalis* genome (from 23 to 16 contigs).
- Krishnendu Guin and I conducted the research, analyzed the data and wrote the manuscript drafts.
- Radha Mishra constructed the engineered aneuploid strains of *C. tropicalis*.
- Siti Rawaidah Binte Mohammad Muzaki generated the 3C library for *C. tropicalis*.
- Bhagya C. Thimmappa performed part of the analysis in contig assembly of *C. tropicalis* genome (from 16 to 12 contigs).
- Caoimhe O'Brien performed the initial step in contig assembly of *C. tropicalis* genome (from 23 to 16 contigs).

Part of chapter 4 is included in Lakshmi Srekumar, Priya Jaitly, Yao Chen, Bhagya C. Thimmappa, Amartya Sanyal, and Kaustuv Sanyal. *Cis*- and *trans*-chromosomal interactions define pericentric boundaries in the absence of conventional heterochromatin. *Genetics*, 2019 Aug;212(4):1121-1132. DOI: 10.1534/genetics.119.302179. PMID: 31142612.

The contributions of the co-authors are as follows:

- Professor Kaustuv Sanyal supervised the study, guided the data analysis and interpretation, revised and edited the manuscript drafts.
- Assistant Professor Amartya Sanyal guided the analysis and interpretation of Hi-C data, revised and edited the manuscript drafts.
- Lakshmi Srekumar designed and performed 5-FOA screening and ChIP experiments, analyzed the data and wrote the manuscript drafts.
- Priya Jaitly performed 5-FOA screening and ChIP experiments and analyzed the data.
- I analyzed Hi-C data and wrote part of the manuscript drafts.
- Bhagya C. Thimmappa analyzed the ChIP-seq data.

This thesis also contains material that was presented in 1 poster in the following conference in which I am listed as an author.

Part of chapter 2 was presented in a poster titled “Alterations in chromatin topology during cancer drug resistance” in Keystone symposia “Chromatin Architecture and Chromosome Organization (X5)” at Whistler, British Columbia, Canada in March 2018. Authors of this poster are Yao Chen, Siti Rawaidah Binte Mohammad Muzaki, Aravinda Chavalmane Krishnappa Naik, Costerwell Khyriem and Amartya Sanyal.

The contributions of the co-authors are as follows:

- Assistant Professor Amartya Sanyal supervised the study and revised the poster.
- I analyzed the 3C-seq data, integrated histone H3K27ac ChIP-seq and gene expression microarray datasets, interpreted the results and prepared the poster.
- Siti Rawaidah Binte Mohammad Muzaki generated the 3C libraries.
- Dr. Aravinda Chavalmane Krishnappa Naik performed CellTier-Glo assay, extracted total RNA for gene expression microarray and generated histone H3K27ac ChIP-seq libraries.
- Costerwell Khyriem analyzed the H3K27ac ChIP-seq and microarray data.

23rd August 2019

.....
Date

Yao Chen

.....
CHEN YAO

Acknowledgements

I would like to thank all the people who have supported me during the four years of study at Nanyang Technological University (NTU), Singapore.

Foremost, I would like to convey my greatest gratitude to my supervisor, Dr. Amartya Sanyal, for his guidance and support in the past four years. His rigorous attitude to science and insights into research have always inspired me to think critically. Thanks for providing a work environment that allows me to explore my strengths and overcome my weaknesses.

I am highly indebted to my thesis advisory committee members, Dr. Mu Yuguang and Dr. Melissa Fullwood, for their valuable advice and critical feedback throughout my study.

I am truly grateful to our collaborators in Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India. I would like to thank Dr. Kaustuv Sanyal for his guidance in our projects. I am also thankful to Mr. Krishnendu Guin, Ms. Lakshmi Sreekumar and Ms. Bhagya C. Thimmappa for the insightful discussions and productive collaboration.

I also wish to express my sincere thanks to all the members and former members in our research group. I would like to give special acknowledgement to Ms. Siti Rawaidah Binte Mohammad Muzaki for her great efforts and teamwork in our common research projects that have been included in this thesis. I am grateful to Dr. Aravinda Chavalmane Krishnappa Naik for the instructions in wet lab experiments and data interpretation. I particularly thank Mr. Costerwell Khyriem for the frequent discussions and kind assistance in bioinformatics algorithms. I am also thankful to Dr. Vijayachandra Kinnimulki, Mr. Ankur Sharma and Ms. Loh Xinpeng for their companionship and support both as colleagues and as friends.

I also thank School of Biological Sciences (SBS), Biosciences Research Center (BRC) and National Supercomputing Centre (NSCC) Singapore for providing research facilities. I am thankful to SBS for the arrangements on academic activities and Nanyang Technological University for the research scholarship.

Last but not least, I would like to thank my family, friends and cats for their constant support.

Table of Contents

Statement of Originality	III
Supervisor Declaration Statement	IV
Authorship Attribution Statement	V
Acknowledgements	VII
Table of Contents	VIII
List of Figures.....	X
List of Tables	XIV
Abbreviations	XVI
Summary.....	1
Chapter 1 Introduction to chromatin and higher-order chromatin organization.....	6
1.1 A brief historical review of chromatin research	6
1.1.1 Structural basis of chromatin function	6
1.1.2 Chromatin organization and its impact on chromatin function	7
1.2 Chromatin research in the era of genomics	10
1.2.1 The comprehensive understanding of the gene regulatory system	10
1.2.2 Modern studies of higher-order chromatin organization	12
1.3 Thesis overview	17
Chapter 2 Drug-induced changes of 3D chromatin interaction network in cancer chemoresistance.....	20
2.1 Introduction	20
2.1.1 A brief introduction to cancer.....	20
2.1.2 Cancer drug resistance.....	21
2.1.3 Drug sensitive and resistant cell line models.....	23
2.1.4 Hypothesis and objectives	25
2.2 Materials and Methods	26
2.2.1 Cells and cell culture.....	26
2.2.2 The 3C library preparation and NGS	26
2.2.3 The 3C-seq data analysis.....	29
2.2.4 Chromatin immunoprecipitation sequencing (ChIP-seq).....	33
2.2.5 Microarray.....	33
2.2.6 Data availability	35
2.3 Results.....	35
2.3.1 Genome-wide chromatin contact maps reveal widespread genetic alterations during the development of drug resistance.....	35
2.3.2 Switching of A/B compartments during the development of drug resistance	42
2.3.3 Alterations in TAD structures during the development of drug resistance	54
2.3.4 Chromatin organization changes in the in-house generated drug-tolerant cells.....	60
2.4 Discussion	73
Work contribution	74
Chapter 3 FANT-C: A rapid and low-input method for genome-wide chromatin interaction study.....	75

3.1 Introduction	75
3.2 Materials and Methods	78
3.2.1 <i>Cells and cell culture.....</i>	78
3.2.2 <i>FANT-C library preparation and NGS.....</i>	78
3.2.3 <i>FANT-C/Hi-C data analysis</i>	80
3.3 Results.....	81
3.3.1 <i>FANT-C: a rapid and low-input protocol to capture genome-wide chromatin interactions</i>	81
3.3.2 <i>FANT-C recapitulates Hi-C data in a proof-of-concept study using K562</i>	83
3.3.3 <i>FANT-C captures abundant cis and short-range interactions</i>	88
3.3.4 <i>Exploring FANT-C library composition.....</i>	91
3.3.5 <i>FANT-C in AML patient samples reveals longitudinal variations in chromatin organization</i>	94
3.3.6 <i>FANT-C uncovers chr16 inversion and chr22 trisomy in one AML patient.....</i>	100
3.4 Discussion	103
Work contribution.....	105
Chapter 4 Scaffolding and characterizing the genome of <i>Candida tropicalis</i> using genome-wide chromatin interaction data	106
4.1 Introduction	106
4.2 Materials and Methods	108
4.2.1 <i>Cells and cell culture.....</i>	108
4.2.2 <i>The 3C library preparation and NGS.....</i>	108
4.2.3 <i>The 3C-seq data analysis.....</i>	109
4.2.4 <i>Synten analysis</i>	110
4.2.5 <i>Data availability</i>	110
4.3 Results.....	111
4.3.1 <i>Constructing genome assembly of <i>C. tropicalis</i></i>	111
4.3.2 <i>Genome-wide contact maps reveal conserved principle of chromatin conformation</i>	117
4.3.3 <i>Structural variations in diploid <i>C. tropicalis</i> genome</i>	124
4.3.4 <i>Centromeres as hotspots for chromosomal recombination during speciation</i>	131
4.4 Discussion	132
Work contribution.....	133
Chapter 5 Conclusions and future work.....	135
References.....	141

List of Figures

Figure 1.1 Cartoon representation of chromatin packaging from DNA double helix to a metaphase chromosome.

Figure 1.2 Schematic diagram showing non-random localization of different chromosomes in a mammalian interphase nucleus.

Figure 1.3 Schematic representation of the different techniques to study chromatin features including FISH, 3C and 3C-based techniques as well as ChIP-seq.

Figure 1.4 Cartoon representation of the hierarchical organization of mammalian interphase chromatin as revealed by chromatin contact heatmaps.

Figure 2.1 Cell viability curves of NCI-H69, H69AR and H69PR.

Figure 2.2 Genome-wide heatmap of NCI-H69.

Figure 2.3 Genome-wide heatmap of H69AR.

Figure 2.4 Genome-wide heatmap of H69PR.

Figure 2.5 Heatmaps of interactions among chromosome 3, 4 and 10 in H69AR and H69PR.

Figure 2.6 Heatmaps of interactions among chromosome 11, 12 and 13 in NCI-H69, H69AR and H69PR.

Figure 2.7 Correlation between PC1 eigenvectors and gene density or H3K27ac ChIP-seq signals in NCI-H69, H69AR and H69PR datasets.

Figure 2.8 Box plots of log₂ gene expression values in compartments with positive and negative PC1 eigenvalues generated for NCI-H69, H69AR and H69PR datasets.

Figure 2.9 The log₂ expression values of genes located in A→B and B→A (NCI-H69→H69AR) regions in NCI-H69 and H69AR.

Figure 2.10 The percentages of up- and down-regulated genes in all DEGs in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR).

Figure 2.11 The percentages of gained and lost peaks in all variant H3K27ac peaks in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR).

Figure 2.12 Example of a B→A region on chromosome 16 from NCI-H69 to H69AR.

Figure 2.13 The behavior of the B→A region (NCI-H69→H69AR) in H69PR.

Figure 2.14 Box plots of PC1 eigenvalues in switched and unswitched compartments before changing of compartments.

Figure 2.15 Box plots of PC1 eigenvalues in switched and unswitched compartments after changing of compartments.

Figure 2.16 The log₂ gene expression values and corresponding PC1 eigenvalues of several DEGs in NCI-H69, H69AR and H69PR.

Figure 2.17 Venn diagram showing the numbers of overlapping and unique TAD boundaries in the three datasets.

Figure 2.18 Example of a TAD changing region in NCI-H69, H69AR and H69PR.

Figure 2.19 Box plot showing the boundary strengths of TAD boundaries that exist in all the three datasets, in two datasets or uniquely in one dataset.

Figure 2.20 Diagram showing the in-house derived drug tolerant and age-matched control cell lines.

Figure 2.21 Cell viability curves of in-house generated H69DT and H69AMC cells.

Figure 2.22 Genome-wide heatmap of H69AMC40.

Figure 2.23 Genome-wide heatmap of H69AMC150.

Figure 2.24 Genome-wide heatmap of H69AMC40.

Figure 2.25 Genome-wide heatmap of H69AMC150.

Figure 2.26 The behavior of the region containing *ABCC1* in H69AMC40, H69DT40, H69AMC150 and H69DT150.

Figure 2.27 The log₂ gene expression values of *ABCC1* and *XYLT1* in H69AMC150 and H69DT150 cells.

Figure 2.28 Example of a TAD changing region in H69AMC40, H69DT40, H69AMC150 and H69DT150.

Figure 2.29 The log₂ gene expression values of *ABCB1* in H69AMC150 and H69DT150 cells.

Figure 3.1 Hierarchical chromatin structures in K562 revealed by FANT-C, conventional Hi-C and *in situ* Hi-C experiments.

Figure 3.2 Pearson correlation coefficients of genome-wide contact probability matrices at different bin sizes.

Figure 3.3 Comparison of PC1 eigenvalues and the identified compartments between K562 datasets.

Figure 3.4 Comparison of insulation score and the identified TAD boundaries between three K562 datasets.

Figure 3.5 Plot of mean *cis* contact probability as a function of genomic distance.

Figure 3.6 Cumulative percentage of *cis* unique valid pairs with increasing genomic distance

Figure 3.7 The percentages of short-range and long-range interactions in all *cis* interactions (*cis* unique valid pairs).

Figure 3.8 The percentages of *cis* and *trans* interactions in all valid interactions (unique valid pairs).

Figure 3.9 The percentage of PCR duplicates in the valid pairs.

Figure 3.10 The percentage of self-circles in the aligned pairs.

Figure 3.11 Diagram of two types of same fragment reads.

Figure 3.12 Workflow of library processing, NGS sequencing and bioinformatics analysis in Hi-C and 3C-seq/FANT-C.

Figure 3.13 The percentage of genomic DNA reads (dangling ends) in the aligned pairs.

Figure 3.14 Percentage of bins in the same and different compartments between pre- and post-therapy AML patient samples.

Figure 3.15 Percentage of bins showing A→B transition and B→A transition (pre-therapy→post-therapy) in all the bins undergoing compartment switching.

Figure 3.16 Examples of bins with A→B transition (pre-therapy→post-therapy), including the regions containing *KRAS* and *SRGN*.

Figure 3.17 Genomic bins ranked by PC1 eigenvalues in pre-therapy samples.

Figure 3.18 Comparison of identified TAD boundaries between datasets of pre- and post-therapy samples.

Figure 3.19 The strength of TAD boundaries that were overlapped or not overlapped with those in the other sample from the same patient.

Figure 3.20 The contact probability heatmap between chr16 p13.3-p12.1 and q12.1-q24.3 in AD598 and AD598d.

Figure 3.21 Genome-wide RD signal in AD519, AD519d, AD598 and AD598d datasets.

Figure 3.22 Zoomed-in comparison of RD signal in chromosome 22 of AD598 and AD598d.

Figure 4.1 The 3C profile anchored on the genomic bin located at 3'-end of contig 6.

Figure 4.2 The 3C profile anchored on the genomic bin located at 3'-end of contig 5.

Figure 4.3 Diagram of the joining of contig 5 and contig 6 through 3'-ends to form a new contig.

Figure 4.4 The 3C profile anchored on genomic bins located in orphan contigs.

Figure 4.5 Genome-wide heatmap of *C. albicans* from a publicly-available Hi-C dataset.

Figure 4.6 Genome-wide heatmap of *C. tropicalis* using the newly-generated genome assembly.

Figure 4.7 The compact chromatin environment of centromere-flanking pericentric domains in *trans* and in *cis*.

Figure 4.8 The contact probability heatmap of regions in chromosome 7 of *C. albicans*.

Figure 4.9 Histogram of non-zero *trans* contact probabilities in *C. tropicalis* 3C-seq dataset.

Figure 4.10 Heatmap of contact probabilities between chromosome 1 and 4.

Figure 4.11 Genome-wide heatmap of *C. tropicalis* using the alternative genome assembly.

Figure 4.12 Heatmap of contact probabilities between chromosome 1 and 4 using the alternative genome assembly of *C. tropicalis*.

Figure 4.13 Diagram of the balanced translocation between chromosome 1 and 4.

Figure 4.14 Read depth signal of the amplified region in chromosome 4.

Figure 4.15 Contact probability heatmaps showing the amplified region in chromosome 4.

Figure 4.16 Read depth signal of the amplified region in chromosome R.

Figure 4.17 Contact probability heatmaps showing the amplified region in chromosome R.

Figure 4.18 The association of synteny blocks in the genomes of *C. albicans* and *C. tropicalis* as revealed by genome-wide synteny analysis.

List of Tables

Table 2.1 Mapping statistics of the 3C-seq datasets of NCI-H69, H69AR and H69PR.

Table 2.2 List of translocations observed from genome-wide heatmaps of NCI-H69, H69AR and H69PR.

Table 2.3 Number of bins in A and B compartments in NCI-H69, H69AR and H69PR datasets.

Table 2.4 Comparison of bins in A/B compartments among NCI-H69, H69AR and H69PR datasets.

Table 2.5 Statistics of genes in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR).

Table 2.6 Statistics of variant H3K27ac peaks in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR).

Table 2.7 Number of TADs and the mean TAD size in NCI-H69, H69AR and H69PR datasets.

Table 2.8 Statistics of TAD calling using technical replicates of NCI-H69 and H69AR.

Table 2.9 Number of overlapping and unique TAD boundaries between technical replicates of NCI-H69 and H69AR.

Table 2.10 Number of overlapping and unique TAD boundaries (allowing one-bin mismatch) between technical replicates of NCI-H69 and H69AR.

Table 2.11 Number of overlapping and unique TAD boundaries between datasets of NCI-H69, H69AR and H69PR.

Table 2.12 Number of overlapping and unique TAD boundaries (allowing one-bin mismatch) between datasets of NCI-H69, H69AR and H69PR.

Table 2.13 Mapping statistics of H69AMC40, H69AMC150, H69DT40 and H69DT150 datasets.

Table 2.14 List of translocations observed from genome-wide heatmaps of NCI-H69, H69AMC40 and H69AMC150, H69DT40 and H69DT150.

Table 2.15 Number of bins in A and B compartments in NCI-H69, H69AMC40, H69AMC150, H69DT40 and H69DT150 datasets.

Table 2.16 Comparison of bins in A/B compartments among NCI-H69, H69DT40 and H69DT150 datasets.

Table 2.17 Comparison of bins in A/B compartments among NCI-H69, H69AMC40 and H69AMC150 datasets.

Table 2.18 Comparison of bins in A/B compartments between H69AMC40 and H69DT40 datasets.

Table 2.19 Comparison of bins in A/B compartments between H69AMC150 and H69DT150 datasets.

Table 2.20 Number of TADs and the mean TAD size in H69AMC40, H69AMC150, H69DT40 and H69DT150 datasets.

Table 3.1 The standard protocols of conventional Hi-C, 3C-seq and FANT-C.

Table 3.2 Mapping statistics of FANT-C, Hi-C and *in situ* Hi-C K562 datasets.

Table 3.3 Number of bins in A and B compartments in the three K562 datasets.

Table 3.4 Number of TADs and the mean TAD size in the three K562 datasets.

Table 3.5 Short-range (< 1 Mb) and long-range (≥ 1 Mb) interactions in all *cis* unique valid pairs.

Table 3.6 Mapping statistics of FANT-C datasets generated using AML patient samples.

Table 3.7 Number of bins in A and B compartments in datasets of patient samples.

Table 3.8 Comparison of compartments between pre- and post-therapy samples.

Table 3.9 Number of TADs and the mean TAD size in datasets of AML patient samples.

Table 4.1 Information about 12 contigs, including eight numbers of Mb-sized long contigs and four numbers of kb-sized small contigs.

Table 4.2 Information about the upgradation of genome assembly from eight long contigs to seven chromosomes.

Table 4.3 Mapping statistics of in-house *C. tropicalis* 3C-seq and publicly-available *C. albicans* Hi-C datasets

Abbreviations

3C	Chromosome conformation capture
4C	Circularized chromosome conformation capture
5C	Chromosome conformation capture carbon copy
5-FOA	5-Fluoroorotic acid
ABC	ATP-binding cassette
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
CFCC	Centromere-flanking compact chromatin
CHEF	Contour-clamped homogeneous electric field gel electrophoresis
CHEF-Southern	CHEF-Southern blotting
ChIA-PET	Chromatin interaction analysis with paired-end-tags
ChIP-qPCR	Chromatin immunoprecipitation quantitative PCR
ChIP-seq	Chromatin immunoprecipitation sequencing
CNV	Copy number variation
CT	Chromosome territory
CTCF	CCCTC-binding factor
DEG	Differentially-expressed gene
DNaseI	Deoxyribonuclease I
DNMT	DNA methyltransferase
dCTP	Deoxycytidine triphosphate
dNTP	Deoxynucleoside triphosphate
EC50	Half-maximal effective concentration
eRNA	Enhancer RNA
FBS	Fetal bovine serum
FISH	Fluorescence in situ hybridization
H3K27ac	Histone 3 Lysine 27 acetylation
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HIR	Homogenized inverted repeat
IC	Interchromatin compartment
MAPQ	Mapping quality
MDR	Multidrug resistance

mRNA	Messenger RNA
MRS	Major repeat sequence
NGS	Next-generation sequencing
NSCLC	Non-small cell lung cancer
PC	Principal component
PC1	The first principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
RD	Read depth
rDNA	ribosomal DNA
RNA-seq	RNA sequencing
RNase	Ribonuclease
RPKM	Reads per kilobase per million
SAM	Sequence alignment/map
SCLC	Small cell lung cancer
SDS	Sodium dodecyl sulphate
SLC	Solute carrier
SMRT	Single-molecule real-time
SNP	Single nucleotide polymorphism
TAD	Topologically associating domain
WGS	Whole-genome sequencing

Summary

Higher-order chromatin organization plays an important role in epigenetic regulation of gene expression. Chromatin topology has been increasingly recognized to define cellular identity. Alterations in chromatin conformation are often associated with phenotypic changes and can even lead to cellular state transitions. The developments of chromosome conformation capture (3C) and its derived techniques have provided molecular approaches to study chromatin interactions based on proximity ligation. Using genome-wide derivatives of 3C such as Hi-C or 3C-seq, three-dimensional chromatin architecture inside the nucleus could be characterized. Comparison of interphase chromatin structures at different cellular states could therefore contribute to the understanding of underlying gene regulatory mechanisms during phenotypic transitions in development, differentiation and disease. Furthermore, the frequency of intra-chromosomal contacts between two chromatin regions decreases with increasing linear genomic distance. This fact can be employed for scaffolding of contigs to facilitate a rapid construction of the chromosome-level genome assembly. In the multiple studies included in this thesis, we explored the role of three-dimensional chromatin structure in defining cellular identity and in achieving phenotypic changes.

Chapter 1 begins with a concise account about chromatin biology from historical perspective followed by a review of a few models of DNA packaging to understand the structural basis of chromatin function. We also discussed the impact of genomics and post-genomics era, which witnessed a huge increment in the genome assembly and annotation of genetic elements and chromatin features in various organisms. All these genome-wide datasets, as well as their large-scale analysis and integration due to tremendous advancement in computational and bioinformatics approaches, provide the necessary foundation for the modern understanding of chromatin function and gene regulatory system. Next, recent researches on higher-order chromatin organization using 3C-based techniques have been elucidated. This chapter concludes with an overview of the studies carried out in the thesis.

In Chapter 2, we studied drug-induced changes of chromatin topology in cancer chemoresistance, one of the major challenges in cancer treatment. Although many genes were found to be involved in cancer drug resistance, the underlying gene-regulatory mechanism remains poorly characterized. Since chromatin organization is known to regulate gene expression during cellular state transition, we hypothesized that chromatin interaction network

changes as cancer cells develop resistance to anticancer drugs. To understand the role of disease-causing chromatin alterations in drug resistance, we performed 3C-seq using doxorubicin-resistance cell line model derived from human small cell lung cancer (SCLC) cells. These cell lines include drug-sensitive NCI-H69, its *in vitro* multidrug resistant variant H69AR as well as the revertant H69PR that is resensitized to doxorubicin after a drug holiday. The genome-wide chromatin contact maps showed different genomic rearrangements in the three cell lines, especially the widespread translocations in H69AR, suggesting karyotype alterations during development of drug resistance. Further analysis revealed the switching of active and repressed compartments between sensitive and resistant cells, which is associated with concomitant changes in gene expression and enhancer profiles. Moreover, the comparison of topologically associating domain (TAD) boundaries suggests a relatively low similarity of TAD structures between cell lines. We found that changes of TAD boundaries are associated with alterations of gene expression patterns and enhancer activities, possibly through rewiring of enhancer-gene interactions. Furthermore, to study the time-resolved changes in chromatin interactions during the development of drug resistance, we have generated in-house drug-tolerant cells by culturing NCI-H69 cells in increasing doses of doxorubicin. The genome-wide chromatin interaction analysis revealed changes in higher-order chromatin organization at different levels between doxorubicin-tolerant H69DT cells and their age-matched control H69AMC cells, suggesting that genomic rearrangements, A/B compartment switching and TAD alterations are frequent in cancer cells like NCI-H69 under drug-selection pressure. Additionally, we found that some new genes are differentially expressed between in-house drug-tolerant and control cells. These genes are not differentially expressed between NCI-H69 and H69AR, suggesting that cancer cells can utilize different sets of genes to develop drug tolerance in response to the same drug challenge.

In Chapter 3, we proposed a rapid and low-input method for genome-wide chromatin interaction study. The current genome-wide variants of the 3C technique have provided the mainstream strategy for characterization of the three-dimensional chromatin organization. However, the extended timeline and complex experimental procedures largely restrict their applicability in clinical settings, where the priority lies in the ease of protocol using a small sample in a limited time-scale. Therefore, we developed FANT-C, a 10-hour simple and straight-forward experimental protocol to prepare 3C library from a low-input sample (as low as 50,000 cells). This library can then be sonicated into shorter fragments and sequenced using next-generation sequencing (NGS) to detect genome-wide chromatin interactions. In addition,

a large proportion of FANT-C reads contain genomic DNA sequences that can provide information about genetic aberrations apart from the intended chromatin interaction information. Currently in clinics, detection of genetic aberrations by NGS is emerging as a primary method for molecular characterization of tumors and clinical decision-making of cancer patients. Hence, FANT-C can be applied to cancer biopsy samples to get additional information about chromatin interactions together with profiles of genetic abnormalities while taking care of patient comfort by avoiding multiple sample collection. In a proof-of-concept study, we performed FANT-C using 50,000 K562 cells and showed that, at different levels of chromatin organization, the FANT-C dataset is comparable with publicly-available conventional Hi-C and *in situ* Hi-C K562 datasets. We next implemented FANT-C in a longitudinal follow-up study using acute myeloid leukemia (AML) patient samples before induction therapy and after complete remission. Switching of chromatin compartments and alterations in TAD structures were detected between pre- and post-therapy samples. In addition, we detected an inversion in chromosome 16 of the patient with AML subtype M4Eo before and after therapy. Using the genomic DNA sequences from FANT-C dataset, we found chromosome 22 trisomy in the same patient after therapy. The ability of FANT-C to simultaneously provide information about chromatin conformation as well as genomic abnormalities of clinical samples within diagnostic timeline empowers its potential application in NGS-based routine cancer test.

In Chapter 4, we incorporated chromatin interaction data for scaffolding and characterizing the genome of *Candida tropicalis*. As a member of the CUG-Ser1 clade of the fungal phylum of Ascomycota, *C. tropicalis* is a common and virulent fungal pathogen in human. The increasing cases of drug resistance in *C. tropicalis* have severely constrained the therapeutic strategy to combat this disease. Previous studies using *Candida albicans*, the well-studied organism from the same clade, revealed large-scale genomic changes during the development of drug resistance. However, the genomic features of *C. tropicalis* remain largely unknown due to the lack of complete genome assembly. Meanwhile, comparison of genomes in *C. albicans* and *C. tropicalis*, the two closely-related species, could potentially indicate the driving force of speciation in their common ancestor. By combining single-molecule real-time (SMRT) sequencing and 3C-seq data, we successfully constructed the complete chromosome-level genome assembly of *C. tropicalis* type strain MYA-3404 based on previously-available contig information. Similar to *C. albicans*, the genome-wide chromatin contact map of *C. tropicalis* revealed strong centromere-centromere interactions as well as telomere-telomere interactions.

We also identified a balanced translocation between one homolog of chromosome 1 and 4 in the diploid genome. Furthermore, the read depth signal generated from the genomic DNA reads in the *C. tropicalis* 3C-seq dataset showed segmental amplifications on chromosomes 4 and R respectively. Orthologs of a few genes in these regions are known to be involved in pathobiology and antifungal drug resistance in *C. albicans*. Finally, genome-wide synteny analysis was performed to study the karyotype changes between seven chromosomes of *C. tropicalis* and eight chromosomes of *C. albicans*, which identified centromeres as possible hotspots for ancient chromosomal rearrangements during karyotype evolution in their common ancestor.

Chapter 4 also includes part of our results in another study, where we analyzed publicly-available Hi-C dataset of *C. albicans* to examine the *cis* and *trans* chromatin interactions in centromere-proximal regions. We identified ~25 kb specialized pericentric domains surrounding the centromeres which form a compact chromatin neighborhood. Segregated from the bulk chromatin, this centromere-flanking compact chromatin (CFCC) was shown to restrict centromere activities, such as neocentromere formation, to the pericentric domains upon native centromere deletion. Inspired by this study, we examined the chromatin organization in pericentric regions of *C. tropicalis*. However, our analysis suggests that CFCC may not be present in *C. tropicalis*.

Chapter 5 is the concluding chapter of the thesis which discusses the underlying impact of the research carried out in this thesis in the context of the results produced by earlier studies in the field. We also explored the future research undertakings that may provide deeper insights into some of the conclusions that remain unclear or unanswered in the present study.

Overall our study on cellular state transitions, *viz.*, drug sensitive vs. resistant SCLC cells and AML patient samples before treatment and after remission, unveiled the important role of chromatin organization in gene regulatory mechanisms. Next, we have put forward a low-input rapid method of generating 3C library which can be adapted in clinical settings to get additional information about genome-wide chromatin interactions along with the intended information about genetic aberrations in clinical samples. Finally, we applied 3C-seq data in constructing chromosome-level genome assembly of pathogenic yeast *C. tropicalis* and deciphering the chromatin organization and genetic alterations in this organism. Additionally, the comparison between two closely-related *Candida* species, *C. tropicalis* and *C. albicans*, suggests

interesting differences in their pericentric chromatin organization as well as the potential mechanism of karyotype evolution in their common ancestor.

Chapter 1 Introduction to chromatin and higher-order chromatin organization

1.1 A brief historical review of chromatin research

1.1.1 Structural basis of chromatin function

As the carrier of genetic information, chromatin plays a central role in the living world. The earliest description of chromatin can be traced back to the 1880s, when biologist Walther Flemming stained this nuclear substance (in the form of a mitotic chromosome) with basophilic dyes and observed it under primitive light microscope¹. His studies on cell division suggested the importance of chromatin in inheritance. Subsequently, the emergence of early genetics further established the chromatin as the carrier of hereditary information. Hugo de Vries, one of the first geneticists, suggested the concept of genes in 1889² and published his work on the rediscovery of Mendelian principles in 1900³. Almost one decade later, Wilhelm Johannsen coined the terms “genotype” and “phenotype” in 1909⁴ based on his investigations into pedigree breeding of plants. The model that genes as linear particles in the chromosome⁵ was later proposed by Thomas Hunt Morgan circa 1910, who developed the theory of genes⁶ and illustrated the principle of genetic linkage in fruit fly⁵. In 1920, Hans Winkler proposed the name genome as the complete haploid set of chromosomes in an organism⁷.

These genetic theories further instigated the research that how the hereditary information is stored in the chromatin and transferred between generations. Moreover, how the chromatin performs its function to convert genotype to phenotype remained unresolved. These questions attracted researchers from different fields to explore the structural basis of chromatin function. The components of chromatin, DNA and histones, were characterized around the same time as the discovery of chromatin. Described in 1871, biochemist Friedrich Miescher first extracted DNA from the nuclei of pus leukocytes⁸. In 1884, another biochemist, Albrecht Kossel, discovered histones in the acidic extracts of avian erythrocyte nuclei⁹. The presence of both nucleic acids and proteins in chromatin led to the debate of physical entity that constitutes the genome. Microbiologist Frederick Griffith’s experiment with pneumococcal bacteria in the 1920s¹⁰, followed by studies published by Oswald Avery, Colin McLeod and Maclyn McCarty in 1944¹¹, eventually proved that DNA, and not the protein, is the hereditary material. In 1953, the proposal of double-helical DNA structure by James Watson and Francis Crick¹² provided

a sophisticated model that explains the heritability of genetic information. The subsequent statement of central dogma by Francis Crick then described the residue-by-residue transfer of sequential information¹³, which explains the execution of chromatin functions through processes like transcription and translation. Meanwhile, the studies on histones were also taking place. Ernest Johns *et al.* first demonstrated the fractionation of histones and characterized different histone proteins^{14,15}. Apart from the role of histones in DNA packaging, histone post-translational modifications have been shown to have a significant impact on transcriptional activity of eukaryotic chromatin¹⁶.

1.1.2 Chromatin organization and its impact on chromatin function

The characterizations of chromatin components have established a fundamental understanding of the chromatin and its functions at the molecular level. Meanwhile, these findings also led to a question that how the long negatively-charged DNA polymer is packed together with histones inside the eukaryotic nucleus. However, little was known about the organization of chromatin except for a few observations that demonstrated the differential packaging of DNA leading to different degrees of chromatin compaction. For example, mitotic chromosomes are much more compact than interphase chromatin, which might allow most DNA-based processes to occur during the long-lasting interphase stage. Additionally, characterization of chromosome bands in fly polytene chromosomes during the 1930s¹⁷⁻¹⁹ suggested the existence of compact heterochromatin and loosely packed euchromatin. Moreover, electron microscopic studies during the 1950s clearly demonstrated that heterochromatin preferentially localizes at the nuclear periphery and around the nucleolus. This packaging of DNA into a small volume is believed to be achieved by the close association of DNA and histone proteins to form nucleohistone complex. In the early 1970s, based on X-ray data, the famous models of nucleohistone were superhelical in character²⁰. It was in the mid-1970s that several studies discovered the existence of the nucleosome as the basic unit of chromatin organization, where DNA is wrapped around a basic core histone octamer that consists of histone proteins, H2A, H2B, H3 and H4²¹⁻²⁴ (**Figure 1.1**). With the help of histone protein H1, nucleosomes are repeatedly linked adjacent to one another to form a 10-nm fiber, which was originally described to resemble the “beads-on-a-string” structure²⁵ (**Figure 1.1**). The discovery of nucleosomes enabled the understanding of DNA-based processes from new perspectives^{26,27}. The coiling of DNA outside of the histone core permits its direct contacts with other nuclear proteins. On the other hand, since two strands of DNA need to be temporarily unwounded and separated during

processes like transcription and replication, the presence of nucleosomes hinders such enzymatic activities²⁸. To regulate the access of polymerases to the DNA template, the local chromatin structure can be modified through epigenetic processes such as 1) covalent enzymatic modifications of DNA or histone proteins^{29,30}, 2) transient modulations of nucleosomes by chromatin remodeling complexes³¹ and 3) substitutions of core canonical histones with certain histone variants³². Therefore, the structure of the nucleosome permits the epigenetic regulation of chromatin functions.

Figure 1.1 Cartoon representation of chromatin packaging from DNA double helix to a metaphase chromosome. The 2-nm DNA double helix is wrapped around core histone octamers to form nucleosomes, which are linked adjacent to each other with the help of H1 histone to form a 10-nm nucleosome fiber. The 10-nm fiber is further folded into higher-order chromatin organization that constitutes the metaphase chromosome of 1,400 nm in size.

While it was clear that the 10-nm nucleosome fiber constitutes the fundamental structure of chromatin packaging, how this fiber is further folded to form the bulk chromatin had been a matter of controversy (**Figure 1.1**). It was further proposed that the 10-nm fiber can be folded into 30-nm chromatin fiber *in vitro*³³. Though studied extensively using various techniques³³⁻⁴³, such as electron microscopy (EM), cryogenic electron microscopy (cryo-EM), X-ray crystallography and small-angle X-ray scattering (SAXS), the definitive organization of the 30-nm fiber remained unclear⁴⁴. The two famous structural models of 30-nm fiber included the solenoid model³³ and the zigzag model that was presented one decade later³⁴. Nevertheless, it had long been assumed that 10-nm fiber is folded into 30-nm fiber inside living cells, which is further packed to eventually form the bulk organization of mitotic chromosomes and interphase chromatin. Well-known models of such higher-order chromatin organization included 1) “hierarchical helical folding” model where 30-nm fiber is folded into larger chromatin fiber in a hierarchical manner⁴⁵⁻⁴⁷, and 2) “radial loop” model where loops formed by 30-nm fibers radially wrap around a scaffold composed of condensin and topoisomerase II α ⁴⁸⁻⁵⁰. However, the first observation of “native” mammalian chromosomes using the cryo-EM technique in

1986 suggested the chromosomes as liquid-like compact aggregations of 10-nm fibers^{51,52}, which contradicted the existing structural models. Subsequent studies in interphase chromatin and mitotic chromosomes further evidenced the lack of 30-nm chromatin fibers *in vivo*⁵³⁻⁵⁶. Based on the experimental results, a “polymer-melt” model was proposed in 2008⁵⁶, which explains the *in vitro* observations of 30-nm fibers as artifacts under low-salt buffer conditions^{57,58}. Furthermore, studies on large-scale chromatin organization in human cells revealed no regular periodic structures up to 1,000 nm in neither interphase chromatin nor mitotic chromosomes⁵⁹. The 10-nm nucleosome fibers are thus believed to directly form the bulk chromatin structures in most eukaryotic cells^{60,61}. However, recent *in situ* imaging of chromatin by electron tomography (ChromEM tomography) revealed the chromatin as a disordered 5 to 24-nm curvilinear polymer chain that is organized at different packing densities in interphase and mitotic cells⁶².

Although the chromatin fiber is locally folded in an irregular manner, the nuclear higher-order chromatin arrangements are not random. Recent studies on mitotic chromosomes suggested a structure similar to the previous radial loop model, where the condensins and topoisomerase II α form a scaffold for the 10-nm fibers to loop around^{60,63}. The organization of interphase chromatin, however, was found to be distinct from that of mitotic chromosomes. Firstly, it was initially proposed in 1885 by Carl Rabl that each chromosome occupies a distinct territory in the animal nucleus during interphase⁶⁴. Based on the observations in roundworms, Theodor Boveri supported this concept and introduced the term chromosome territory (CT) in 1909⁶⁵. However, it was rarely considered for many decades⁶⁶ until revisited by a few studies in *Allium cepa* root tips and Chinese hamster cells in the 1970s-1980s⁶⁷⁻⁷¹. The subsequent development of fluorescence *in situ* hybridization (FISH)⁷² technique since the 1980s enabled the direct evaluation of the spatial relationship among genomic loci. Using FISH, CTs in the mammalian interphase nucleus were unequivocally proven⁷³⁻⁷⁷ (**Figure 1.2**). Although CTs are mostly discrete, it was found that the small, gene-rich chromosomes are generally clustered at the nuclear center, while the large, gene-poor chromosomes are located near the nuclear periphery^{78,79}. Further studies revealed the localization preferences of active or inactive genes with respect to CT boundary⁸⁰⁻⁸². For example, actively co-transcribing genes were found to loop out of the CTs^{83,84} and co-associated spatially at interchromatin compartments (ICs), that are enriched with RNA polymerases, transcription factors and splicing speckles^{85,86} (**Figure 1.2**). These studies suggested the significant impact of interphase chromatin organization on chromatin function.

Figure 1.2 Schematic diagram showing non-random localization of different chromosomes (shown with different colors) in a mammalian interphase nucleus. For transcription, genes loop out of the respective chromosome territories (CTs) to associate with transcriptional machinery in the interchromatin compartments (ICs).

1.2 Chromatin research in the era of genomics

So far, we have reviewed the studies on the components and higher-order organization of the chromatin from the perspective of its function. These studies have made great strides in the understanding of genetic activities and epigenetic modulations. However, they were largely based on low-throughput methods like imaging techniques. The recent developments in genomics and epigenomics fields have allowed systemic characterizations of chromatin function and regulation through high-throughput analysis.

1.2.1 The comprehensive understanding of the gene regulatory system

The mystery of the genome and its biological functions have been gradually uncovered since the first complete sequencing of viral DNA genome by Frederick Sanger in 1977⁸⁷. The successful approaches in sequencing several viral genomes⁸⁷⁻⁹⁰, though with the aid of existing information such as physical genomic maps, proved the feasibility of genome assembly from small sequencing reads. Following the first DNA sequencing procedures developed by Sanger⁹¹, various sequencing techniques and bioinformatic algorithms have been developed and continuously improved to profile the genomes of numerous organisms. For example, the first complete bacterial genomes were assembled entirely based on reads from whole-genome shotgun sequencing in 1995^{92,93}. It was one year later that the genome sequencing of yeast *Saccharomyces cerevisiae* marked the completion of the first eukaryotic genome⁹⁴. These achievements set the stage for genome sequencing of larger multicellular organisms, including human. With the great efforts from laboratories around the world, the Human Genome Project

(HGP), formally launched in 1990, was finally declared complete in 2003^{95,96}. The human genome, with approximately 3 Gb in size, contains > 20,000 protein-coding genes⁹⁶. However, each cell type expresses only a subset of these genes in a systematic manner. With the complete genome assembly, the gene expression profiles can be obtained comprehensively by recent techniques such as RNA sequencing (RNA-seq) or microarray.

The activities of protein-coding genes, on the other hand, are regulated through various epigenetic processes such as DNA methylation, histone modifications and chromatin remodeling in response to environmental stimuli and cellular signals. Apart from protein-coding genes, the genome sequences of eukaryotes show large proportions of non-coding regions that comprise introns and intergenic sequences. In human, such regions occupy approximately ~98% of the entire genome⁹⁷. Originally called the “junk DNA”, these regions were later found to harbor the majority of genetic polymorphisms associated with human diseases⁹⁸. Regulatory sequences, located in non-coding regions, have been shown to control gene expression in many cellular activities. Similar to genic sequences, the activities of regulatory sequences can also be modulated by different types of epigenetic modifications. For example, among millions of potential enhancer regions in the human genome, only a small subset is selectively active in a certain cell type⁹⁹. In response to cellular signals, transcription factors bind to enhancer regions, which usually leads to the recruitment of co-regulators like histone acetyltransferase (HAT) p300/CBP. Acetylation of histone tails in enhancer-associated nucleosomes, particularly H3K27 acetylation (H3K27ac), then results in active transcription of enhancer RNAs (eRNAs) by RNA polymerase II. The eRNA transcription was found to positively correlate with enhancer-specific activities¹⁰⁰, suggesting the functional roles of eRNAs in gene regulation. The histone H3K27ac status of enhancers can thus be used to infer their activities in different cell types^{101,102}. Similarly, other types of epigenetic modifications on genomic elements can provide information about the activities of regulatory sequences or genes. The epigenetic modifications of histones, as well as transcription factor binding, can be successfully annotated in a genome-wide manner. For example, chromatin immunoprecipitation sequencing (ChIP-seq), a combination of chromatin immunoprecipitation and next-generation sequencing (NGS), allows global mapping of DNA loci that interact with the given protein of interest¹⁰³ (**Figure 1.3**). Genome-wide epigenetic marking data from a large collection of cell lines and primary tissues are available in many databases and consortia, including Encyclopedia of DNA Elements (ENCODE)¹⁰¹ and RoadMap Epigenomics¹⁰⁴. These marks allow the identification of different genomic elements.

The regulatory elements, however, may locate upstream or downstream of the corresponding gene promoters. Interestingly, cell-type-specific gene transcription was found to be governed by distal regulatory sequences¹⁰⁵. Though far from the target genes, these regulatory sequences can exert their effects to modulate gene expression. It was demonstrated that the long-range gene regulation is achieved by physical interactions between promoters and regulatory sequences through the formation of chromatin loops¹⁰⁶. These findings suggested the significant impact of higher-order chromatin organization on gene regulation.

1.2.2 Modern studies of higher-order chromatin organization

As discussed in **1.1.2**, the studies using the FISH technique have established many aspects of chromatin organization. However, FISH allows only a handful of genomic loci to be simultaneously examined in a few hundreds of cells, making it low-throughput in detecting chromatin interactions¹⁰⁷ (**Figure 1.3**). As an imaging technique, the spatial resolution of FISH is intrinsically limited by diffraction barrier¹⁰⁸. In contrast, the recent developments of chromosome conformation capture (3C) technique¹⁰⁹ and its variants have provided molecular approaches to detect spatial proximity between chromatin regions (**Figure 1.3**). In a 3C experiment¹⁰⁹, physically interacting chromosomal regions are first crosslinked by formaldehyde. The chromatin is subsequently digested by a restriction enzyme, and the resulting restriction fragments are ligated through overhangs in a dilute condition to facilitate intramolecular ligation. The 3C library that contains the information about genome-wide chromatin interactions is then obtained after reverse crosslinking and DNA purification. Since the invention of 3C in 2002, the original protocol has been modified to study chromatin interactions at different resolutions for different research purposes. While the incorporation of polymerase chain reaction (PCR) in conventional 3C enables detection of pairwise interactions, it requires prior knowledge of the regions involved. To identify all the interactions of a specific genomic locus, circularized chromosome conformation capture (4C) was then developed in 2006 with the utilization of inverse PCR on circularized DNA products^{110,111} (**Figure 1.3**). Studies using 4C showed higher frequencies of intra-chromosomal (*cis*) interactions than inter-chromosomal (*trans*) interactions¹¹¹⁻¹¹⁴, supporting the previous CT model. Another derivative of 3C published in the same year, chromosome conformation capture carbon copy (5C), allows high-throughput mapping of local chromatin organization by quantitatively measuring a large number of ligation products of many loci from different chromosomes simultaneously¹¹⁵

(**Figure 1.3**). Subsequently in 2009, two genome-wide approaches were developed, namely, Hi-C¹¹⁶ and chromatin interaction analysis with paired-end-tags (ChIA-PET)¹¹⁷ (**Figure 1.3**). In Hi-C, digested and ligated restriction fragments are labelled with biotin and enriched using streptavidin beads. The sequence information of all the ligation products is then obtained by using paired-end NGS. On the other hand, ChIA-PET incorporates immunoprecipitation technique and biotin affinity purification for genome-wide detection of chromatin interactions mediated by a specific protein. With the recent developments in sequencing techniques, 3C-seq, the direct combination of 3C library preparation followed by detection using paired-end NGS, has emerged since 2012¹¹⁸. Similar to Hi-C, 3C-seq also captures “all-to-all” interactions, which allows high-throughput analysis of global chromatin architecture. Taken together, these 3C-based techniques have the potential to detect interactions across millions of cells at the genome-wide scale or in a localized manner with high resolution. Recently, single-cell approaches of Hi-C¹¹⁹ and ChIA-PET (known as ChIA-Drop)¹²⁰ were also developed. The single-cell Hi-C study revealed stochasticity of chromatin structures across the cell population¹¹⁹. On the other hand, ChIA-Drop captured multiplex chromatin interactions with high heterogeneity in *Drosophila* cells¹²⁰.

Figure 1.3 Schematic representation of the different techniques to study chromatin features including FISH (*top*), 3C and 3C-based techniques (*middle*) as well as ChIP-seq (*bottom*). The 3C and 3C-based techniques share the starting steps, i.e., formaldehyde crosslinking and restriction enzyme digestion. The various 3C-based methods allow researchers to address specific research questions.

In fact, Hi-C has been widely used for the characterizations of chromatin structures in many species. The genome-wide contact maps constructed from mammalian Hi-C data confirmed the presence of chromosome territories and the tendency of small, gene-rich chromosomes to

be clustered together in the nucleus¹¹⁶. The analysis on large regions of enriched and depleted interactions between intra-chromosomal loci revealed that the entire genome can be divided into two compartments¹¹⁶ (**Figure 1.4**). One of these compartments, “A”, was observed to strongly correlate with genomic and epigenomic features such as gene density, gene expression (mRNA expression), accessible chromatin (DNaseI sensitivity) and active chromatin (histone H3K36 trimethylation) marks. In contrast, the “B” compartment strongly correlates with repressive chromatin markers like H3K9 and H3K27 trimethylation. This connects A and B compartments with previous concepts of euchromatin and heterochromatin respectively. Switching of A/B compartments has been shown in diseases like B-cell lymphoma¹²¹ and Hutchinson-Gilford progeria syndrome (HGPS)¹²². These compartments can be further divided into megabase-sized topologically associating domains (TADs) that serve as basic units of chromatin within which genomic loci interact with each other at high frequencies^{123,124} (**Figure 1.4**). The interactions between adjacent TADs are restrained by their boundaries, which are usually associated with cohesin and CCCTC-binding factor (CTCF)^{123,124}. Those TAD boundaries were found to be highly conserved across different cell types and species, indicating a stable chromatin topology in mammalian genomes^{123,125}. Alterations in TAD boundaries usually occur during transitions of cellular states such as stem cell differentiation¹²³, somatic cell reprogramming¹²⁶ and development of diseases¹²⁷ or disorders¹²⁸. For example, a recent study on TADs in T cell acute lymphoblastic leukemia (T-ALL) revealed recurrent deletions of boundary sites between neighboring TADs which contain proto-oncogenes¹²⁹. Disruption of these boundaries in normal cells led to the activation of the proto-oncogenes¹²⁹. Similarly, reorganization of TADs was observed during replicative senescence¹³⁰. Such changes in TAD structures were found to regulate gene expression through rewiring of enhancer-gene interactions¹³¹. At a finer scale, the long-range looping interactions between regulatory elements and genes (**Figure 1.4**) were found to control the gene expression levels as demonstrated at beta globin¹³² and *IFNG*¹³³ loci. In summary, these Hi-C studies established a comprehensive understanding of hierarchical chromatin structures inside the mammalian interphase nucleus, which play significant roles in maintaining regular gene expressions and a stable cellular state. In mitotic cells, polymer simulation of Hi-C data revealed that metaphase chromosomes are organized in a linear compressed longitudinal array of consecutive chromatin loops¹³⁴. Over the years, genome-wide chromatin interaction data have also provided valuable insights into the three-dimensional genome organization of different non-mammalian organisms, such as plant¹³⁵, bacteria¹³⁶, yeast¹³⁷, nematode¹³⁸, fruit fly¹¹⁸ and zebrafish¹³⁹.

Figure 1.4 Cartoon representation of the hierarchical organization of mammalian interphase chromatin (*bottom panel*) as revealed by chromatin contact heatmaps (*top panel*). *Left*: A chromosome-wide heatmap with “plaid” patterns suggesting the division of genome into A (active) and B (repressed) compartments. *Middle*: Zoomed-in heatmap showing the further division of each compartment into TADs, whose boundaries are usually associated with cohesin and CTCF. *Right*: The long-range looping interactions between an enhancer and a gene within a sub-TAD structure.

Taken together, the packaging of DNA into chromatin in the interphase nucleus is essential in the regulation of all genomic processes. The organization and partitioning of the genome into functional domains and compartments of different length scales further help in fine-tuning gene regulatory mechanisms. Maintenance and safeguarding this organization is central to preserve the cellular identity and a disease-free cellular state. Any alteration in the organization may lead to a disease state. Therefore, understanding the cell state transition-related fundamental changes in chromatin organization may pave the way for the development of novel therapeutic interventions that can revert the chromatin structure to healthy cellular state.

1.3 Thesis overview

Inspired by recent studies on higher-order chromatin architecture, we aim to investigate the role of chromatin conformation as a strong predictor of cell identity, especially during phenotypic transitions. In this thesis, we first examined the changes in 3D chromatin topology when cancer cells develop resistance to anti-cancer drugs. As one of the major obstacles in modern cancer therapy, drug resistance has been attributed to a broad spectrum of mechanisms resulting in transcriptional reprogramming of the cancer cell¹⁴⁰. We thus hypothesized that chromatin organization changes during the development of drug resistance. By performing 3C-seq on doxorubicin-sensitive and resistant human small cell lung cancer (SCLC) cell lines, we observed A/B compartment switching and TAD structure changing during the phenotypic transition of cancer cells from drug-sensitive to resistant states. These alterations in chromatin organization were found to correlate with changes in enhancer activities as well as gene expression. These findings may contribute to the understanding of the underlying gene regulatory mechanisms behind chemoresistance.

Nevertheless, the above-mentioned study was conducted using the established cell line model, which may not recapitulate the practical scenario of patient samples. Indeed, each cancer has its unique genomic and epigenomic profiles that are dynamic and can evolve quickly over time upon therapy. To obtain a personalized chromatin conformation profile, it is highly-desired to directly apply 3C-seq or other Hi-C like technique in the clinical diagnosis of cancer patients. However, current high-throughput 3C-based methods involve a series of complex experimental procedures, resulting in general requirements of ~3 days for library preparation and millions of cells as input^{109,116,141}. These factors intrinsically limit the applicability of such techniques in clinical settings. Therefore, we developed a 10-hour simplified 3C library preparation protocol, termed as FANT-C, for detection of genome-wide chromatin interactions using as low as 50,000 cells. By implementing FANT-C to bone marrow samples from acute myeloid leukemia (AML) patients, we detected alterations in 3D chromatin structures before and after induction therapy (in remission). We also applied FANT-C to our in-house developed doxorubicin-tolerant SCLC cell lines, which show time-resolved changes in chromatin conformation when the cells were exposed to increasing concentration of the drug. Our analysis reveals the regulation of different arrays of genes in the cell line model and in-house generated cell lines. In addition to chromatin organization data, FANT-C generates a majority of reads

with genomic DNA sequences which can provide additional information about genetic abnormalities such as DNA copy number changes.

Apart from cancers, fungal infections are another kind of diseases that threaten human health, especially in immunocompromised individuals like cancer patients and organ transplant recipients¹⁴². The CUG-Ser1 clade of the fungal phylum of Ascomycota comprises many pathogenic species that can infect human, including the most frequently-isolated fungal pathogen from patients in the world, *Candida albicans*¹⁴³. On the other hand, in tropical countries, the most common *Candida* species isolated from patients is *Candida tropicalis*, another disease-causing species of the same clade. *C. tropicalis* frequently leads to systemic ICU-acquired candidiasis^{144,145} and is increasingly associated with drug resistance¹⁴⁴⁻¹⁴⁶. However, in contrast to *C. albicans* that has been studied extensively, the genome-wide study of *C. tropicalis* is limited by the lack of a complete genome assembly. To explore the genomic features of *C. tropicalis*, we constructed a chromosome-level genome assembly of this organism from the previous contig-level fragmented assembly by utilizing single-molecule real-time (SMRT) sequencing reads as well as chromatin interaction information from 3C-seq data. The genome-wide chromatin contact maps indicate clustering of centromeres in *C. tropicalis* as previously observed in *C. albicans*. Together with synteny analysis that reveals enrichment of ICSBs near centromeres in both organisms, we speculate their centromeres as possible hotspots for chromosomal recombination during karyotype evolution in the common ancestor of these two closely-related species. Moreover, since chromosomal and segmental aneuploidy are known to be associated with drug resistance in *C. albicans*¹⁴⁷⁻¹⁴⁹, we searched the genome of *C. tropicalis* for such abnormalities. Using chromatin interaction data as well as genomic DNA reads in the 3C-seq dataset, we successfully identified genetic variations in the diploid genome, including heterozygous loci, balanced translocation as well as segmental amplifications. These results would facilitate future genome-wide research in *C. tropicalis*, such as the understanding of its pathogenicity and drug resistance.

The rest of this thesis is structured as follows. Chapter 2 describes the study of drug-induced changes in chromatin organization, at the level of A/B compartments and TADs, and their impact on enhancer activity and gene expression during the development of cancer drug resistance. Chapter 3 presents the FANT-C protocol, its comparison to existing high-throughput 3C-based techniques and its implementation in AML patient samples. Chapter 4 demonstrates the application of chromatin interaction data in constructing the chromosome-

level genome assembly of *C. tropicalis*. This chapter also includes the comparative analysis between *C. albicans* and *C. tropicalis* from the aspects of synteny and genome-wide chromatin organization, as well as the characterizations of genetic variations in *C. tropicalis*. Chapter 5 concludes the research work presented in this thesis and discusses future research directions.

Chapter 2 Drug-induced changes of 3D chromatin interaction network in cancer chemoresistance

2.1 Introduction

2.1.1 A brief introduction to cancer

Cancer is a leading cause of mortality in the world. In 2018, an estimated 9.6 million deaths were recorded due to cancer, accounting for approximately 1 in 6 deaths worldwide¹⁵⁰. The earliest known description of cancer was found around 1500-1600 BC in ancient Egyptian manuscripts¹⁵¹. Puzzling over the causes of cancer since the earliest times, physicians proposed various theories that evolved over time with the development of knowledge and new technologies. It is now widely accepted that cancer arises from a series of genetic and epigenetic changes which upset the normal balance between cell proliferation and programmed apoptosis¹⁵²⁻¹⁵⁴. Driven by intrinsic and/or environmental factors¹⁵⁵⁻¹⁶², these genetic variations usually lead to 1) activation of oncogenes, which promotes the malignant features of cancer cells¹⁶³, and 2) disabled functions of tumor suppressor genes, which fail to arrest cell cycle¹⁶⁴ or carry out DNA repair¹⁶⁵. Epigenetic modifications, such as DNA methylation, histone modifications and chromatin remodeling, have been found to regulate gene expression during cancer formation and progression¹⁶⁵⁻¹⁷⁰. Mutations in epigenetic regulators like chromatin modifying enzymes and chromatin remodelers have been recurrently observed in many cancer genomes^{166,171,172}. These genetic and epigenetic changes trigger the progression of normal cells to cancer cells and promote the growth of tumorous tissues¹⁷³⁻¹⁷⁶. With the ability to rapidly evolve and evade the usual defenses of the body, these cancer cells can uninhibitedly grow beyond their primary sites and spread to other organs during the process of metastasis¹⁷⁷. The impairment of organ functions affected by cancer can cause failure in the maintenance of normal physiology and immunity, which ultimately can lead to mortality¹⁷⁸.

The history of cancer treatment can be traced back to the time of cancer discovery, where the ancient Egyptian manuscripts also described a procedure to remove breast tumors through cauterization¹⁵¹. Over the years, multiple therapeutic modalities have been developed in order to cure cancer or to prolong the lives of cancer patients¹⁷⁹. Different types of cancer treatment in modern times include local therapies such as surgery and radiation therapy as well as systemic therapies such as chemotherapy and immunotherapy. Depending on the

characteristics of cancer and conditions of the patient, the therapeutic strategy for a particular individual usually encompasses one or a combination of treatments.

2.1.2 Cancer drug resistance

As one of the major categories in cancer therapeutics, chemotherapy has been gradually improved and becomes increasingly sophisticated. Over the past few decades, hundreds of anticancer drugs have been designed and produced for clinical therapy. However, the effectiveness of cancer therapy is seriously limited by drug resistance, which affects a variety of blood cancers as well as solid tumors¹⁷⁹. While some cancers are inherently tolerant to potent anticancer agents commonly used in chemotherapy, other cancers that are initially sensitive to the treatment may also acquire resistance over time through various adaptive mechanisms¹⁸⁰. Moreover, such cancers may not only fail to respond to the specific therapeutic drug but can simultaneously become resistant to other drugs with distinct structures and/or modes of action, resulting in multidrug resistance (MDR)¹⁷⁹. As a consequence, drug resistance can cause cancer recurrence and dissemination¹⁸¹. It is usually associated with a poor prognosis and thus remains the major obstacle in cancer therapy.

In order to understand drug resistance in cancer, researchers have proposed a broad range of mechanisms^{140,182}. Some of the mechanisms are specific to the drug used in therapy¹⁸³, including decreased expression of specific drug transporter, up-regulation or mutation of the drug targets, drug metabolism, etc. These types of drug resistance can possibly be overcome by adopting combination therapy, where more than one drugs are used in the treatment¹⁸⁴. However, in many cases, cancers develop MDR and become resistant to a broad spectrum of structurally and functionally unrelated drugs¹⁷⁹. The possible mechanisms of MDR include, but not limited to, reduced accumulation of drugs in the cancer cells by overexpressing efflux pumps on plasma membrane^{185,186}, decrease of drug intake, alteration of the lipid composition of cell membrane¹⁸⁷, etc. Other mechanisms of drug resistance have also been proposed, such as compartmentalization of drugs, deregulation of apoptosis¹⁸⁸, induction of emergency response genes that repair DNA damage¹⁸⁹ and alteration of cell cycle checkpoints.

Among all the mechanisms, the increase of drug efflux and decrease of drug intake are considered as the major causes of MDR. Significantly enhanced expression of efflux pumps has been found in most multidrug resistant cell lines as well as patient samples, where a broad

range of anticancer drugs are exported to extracellular space through ATP-dependent or ATP-independent manners¹⁹⁰. Most ATP-dependent pumps in the plasma membrane belong to ATP-binding cassette (ABC) transmembrane transporter superfamily. These transporters include ABCB family, ABCC family, ABCG2 (also known as MXR, BCRP or ABC-P), etc. The first-discovered multidrug-resistant gene¹⁹¹, *ABCB1* (encodes a member of ABCB family, also known as P-gp or MDR1), is widely expressed in many cancers. ABCB1 transporter consists of 2 ATP binding cassettes and 12 transmembrane domains¹⁹² and has the ability to bind to a large variety of hydrophobic drugs¹⁴⁰ and export them outside the cell via a conformation change caused by ATP hydrolysis¹⁹³. After the discovery of *ABCB1*, another important multidrug-resistant gene *ABCC1* (encodes a member of ABCC family, also known as MRP1) was discovered in a multidrug-resistant lung cancer cell line named H69AR¹⁹⁴, which will be further described in the following sections. Other members of ABCB and ABCC families, as well as ABCG2, were later found to be overexpressed in various drug-resistant cancers. In contrast to ABC transporters, solute carrier (SLC) transporters, such as organic cation transporters (OCTs), organic anion transporters (OATs) and organic anion transporters polypeptides (OATPs), are responsible for uptake of nutrients in normal cells. Decreased expression of SLC transporters involved in the drug uptake in cancer cells can result in their resistance to anticancer drugs. In certain cases, the increased expression of SLC transporters involved in nutrient uptake may also lead to drug resistance through increased survival chance of cancer cells under limited nutrient supply¹⁹⁵. Besides, a vault protein named lung resistance-related protein (LRP) also plays an important role in the regulation of nuclear entry of drugs and has been found to be expressed at high levels in many drug-resistant cell lines^{196,197}.

Inspired by these studies on drug resistance, many inhibitors that interfere with the function of drug-resistance proteins have been developed. However, the treatment using inhibitors often results in enhanced expression of targeted genes that counteract the inhibitors. For example, the clinical use of ABCB1 inhibitors showed no significant evidence that supports the change of drug response rates in any cancer¹⁹⁸. The ineffectiveness of these inhibitors suggests the presence of additional mechanisms in drug resistance. On the other hand, it is increasingly recognized that cancers are comprised of a population of cells with high genetic and epigenetic heterogeneity. Unlike intrinsic chemoresistance that is usually due to genetic variations, acquired chemoresistance is often believed to be caused by epigenetic changes that allow cancer cells to gradually adapt to anticancer drugs¹⁴⁰. For example, a study in glioblastoma stem cell showed the widespread redistribution of repressive histone methylation upon drug

treatment, which could contribute to the up-regulation of primitive developmental programs¹⁹⁹. KMT2D, a histone methyltransferase, was found to regulate the expression of estrogen receptor (ER) when Phosphoinositide-3-kinase (PI3K) pathway is blocked by specific inhibitors^{200,201}. Similarly, other types of epigenetic processes like DNA methylation^{202,203} and chromatin remodeling²⁰⁴ were also found to be involved in the development of drug resistance by cancer cells. These findings suggest that epigenetic modulators could be the targets in cancer therapy. In fact, epigenetic drugs such as inhibitors of DNA methyltransferase (DNMT) and histone deacetylase (HDAC) have been successfully used in chemotherapy¹⁶⁷. Studies also showed resensitization or increased drug efficiency following treatment with inhibitors targeting DNMT and/or HDAC²⁰⁵⁻²⁰⁸. However, epigenetic drugs targeting DNA and histone modifiers are promiscuous and may revert epigenetic marks needed to maintain cell survival in normal tissues. A more targeted therapeutic approach towards specific gene regulatory pathway involved in drug resistance is thus highly desired to achieve effective therapy with minimal toxicity. Higher-order chromatin organization plays an important role in governing cell type-specific gene expression pattern and thereby specifying cellular identity. Consequently, cancer cells are thought to be associated with specific chromatin configuration²⁰⁹. We thus believe that the study of chromatin interactions in drug-resistant cancers could provide a more comprehensive understanding of the underlying gene regulatory mechanisms during the development of drug resistance.

2.1.3 Drug sensitive and resistant cell line models

In 2012, World Cancer Research Fund International reported that 13% of new cancer cases diagnosed in that year were resulted from lung cancer, making it the most common cancer worldwide²¹⁰. In Singapore, lung cancer contributed to the highest and the second-highest cancer deaths in males and females, respectively²¹¹. Smoking is the main cause of small cell lung cancer (SCLC)²¹² and accounts for ~80% of lung cancer deaths¹⁵⁵. Therefore, the increased number of smokers²¹³, especially young smokers, might have attributed to the increment in younger SCLC patients in Singapore²¹¹. SCLC is one of the major types of lung cancer that accounts for 10%-15% of all lung cancer cases. Combination chemotherapy including cisplatin and etoposide along with doxorubicin, cyclophosphamide or vincristine is usually recommended in SCLC²¹⁴. However, SCLC has a high relapse rate and a poor prognosis because of acquired drug resistance during therapy. The 5-year survival rate of SCLC patients ranges from 2%-31% depending on the stage of cancer²¹⁵. Therefore, we focused on

SCLC in this study as it can serve as a model for the research on drug resistance. Since surgery is usually not recommended in SCLC, surgical resection is not a viable option. Besides, there is no suitable murine model available²¹⁶. Thus, this study uses a series of cell line models derived from the neuroendocrine subtype of SCLC. The cell lines include a drug-sensitive cell line NCI-H69, its multidrug-resistant derivative H69AR as well as the re-sensitized revertant H69PR.

NCI-H69 was derived from a 55-year-old male SCLC patient²¹⁷. H69AR ('AR', adriamycin resistant), the multidrug resistant variant of NCI-H69, was generated by culturing NCI-H69 cells in the presence of gradually increasing doses of doxorubicin (also known as adriamycin) for 14 months²¹⁸. H69PR ('PR', previously resistant) was derived from H69AR after continuously culturing the cells for more than 37 months in drug-free medium²¹⁹. Cell viability assay (CellTiter-Glo) following treatment with different concentrations of doxorubicin for 5 days shows the half-maximal effective concentration (EC₅₀) of H69AR is ~75 fold more than that of NCI-H69 (**Figure 2.1**). On the other hand, H69PR is almost as sensitive as NCI-H69 to doxorubicin. Besides doxorubicin, H69AR is also cross-resistant to a variety of other anti-cancer drugs such as etoposide and vincristine, etc.²¹⁸. Accordingly, H69PR has also regained sensitivity to these drugs (VP-16 and vincristine)²¹⁹. Therefore, this H69 series of cell lines, especially the pair of NCI-H69 and H69AR, have been widely used to study the differences between sensitive and resistant cells.

Figure 2.1 Cell viability curves (CellTiter-Glo assay after 5 days of drug treatment) of NCI-H69, H69AR and H69PR. The EC₅₀ values of the cell lines are shown on the top right.

The major change between NCI-H69 and H69AR at gene expression level is the overexpression of *ABCC1* in H69AR¹⁹⁴, as mentioned in 2.1.2. Apart from *ABCC1*, many other

genes have also been found to be differentially expressed in NCI-H69 and H69AR. For example, *NOTCH1*, a gene that plays important roles in cell fate decisions and a variety of developmental processes, is known to be up-regulated in H69AR. It controls cell invasion and metastasis by switching off the tumor epithelial-mesenchymal transition²²⁰. The neuroendocrine markers INSM1 and ASCL1 are downregulated in H69AR, suggesting the occurrence of neuroendocrine differentiation from NCI-H69 to H69AR²²¹. The expression levels of a few other genes or proteins, including fibroblast growth factor-inducible 14 (FN14), B-cell lymphoma-extra large (BCL-XL)²²², ER membrane calcium pump NERCA2²²³, etc., have also been found to be changed in H69AR compared to NCI-H69. Apart from protein-coding genes, a long intergenic noncoding RNA named HOTAIR (Homeobox (HOX) transcript antisense RNA) was reported to promote chemoresistance by regulating DNA methylation of a HOX gene *HOXA1*²²⁴. In addition, some miRNAs also play important roles in regulating drug resistance in H69AR²²⁵. These observations suggest that the H69AR cells have undergone a transcriptional reprogramming during the process of drug resistance.

2.1.4 Hypothesis and objectives

As described above, drug resistance is the most challenging problem in cancer therapy. While several mechanisms of drug resistance have been established, how drug-resistant genes are regulated by epigenetic factors remains unclear. We hypothesized that the 3D chromatin organization changes in response to anticancer drug treatment. These changes lead to transcriptional reprogramming which results in the transition of cancer cells from sensitive to drug-resistant state. To test our hypothesis, we use the H69 series of SCLC cell line. In order to decipher the role of 3D chromatin organization in cancer drug resistance, we have set the following objectives for this study: 1) To perform 3C-seq in sensitive and resistant cells and discover the changes in 3D chromatin organization during the development of drug resistance. 2) To interrogate the impact of alteration of chromatin structure on gene regulation by integrating gene expression (microarray) and active enhancer (marked by histone H3K27ac modification²²⁶) profiling datasets. 3) To study the temporal changes in chromatin conformation during the development of drug resistance by using in-house generated drug-tolerant cells with varying degree of drug sensitivity.

2.2 Materials and Methods

2.2.1 Cells and cell culture

NCI-H69 (HTB-119), H69AR (CRL-11351) and H69PR (CRL-11350) cells were obtained from American Type Culture Collection (ATCC). In-house drug-tolerant H69DT (“DT”, doxorubicin-tolerant) cell lines were derived by culturing NCI-H69 cells in increasing doxorubicin concentration starting from 10 nM doxorubicin concentration till 150 nM. At the same time, age-matched controls H69AMC (“AMC”, age-matched control) without doxorubicin treatment were generated for each drug-tolerant stage. All the NCI-H69 derived cell lines, except H69AR, were grown in high glucose RPMI-1640 media containing 10% (v/v) fetal bovine serum (FBS) and supplemented with penicillin (1000 U/mL) /streptomycin (1000 µg/mL) and 2 mM L-glutamine. H69AR cells were grown in the same media with 20% (v/v) FBS with occasional selection with 0.8 µM doxorubicin (Fisher Scientific Catalog # BP251610) for 3 days. All cells were cultured in an incubator with 5% CO₂ and 95% relative humidity.

2.2.2 The 3C library preparation and NGS

The 3C libraries of NCI-H69, H69AR and H69PR were prepared based on previously described method¹⁴¹, including formaldehyde crosslinking, cell membrane lysis and chromatin solubilization, restriction enzyme (*DpnII*) digestion, sticky-end ligation, reverse crosslinking, and DNA purification and desalting. The library was then sonicated, size-selected, ligated with NGS adapters and submitted for paired-end sequencing. The detailed steps will be described below. On the other hand, the 3C libraries of our in-house derived H69DT and H69AMC cells were prepared using our in-house developed FANT-C protocol (3.2.2).

(i) Formaldehyde crosslinking The cell pellets containing 2×10^7 cells were resuspended with 45 mL of fresh RPMI medium. Crosslinking was performed by quick addition of 1.25 mL of 37% formaldehyde to obtain a 1% final concentration and incubated at room temperature for exactly 10 minutes. The reaction was stopped by addition of 2.5 mL of 2.5 M glycine (to a final concentration of 125 mM) and incubated at room temperature for at least 10 minutes and later on ice for at least 15 minutes. The crosslinked cells were pelleted at 200 g for 10 minutes and stored at -80 °C freezer.

(ii) Cell membrane lysis and chromatin solubilization Frozen crosslinked cell pellets were thawed on ice and resuspended in a mixture of 2 mL of ice-cold hypotonic cell lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM sodium chloride and 0.2% (v/v) Igepal CA-630) and 20 μ L of protease inhibitor cocktail (Sigma-Aldrich, product # P8340). The cell suspension was incubated on ice for at least 15 minutes and homogenized using 20 strokes of a 2 mL loose pestle in a Dounce homogenizer. The lysates were centrifuged at 2,000 g for 5 minutes and washed twice with cold 1X *DpnII* buffer (50 mM Bis-Tris-HCl, 100 mM NaCl, 10 mM MgCl₂, 1 mM DTT, pH 6.0). Pellets were resuspended in a total 250 μ L of *DpnII* buffer and the pooled cell suspension was divided into 50 μ L aliquots. 312 μ L of *DpnII* buffer was added to each 50 μ L aliquot. To solubilize the chromatin, 38 μ L of 1% SDS (sodium dodecyl sulphate) was added to each tube, followed by incubation at 65 °C for exactly 10 minutes. SDS was quenched immediately by addition of 44 μ L of 10% Triton X-100.

(iii) Restriction enzyme digestion To digest the chromatin, 1,000 U of *DpnII* (50 U/ μ L, New England Biolabs, catalog # R0543M) was added to digest the DNA overnight at 37 °C. *DpnII* was inactivated the next day by incubating at 65 °C for 20 minutes and the digestion reaction was further processed for ligation.

(iv) Sticky-end ligation A ligation master mix containing 745 μ L of 10X T4 ligation buffer (500 mM Tris-HCl, pH 7.5, 100 mM magnesium chloride and 100 mM DTT), 80 μ L of 10 mg/mL BSA, 80 μ L of 100 mM ATP, and 6,705 μ L of water for each tube was prepared and placed on ice prior to use. After the inactivation step, the digestion products were transferred to tubes containing 7.61 mL ligation master mix. The ligation reaction was carried out by adding 50 μ L of T4 DNA ligase (1 U/ μ L, ThermoFisher Scientific catalog # 15224025) followed by incubation at room temperature for 4 hours on a HulaMixer with low rotation.

(v) Reverse crosslinking To reverse the crosslinking, 100 μ L of 10 mg/mL proteinase K (ThermoFisher Scientific catalog # 25530031) (dissolved in 1X TE Buffer, pH 8.0) was added to each tube. The tubes were incubated overnight at 65 °C.

(vi) DNA purification and desalting The ligation mixtures were pulled and aliquoted into 20-25 mL in each 50 mL falcon tube. An equal volume of 1:1 phenol-chloroform, pH 8.0, was added to each ligation mixture. The tubes were vortexed for 2 minutes and centrifuged at 3,000 g for 10 minutes at room temperature. The aqueous phase was transferred to a new 50 mL

falcon tube. Further, an equal volume of phenol-chloroform was added. The tubes were again vortexed for 2 minutes and centrifuge at 3,000 g for 10 minutes at room temperature. Each 50 mL of the aqueous phase was transferred to a 250 mL screw-cap centrifuge tube. 10% (v/v) 3 M sodium acetate, pH 5.2, and 125 mL of ice-cold absolute ethanol (2.5X of the volume of aqueous solution) were then added to each tube, and DNA was precipitated by incubating at -80 °C for at least 1 hour. The tubes were centrifuged in a high-speed centrifuge at 12,000 g for 20 minutes at 4 °C. Each pellet was then dissolved in 500 µL 1X IDTE Buffer (Integrated DNA Technologies, catalog # 97868716), pH 8.0, transferred into an Amicon Ultra 0.5 mL 30K column (Millipore, catalog # UFC903024), and centrifuged at 18,000 g for 5 minutes. The pellet was washed 4 times with 450 µL of 1X TE Buffer each time to remove any traces of salts. The column was then inverted and placed in a fresh collection tube and centrifuged at 1,000 g for 2 minutes. Another 50 µL of 1X TE Buffer was added to the column. The column was placed in the collection tube and centrifuged at 1,000 g for another 2 minutes. The DNA sample was diluted with 100 µL of 1X IDTE buffer. To degrade RNA, 1 µL of 20 mg/mL of DNase-free RNase A (Life Technologies, catalog # 12091021) was added and incubated at 37 °C for 15 minutes. The 3C library was stored at -20 °C.

(vii) Sonication 12 µg of each library were topped up to 130 µL with 1X IDTE buffer, transferred to Covaris AFA tubes (Axil Scientific, catalog # 520045) and sonicated with a Covaris M220 using optimized parameters (Intensity: 3, Cycles per burst: 1000, Duty factor: 10%, Time: 50 seconds).

(viii) Library size selection The sonicated libraries were size-selected for fragments equivalent to the size of two ligated *DpnII* restriction fragments, ~500 bp. 65 µL cold Ampure XP beads were mixed with the libraries by pipetting at least 10 times. The reaction is incubated at room temperature for 5 minutes. Tubes were placed on a magnetic stand for beads to separate. Cleared supernatant was transferred to a clean microfuge tube and a further 26 µL of cold beads were mixed. After 5 minutes incubation at room temperature, tubes were placed back onto the magnetic stand and the supernatant was discarded. To wash bound DNA, 200 µL of freshly-prepared 70% ethanol was dispensed onto the beads, and carefully removed by pipetting. The wash step was carried out once more. Beads were air-dried and 50 µL 1X IDTE buffer pH 8.0 was added when it is still glossy and dark brown. After mixing, the reaction mix was incubated for 2 minutes at room temperature. Tubes were transferred back to the magnetic stand and supernatant was pipetted into a new microfuge tube.

(ix) NGS library preparation The NGS libraries of NCI-H69, H69AR and H69PR were prepared using in-house standardized protocol. The steps included end-repair, 3' adenylation, adapter ligation and PCR amplification of adapter-ligated 3C-library. The NGS libraries of H69DT40, H69DT150, H69AMC40 and H69AMC150 were performed using NEBNext Ultra II DNA library prep kit for Illumina (catalog # E7645S) according to manufacturer's protocol. The libraries were then quantified using KAPA (catalog # KK4824) library quantification methods and purified with AMPure XP beads to remove PCR primers and adapters. The resulting pellet was then resuspended in 50 μ L of 1X TE buffer, pH 8. The supernatant was transferred into a DNA LoBind tube (Eppendorf, catalog # EPPE0030108.051) and stored at -20 °C.

(x) Paired-end sequencing Paired-end sequencing of the library on Illumina HiSeq 2500 platform with 2×101 bp multiplex protocol was performed by a third-party service provider.

2.2.3 The 3C-seq data analysis

The analysis of 3C-seq data was based on a few packages, tools and methods, especially a Python package *hiclib*²²⁷, which was designed for Hi-C data analysis but can be applied for 3C-seq data as well. Functions in *hiclib* have been integrated into scripts for many steps, including iterative mapping, fragment assignment, fragment filter, binning, bin filter, iterative bias correction and compartment analysis.

(i) Iterative mapping To obtain the genomic information of restriction fragments from both sides of ligation junction, two sides of the paired-end reads were separated and aligned to human reference genome hg19⁹⁶. The genome assembly file was obtained from UCSC genome browser (<https://genome.ucsc.edu>). The unlocalized sequences starting with "chrUn" or ending with "random" or "hap" were excluded from the analysis. The remaining sequences were used for index building with *bowtie2-build* command in bowtie2 package²²⁸. When the sequence of a read covers the 3C ligation junction, it is composed of sequences from different restriction fragments, which can result in a failure in aligning this read to reference genome. To improve the alignment efficiency of such reads, the mapping was performed iteratively as previously described²²⁷. The truncated reads were aligned over increasing truncation lengths by bowtie2 with *--very-sensitive* option until they were uniquely mapped with mapping quality scores (MAPQ) ≥ 1 or the iteration reached the end of a read. Each iteration generated one sequence

alignment/map (SAM) file²²⁹ containing the alignment results. After iterative mapping, all the SAM files were converted to BAM files by SAMtools²²⁹ to save disk space and reduce the time duration required in downstream steps. For NCI-H69 and H69AR, the BAM files from two sequencing runs were pooled respectively.

(ii) *Fragment assignment and fragment filter* The reads that were uniquely aligned with ≥ 1 MAPQ were parsed to fragment assignment and fragment filter. Reads from chromosome Y and mitochondrial chromosome were excluded from the analysis. The alignment results from both sides were paired based on their sequence IDs in the BAM files, keeping those reads which had at least one side aligned. The reads were assigned to restriction fragments according to the aligned positions. The chromosome IDs, positions, strand information, fragment IDs and restriction sites of one side or both sides of reads were stored in an HDF5 dictionary file. The fragment filter then removed reads that had 1) only one side aligned; 2) both sides aligned to same restriction fragment, including self-circles, dangling ends, and error pairs; 3) two sides too close to each other, which was possibly a result of incomplete digestion or re-ligation of neighboring restriction fragments. Reads with the same chromosome IDs and positions on both sides were considered as duplicates, which were also removed from the remaining valid pairs.

(iii) *Binning and bin filter* All the unique valid pairs were binned into equally sized non-overlapping genomic intervals (bin size = 100 kb, 200 kb, 500 kb or 1 Mb), generating a symmetric genome-wide interaction matrix. The bin filter first removed bins where more than 50% of the bases were undetermined in the reference genome (“N” in sequence, usually near centromeres and telomeres). Next, the interaction counts in each bin were summed up and bins with the lowest 1% sums (excluding zero-summed bins) were removed.

(iv) *Iterative bias correction* Different factors in experimental procedures or upstream bioinformatics processing may introduce biases to the system and thus affect the values in each genomic bin. In this analysis, the biases were corrected implicitly based on the assumption that each bin should theoretically have equal visibility in the experiment. Given the bias (B_{ij}) of interaction between two bins, namely, i and j , it is assumed to be represented as the product of the individual biases of both bins (B_i and B_j), as in Formula 1. Thus, the observed interaction frequency (O_{ij}) can be written as a product of B_i , B_j , and the true relative interaction frequency (T_{ij}) (Formula 2). To ensure a uniform coverage in total N bins throughout the genome, the bin sums in the matrix are expected to be equal (Formula 3).

$$B_{ij} = B_i B_j \text{ (Formula 1)}$$

$$O_{ij} = B_i B_j T_{ij} \text{ (Formula 2)}$$

$$\sum_j T_{ij} = \frac{\sum_{i,j} O_{ij}}{N}, \text{ for any } i \text{ in the range of 1 to } N \text{ (Formula 3)}$$

To obtain the expected matrix, bias correction was performed iteratively as previously described²²⁷. The corrected genome-wide interaction matrix was then used for downstream analysis and interpretation.

(v) Heatmap generation To compare datasets with different read coverage, the corrected interaction matrix (E_{ij}) was first normalized to a contact probability matrix (T_{ij}) where the sum of each row/column approximates to 1 (Formula 4).

$$C_{ij} = \frac{T_{ij} * N}{\sum_{i,j} T_{ij}}, \text{ where } N \text{ is the total number of bins (Formula 4)}$$

The visualization was achieved by heatmaps in log2 scale, using python package *matplotlib*²³⁰. The resolution of heatmap was dependent on the bin size of the matrix. Nearby interactions are expected to have higher frequencies than far-away interactions, typically appearing as a distinguishable diagonal in the heatmap.

(vi) Compartment analysis The corrected interaction matrix (T_{ij}) was used to identify A and B compartments. For each chromosome-wide matrix, an observed-expected ratio matrix was first generated, where the expected matrix (E_{ij}) consisted of the mean interaction frequencies at certain distances. Pearson correlation was then conducted on the observed-expected ratio matrix to compute the correlation coefficients of interactions profiles between genomic bins, resulting in a Pearson correlation matrix. Next, principal component analysis (PCA) was performed based on the assumption that the interaction matrix can be decomposed into independent principal components (eigenvectors) which represent interaction preferences of each bin¹¹⁶. These eigenvectors can be written as in Formula 5, where E_i^k and E_j^k are the eigenvector k at bin i and bin j , respectively, and λ_k is the eigenvalue of k .

$$E_{ij} = \sum_k \lambda_k E_i^k E_j^k + \text{const} \quad \text{(Formula 5)}$$

The leading eigenvector/principal component (PC1) with the highest magnitude of eigenvalues is considered as the most important interaction preference and is usually correlated with many genomic features. After PCA, the sign of PC1 in each chromosome was oriented to obtain a positive correlation (Spearman's correlation) between positive values in PC1 and GC-content.

Bins with positive and negative eigenvalues were then assigned to A and B compartments respectively. Spearman's correlation was performed between PC1 eigenvector and gene density signal (bin size = 1 Mb) generated from RefSeq database²³¹, as well as between PC1 eigenvector and H3K27ac ChIP-seq signal (bin size = 1 Mb). Mann-Whitney U test was performed for log2 expression values of genes in A and B compartments respectively. To compare compartments in different datasets, bins in A and B compartments were separately saved to two files in BED format and the comparison was then achieved by BEDTools²³² command *intersect*. Next, to examine the numbers of differentially expressed genes (DEGs) and variant H3K27ac ChIP-seq peaks in different regions, BEDTools command *intersect* was further used. Hypergeometric test was performed to compare all DEGs or variant peaks from genome-wide with those in specific regions.

(vii) TAD analysis The identification of TADs was based on insulation score analysis¹³⁸. For a corrected chromosome-wide matrix with a bin size of 100 kb, an 800 kb × 800 kb sliding window (i.e. square size = 800 kb) along the matrix diagonal was used for the calculation of the average number of interactions across each bin, which was then assigned to the central bin as the insulation score. The log2 ratio of insulation score of each bin and the mean scores of all the bins in the corresponding chromosome was calculated and used as a numerator to normalize the insulation score. A delta vector representing the difference of insulation change between left 400 kb and right 400 kb of the central bin (delta span = 800 kb) was used for the identification of valleys in insulation scores, which were usually considered as potential TAD boundaries. The boundary strength of valleys was calculated as the difference of delta vector between the left local maximum value and the right local minimum value, and those valleys with < 0.1 boundary strength were removed from the final TAD boundaries. TADs were later found by subtracting TAD boundaries from the entire genome, with the exclusion of the ones encompassing centromeres. The overlapping or unique TAD boundaries between datasets were obtained by BEDTools command *intersectBed* without or with -v option, respectively. Jaccard indices were computed using another command *jaccardBed*. To allow one-bin mismatch, command *windowBed* with -w 100000 was applied. Similarly, the unique boundaries in one dataset were found by using extra -v option, while those overlapping with boundaries in another dataset were obtained by using extra -u option. Thus, the number of overlapping boundaries when allowing one-bin mismatch were computed for each dataset separately as the number of boundaries in one dataset that overlap with the other dataset. For calculation of Jaccard index, the average number of overlapping boundaries in two datasets were taken.

2.2.4 Chromatin immunoprecipitation sequencing (ChIP-seq)

As mentioned in Chapter 1, the histone H3K27ac is a well-recognized epigenetic marker for active enhancers^{101,102}. Therefore, the H3K27ac ChIP-seq signal was used to indicate the enhancer activities throughout the genome. For NCI-H69 and H69AR, ChIP-seq experiments were performed using two biological replicates. For H69PR and the in-house generated H69AMC and H69DT cells, only one replicate was used.

(i) H3K27ac ChIP-seq The ChIP protocol was adapted from ENCODE project²³³ with slight modifications: i) the antibody and chromatin incubation was performed for 5 hours at 4 °C, and ii) after reverse crosslinking, the ChIP chromatin was purified using Amicon Ultra 30K column followed by AMPure XP beads. NGS libraries were generated from both ChIP and input DNA (1% of total) using ThruPLEX DNA-seq single index kit (Rubicon Genomics). Libraries were pooled, size selected and sequenced using Illumina HiSeq 2500 1 × 100 bp platform.

(ii) Read mapping and peak calling FASTQ files containing single-end reads were mapped to the human reference genome hg19 using bowtie2. After read filtering, the resultant BAM files (for ChIP and the corresponding input DNA) were processed using MACS2^{234,235} to identify the H3K27ac ChIP-seq peaks with *--broad* option. For NCI-H69 and H69AR, peak calling was separately performed for the two biological replicates. The overlapping peaks between replicates were taken as consensus peaks in the cell line and used in the downstream analysis.

(iii) H3K27ac ChIP-seq signal calling The BAM files generated from the above analysis were sorted and indexed by SAMtools. To generate the input subtracted signal, the bamCompare utility in deepTools²³⁶ package was used with options *--scaleFactorsMethod None*, *--operation subtract* and *--normalizeUsing RPKM*, where the BAM files for ChIP and input DNA were used as treatment and control data respectively. The signal was generated at different bin sizes and normalized to reads per kilobase per million (RPKM). For NCI-H69 and H69AR, the signal from one biological replicate was selected and used in the analysis throughout the study.

2.2.5 Microarray

(i) RNA extraction Total RNA was extracted using RNeasy Plus Mini Kit (Qiagen catalog # 74134) according to the manufacturer's instructions with modifications. Briefly, for each cell line, two biological replicates containing 4×10^6 cells were lysed in RLT plus buffer

supplemented with β -mercaptoethanol and passed through 1 mL sterile syringe 5-6 times. The lysates were passed through a gDNA eliminator column to remove genomic DNA. The flow-through was mixed with an equal amount of 100% ethanol and loaded to the RNeasy Plus Mini column. After washing with 350 μ L of RW1 buffer, on-column RNase free DNase I (ThermoFisher Scientific) treatment was performed for 15min according to manufacturer's instruction. The column was washed with 350 μ L of RW1 buffer. The subsequent column washing steps were performed according to the manufacturer's instruction manual and the RNA was eluted in 50 μ L of nuclease-free water. The RNA concentration was measured using a Qubit RNA HS assay kit (ThermoFisher Scientific).

(ii) *Microarray* Microarray was performed using Affymetrix Human Transcriptome 2.0 array containing > 6 million probes covering both coding and non-coding transcripts by a third-party service provider. Briefly, the extracted total RNA from each cell lines were subjected to quality control using Agilent Bioanalyser. RNA samples with RIN (RNA Integrity Number) of 9-10 were used for the experiments. For microarray probe preparation, 100 ng of total RNA was reverse transcribed to produce cDNA/mRNA hybrid molecule that was subsequently used as a template to make double-stranded cDNA. This double-stranded cDNA was amplified using *in vitro* transcription to make cRNA and purified. It was again subjected to the second cycle of single-stranded sense cDNA synthesis. This sense cDNA was fragmented, labeled and hybridized to Affymetrix Human Transcriptome 2.0 array for 16 hours at 45 °C with 60 rpm rotation. The arrays were washed and stained using the Affymetrix Fluidics station with FS450_0001 fluidics protocol and scanned using Affymetrix 3000 7G scanner.

(iii) *Microarray data analysis* The CEL files of NCI-H69, H69AR and H69PR were processed in a single session using Affymetrix® Expression Console™ v1.4.1.46 (EC) to obtain normalized probe intensity values. Similarly, the CEL files of H69AMC150 and H69DT150 were also processed in another session. Following Gene Level SST-RMA normalization, the resulting CHP files were then loaded into the Affymetrix® Transcriptome Analysis Console (TAC)™ Software v3.0 for differential gene expression analyses. The results from TAC analysis were exported as a tab-delimited file which was then used to create a consolidated sheet with non-redundant RefSeq gene names. In order to make the consolidated sheet, the expression values of 2 biological replicates were first combined using Tukey's biweight average. Then, we retained only those Transcript Cluster (TC) IDs that have been assigned with RefSeq IDs. Multiple TC IDs having same RefSeq ID were then averaged (Tukey's

biweight) and collapsed into a list containing the unique entry for each RefSeq ID. The RefSeq IDs were then assigned to their common gene name using RefSeq Gene tracks downloaded from UCSC table browser. Deprecated RefSeq IDs, as well as IDs from chromosome Y, HapMap and pseudogenes, were removed from the final list. The differentially expressed genes (DEGs) between NCI-H69 and H69AR as well as between H69AMC150 and H69DT150 were defined by RefSeq IDs whose log2 fold change of ≥ 2 between cell types.

2.2.6 Data availability

The H3K27ac ChIP-seq data of all the cell lines used in this study as well as the microarray data of NCI-H69, H69AR and H69PR have been submitted to GEO accession number GSE132087 (currently private).

2.3 Results

2.3.1 Genome-wide chromatin contact maps reveal widespread genetic alterations during the development of drug resistance

To investigate the changes in chromatin structures during the development of drug resistance, we have used NCI-H69 and its doxorubicin-resistant variant H69AR as a model of MDR in cancer. We have also used revertant H69PR cells which were derived from H69AR after prolonged culture without drug challenge. In order to study their chromatin conformations, we first performed 3C-seq using NCI-H69, H69AR and H69PR cells. The NCI-H69 and H69AR 3C-seq libraries were sequenced twice (i.e., two technical replicates) to achieve a total of 139 million and 199 million paired-end reads respectively. The H69PR library was later sequenced to produce 74 million read pairs. All the reads were processed through the same pipeline as described in 2.2.3, generating 29 million, 25 million and 11 million unique valid pairs in NCI-H69, H69AR and H69PR datasets respectively (**Table 2.1**).

Cell line	NCI-H69	H69AR	H69PR
Number of replicates	2	2	1
Total read pairs	139,047,479	199,423,323	73,858,624
Bases per read pair	101 X 2	101 X 2	101 X 2
Aligned pairs	98,488,665	143,360,332	54,045,310
Cis aligned pairs	89,452,757	133,582,974	50,183,313
Self-circles	12,655	13,468	7,713
Dangling ends	55,821,728	84,053,340	33,655,995
Valid pairs	25,186,918	29,366,202	11,392,624
Duplicates	350,689	467,177	153,240
Unique valid pairs	24,836,229	28,899,025	11,239,384
Cis unique valid pairs	15,914,476	19,248,035	7,093,452

Table 2.1 Mapping statistics of the 3C-seq datasets of NCI-H69, H69AR and H69PR.

We first assessed the overall chromatin structures of the three cell lines by generating genome-wide chromatin contact heatmaps from corresponding contact probability matrices (**Figure 2.2-2.4**). Consistent with previous Hi-C contact maps¹¹⁶, intra-chromosomal (*cis*) interactions are found to be much stronger than inter-chromosomal (*trans*) interactions (**Figure 2.2-2.4**), which is believed to be a feature of chromatin territories (CTs) in interphase nucleus⁶⁵. Strikingly, in all the three datasets, we observed several *trans* regions with high contact probabilities that can be visually distinguished from the background (**Figure 2.2-2.4**). These contacts are as strong as interactions between neighboring *cis* loci, suggesting the presence of translocations in these *trans* regions. All the identified translocations are listed in **Table 2.2**. While many translocations in NCI-H69 are also observed in H69AR, the latter has many additional translocations in other *trans* regions (**Figure 2.3; Table 2.2**). Some of these translocations in H69AR are carried over in H69PR, while others are not detected in the resensitized cell line (**Figure 2.4; Table 2.2**). In fact, a few new translocations are gained in H69PR (**Figure 2.4; Table 2.2**). Many of these findings are consistent with earlier karyotyping studies²³⁷⁻²³⁹, suggesting the ability of genome-wide heatmap to detect translocations. Notably, although most potential translocations that are conserved in two or three cell lines show similar patterns of interactions, in some cases we also observed changes of breakpoints. For instance, the strong interactions between chromosome 4 and 10 are found in both H69AR and H69PR but the affected region is extended to entire chromosome 4 in H69PR as compared to H69AR (**Figure 2.5**). Meanwhile, the translocated region between chromosome 3 and 4 in H69AR is not detected in H69PR (**Figure 2.5**). The complex translocations between chromosome 11, 12 and 13 in NCI-H69 are mostly retained in H69AR, except for an extra intra-chromosomal

inversion in chromosome 12 of H69AR (**Figure 2.6**). However, only one of the translocations is carried over in H69PR (**Figure 2.6**). It is important to note that we have performed 3C-seq which provides population-average data. Hence, these rearrangements may not be homogeneously present in all the cells of the population (genetic heterogeneity). All these data suggest that the parental NCI-H69 cells possibly have clonal heterogeneity and subsequently, some subclonal populations with specific genetic rearrangements get selected during drug challenge. These cells may further acquire additional drug-induced translocations over time (clonal evolution). In the absence of drug-selection pressure, subclonal populations with proliferation advantages may gradually take over the revertant cell population and many subclones present in drug resistant variant is lost over time. Nevertheless, the massive genomic rearrangements between sensitive and resistant cells suggest pervasive genome instability in these cells in response to drug challenge. Though heterogeneous in nature, these translocations have possibly provided growth advantage to the cells, which might contribute to the development of drug tolerance.

Figure 2.2 Genome-wide heatmap (bin size = 1 Mb) of NCI-H69. Chromosomes are labeled on the axes of the heatmap. Blue arrows point to the translocations observed in inter-chromosomal regions. Color bar denotes the scale of contact probability in the heatmap.

Figure 2.3 Genome-wide heatmap (bin size = 1 Mb) of H69AR. Chromosomes are labeled on the axes of the heatmap. Arrows point to the translocations that are 1) present in both NCI-H69 and H69AR (blue) and 2) additional in H69AR (red). Color bar denotes the scale of contact probability in the heatmap.

Figure 2.4 Genome-wide heatmap (bin size = 1 Mb) of H69PR. Chromosomes are labeled on the axes of the heatmap. Arrows point to the translocations that are 1) present in NCI-H69, H69AR and H69PR (blue), 2) present in H69AR and H69PR (red) and 3) additional in H69PR (green). Color bar denotes the scale of contact probability in the heatmap.

First side (chr)	Second side (chr)	NCI-H69	H69AR	H69PR
1q	20	y	y*	
3q	13q	y		
3q	20	y	y	y
5q	16p	y	y	y
11q	12	y	y	y
12	13q	y	y	
12q	13q	y	y	y*
12	19	y	y	y
12	22q	y	y	y
13q	17q	y	y	y
1p	3q		y	
1p	6q		y	y
1p	7q		y	y
1p	12		y	
1p	17		y	y
1p	20		y	
1p	21q		y	
2q	7		y	y
2p	20		y	
3	4q		y	
3p	5q		y	
3p	16p		y	
4	10		y	y*
5p	16		y	
5q	17q		y	y
6q	11q		y	
7q	Xp		y	
7q	Xq		y	
8q	12q		y	
9p	21q		y	y
9q	21q		y	
10q	18p		y	y
10q	20		y	y
13q	18q		y	
15q	Xq		y	y
1q	12q			y
1p	16			y
7q	11q			y
13q	Xq			y

Table 2.2 List of translocations observed from genome-wide heatmaps of NCI-H69, H69AR and H69PR. The presence of translocation in a cell line is marked as “y” (*: shifted breakpoints).

Figure 2.5 Heatmaps (bin size = 1 Mb) of interactions among chromosome 3, 4 and 10 in H69AR (*left*) and H69PR (*right*). Red arrows point to the location where breakpoint of the translocation changes between H69AR and H69PR. Color bars denote the scale of contact probabilities in the heatmaps.

Figure 2.6 Heatmaps (bin size = 1 Mb) of interactions among chromosome 11, 12 and 13 in NCI-H69 (*left*), H69AR (*middle*) and H69PR (*right*). Blue arrows point to the location where the translocation changes between cell lines. The black arrow in H69AR heatmap points to the intra-chromosomal inversion in H69AR. Color bars denote the scale of contact probabilities in the heatmaps.

2.3.2 Switching of A/B compartments during the development of drug resistance

In order to investigate the changes in higher-order chromatin organization, we next performed A/B compartment analysis for all the three 3C-seq datasets. For each chromosome-wide matrix with a bin size of 1 Mb, the PC1 eigenvalue at each genomic bin was computed as described in 2.2.3. The sign of the eigenvector was oriented to positively correlate with GC-content. The genome was then partitioned into two compartments which have positive and negative eigenvalues respectively. To assess whether the two compartments are associated with other

genetic and epigenetic features, we compared PC1 eigenvector with other datasets, including gene density signal computed from Refseq gene database, in-house generated H3K27ac ChIP-seq signal (active enhancer profile) and gene expression (microarray) data. In all three 3C-seq datasets, the PC1 eigenvectors are found to positively correlate with both gene density (Spearman's $\rho > 0.225$, p-value $< 10^{-31}$) and H3K27ac ChIP-seq signal (input subtracted) (Spearman's $\rho > 0.410$, p-value $< 10^{-114}$) (**Figure 2.7**). Genes in compartments with positive eigenvalues have significantly higher log2 expression values than those in the other compartments (Mann-Whitney U test p-value $< 10^{-45}$) (**Figure 2.8**). These results are consistent with previous studies showing the association between A compartment with active chromatin and B compartment with inactive/repressed chromatin¹¹⁶. We therefore denote the compartment with positive eigenvalues as A compartment and the other as B compartment.

Figure 2.7 Correlation between PC1 eigenvectors and gene density (bin size = 1Mb; *left panel*) or H3K27ac ChIP-seq (bin size = 1 Mb; *right panel*) signals in NCI-H69 (*top*), H69AR (*middle*) and H69PR (*bottom*) datasets. Spearman's correlation coefficients (ρ) and p-values are shown on the top of each panel (*: p-value < 0.05).

Figure 2.8 Box plots of log2 gene expression values in compartments with positive and negative PC1 eigenvalues generated for NCI-H69 (*left*), H69AR (*middle*) and H69PR (*right*) datasets. The median log2 gene expression values are marked as orange lines in the boxes. The statistics and p-values from Mann-Whitney U tests are shown on the top of each panel (*: p-value < 0.05).

We then sought to compare A and B compartments between datasets. Since the upstream bin filter steps remove bins with low read coverage, a few bins might be removed in one dataset and not in another. We, therefore, chose only the bins that have values in all the three datasets (common bins) for downstream compartment analysis. In NCI-H69 dataset, 1,422 out of 2,812 common bins are found in A compartment, whereas 1,389 bins are in B compartment (**Table 2.3**). Interestingly, in H69AR dataset, the number of bins in A compartment decreases to 1,329 and that in B compartment increases to 1,482 (**Table 2.3**). Analysis in H69PR dataset shows 1,378 and 1,433 bins in A and B compartments respectively (**Table 2.3**). To investigate the compartment change during the development of drug resistance, we first compared the compartments in NCI-H69 and H69AR datasets. This comparison reveals 474 bins in NCI-H69 switching from A to B compartment in H69AR (A→B) and 381 bins in NCI-H69 from B to A compartment (B→A) in H69AR (**Table 2.4**). The log2 gene expression values in A→B (NCI-H69→H69AR) regions are significantly lower in H69AR compared to NCI-H69 (Mann-Whitney U test p-value < 0.008) (**Figure 2.9**). In B→A (NCI-H69→H69AR) regions, though not statistically significant, the gene expression values show a slight increase from NCI-H69 to H69AR (Mann-Whitney U test p-value < 0.063) (**Figure 2.9**). By setting a cutoff of log2 fold change ≥ 2 , we defined 1,015 DEGs between NCI-H69 and H69AR. We then examined these DEGs in different groups of regions, including A→B, B→A regions (NCI-H69→H69AR) as well as the regions that are in the same compartments in NCI-H69 and H69AR (A→A and B→B, respectively) (**Table 2.5**). The percentages of DEGs in each of these regions (A→B, B

→A, A→A and B→B) are almost comparable (**Table 2.5**). However, the percentages of up-regulated/down-regulated DEGs vary in different regions. For example, 145/171 (84.8%) DEGs in A→B regions (NCI-H69→H69AR) are down-regulated in H69AR (hypergeometric test p-value < 10^{-15} as compared to all down-regulated DEGs genome-wide), while 93/145 (64.1%) DEGs in B→A regions (NCI-H69→H69AR) are up-regulated in H69AR (hypergeometric test p-value < 10^{-8} as compared to all up-regulated DEGs genome-wide) (**Table 2.5; Figure 2.10**). Similarly, we explored the H3K27ac ChIP-seq peaks in different regions. Total 32,628 peaks were identified as variant peaks that are not overlapping between NCI-H69 and H69AR. Out of these peaks, 18,206 (55.8%) of them exist in H69AR but not in NCI-H69 (gained peaks) and 14,422 (44.2%) of them exist in NCI-H69 but not in H69AR (lost peaks) (**Table 2.6**). In A→B regions (NCI-H69→H69AR), we found 2,700 out of 4,355 variant peaks (62.0%) are lost peaks (hypergeometric test p-value < 10^{-141} as compared to all lost peaks genome-wide) (**Table 2.6; Figure 2.11**). In contrast, in B→A regions (NCI-H69→H69AR), 4,186 out of 6,022 variant peaks (69.5%) are gained peaks (hypergeometric test p-value < 10^{-127} as compared to all gained peaks genome-wide) (**Table 2.6; Figure 2.11**). **Figure 2.12** shows an example of a B→A region (NCI-H69→H69AR) in chromosome 16. Genes in this region include *ABCC1* and *XYLT1*. The former is a known multidrug resistance gene as described in **2.1.2**. The latter encodes a xylosyltransferase enzyme known to be involved in the biosynthesis pathway of heparan sulfate²⁴⁰, a proteoglycan that contributes to chemoresistance²⁴¹. In H69AR, both *ABCC1* and *XYLT1* show enhanced H3K27ac ChIP-seq signal as well as increased gene expression levels (**Figure 2.12, 2.16**). Taken together, these results demonstrate the switching of compartments between NCI-H69 and H69AR in many regions. These regions show concomitant changes in gene expression and gain/loss of active enhancers, indicating that the compartment changing may contribute to the differential regulation of genes in response to drug challenge.

Cell line	NCI-H69	H69AR	H69PR
# of bins in A compartment	1,422	1,329	1,378
# of bins in B compartment	1,389	1,482	1,433

Table 2.3 Number of bins (bin size = 1 Mb) in A and B compartments in NCI-H69, H69AR and H69PR datasets.

NCI-H69	H69AR	H69PR
A (1,422)	A (948)	A (752)
	B (474)	B (196)
B (1,389)	A (381)	A (218)
	B (1,008)	B (163)
		A (217)
		B (791)

Table 2.4 Comparison of bins (number of bins shown in brackets; bin size = 1 Mb) in A/B compartments among NCI-H69, H69AR and H69PR datasets.

Figure 2.9 The log₂ expression values of genes located in A→B (*left*) and B→A (*right*) (NCI-H69→H69AR) regions in NCI-H69 and H69AR. Medians of log₂ gene expression values are marked as orange lines inside the boxes. The statistics and p-values from Mann-Whitney U tests are shown on the top of each panel (*: p-value < 0.05).

Region	Tested genes	DEGs	DEGs (%)	Up-regulated	Up-regulated (%)	Down-regulated	Down-regulated (%)
A→B	3,001	171	5.7	26	15.2	145	84.8
B→A	2,250	145	6.4	93	64.1	52	35.9
A→A	12,935	422	3.3	221	52.4	201	47.6
B→B	4,880	277	5.7	90	32.5	187	67.5
All	23,066	1,015	4.4	430	42.4	585	57.6

Table 2.5 Statistics of genes in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR). The percentages of DEGs in all tested genes are shown in the column “DEGs (%)”. The percentages of up- or down-regulated genes in all DEGs are shown in the columns “Up-regulated (%)” and “Down-regulated (%)” respectively.

Figure 2.10 The percentages of up- and down-regulated genes in all DEGs in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR).

Region	Variant H3K27ac peaks	Gained peaks	Gained peaks (%)	Lost peaks	Lost peaks (%)
A→B	4,355	1,655	38.0	2,700	62.0
B→A	6,022	4,186	69.5	1,836	30.5
A→A	14,666	8,875	60.5	5,791	39.5
B→B	7,585	3,490	46.0	4,095	54.0
All	32,628	18,206	55.8	14,422	44.2

Table 2.6 Statistics of variant H3K27ac peaks in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR). The percentages of gained or lost peaks in all variant peaks are shown in the columns “Gained peaks (%)” and “Lost peaks (%)” respectively.

Figure 2.11 The percentages of gained and lost peaks in all variant H3K27ac peaks in A→B, B→A, A→A and B→B regions (NCI-H69→H69AR).

Figure 2.12 Example of a B→A region on chromosome 16 from NCI-H69 (*left*) to H69AR (*right*). *Top*: Pearson correlation heatmaps of chromosome 16 along with the corresponding PC1 eigenvalues and H3K27ac ChIP-seq signal (bin size = 100 kb). Color bars denote the scale of Pearson correlation coefficient. *Bottom*: Zoomed-in view of PC1 eigenvalues and H3K27ac signal (bin size = 100 bp) in a region where two genomic bins show B→A (NCI-H69→H69AR) transition. Arrows at the bottom represent the locations and orientations of *ABCC1* and *XYLT1* in this region.

In H69PR, we found that 191 (40.3%) of all A→B (NCI-H69→H69AR) bins are switched back to A compartment (A→B→A), while 163 (42.8%) of all B→A bins (NCI-H69→H69AR) are switched back to B compartment (B→A→B) (**Table 2.4**). In A→A and B→B regions (NCI-H69→H69AR), the percentage of bins being in switched compartments in H69PR are 20.7% (A→A→B) and 21.5% (B→B→A) respectively (**Table 2.4**), which are much lower than the percentages in A→B and B→A regions (NCI-H69→H69AR). This indicates that bins in switched compartments between NCI-H69 and H69AR may have a higher chance of being switched back to the original compartments in H69PR. In the B→A region (NCI-H69→H69AR) discussed above (**Figure 2.12**), the bin containing *XYLT1* is switched back to

B compartment in H69PR, showing B→A→B transition (**Figure 2.13**). The other B→A (NCI-H69→H69AR) bin containing the gene *ABCC1* remains in A compartment in H69PR but shows a low PC1 eigenvalue (**Figure 2.13**). Concomitantly, H3K27ac ChIP-seq signal in this region and the gene expression values of these two genes both decrease in H69PR compared to H69AR (**Figure 2.13, 2.16**).

Figure 2.13 The behavior of the B→A region (NCI-H69→H69AR; **Figure 2.12**) in H69PR. *Top*: Pearson correlation heatmap of chromosome 16 along with the corresponding PC1 eigenvalues and H3K27ac ChIP-seq signal (bin size = 100 kb). Color bar denotes the scale of Pearson correlation coefficient. *Bottom*: Zoomed-in view of PC1 eigenvalues and H3K27ac signal (bin size = 100 bp) in the region containing *ABCC1* and *XYLT1*. Arrows at the bottom represent the locations and orientations of *ABCC1* and *XYLT1* in this region.

Next, we aimed to interrogate if behavior of compartment switching is affected by PC1 eigenvalues. By comparing the regions in switched and invariant compartments, we found that the magnitudes of the PC1 eigenvalues in NCI-H69 are significantly lower in A→B and B→

A regions (NCI-H69→H69AR) than invariant ($A \rightarrow A$ and $B \rightarrow B$) regions (Mann-Whitney U test p-value $< 10^{-14}$) (**Figure 2.14**). Similarly, switched compartments from H69AR to H69PR ($A \rightarrow A \rightarrow B$, $A \rightarrow B \rightarrow A$, $B \rightarrow A \rightarrow B$, $B \rightarrow B \rightarrow A$ w.r.t NCI-H69→H69AR→H69PR) also show significantly lower eigenvalue magnitudes in H69AR compared to corresponding unswitched compartments ($A \rightarrow A \rightarrow A$, $A \rightarrow B \rightarrow B$, $B \rightarrow A \rightarrow A$, $B \rightarrow B \rightarrow B$ w.r.t NCI-H69→H69AR→H69PR) (Mann-Whitney U test p-value < 0.001) (**Figure 2.14**). These results indicate that regions with low magnitudes of eigenvalues may be more vulnerable to compartment switching during cellular state transitions such as the development of resistance and resensitization to anticancer drugs.

In the previous section, we have shown that switched compartments have low eigenvalue magnitudes compared to invariant compartments. We then interrogated if the magnitude of eigenvalues remains low after compartment switching. Interestingly, we found that the switched compartments ($A \rightarrow B$ and $B \rightarrow A$ w.r.t. NCI-H69→H69AR) have low eigenvalue magnitude in H69AR compared to invariant compartments ($A \rightarrow A$ and $B \rightarrow B$ w.r.t. NCI-H69→H69AR) (Mann-Whitney U test p-value $< 10^{-6}$) (**Figure 2.15**). This could possibly explain the high proportions of $A \rightarrow B \rightarrow A$ and $B \rightarrow A \rightarrow B$ transitions (NCI-H69→H69AR→H69PR) among $A \rightarrow B$ and $B \rightarrow A$ regions (NCI-H69→H69AR).

In H69PR, after changing from H69AR, the eigenvalue magnitudes are also lower in switched (w.r.t. H69AR) compartments ($A \rightarrow A \rightarrow B$, $A \rightarrow B \rightarrow A$, $B \rightarrow A \rightarrow B$, $B \rightarrow B \rightarrow A$ w.r.t. NCI-H69→H69AR→H69PR) than unswitched (w.r.t H69AR) compartments ($A \rightarrow A \rightarrow A$, $A \rightarrow B \rightarrow B$, $B \rightarrow A \rightarrow A$, $B \rightarrow B \rightarrow B$ w.r.t. NCI-H69→H69AR→H69PR) between H69AR and H69PR (Mann-Whitney U test p-value < 0.001) (**Figure 2.15**). These results suggest that, after switching of compartments, the magnitudes of PC1 eigenvalues in compartment switching regions remain low, making them possible targets for a further compartment switch.

Figure 2.14 Box plots of PC1 eigenvalues in switched and unswitched compartments before changing of compartments. *Left:* The eigenvalues in NCI-H69 in A→B, B→A, A→A and B→B (NCI-H69→H69AR) regions. *Right:* The eigenvalues in H69AR in A→A→A, A→A→B, A→B→A, A→B→B, B→A→A, B→A→B, B→B→A and B→B→B (NCI-H69→H69AR→H69PR) regions. Medians of PC1 eigenvalues are marked as orange lines in the boxes.

Figure 2.15 Box plots of PC1 eigenvalues in switched and unswitched compartments after changing of compartments. *Left:* The eigenvalues in H69AR in A→B, B→A, A→A and B→B (NCI-H69→H69AR) regions. *Right:* The eigenvalues in H69PR in A→A→A, A→A→B, A→B→A, A→B→B, B→A→A, B→A→B, B→B→A and B→B→B (NCI-H69→H69AR→H69PR) regions. Medians of PC1 eigenvalues are marked as orange lines in the boxes.

Apart from *ABCC1* and *XYLT1* in the above example, we further checked the PC1 eigenvalues and gene expression profiles of a few other DEGs between NCI-H69 and H69AR (**Figure 2.16**). We have shown that the eigenvalue of the genomic bin containing *ABCC1* dramatically increases from NCI-H69 to H69AR (which led to a B→A transition) and decreases in H69PR (**Figure 2.12, 2.13**). *NOMO3*, another DEG located in the same bin as *ABCC1*, is also up-regulated in H69AR and down-regulated in H69PR (**Figure 2.16**). *NOMO3* encodes one of the three highly similar modulators in Nodal signaling pathway²⁴², which participates in early vertebrate development²⁴³. Consistent with the previous study²²¹, we found that *INSML* and *ASCL1*, the two neuroendocrine marker genes, are down-regulated from NCI-H69 to H69AR (**Figure 2.16**). Their corresponding PC1 eigenvalues also decrease from NCI-H69 to H69AR. Similarly, we found some other DEGs that have concomitant changes of gene expression levels and PC1 eigenvalues from NCI-H69 to H69AR (**Figure 2.16**). These DEGs include 1) *VIM*, which encodes vimentin, a type III intermediate filament protein that is usually differentially expressed in different cell types (e.g. vimentin was found in early developing muscle but not terminally differentiated muscle)²⁴⁴⁻²⁴⁶, 2) *HES1*, which encodes HES1, a basic helix-loop-helix (bHLH) transcriptional repressor that has been shown to be involved in cell proliferation and differentiation²⁴⁷⁻²⁵¹, cancer progression^{249,252,253} and multidrug resistance²⁵⁴, 3) *CDH1* (encodes cadherin 1 or epithelial cadherin) and *CDH2* (encodes cadherin 2 or neural cadherin), whose opposite behaviors have been associated with cell motility and invasiveness in cancer²⁵⁵⁻²⁵⁷, and 4) *NCAM1* (encodes Neural cell adhesion molecule), whose down-regulation has been linked to poor prognosis in certain cancer types²⁵⁸⁻²⁶¹.

However, we have noticed that the compartment change does not necessarily guaranty a concomitant gene expression change. For example, *INSML*, *ASCL1*, *VIM*, *HES1*, *CDH2*, *NCAM1* genes do not change expression between H69AR and H69PR although we observed changes in their corresponding PC1 eigenvalues (**Figure 2.16**). Therefore, though compartment switching from NCI-H69 to H69AR results in concomitant changes in gene expression levels, the same trend is not observed during compartment switching from H69AR to H69PR. This suggests that compartment switching may provide a conducive chromatin environment where gene expression can be easily regulated. However, additional factors, such as enhancer activation and transcription factor binding, may be necessary for transcriptional activation or repression. NCI-H69 to H69AR transition happens under drug-selection pressure. This triggers an adaptive evolution where additional factors may facilitate gene activation/repression upon compartment switching that allows sensitive cells to evade drug

challenge. In contrast, the transition from H69AR to H69PR occurs under no selection pressure, hence the gene regulatory factors are not positively selected.

Figure 2.16 The log2 gene expression values (*top*) and corresponding PC1 eigenvalues (*bottom*) of several DEGs in NCI-H69 (blue), H69AR (red) and H69PR (green). As defined previously, the DEGs are genes with log2 fold change ≥ 2 between NCI-H69 and H69AR.

2.3.3 Alterations in TAD structures during the development of drug resistance

To investigate the TAD organization in sensitive and resistant cells, we identified an average of 2,194 TADs with a mean size of 1.2 Mb in NCI-H69, H69AR and H69PR datasets (**Table 2.7**). Jaccard index (size of the intersection divided by size of the union) is used to quantify the similarity of TAD boundaries in different datasets. It was previously reported that TAD boundaries identified using insulation score method show average Jaccard indices of ~ 0.3 (range of 0.25-0.35) between Hi-C replicates from the same study using the same cell line²⁶². Our analysis using two technical replicates of NCI-H69 and H69AR show Jaccard indices of 0.343 and 0.329 respectively (**Table 2.8, 2.9**), consistent with previous assessment. We suspect

that the low Jaccard indices could be due to randomness or inconsistent TAD boundaries. In fact, single-cell Hi-C study showed that TADs vary considerably across the cell population compared to A/B compartments²⁶³. Therefore, we tested the Jaccard indices allowing one-bin (bin size = 100 kb) mismatch. We found a dramatic increase of boundary overlapping between technical replicates as evidenced by the increasing of Jaccard indices to 0.686 (previously 0.343) and 0.676 (previously 0.329) between NCI-H69 replicates and between H69AR replicates respectively (**Table 2.10**).

Cell line	NCI-H69	H69AR	H69PR
# of TADs	2,162	2,196	2,225
mean TAD size (bp)	1,197,828	1,181,286	1,166,564

Table 2.7 Number of TADs and the mean TAD size in NCI-H69, H69AR and H69PR datasets.

Cell line	NCI-H69		H69AR	
Replicate	Rep1	Rep2	Rep1	Rep2
# of TADs	2,191	2,144	2,194	2,212
mean TAD size (bp)	1,182,567	1,211,709	1,181,497	1,169,577

Table 2.8 Statistics of TAD calling using technical replicates of NCI-H69 and H69AR.

Dataset #1	NCI-H69 Rep1	H69AR Rep1
Dataset #2	NCI-H69 Rep2	H69AR Rep2
Common TAD boundaries in both datasets	1,133	1,122
Unique TAD boundaries in dataset #1	1,115	1,126
Unique TAD boundaries in dataset #2	1,058	1,158
Jaccard index	0.343	0.329

Table 2.9 Number of overlapping and unique TAD boundaries between technical replicates of NCI-H69 and H69AR.

Dataset #1	NCI-H69 Rep1	H69AR Rep1
Dataset #2	NCI-H69 Rep2	H69AR Rep2
TAD boundaries in dataset #1 that overlap with those in dataset #2	1,807	1,826
TAD boundaries in dataset #2 that overlap with those in dataset #1	1,806	1,828
Unique TAD boundaries in dataset #1	441	422
Unique TAD boundaries in dataset #2	385	452
Jaccard index (taking average of row 3 and 4 as common boundaries)	0.686	0.676

Table 2.10 Number of overlapping and unique TAD boundaries (allowing one-bin mismatch) between technical replicates of NCI-H69 and H69AR.

Next, we compared the TAD boundaries in the merged datasets of NCI-H69 and H69AR. Direct intersection shows 769 overlapping boundaries between the two datasets, whereas 1,446 and 1,492 boundaries are unique in NCI-H69 and H69AR respectively (**Table 2.11**). Allowing one-bin mismatch, we found 1,517 boundaries in NCI-H69 overlapping with 1,514 boundaries in H69AR, leaving 698 and 747 unique boundaries in NCI-H69 and H69AR respectively (**Table 2.12**). In both cases, the Jaccard indices (0.207 in the direct intersection; 0.512 when allowing one-bin mismatch) are much lower than those between technical replicates (**Table 2.9, 2.10**), indicating notable alterations of TAD structures from NCI-H69 to H69AR. A similar comparison was conducted for other pairs of datasets (i.e. between NCI-H69 and H69PR and between H69AR and H69PR), as well as for all the three datasets (**Table 2.11, 2.12, Figure 2.17**). Interestingly, the Jaccard indices between H69AR and H69PR (0.261 in the direct intersection; 0.581 when allowing one-bin mismatch) are the highest among all the pairs, indicating relatively higher similarity between these two cell types at TAD level. Nevertheless, the high proportion of unique boundaries in each dataset suggests that TADs are largely variant among three datasets. This suggests dynamic losing and gaining of TAD boundaries upon drug treatment or after drug removal.

Dataset #1	NCI-H69	H69AR	NCI-H69
Dataset #2	H69AR	H69PR	H69PR
Common TAD boundaries in both datasets	769	941	798
Unique TAD boundaries in dataset #1	1,446	1,320	1,417
Unique TAD boundaries in dataset #2	1,492	1,340	1,483
Jaccard index	0.207	0.261	0.216

Table 2.11 Number of overlapping and unique TAD boundaries between datasets of NCI-H69, H69AR and H69PR.

Dataset #1	NCI-H69	H69AR	NCI-H69
Dataset #2	H69AR	H69PR	H69PR
TAD boundaries in dataset #1 that overlap with those in dataset #2	1,517	1,669	1,500
TAD boundaries in dataset #2 that overlap with those in dataset #1	1,514	1,669	1,506
Unique TAD boundaries in dataset #1	698	592	715
Unique TAD boundaries in dataset #2	747	612	775
Jaccard index (taking average of row 3 and 4 as common boundaries)	0.512	0.581	0.502

Table 2.12 Number of overlapping and unique TAD boundaries (allowing one-bin mismatch) between datasets of NCI-H69, H69AR and H69PR.

Figure 2.17 Venn diagram showing the numbers of overlapping and unique TAD boundaries in the three datasets.

Consequently, alterations of TAD structures are found at regions with these unique TAD boundaries. For example, **Figure 2.18** shows a region on chromosome 5 where a TAD boundary change can be observed between NCI-H69 and H69AR. Interestingly in H69PR, the TAD boundaries in the same region show resemblance to both NCI-H69 (between TAD No. 1 and 2 of H69PR) and H69AR (between TAD No. 2 and 3 of H69PR) (**Figure 2.18**). Accordingly, we observed changes in chromatin interaction patterns as well as redistribution of H3K27ac ChIP-seq signal in these cells (**Figure 2.18**). Compared to NCI-H69, the H3K27ac signal greatly increases in TAD No. 2 of H69AR, suggesting a gain of active enhancer loci inside this TAD. On the contrary, the enhancer loci in the neighboring regions (TAD No. 1 and 3 of H69AR) are nearly lost in H69AR. In H69PR, the H3K27ac signal largely decreases at the active enhancer loci observed in H69AR, leading to a status that is similar to NCI-H69 (TAD No. 3 of H69PR). Meanwhile, a subtle increase of the H3K27ac signal is observed in the neighboring regions (TAD No. 2 and 4 of H69PR). In agreement to these changes, the two genes in this region, *GRIA1* and *GALNT10*, are down- and up-regulated respectively in H69AR as compared to NCI-H69, which are then regulated oppositely in H69PR (**Figure 2.18**). *GRIA1* encodes a glutamate receptor that can form ligand-gated ion channel with other glutamate receptors, which acts as neurotransmitter receptors in the mammalian brain and is associated with normal neurophysiological processes. The other gene, *GALNT10*, encodes a member of GalNAc polypeptide N-acetylgalactosaminyltransferases that catalyses the synthesis of mucin-type oligosaccharides. The differential gene regulation in this region is possibly achieved by remodeling of enhancer-gene connectivity through changing of TAD or sub-TAD structures in

response to drug challenge. The TAD analysis was performed based on 100-kb binning of 3C-seq data. A higher-resolution contact map can provide better insights into such enhancer-gene interactions.

Figure 2.18 Example of a TAD changing region (chr5:150,800,000-154,900,000; bin size = 100 kb) in NCI-H69 (*top left*), H69AR (*top right*) and H69PR (*bottom left*). Heatmaps of the region are shown along with TADs (numbered from upstream to downstream), H3K27ac ChIP-seq signals (bin size = 1 kb) as well as the locations and orientations of the two genes in this

region (*GRIA1* and *GALNT10*; arrows). Color bars denote scale of contact probabilities in the heatmaps. *Bottom right*: The log₂ gene expression values of *GRIA1* and *GALNT10* in NCI-H69 (blue), H69AR (red) and H69PR (green).

Each TAD boundary is assigned a boundary strength score representing the difference of signal intensity between local maximum (TAD center) and local minimum (TAD boundary) (2.2.3). We therefore hypothesized that boundary strength could be a factor that affects boundary stability. Interestingly, we found that the overlapping boundaries in all three datasets have significantly higher boundary strengths than those existing in two datasets (Mann-Whitney U test p-value < 10⁻⁶) (**Figure 2.19**). Similarly, the boundary strengths of the latter are significantly higher than those of unique boundaries in each dataset (Mann-Whitney U test p-value < 10⁻⁶) (**Figure 2.19**). These results hint that the TADs with low boundary strengths could possibly be more vulnerable to undergo topological changes during transitions of drug response states.

Figure 2.19 Box plot showing the boundary strengths of TAD boundaries that exist in all the three datasets (NCI-H69, H69AR and H69PR), in two datasets (NCI-H69 and H69AR; H69AR and H69PR; NCI-H69 and H69PR) or uniquely in one dataset (NCI-H69; H69AR; H69PR). Median boundary strengths are marked as orange lines in the boxes.

2.3.4 Chromatin organization changes in the in-house generated drug-tolerant cells

The 3C-seq datasets of NCI-H69, H69AR and H69PR reveal alterations in chromatin organization at various levels between drug-sensitive and resistant cell lines, which are further found to be associated with differential gene expression and variant enhancer loci. However, these cell lines were derived from extreme ends of the treatment (i.e. drug exposure or drug holiday) and they have distinct drug response behaviors (2.1.3; **Figure 2.1**). In addition, the long-time cell culture to generate drug-resistant and resensitized cell lines may be responsible for the chromatin topology changes observed in these cell lines. In the absence of proper age-matched controls, we cannot interrogate this notion. Therefore, to study the temporal changes of chromatin organization during the development of drug resistance, we cultured NCI-H69 cells in increasing doses of doxorubicin. The doxorubicin-tolerant cells (H69DT) of varying drug sensitivities (40 nM and 150 nM) were collected in-house at different time points (**Figure 2.20**). Simultaneously, their age-match controls (H69AMC) were also collected at the corresponding time points by culturing cells without doxorubicin treatment (**Figure 2.20**). H69DT cells selected at the drug concentration of 150 nM (H69DT150) show increased tolerance to doxorubicin compared to all other cells (H69DT40 and age-matched controls, H69AMC40 and H69AMC150 cells) (**Figure 2.21**). 3C libraries were then generated from H69DT40, H69DT150, H69AMC40 and H69AMC150, and deep-sequenced. Total 50-82 million paired-end reads were generated from NGS and processed through the data analysis pipeline (**Table 2.13**).

Figure 2.20 Diagram showing the in-house derived tolerant and age-matched control cell lines.

Figure 2.21 Cell viability curves (CellTiter-Glo assay after 5 days of drug treatment) of in-house generated H69DT and H69AMC cells. The EC₅₀ values of the cell lines are shown on the top right.

Cell line	H69AMC40	H69AMC150	H69DT40	H69DT150
Number of replicates	1	1	1	1
Total read pairs	51,175,368	49,673,084	64,673,402	81,527,643
Bases per read pair	101 X 2	101 X 2	101 X 2	101 X 2
Aligned pairs	34,834,302	34,074,144	44,468,200	56,749,697
Cis aligned pairs	32,056,369	32,425,774	41,030,136	53,520,770
Self-circles	9,100	4,587	11,478	10,461
Dangling ends	18,328,150	19,042,178	22,046,876	34,135,134
Valid pairs	9,890,631	7,589,035	13,698,407	11,755,120
Duplicates	85,587	72,831	105,801	121,129
Unique valid pairs	9,805,044	7,516,204	13,592,606	11,633,991
Cis unique valid pairs	7,053,475	5,889,441	10,182,351	8,448,381

Table 2.13 Mapping statistics of H69AMC40, H69AMC150, H69DT40 and H69DT150 datasets.

The genome-wide heatmaps show overall similar chromatin structures in all the four cell types (**Figure 2.22-2.25**). Derived from NCI-H69, as expected, H69AMC and H69DT cells retain most of the translocations in NCI-H69 (**Figure 2.22-2.25; Table 2.14**). However, the translocation between chromosome 3q and 13q is only detected in the two H69DT cell lines but not in the two H69AMC cell lines (**Figure 2.22-2.25; Table 2.14**). This further evidences that NCI-H69 is composed of different clonal populations with heterogeneity in chromatin organization. Apart from existing translocations, we observed one extra translocation between chromosome 2 and 3 in H69DT150 but not in any other cell types (**Figure 2.22-2.25; Table 2.14**). Interestingly, this translocation is also absent in NCI-H69, H69AR and H69PR (**Figure 2.2-2.4; Table 2.2**). This suggests the rapid karyotype evolution of cancer cells upon drug treatment.

Figure 2.22 Genome-wide heatmap (bin size = 1 Mb) of H69AMC40. Chromosomes are labeled on the axes of the heatmap. Blue arrows point to the translocations found in H69AMC40, all of which are also present in NCI-H69 (**Figure 2.2**). Color bar denotes the scale of contact probability in the heatmap.

Figure 2.23 Genome-wide heatmap (bin size = 1 Mb) of H69AMC150. Chromosomes are labeled on the axes of the heatmap. Blue arrows point to the translocations found in H69AMC150, all of which are also present in NCI-H69 and H69AMC40. Color bar denotes the scale of contact probability in the heatmap.

Figure 2.24 Genome-wide heatmap (bin size = 1 Mb) of H69DT40. Chromosomes are labeled on the axes of the heatmap. Blue arrows point to the translocations found in H69DT40, all of which are also present in NCI-H69. Color bar denotes the scale of contact probability in the heatmap.

Figure 2.25 Genome-wide heatmap (bin size = 1 Mb) of H69DT150. Chromosomes are labeled on the axes of the heatmap. Arrows point to the translocations that are 1) present in NCI-H69, H69DT40 and H69DT150 (Blue) and 2) additional in H69DT150 (Red). Color bar denotes the scale of contact probability in the heatmap.

First side (chr)	Second side (chr)	NCI-H69	H69AMC40	H69AMC150	H69DT40	H69DT150
1q	20	y	y	y	y	y
3q	13q	y			y	y
3q	20	y	y	y	y	y
5q	16p	y	y	y	y	y
11q	12	y	y	y	y	y
12	13q	y	y	y	y	y
12q	13q	y	y	y	y	y
12	19	y	y	y	y	y
12	22q	y	y	y	y	y
13q	17q	y	y	y	y	y
2	3					y

Table 2.14 List of translocations observed from genome-wide heatmaps of NCI-H69, H69AMC40 and H69AMC150, H69DT40 and H69DT150. The presence of translocation in a cell line is marked as “y”.

We then performed A/B compartment analysis for all the four datasets. Common bins that exist in four datasets, as well as in NCI-H69 dataset, are chosen. Compared to NCI-H69, the number of bins in the A compartment increases from 1,423 to 1,459 in H69DT40, and then dramatically decreases to 1,404 in H69DT150 (**Table 2.15**). Accordingly, the number of bins in the B compartment shows the opposite trend in H69DT cells, from the parental NCI-H69 to H69DT40, and from H69DT40 to H69DT150. In contrast, the control cells, H69AMC40 and H69AMC150, retain similar numbers of bins in A and B compartments (**Table 2.15**). Analysis of the compartment switching reveal that 265/1,423 (18.6%) bins in A compartment in NCI-H69 are switched to B compartment in H69DT40 (A→B), while 301/1,391 (21.6%) bins in B compartment in NCI-H69 show B to A switch in H69DT40 (B→A) (**Table 2.16**). Similar to three model cell lines, the percentage of bins showing compartment switching from H69DT40 to H69DT150 is much higher in A→B (36.2%) and B→A (46.2%) regions (NCI-H69→H69DT40) compared to invariant A→A (11.1%) and B→B (10.7%) regions (NCI-H69→H69DT40) (**Table 2.16**). Although the total number of bins in A and B compartments remain similar between H69AMC cell lines and NCI-H69, we found 23.1% of all A compartment bins in NCI-H69 showing A→B transitions in H69AMC40 and 24.2% of all B compartment bins in NCI-H69 showing B→A transitions in H69AMC40 (**Table 2.17**). The percentages of switched compartments from H69AMC40 to H69AMC150 are also higher in A→B (53.9%) and B→A (59.2%) regions (NCI-H69→H69AMC40) than the invariant A→A (13.6%) and B→B (16.8%) regions (NCI-H69→H69AMC40) (**Table 2.17**). These results demonstrate dynamic compartment switching due to long-term cell culture. Moreover, the percentages of bins in A and B compartments in H69AMC40 and H69AMC150 are similar to NCI-H69. On the other hand, H69DT150 shows the highest percentage of bins in B compartment among all the cell types.

Cell line	NCI-H69	H69AMC40	H69AMC150	H69DT40	H69DT150
# of bins in A compartment	1,423	1,429	1,436	1,459	1,404
# of bins in B compartment	1,391	1,385	1,378	1,355	1,410

Table 2.15 Number of bins in A and B compartments in NCI-H69, H69AMC40, H69AMC150, H69DT40 and H69DT150 datasets.

NCI-H69	H69DT40	H69DT150
A (1,423)	A (1,158)	A (1,029)
		B (129)
	B (265)	A (96)
		B (169)
B (1,391)	A (301)	A (162)
		B (139)
	B (1,090)	A (117)
		B (973)

Table 2.16 Comparison of bins (number of bins shown in brackets; bin size = 1 Mb) in A/B compartments among NCI-H69, H69DT40 and H69DT150 datasets.

NCI-H69	H69AMC40	H69AMC150
A (1,423)	A (1,093)	A (944)
		B (149)
	B (330)	A (178)
		B (152)
B (1,391)	A (336)	A (137)
		B (199)
	B (1,055)	A (177)
		B (878)

Table 2.17 Comparison of bins (number of bins shown in brackets; bin size = 1 Mb) in A/B compartments among NCI-H69, H69AMC40 and H69AMC150 datasets.

To investigate the effects of drug treatment on drug-tolerant cells, we directly compared the H69DT cell lines to their corresponding age-matched controls. Among all the 1,429 A compartment bins in H69AMC40, 294 (20.6%) of them are in B compartment in H69DT40 (A→B) (**Table 2.18**). On the other hand, 324 (23.4%) of 1,385 B compartment bins in H69AMC40 are in A compartment in H69DT40 (B→A) (**Table 2.18**). These percentages are similar to A→B (23.7%) and B→A (22.4%) bin transitions from H69AMC150 to H69DT150 (**Table 2.19**). When we compared the behaviour of compartment switching in drug-tolerant H69DT150 with the established drug-resistant H69AR, we found that drug resistance phenotype is associated with an expansion of repressive chromatin as evidenced by large number of bins showing A→B transition (sensitive→resistant).

H69AMC40	H69DT40
A (1,429)	A (1,135)
	B (294)
B (1,385)	A (324)
	B (1,061)

Table 2.18 Comparison of bins (number of bins shown in brackets; bin size = 1 Mb) in A/B compartments between H69AMC40 and H69DT40 datasets.

H69AMC150	H69DT150
A (1,436)	A (1,095)
	B (341)
B (1,378)	A (309)
	B (1,069)

Table 2.19 Comparison of bins (number of bins shown in brackets; bin size = 1 Mb) in A/B compartments between H69AMC150 and H69DT150 datasets.

Next, we investigated if compartment switching in in-house derived drug-tolerant cells is associated with differential gene expression. We performed microarray analyses of H69AMC150 and H69DT150 cells and found 76 DEGs ($\log_2$ fold change ≥ 2) between these two cell lines. We observed time-resolved compartment switching in response to increasing concentration of drug leading to gene transcription. For example, *ABCC1* locus resides in B compartment in H69AMC40 and H69AMC150 cells which then loses strength in H69DT40 and is ultimately switched to A compartment in H69DT150 (**Figure 2.26**). The H3K27ac ChIP-seq signal and gene expression level of *ABCC1* show concomitant changes in these cells (**Figure 2.26, 2.27**). Taken together, these results associate compartment switching to changes in H3K27ac ChIP-seq signal and gene expression between H69AMC and H69DT cells. In contrast, the neighboring *XYLT1* locus remains in B compartment throughout in all these H69AMC and H69DT cells with little change in H3K27ac ChIP-seq signal and gene expression (**Figure 2.26, 2.27**). This suggests that the *ABCC1* activation is an early event in the development of drug tolerance.

Figure 2.26 The behavior of the region containing *ABCC1* in H69AMC40 (top left), H69DT40 (top right), H69AMC150 (bottom left) and H69DT150 (bottom right). The Pearson correlation heatmaps of chromosome 16 are shown along with the corresponding PC1 eigenvalues and H3K27ac ChIP-seq signal (bin size = 100 kb). Color bars denote the scale of Pearson correlation coefficient. Red boxes show zoomed-in views of PC1 eigenvalues and H3K27ac signal (bin size = 100 bp) in the region containing *ABCC1* and *XYLT1*. Arrows at the bottom represent the locations and orientations of *ABCC1* and *XYLT1* in this region.

Figure 2.27 The log2 gene expression values of *ABCC1* (left) and *XYLT1* (right) in H69AMC150 and H69DT150 cells.

Next, an average number of 2,194 TADs with a mean size of 1.2 Mb were identified in H69AMC and H69DT datasets (**Table 2.20**). Intersections of TAD boundaries show Jaccard indices larger than 0.3 between all pairs of datasets (**Table 2.21**), which are comparable to, if not larger than, those between replicates (**Table 2.9**). Although the similarities of TAD boundaries in these datasets are relatively high, boundary changing is still observed at some loci. For example, *ABCB1*, one of the DEGs between H69AMC150 and H69DT150, resides in a TAD changing region in these in-house derived cell lines (**Figure 2.28**). In both H69AMC40 and H69AMC150, this region in chromosome 7 comprises three TADs separated by two TAD boundaries, where *ABCB1* is located inside the TAD No. 2. However, the boundary between TAD No. 2 and 3 is absent in both H69DT40 and H69DT150, resulting in “fusion” of these two adjacent TADs. Interestingly, the heatmaps of H69AMC40, H69AMC150 as well as H69DT40 show similar interaction patterns (**Figure 2.28**). In H69DT40, though the above-mentioned TAD boundary is absent, the heatmap retains a clear boundary-like pattern at the same locus as the TAD boundary in H69AMC cells (between TAD No. 2 and 3 of H69AMC cells), suggesting the presence of sub-TAD boundary at this locus in this cell line. In H69DT150, however, we observed a subtle decay of the boundary-like pattern at this locus. On the other hand, the boundary between TAD No. 1 and 2 of H69DT150 is shifted to the bin encompassing the *ABCB1* locus (**Figure 2.28**). Although the overall distribution of H3K27ac markers remains similar in all the four cell lines, we found an increase in the signal at the TAD boundary in H69DT150 where *ABCB1* is also located (between TAD No. 1 and 2 of H69DT150) (**Figure 2.28**). Meanwhile, the expression of *ABCB1* is dramatically up-regulated in H69DT150 compared to H69AMC150 (**Figure 2.29**). These results hint a possible mechanism of gene regulation at *ABCB1* locus in H69AMC150: bringing active enhancer loci and *ABCB1* in contact by removing a TAD/sub-TAD boundary.

Cell line	H69AMC40	H69AMC150	H69DT40	H69DT150
# of TADs	2,195	2,175	2,211	2,196
mean TAD size (bp)	1,180,822	1,190,760	1,171,282	1,176,322

Table 2.20 Number of TADs and the mean TAD size in H69AMC40, H69AMC150, H69DT40 and H69DT150 datasets.

Dataset #1	NCI-H69	NCI-H69	NCI-H69	NCI-H69	H69DT40	H69AMC40
Dataset #2	H69DT40	H69AMC40	H69DT150	H69AMC150	H69DT150	H69AMC150
Common TAD boundaries in both datasets	1,225	1,136	1,123	1,064	1,213	1,042
Unique TAD boundaries in dataset #1	990	1,079	1,092	1,151	1,056	1,216
Unique TAD boundaries in dataset #2	1,044	1,122	1,137	1,184	1,047	1,206
Jaccard index	0.376	0.340	0.335	0.313	0.366	0.301

Table 2.21 Number of overlapping and unique TAD boundaries between datasets of in-house derived cell lines.

Figure 2.28 Example of a TAD changing region (chr7:85,700,000-89,900,000) in H69AMC40 (top left), H69DT40 (top right), H69AMC150 (bottom left) and H69DT150 (bottom right).

Heatmaps of the region are shown along with TADs (numbered from upstream to downstream), H3K27ac ChIP-seq signals (bin size = 1 kb) as well as the locations and orientations of the gene *ABCB1* in this region (arrows). Color bars denote scale of contact probabilities in the heatmaps.

Figure 2.29 The log2 gene expression values of *ABCB1* in H69AMC150 and H69DT150 cells.

The above examples demonstrate the changes in chromatin conformation and gene expression levels of *ABCC1* and *ABCB1*, the two multidrug resistance genes, between drug-tolerant and control cells (**Figure 2.26-2.29**). As described in 2.1.2, both *ABCC1/MRP1* and *ABCB1/MDR1* encode ABC transmembrane transporters that can act as efflux pumps to transport drugs from inside the cell to extracellular space^{140,193}. While *ABCB1* is known to be lowly-expressed in both NCI-H69 and H69AR, *ABCC1* is the most differentially regulated gene between NCI-H69 and H69AR¹⁹⁴. These are validated by our gene expression results of *ABCB1* (data not shown) and *ABCC1* (**Figure 2.16**) in NCI-H69 and H69AR cells. As expected, *ABCC1* also shows an increased expression level in H69DT150 compared to H69AMC150 (**Figure 2.27**). However, we found that *ABCB1* and not *ABCC1* is the most differentially expressed gene between H69AMC150 and H69DT150. Read depth analysis using the input control data of H3K27ac ChIP-seq experiments shows copy number amplification of *ABCC1* region in the resistant H69AR cells and only in H69DT150 (data not shown). This may explain the gene expression change of *ABCC1* in H69DT150. However, *ABCB1* region remains copy number neutral in all cells, suggesting the involvement of enhancer activation or transcriptional activator in its higher expression in H69DT150 cells. These results indicate that ABC transporters are frequently up-regulated by NCI-H69 cells in response to doxorubicin treatment. However, the up-regulated members in ABC transporter family cells may be determined based on random selection.

2.4 Discussion

Drug resistance is one of the major challenges in cancer treatment. However, the underlying gene regulatory mechanisms that control transcriptional reprogramming remain unclear. To understand the role of chromatin organization in drug resistance, we first performed 3C-seq experiments for NCI-H69, H69AR and H69PR cells. Pervasive genomic rearrangements are detected in the 3D contact heatmap of these three cell lines, especially in H69AR (**Figure 2.2-2.4; Table 2.2**). Next, we observed A/B compartment switching as well as TAD alterations between datasets, which in many cases, are accompanied by concomitant changes in gene expression and H3K27ac ChIP-seq signal. These results support our hypothesis that the chromatin interaction network changes during the development of drug resistance. Moreover, they also suggest the contribution of three-dimensional chromatin organization in enhancer-driven gene regulation during changes in drug-response states. Furthermore, we found the magnitudes of PC1 eigenvalues are significantly lower in switched compartments than invariant compartments (**Figure 2.14, 2.15**). Similarly, the boundary strengths of unique TAD boundaries in each dataset are significantly lower than overlapping boundaries between datasets (**Figure 2.19**). These results hint that compartments with low eigenvalue magnitudes and TADs with low boundary strengths may be more vulnerable to organizational changes under environmental pressure.

To investigate the time-resolved changes in chromatin organization during the development of drug resistance, we generated in-house H69AMC and H69DT cell lines. Genome-wide chromatin contact maps show one translocation that only exists in H69DT150 and not in any other cell types in this study (**Figure 2.22-2.25; Table 2.14**), suggesting karyotype evolution in response to drug challenge. The A/B compartment switching is found to be dynamic during cell culture, as suggested by the comparison between NCI-H69, H69AMC40 and H69AMC150 datasets (**Table 2.15, 2.17**). We observed high A→B compartment transitions in both H69AR and H69DT150 with respect to their sensitive counterparts, indicating the expansion of repressive chromatin may be possible driving events for the development of drug resistance. At TAD level, we also detected changes of boundaries in H69AMC and H69DT cells (**Table 2.21, Figure 2.28**). These findings suggest that changes in higher-order chromatin structure may be recurrent features in cancer cells like NCI-H69 upon drug treatment. Moreover, we found that *ABCB1* is the leading DEG between H69AMC150 and H69DT150. However, it is

lowly expressed in both NCI-H69 and H69AR cells. This suggests that the cells can regulate different sets of genes to develop drug tolerance when exposed to the same drug challenge.

In summary, using the available SCLC drug resistance cell line model and in-house generated drug-tolerant cell lines, we found changes in chromatin structures at different levels between sensitive and resistant cells, which may contribute to enhancer-mediated activation of drug-resistant genes. It may shed light on the future development of therapeutic strategies against epigenetic regulators involved in drug resistance. However, observations from cell line models may not be applicable for *in vivo* cancers. Besides, due to the high heterogeneity of cancer cells, each cancer may express a different array of drug-resistant genes through different epigenetic regulatory mechanisms. 3C-seq in patient samples is thus needed to provide personalized interaction profile, which can possibly lead to an era of personalized medicine.

Work contribution

This work is carried out exclusively in the Sanyal laboratory at Nanyang Technological University, Singapore. Under the guidance of Assistant Prof. Amartya Sanyal, Siti Rawaidah Binte Mohammad Muzaki prepared the 3C libraries of NCI-H69, H69AR and H69PR cells, and I prepared the 3C libraries of H69DT40, H69DT150, H69AMC40 and H69AMC150 cells. I have also performed all the 3C-seq data analysis presented in this chapter. In this study, I have conducted all the integrative analyses using 3C-seq, histone H3K27ac ChIP-seq and gene expression microarray datasets. In house generation of drug-tolerant H69DT and age-matched control H69AMC cell lines, as well as CellTier-Glo assay for all the cell lines presented in this chapter, were carried out by Dr. Aravinda Chavalmane Krishnappa Naik. Dr. Aravinda and also prepared the histone H3K27ac ChIP-seq libraries for all the cell lines and extracted the total RNA samples of NCI-H69, H69AR and H69PR cells for microarray. I performed the total RNA extraction of H69DT and H69AMC cells under the guidance of Dr. Aravinda. The bioinformatics analysis of H3K27ac ChIP-seq data, including read mapping and peak calling, was carried out by Costerwell Khyriem. I generated all the input subtracted H3K27ac ChIP-seq signals used in this study. The analysis of microarray data for NCI-H69, H69AR and H69PR was performed by Costerwell Khyriem and that for H69DT and H69AMC cell lines was performed by me.

Chapter 3 FANT-C: A rapid and low-input method for genome-wide chromatin interaction study

3.1 Introduction

As discussed in Chapter 1, the higher-order chromatin organization in interphase nucleus is increasingly recognized as a key epigenetic regulator of gene expression. Alterations in chromatin structures have been associated with transcriptional reprogramming that leads to cellular state transition^{123,126-128}. The development of 3C¹⁰⁹ and its derived techniques, especially the genome-wide variant Hi-C¹¹⁶, have enabled molecular-level characterization of three-dimensional chromatin organization (**Figure 1.3**). Therefore, it is highly-anticipated to include Hi-C like technique in clinical diagnosis as it can provide information on the chromatin basis of the gene-regulatory mechanisms during disease progression. However, the complex multi-step experimental procedure and the elaborate timeline to produce the Hi-C library make it less adaptive to clinical settings in a diagnostic timescale.

In a conventional 3C experiment¹⁰⁹ (**Figure 1.3**), chromatin inside the cell nucleus is first treated with formaldehyde to covalently crosslink the physically-interacting genomic regions through DNA-bound proteins or other nuclear proteins like coactivators and corepressors. The degree of crosslinking is based on spatial proximity, suggesting a higher frequency of crosslinking between spatially-proximal regions than between spatially-distant regions. To quantify the frequency of crosslinking, a restriction enzyme that can produce single-stranded DNA overhangs is applied to digest the crosslinked chromatin into restriction fragments. While still crosslinked, these fragments are ligated by the enzyme T4 DNA ligase in dilute condition, to favor the intramolecular ligation over intermolecular ligation. The restriction fragments that are far apart in linear genome sequence but physically interacting inside the nucleus can thus be ligated together to form a single contiguous DNA molecule. These ligation products that contain information about chromatin interactions between genomic regions compose the so-called 3C library. To obtain the 3C library DNA, the reverse crosslinking is performed by treating the ligated chromatin with Proteinase K and heat, and DNA is purified using phenol-chloroform extraction method followed by RNase treatment. In conventional 3C, PCR is used for the pairwise detection of individual ligation product. The quantity of the PCR-amplified ligation products between two restriction fragments represents the frequency at which they interacted with each other inside the nucleus.

To map the genome-wide “all-to-all” chromatin interactions, the ideal and straight-forward method is to first generate short inserts of conventional 3C library by sonication and then detect all ligation products using paired-end NGS. If sequenced comprehensively and thoroughly, these ligation products can theoretically provide information about all chromatin interactions inside the cell population. However, given the fact that commonly-used restriction enzymes usually recognize 6 or 4 bases of DNA sequences, the total numbers of corresponding restriction fragments generated in the human genome are estimated to be more than millions, if not more. Detecting all possible pairwise combinations of these fragments thus requires trillions of sequencing reads to achieve a 1X coverage. Although binning of the genome into intervals could largely reduce the required number of reads by aggregating data points, the sequencing depth is still a major limiting factor for high-resolution study. Furthermore, the 3C library contains a large proportion of short inserts of genomic sequences following sonication, which are non-informative regarding chromatin interactions. Therefore, the 3C library needs to be deep sequenced in order to obtain sufficient reads for constructing a high-resolution contact map.

Hi-C technique, introduced in 2009 when NGS was at its infancy, provides an approach to circumvent the problem of undigested restriction fragments in 3C library by enriching the DNA sequences with ligation junctions¹¹⁶ (**Figure 1.3**). In Hi-C, after restriction enzyme digestion, the sticky ends are filled up with dNTPs, among which dCTPs are labeled with biotin. This step is followed by blunt-end ligation, reverse crosslinking and DNA purification. The biotin labels at unligated ends of restriction fragments are then removed by T4 DNA polymerase with its exonuclease activity. Next, the DNA is sheared, and the ligation products with biotin labels are pulled down (precipitated) by streptavidin beads via specific binding to biotin. These steps allow the enrichment of digested and ligated fragments in the Hi-C library. After NGS, the high percentage of informative reads can thus be used to generate the chromatin contact map at a relatively high resolution.

In recent times, the boost in data output and cheaper cost of NGS runs allow one to revisit the original idea to directly sequence the 3C library. In fact, the 3C-seq technique which combines 3C library preparation with NGS has been successfully performed in several studies from flies to mammals^{118,126,138}. Meanwhile, the rapid evolution and widespread use of whole-genome sequencing (WGS) in clinical laboratories have allowed incredible progress in the diagnosis of genetic alterations of several inherited disorders and cancer. In the clinical setting, the most

important factor for developing a diagnostic tool is the ease of experimental protocol with a small input sample in a limited time-scale. However, both 3C-seq and Hi-C protocols incorporate most steps of conventional 3C library preparation. These multiple complex procedures lead to 1) lengthy processing time and high cost in the library preparation, 2) loss of sample which affects the final library yield and, 3) introduction of biases to the library composition. These factors largely restrict the applicability and versatility of both 3C-seq and Hi-C, especially in low cell number scenarios such as clinical samples, tissue biopsies and rare cells. Moreover, in the clinical scenario, minimal invasive sample collection becomes the prerogative of the patient. In order to introduce a new method in regular genetic screening, it is necessary to adapt this new method to the current diagnostic test without sacrificing patient comfort. Since WGS is widely used for the discovery of genetic abnormalities in clinical diagnosis nowadays, the ideal situation of incorporating Hi-C like method is to simultaneously obtain information about chromatin interactions and genetic abnormalities without extra sample collection. Thus, in cases where WGS is recommended for diagnostics, such Hi-C like method could capture additional information from the same NGS run using the same collection procedure and amount of input cells.

Here, we introduce FANT-C (Fast 'N' Tiny sample-C), a rapid (10-hour) and low-input (as low as 50,000 cells) 3C library preparation protocol which can be combined with high-throughput sequencing to construct the genome-wide chromatin contact map. With reactions conducted within a microcentrifuge tube, FANT-C maximizes the DNA recovery from a small sample with minimal cost. In addition to chromatin interaction data, FANT-C library composes numerous reads from genomic DNA, which can provide information about genetic abnormalities. We have benchmarked our FANT-C results against the publicly-available chromatin interaction data generated using conventional Hi-C and *in situ* Hi-C protocols. We found that FANT-C can recapitulate the A/B compartments and topologically associating domain (TAD) profiles of K562 cells. Next, we applied FANT-C to bone marrow samples from two AML patients with different FAB subtypes (M5a and M4Eo) and discovered variations in chromatin organization before and after chemotherapy treatment. Using genomic DNA reads in FANT-C libraries, inversion of chromosome 16 region and chromosome 22 trisomy are found in AML M4Eo (M4 with bone marrow eosinophilia) patient. After successful implementation of FANT-C protocol using clinical samples, we envisage FANT-C has the potential to become a powerful diagnostic protocol for capturing genome-scale information

about chromatin connectivity and genetic alterations in a standard biopsy sample within a clinical timeline.

3.2 Materials and Methods

3.2.1 Cells and cell culture

Bone marrow blast cells from newly diagnosed AML patients were obtained from National University Hospital (NUH) in Singapore with informed consent. The institutional review board (IRB) approval of this study was granted by NHG Domain Specific Review Board to our collaborator, Prof. (Dr.) Chng Wee Joo of Department of Hematology and Oncology, NUH (Ref # DSRB-B/09/495; Transcriptional Studies in Haematopoietic Stem Cells). We have procured four samples *viz.*, AD519, AD519d, AD598 and AD598d. AD519 was taken from a female patient with AML subtype M5a (85% blasts), and AD598 was from another female patient with AML subtype M4Eo (40% blasts). AD519d and AD598d samples were from the same patients respectively after remission (< 5% blasts) following induction therapy (daunorubicin and Ara-C in standard 3+7 followed by high dose Ara-C for consolidation). Human chronic myelogenous leukemia cell line K562 was cultured in RPMI-1640 medium supplemented with 10% fetal bovine serum (FBS), 1X penicillin-streptomycin solution (Hyclone catalog # SV30010) and 2 mM L-glutamine.

3.2.2 FANT-C library preparation and NGS

Modified from 3C-seq, the FANT-C protocol incorporates a similar workflow from 3C library preparation (2.2.2). The procedures in FANT-C include formaldehyde crosslinking, chromatin solubilization, restriction enzyme (*DpnII*) digestion, sticky-end ligation, reverse crosslinking, and DNA purification and desalting. The library is then sonicated, size-selected, ligated with Illumina sequencing adapters to make NGS library and submitted for paired-end sequencing.

(i) Formaldehyde crosslinking Frozen vials of AML patient cells were quickly thawed and resuspended in RPMI-1640 media. Cells were harvested and counted using a hemocytometer. One million cells were aliquoted in DNA LoBind 1.5 mL tube (Eppendorf). Cells were centrifuged and resuspended in 1 mL of fresh RPMI-1640 medium and processed for formaldehyde crosslinking. Similarly, 50,000 K562 cells were collected in 1 mL RPMI-1640 medium in a DNA Lo-Bind 1.5 mL microcentrifuge tube. Crosslinking was then performed by

quick addition of 28 μL of 37% formaldehyde to the cells to obtain a 1% final concentration and the tube was incubated at room temperature for exactly 10 minutes on a HulaMixer (ThermoFisher Scientific). The reaction was stopped by addition of 54 μL of 2.5 M glycine (to a final concentration of 125 mM), and the tube was incubated first at room temperature for 5 minutes and later on ice for 15 minutes. The crosslinked cells were pelleted at 450 g for 10 minutes at 4 °C and stored frozen at -80 °C till further use.

(ii) Chromatin solubilization Frozen crosslinked cells were thawed on ice, washed with 1 mL chilled 1X *DpnII* buffer (New England Biolabs, Cat#B7006S) and centrifuged at 2,000 g for 5 minutes at room temperature. The supernatant was carefully aspirated and the pellet was resuspended in 50 μL of 1.25X *DpnII* buffer by gentle tapping. Next, 5.25 μL of 1% SDS (sodium dodecyl sulphate) was added for chromatin solubilization. The tube was mixed well and incubated at 65 °C for exactly 10 minutes. SDS was quenched immediately by addition of 6.08 μL of 10% Triton X-100.

(iii) Restriction enzyme digestion To digest the chromatin, 500 U of *DpnII* (50 U/ μL , New England Biolabs, catalog # R0543M) was added to the tube and incubated at 37 °C for 2 hours. *DpnII* was then heat-inactivated by incubating at 65 °C for 20 minutes and the digestion reaction was further processed for ligation.

(iv) Sticky-end ligation A ligation master mix containing 90 μL of 10X T4 DNA ligase buffer (500 mM Tris-HCl, pH 7.5, 100 mM magnesium chloride and 100 mM DTT), 9.67 μL of 10 mg/mL BSA and 9.67 μL of 100 mM ATP was added to the digested chromatin. Water was added to a final volume of 1 mL. The ligation reaction was carried out by adding 20 μL of T4 DNA ligase (1 U/ μL , ThermoFisher Scientific, catalog # 15224025) followed by incubation at room temperature for 4 hours on a HulaMixer with slow rotation.

(v) Reverse crosslinking Fifty μL of 20 mg/mL Proteinase K (ThermoFisher Scientific, catalog # 25530031) (dissolved in 1X TE Buffer, pH 8.0) was added to the tube and incubated at 65°C for 30 minutes.

(vi) DNA purification and desalting DNA was purified using a 0.5 mL Amicon Ultra centrifugal filter (MWCO 30kDa) (Merck, catalog # C82301). Desalting was accomplished by concentrating the sample and reconstituting the original sample to 0.5 mL with 1X TE buffer.

This step was repeated 3 times to remove any traces of salt. Purified DNA was collected by inverting the column into a clean collection tube and centrifuging at 1,000 g for 2 minutes. The eluate (which contained the FANT-C library) was transferred into a fresh DNA Lo-Bind 1.5 mL microcentrifuge tube. The RNA was degraded by adding 1 μ L of 20 mg/mL PureLink RNase A (Life Technologies, catalog # 12091021) and the tube was incubated for 15 minutes at 37 °C. The FANT-C library was concentrated using Agencourt AMPure XP beads (Beckman Coulter, catalog # A63881) and eluted in 50 μ L of 1 X TE buffer.

(vii) Sonication and size selection Same as 2.2.2.

(viii) NGS library preparation and paired-end sequencing The NGS libraries were prepared using NEBNext Ultra II DNA library prep kit for Illumina (catalog # E7645S) according to the manufacturer's protocol.

(ix) Paired-end sequencing Same as 2.2.2.

3.2.3 FANT-C/Hi-C data analysis

(i) Public data downloading Previously published K562 Hi-C and *in situ* Hi-C datasets were obtained from GEO database (<http://www.ncbi.nlm.nih.gov/geo/>) with accession numbers GSE18199 and GSE63525 respectively.

(ii) Data processing The FANT-C/Hi-C raw reads were processed through the same pipeline as described in 2.2.3, including iterative mapping, fragment assignment and fragment filter, binning and bin filter, iterative bias correction, heatmap generation, compartment analysis and TAD analysis. An additional step of bin filtering was applied to reduce experimental noises in *trans* regions. After the existing bin filter steps in 2.2.3, the values in *trans* regions were ranked, and those with very high signals (> 99.95 percentile) were truncated to 99.95 percentile in the ranking.

(iii) Pearson correlation Pearson correlation was performed to compare the genome-wide matrices, PC1 eigenvectors and insulation scores, respectively, across the datasets. Prior to Pearson correlation, the genome-wide matrix was converted into a 1-dimensional array, with the exclusion of diagonal values. For PC1 and insulation scores, the values in each chromosome

were concatenated into a single array. The Pearson correlation was computed by using *stats.pearsonr* in Python package *scipy*²⁶⁴.

(iv) Distance-dependent contact probability plot The contact probabilities in all *cis* regions were taken from the genome-wide matrix (bin size = 100 kb). The mean contact probability between pairs of loci separated at a given genomic distance was computed and plotted as a function of distance in a log-log scale.

(v) Distance-dependent interaction counts The *cis* unique valid pairs separated at a given genomic distance were counted from genome-wide interaction matrix before bin filter (bin size = 100 kb). Cumulative percentage of such interaction counts was plotted with increasing distance. The percentages of short-range (< 1 Mb) and long-range ($\geq$ 1 Mb) interactions were calculated for each dataset.

(vi) Read depth analysis To calculate read coverage throughout the genome, two sides of paired-end reads were separated and aligned to hg19 using bowtie2²²⁸, excluding the reads that contain the sequence of *DpnII* ligation junction. The alignment results were then processed through CNATra tool²⁶⁵ (<https://github.com/AISKhalil/CNATra>) to compute read depth (RD) signal and subsequently to identify the DNA copy number variations (CNVs).

3.3 Results

3.3.1 FANT-C: a rapid and low-input protocol to capture genome-wide chromatin interactions

The conventional 3C-seq and Hi-C protocols consist of a series of complicated and time-consuming steps such as formaldehyde crosslinking, restriction enzyme digestion, intramolecular ligation, reverse crosslinking and DNA purification^{109,116,141}. To shorten the processing time and reduce the experimental complexity, we developed FANT-C by modifying many steps of the existing 3C library preparation protocol (**Table 3.1**). Firstly, conventional 3C-seq technique and Hi-C both require ~3 days for library generation. overnight reactions are often necessary during restriction enzyme digestion as well as reverse crosslinking to allow sufficient enzyme catalysis in a large sample. In FANT-C, with small cell input, these two steps are narrowed to 2 hours and 30 minutes respectively, allowing rapid library preparation within

one day (~10 hours) (**Table 3.1**). Secondly, many steps involved in conventional 3C and Hi-C library preparation intrinsically affect library yield. For example, cell lysis using mechanical disruption methods like Dounce homogenization can result in loss of sample by sticking to the surface of the homogenizer. Multiple times of sample transferring between tubes can also cause unavoidable mass loss. Particularly in Hi-C, the low-efficient biotin incorporation leaves 70-80% of digested fragments unlabeled²⁶⁶. The filling-in of overhangs lead to blunt-end ligation whose efficiency is much lower than sticky-end ligation at same ligase concentration. Moreover, not all the ligation junction products can be selectively isolated and recovered using streptavidin beads. As a consequence, few to hundred million cells are usually required in 3C or Hi-C library preparation to achieve sufficient library complexity with relatively low experimental noise^{109,116,141,267}. Though single-cell Hi-C has been developed^{119,263,268}, however, the sparsity of data precludes the identification of higher-order structures at sufficient resolution. The multi-step experimental procedure and the requirement of sophisticated technical expertise also limit their use in a regular clinical laboratory. Nevertheless, few Hi-C methods using low starting cell number have been reported²⁶⁹⁻²⁷¹ that circumvent the small sample problem. However, these methods, all modified from elaborate *in situ* Hi-C protocol, still require lengthy sample preparation time. Therefore, to minimize unnecessary sample loss and lower the rate of systemic biases, we further modified multiple steps in FANT-C protocol:

- 1) Removal of cell lysis step. We found that the addition of SDS followed by Triton X-100 is sufficient to permeabilize the cell and nuclear membranes for efficient entry of restriction enzyme as reported earlier²⁷².
- 2) *DpnII* is used for chromatin digestion. Since *DpnII* enzyme can be heat-inactivated, therefore the addition of SDS followed by Triton X-100 detergent during enzyme inactivation and SDS quenching steps are avoided.
- 3) Amicon Ultra centrifugal filter and Agencourt AMPure XP beads are employed for DNA purification, replacing the phenol-chloroform extraction method in conventional Hi-C and 3C experiments that is potentially corrosive (**Table 3.1**). These modifications enable the reduction of starting cell number in FANT-C to as low as 50,000.

Thirdly, we have also modified the protocol to minimize the reaction volume at every step, which allow the reactions to perform in a 1.5 mL microcentrifuge tube. Moreover, the complex experimental procedures to produce a Hi-C library can introduce intrinsic biases and experimental artifacts to the products^{273,274}. FANT-C bypasses biotin-based enrichment procedures and incorporates high-efficient sticky-end ligation (**Table 3.1**). Taken together, FANT-C is a fast and simple protocol to capture genome-wide chromatin interactions compatible with clinical timescale and small cell number.

<i>Hi-C</i> Cell number: 10 ⁷ - 10 ⁸	<i>3C-seq</i> Cell number: 10 ⁷ - 10 ⁸	<i>FANT-C</i> Cell number: 10 ⁵
Formaldehyde crosslinking Cell lysis <i>(Lysis buffer, Dounce homogenization)</i> Chromatin solubilization <i>(SDS, Triton X-100)</i> Restriction enzyme digestion <i>(37 °C, overnight)</i>	Formaldehyde crosslinking Cell lysis <i>(Lysis buffer, Dounce homogenization)</i> Chromatin solubilization <i>(SDS, Triton X-100)</i> Restriction enzyme digestion <i>(37 °C, overnight)</i>	Formaldehyde crosslinking Chromatin solubilization <i>(SDS, Triton X-100)</i> Restriction enzyme digestion <i>(37 °C, 2 hours)</i> Enzyme inactivation <i>(65 °C)</i>
Enzyme inactivation <i>(SDS, 65 °C)</i> Biotin incorporation Dilution <i>(ligation mix, with Triton X-100)</i> Blunt-end ligation <i>(T4 DNA ligase)</i> Reverse crosslinking <i>(Proteinase K, overnight)</i>	Enzyme inactivation <i>(SDS, 65 °C)</i> Dilution <i>(ligation mix, with Triton X-100)</i> Sticky-end ligation <i>(T4 DNA ligase)</i> Reverse crosslinking <i>(Proteinase K, overnight)</i>	Dilution <i>(ligation mix, no Triton X-100)</i> Sticky-end ligation <i>(T4 DNA ligase)</i> Reverse crosslinking <i>(Proteinase K, 30 minutes)</i>
DNA purification and desalting <i>(Phenol-chloroform purification, ethanol precipitation)</i> Biotin removal from unligated ends <i>(T4 DNA polymerase)</i> Shearing and size selection Biotin pull-down <i>(Streptavidin beads)</i>	DNA purification and desalting <i>(Phenol-chloroform purification, ethanol precipitation)</i>	DNA purification and desalting <i>(Column and magnetic beads)</i>
NGS	NGS	NGS
NGS library preparation Paired-end NGS	NGS library preparation Paired-end NGS	NGS library preparation Paired-end NGS

Table 3.1 The standard protocols of conventional Hi-C, 3C-seq and FANT-C. The library preparation procedures are demarcated by days of experiment and highlighted in different colors.

3.3.2 FANT-C recapitulates Hi-C data in a proof-of-concept study using K562

To test the feasibility of FANT-C, we prepared the libraries for human chronic myelogenous leukemia cell line K562. Following the protocol, the experiment was carried out with 50,000 K562 cells as input and *DpnII* was used for chromatin digestion. Total 105 million paired-end reads were produced from NGS and processed through the data analysis pipeline (3.2.3), generating 24 million unique valid pairs (Table 3.2). To evaluate the performance of FANT-C, we compared it to previously reported K562 Hi-C datasets, where approximately 100 million and 2-5 million K562 cells were subjected to Hi-C¹¹⁶ and *in situ* Hi-C²⁷⁵ protocols respectively. The published datasets that have similar numbers of unique valid pairs are selected for comparison (Table 3.2).

Protocol	FANT-C	Hi-C	<i>in situ</i> Hi-C
Data source	In house	GSE18199 (SRR027962 SRR027963)	GSE63525 (SRR1658695)
Starting cell number	50 k	100 M	2-5 M
Restriction enzyme	<i>Dpn</i> II	<i>Hind</i> III	<i>Mbo</i> I
Total read pairs	104,623,539	36,823,509	40,697,717
Bases per read	101 X 2	76 X 2	101 X 2
Aligned pairs	74,322,253	19,957,318	26,353,487
<i>Cis</i> aligned pairs	70,946,865	9,346,746	16,848,098
<i>Cis</i> aligned pairs (%)	95.46%	46.83%	63.93%
Self-circles	24,300	146,900	8,341
Self-circles (%)	0.03%	0.74%	0.03%
Dangling ends	35,988,358	2,190,640	157,331
Dangling ends (%)	48.42%	10.98%	0.60%
Valid pairs	23,846,047	17,272,744	25,888,325
Duplicates	239,927	179,506	65,070
Duplicates (%)	1.01%	1.04%	0.25%
Unique valid pairs	23,606,120	17,093,238	25,823,255
<i>Cis</i> unique valid pairs	20,263,958	6,587,698	16,340,927
<i>Cis</i> unique valid pairs (%)	85.84%	38.54%	63.28%

Table 3.2 Mapping statistics of FANT-C, Hi-C and *in situ* Hi-C K562 datasets.

Overall, the genome-wide and chromosome-level heatmaps reveal similar chromatin interaction patterns in the three datasets (**Figure 3.1**), despite large variations in experimental procedures and number of input cells. Several translocations, such as the *trans* region between chr3p and chr10q, are found in all the three genome-wide heatmaps (**Figure 3.1**). However, the heatmap of FANT-C dataset shows a few translocations that are not visible in the Hi-C datasets, possibly due to the strain differences of K562 cell line between labs or due to cell line evolution over time during the prolonged cell culture. Nevertheless, the genome-wide matrices show strong correlation between any two datasets (Pearson's $r > 0.95$, p negligible, at 1 Mb bin size), although the correlation coefficients decrease slightly with increased resolution (Pearson's $r > 0.82$, p negligible, at 200 kb bin size) (**Figure 3.2**). It is interesting to note that at all resolutions, the FANT-C and *in situ* Hi-C data show higher correlations compared to FANT-C/Hi-C and *in situ* Hi-C/Hi-C datasets (**Figure 3.2**). This may be due to the fact that the FANT-C and *in situ* Hi-C libraries were prepared with 4-base cutter enzyme *Dpn*II and *Mbo*I respectively which are isoschizomers whereas the Hi-C library used 6-base cutter *Hind*III enzyme.

Figure 3.1 Hierarchical chromatin structures in K562 revealed by FANT-C (*left panel*), conventional Hi-C (*middle panel*) and *in situ* Hi-C (*right panel*) experiments. *Top*: The genome-wide heatmaps at bin size of 1 Mb. *Middle*: Pearson correlation heatmaps (bin size = 1 Mb) along with PC1 eigenvector of chromosome 14. *Bottom*: Zoomed-in heatmap (bin size = 200 kb) along with TADs in a specific region (chr14:88,600,000-96,000,000).

Figure 3.2 Pearson correlation coefficients of genome-wide contact probability matrices at different bin sizes. Diagonal values are excluded from this analysis.

Next, we compared the datasets at A/B compartment level (**Table 3.3**). The Pearson correlation heatmaps of chr14, as an illustration, reveal similar “plaid” patterns in the three datasets (**Figure 3.1**). Accordingly, the PC1 eigenvectors from PCA behave almost similarly across three datasets, leading to almost identical regions of A and B compartments. The eigenvalues show strong correlations between datasets throughout the genome (Pearson’s $r > 0.91$, p negligible) (**Figure 3.3**). Out of 2,814 bins that are common in all datasets, 2,581 (91.7%) bins are in the same compartments between FANT-C and Hi-C dataset, and 2,608 (92.7%) bins are in the same compartments between FANT-C and *in situ* Hi-C, both of which are comparable to 2,665 (94.7%) same compartment bins between the two Hi-C datasets (**Figure 3.3**). Furthermore, total 2,148 TADs were identified in FANT-C dataset, with an average length of 1.2 Mb (**Table 3.4**). This number of TADs is slightly higher than those in Hi-C (2,049 TADs) and *in situ* Hi-C (2,010 TADs) datasets. The signal intensity along the matrix diagonal, quantified as insulation score¹³⁸, shows the highest similarity between FANT-C and *in situ* Hi-C datasets (Pearson’s $r > 0.85$, p negligible) and the lowest between FANT-C and Hi-C datasets (Pearson’s $r > 0.72$, p negligible) (**Figure 3.4**). On the other hand, we found that the Jaccard index between FANT-C and *in situ* Hi-C datasets is 0.274, which is higher than those between the two Hi-C datasets (0.240) as well as between FANT-C and Hi-C (0.213) (**Figure 3.4**). This indicates comparable TAD calling results between FANT-C and *in situ* Hi-C. Overall, these comparisons suggest that FANT-C has the ability to reveal hierarchical levels of higher-order chromatin structures despite the low starting cell number.

Dataset	FANT-C	Hi-C	<i>in situ</i> Hi-C
# of bins in A compartment	1,383	1,344	1,356
# of bins in B compartment	1,433	1,472	1,460

Table 3.3 Number of bins (bin size = 1 Mb) in A and B compartments in the three K562 datasets. Only the common bins existing in three datasets are taken.

Figure 3.3 Comparison of PC1 eigenvalues and the identified compartments between K562 datasets. *Left*: Pearson correlation coefficients of PC1 eigenvectors. *Right*: Percentage of bins in the same compartments, taking only the common bins existing in three datasets.

Dataset	FANT-C	Hi-C	<i>in situ</i> Hi-C
# of TADs	2,148	2,049	2,010
Mean TAD size (bp)	1,207,451	1,264,082	1,301,295

Table 3.4 Number of TADs and the mean TAD size in three K562 datasets.

Figure 3.4 Comparison of insulation score and the identified TAD boundaries between K562 datasets. *Top left*: Pearson correlation coefficients of insulation scores. *Top right*: Jaccard index of TAD boundaries. *Bottom*: Venn diagrams of TAD boundaries between FANT-C and Hi-C (*left*), FANT-C and *in situ* Hi-C (*middle*), and Hi-C and *in situ* Hi-C (*right*) datasets. Numbers in Venn diagrams denote the number of intersected and unique TAD boundaries respectively.

3.3.3 FANT-C captures abundant *cis* and short-range interactions

We further characterized the properties of FANT-C library by statistical examination of the captured interactions. Generally, the contact probability between two *cis* loci is inversely proportional to their linear distance on the genome¹⁰⁹. Interestingly, the plot of mean contact probability as a function of genomic distance shows a slightly steeper slope in FANT-C dataset than Hi-C, especially at distances below 1 Mb (**Figure 3.5**). We thus hypothesized that FANT-C captures more short-range *cis* interactions compared to Hi-C. The analysis of the linear genomic distances between interacting bins reveals that the cumulative percentage of short-range interactions is much higher in FANT-C compared to Hi-C (**Figure 3.6**). Particularly, short-range interactions separated by < 1 Mb distance account for 84.74% of total *cis* interactions in FANT-C, whereas the percentages are 48.60% and 53.38% in the Hi-C and *in situ* Hi-C datasets respectively (**Table 3.5**; **Figure 3.7**). Similarly, in a genome-wide matrix,

trans interactions are expected to be less frequent compared to *cis* interactions¹¹⁶, whereas a large proportion of *trans* interactions usually indicates a high rate of random ligations. We then calculated the percentages of *cis* and *trans* interactions in the entire datasets (unique valid pairs). The FANT-C dataset comprises 85.84% of *cis* and 14.16% of *trans* interactions (**Table 3.2; Figure 3.8**). This percentage of *cis* interactions in FANT-C is much higher than those in Hi-C (38.54%) and *in situ* Hi-C (63.28%) datasets (**Table 3.2; Figure 3.8**). A possible explanation for this discrepancy is that the ligation junction products are enriched in Hi-C experiments allowing them to capture and amplify rare contacts, thus causing a bias towards *trans* and long-range *cis* interactions. These results suggest that in the absence of the enrichment step, FANT-C probably captures the chromatin interactions in an unbiased manner.

Figure 3.5 Plot (bin size = 100 kb) of mean *cis* contact probability as a function of genomic distance. The maximum genomic distance between two *cis* loci is corresponding to the size of the largest chromosome in the genome.

Figure 3.6 Cumulative percentage of *cis* interactions (unique valid pairs) with increasing genomic distance (bin size = 100 kb). Horizontal dotted line represents 100% of the interaction counts. The maximum genomic distance between *cis* unique valid pairs is corresponding to the size of the largest chromosome in the genome.

Dataset	FANT-C	Hi-C	<i>in situ</i> Hi-C
short-range interactions	17,171,575	3,201,458	8,723,193
short-range interactions (%)	84.74%	48.60%	53.38%
long-range interactions	3,092,383	3,386,240	7,617,734
long-range interactions (%)	15.26%	51.40%	46.62%

Table 3.5 Short-range (< 1 Mb) and long-range ($\geq$ 1 Mb) interactions in all *cis* unique valid pairs.

Figure 3.7 The percentages of short-range and long-range interactions in all *cis* interactions (*cis* unique valid pairs).

Figure 3.8 The percentages of *cis* and *trans* interactions in all valid interactions (unique valid pairs).

3.3.4 Exploring FANT-C library composition

To ensure that low cell number in FANT-C does not affect library composition, we examined different types of reads in the datasets. Firstly, a low percentage of PCR duplicates is known to indicate a high library complexity²⁶⁷. The FANT-C dataset has 239,000 (1.01%) PCR duplicates in the valid pairs, similar to those in conventional Hi-C (1.04%) and *in situ* Hi-C (0.25%) (**Figure 3.9**). Secondly, same-fragment ligation products such as self-circles can usually indicate systemic biases (**Figure 3.11**). Out of 74 million aligned read pairs in the FANT-C dataset, 24,000 (0.03%) are self-circles (**Figure 3.10**). This number is much smaller than that of conventional Hi-C (0.74%) and similar to that of *in situ* Hi-C (0.03%) (**Figure 3.10**). Taken together, these results suggest that low-input FANT-C library has comparable complexity and systemic biases as observed in conventional and *in situ* Hi-C libraries.

Figure 3.9 The percentage of PCR duplicates in the valid pairs.

Figure 3.10 The percentage of self-circles in the aligned pairs.

Furthermore, apart from self-circles, dangling ends are another type of same-fragment ligation products where the single restriction fragments are digested but remain unligated (**Figure 3.11**). Since such reads contain only genomic sequences and do not provide information about chromatin interaction, they are usually removed in Hi-C data analysis. In the absence of enrichment of ligation junctions, FANT-C produces a large proportion of genomic DNA reads with same read orientation as dangling ends (**Figure 3.12**). As expected, we found that 48.4% of the reads in our K562 FANT-C dataset are from genomic DNA (dangling ends), which is much higher than those in the two Hi-C datasets (**Figure 3.13**). These reads can be utilized for calculating read depth signal and subsequent detection of genetic aberrations, such as single nucleotide polymorphisms (SNPs), CNVs and genomic rearrangements (**Figure 3.12**). Hence, FANT-C can simultaneously provide different types of genomic information without extra cost or effort of sample collection. Previously, Hi-C data has been employed for the detection of CNVs^{276,277}. However, the read coverage used for CNV detection may be biased towards genomic regions containing restriction sites. The poor resolution (~1 Mb) of their CNV calling fails to detect focal (< 1 Mb) copy number alterations, which limits the application of Hi-C in CNV detection for diagnostic purpose. Though high-resolution data for CNV and even SNP detection can be produced with as few as 100,000 cells using Capture-C²⁷⁸, it is only targeted at specific regions and cannot be used for genome-wide analysis. Moreover, the detection of chromosomal rearrangements in clinical WGS test is usually achieved by investigating the reads that compose sequences from two different chromosomal regions. Such rearrangements can also be inferred from strong Hi-C contacts between translocation partners. Current approaches of translocation detection from Hi-C data include: 1) setting anchor at genomic loci of a particular chromosome and detect high signal in other chromosomes²⁶⁹, 2) discovering

change points by scanning each row and column of an inter-chromosomal matrix²⁷⁷ and 3) differentiating high *trans* contacts in translocated regions from the background²⁷⁹. However, distinguishing a valid interaction and an actual translocation remains ambiguous. A normal control with no translocation is often needed for comparison of Hi-C contacts in translocated regions. Nevertheless, in FANT-C, the large proportion of genomic DNA reads can achieve high read coverage throughout the genome, allowing precise detection of structural variations at high resolution. This equips FANT-C to become a powerful tool that can be potentially applied in clinical genetic testing of cancer patients.

Figure 3.11 Diagram of two types of same fragment reads. When physically proximal restriction fragments in the digested chromatin (*left*) are treated with DNA ligase, inter-fragment ligation may not always occur. The overhangs (red) of a single fragment could either ligate with each other (self-circle) or remain unligated (dangling end). After sonication, two strands (yellow and blue) of the DNA molecule could thus be orientated towards outside (self-circle) or inside (dangling end). Yellow and blue arrows represent the 5'→3' direction of the respective strands.

Figure 3.12 Workflow of library processing, NGS sequencing and bioinformatics analysis in Hi-C and 3C-seq/FANT-C.

Figure 3.13 The percentage of genomic DNA reads (dangling ends) in the aligned pairs.

3.3.5 FANT-C in AML patient samples reveals longitudinal variations in chromatin organization

To evaluate the application of FANT-C in the clinical scenario, we prepared FANT-C libraries using 1 million bone marrow cells obtained from two acute myeloid leukemia (AML) patients,

AD519 and AD598, with 85% and 40% blast cells respectively at the time of diagnosis. FANT-C libraries were also generated using 1 million bone marrow cells from the same patients after complete remission (< 5% blast cells) following induction therapy (AD519d, AD598d). The crosslinking of cells was carried out after thawing of frozen samples. For each library, an average of 75 million reads were produced from NGS and processed to generate 9 to 16 million unique valid pairs (**Table 3.6**).

Patient sample	AD519	AD519d	AD598	AD598d
Therapy status	Before	After	Before	After
Starting cell number	1 M	1 M	1 M	1 M
Restriction enzyme	<i>Dpn II</i>	<i>Dpn II</i>	<i>Dpn II</i>	<i>Dpn II</i>
Total read pairs	72,355,728	74,406,434	87,274,338	65,169,402
Bases per read	101 X 2	101 X 2	101 X 2	101 X 2
Aligned pairs	51,548,315	52,554,589	62,628,393	46,155,856
<i>Cis</i> aligned pairs	45,913,729	48,036,206	58,219,837	43,287,097
<i>Cis</i> aligned pairs (%)	89.07%	91.40%	92.96%	93.78%
Self-circles	24,167	363,658	28,308	14,443
Self-circles (%)	0.05%	0.69%	0.05%	0.03%
Dangling ends	33,811,167	35,852,126	34,514,430	25,904,492
Dangling ends (%)	65.59%	68.22%	55.11%	56.12%
Valid pairs	11,362,401	9,727,494	15,700,399	10,998,631
Duplicates	139,512	313,174	97,515	57,920
Duplicates (%)	1.23%	3.22%	0.62%	0.53%
Unique valid pairs	11,222,889	9,414,320	15,602,884	10,940,711
<i>Cis</i> unique valid pairs	5,656,141	5,041,510	11,221,170	8,090,118
<i>Cis</i> unique valid pairs (%)	50.40%	53.55%	71.92%	73.95%

Table 3.6 Mapping statistics of FANT-C datasets generated using AML patient samples.

We first compared the chromatin organization in samples from the same patient before and after therapy (**Table 3.7**). Compartment analysis reveals 93.3% and 94.4% of bins (bin size = 1 Mb) remaining invariant between AD519 and AD519d, and between AD598 and AD598d, respectively (**Figure 3.14; Table 3.8**). The remaining bins which exhibit compartment switching show a relatively higher percentage of A→B transition rather than B→A transition (pre-therapy→post-therapy), especially from AD598 to AD598d (**Figure 3.15; Table 3.8**). **Figure 3.16** shows examples of A→B compartment transition in regions containing *KRAS* (AD519→AD519d) and *SRGN* (AD598→AD598d) genes. The gene *KRAS* is a well-characterized oncogene that plays an essential role in cancer formation and progression²⁸⁰. The

other gene *SRGN* encodes the proteoglycan serglycin, which is a marker for immature myeloid cells and often used to distinguish AML from acute lymphoblastic leukemia (ALL)²⁸¹. Both *KRAS* and *SRGN* are known to be over-expressed in AML patients²⁸². The compartment switching from A (active) to B (repressed) in these regions suggest that cancer remission following induction therapy possibly downregulates *KRAS* and *SRGN* genes. It is important to note here that the FANT-C libraries were prepared from pre- and post-therapy samples that contain heterogeneous bone marrow cell populations, where we did not purify the blast cells exclusively. The induction therapy caused a dramatic reduction in the blast cell population in post-therapy samples compared to pre-therapy samples. Therefore, the switching of A/B compartments presented here may actually be the results of the blast cell reduction from pre- to post-therapy samples. Nevertheless, these results indicate a possible trend towards chromatin inactivation from pre- to post-therapy samples. In fact, a previous study suggested the expansion of repressive chromatin during stem cell differentiation¹²³. In AML, the differentiation block leads to immature cell (blasts) accumulation. Chromatin repression by compartment switching possibly suggests resumption of normal myeloid differentiation after therapy.

Moreover, as compartments are predicted by the sign of PC1 eigenvector, we hypothesized that the magnitude of its eigenvalue could affect the sign switch. To test this hypothesis, we ranked all the genomic bins by the PC1 eigenvalues in pre-therapy samples and plotted their eigenvalues in both pre- and post-therapy samples (**Figure 3.17**). We found that those bins with lower magnitudes of eigenvalues were more likely to have an opposite sign in post-therapy samples (**Figure 3.17**).

Sample	AD519	AD519d	AD598	AD598d
# of bins in A compartments	1,423	1,413	1,347	1,316
# of bins in B compartments	1,392	1,402	1,469	1,500

Table 3.7 Number of bins (bin size = 1 Mb) in A and B compartments in datasets of patient samples.

Region	AD519→519d	AD598→598d
A→A	1,323	1,252
B→B	1,302	1,405
A→B	100	95
B→A	90	64

Table 3.8 Comparison of compartments between pre- and post-therapy samples.

Figure 3.14 Percentage of bins in the same and different compartments between pre- and post-therapy AML patient samples.

Figure 3.15 Percentage of bins showing A→B transition and B→A transition (pre-therapy→post-therapy) in all the bins undergoing compartment switching.

Figure 3.16 Examples of bins with A→B transition (pre-therapy→post-therapy), including the regions containing *KRAS* (top; chr12:20,000,000-30,000,000) and *SRGN* (bottom; chr10:63,000,000-73,000,000). The PC1 eigenvalue of those bins (blue bars; bin size = 1 Mb) are positive in pre-therapy sample and negative in post-therapy sample.

Figure 3.17 Genomic bins (bin size = 1 Mb) ranked by PC1 eigenvalues in pre-therapy samples. Left: PC1 eigenvalues in AD519 (top) and AD519d (bottom) are plotted for bins that are ranked by the eigenvalues in AD519. Bins with positive and negative values are shown in red and blue respectively, and the background colors represent the corresponding sign of PC1 in AD519.

Right: PC1 eigenvalues in AD598 (*top*) and AD598d (*bottom*) are plotted similar to the *left* for bins that are ranked by the eigenvalues in AD598.

Next, we performed TAD identification in all the four samples (**Table 3.9**). The numbers of TADs vary less between AD519 and AD519d than between AD598 and AD598d. Interestingly, TAD boundaries in the former pair show less overlapping and more divergence, as indicated by Jaccard index (**Figure 3.18**). The unique TAD boundaries in each sample are then found to have significantly lower boundary strength than overlapped boundaries (**Figure 3.19**). Taken together, these results suggest that, though overall similar, higher-order chromatin structures slightly differ between pre- and post-therapy samples, especially for regions with low PC1 magnitudes or weak TAD boundaries.

Sample	AD519	AD519d	AD598	AD598d
# of TADs	2,016	2,023	1,953	2,079
Mean TAD size (bp)	1,295,736	1,284,036	1,335,383	1,250,796

Table 3.9 Number of TADs and the mean TAD size in datasets of AML patient samples.

Figure 3.18 Comparison of identified TAD boundaries between datasets of pre- and post-therapy samples. *Top:* Venn diagram of intersected and unique TAD boundaries between pre- and post-therapy samples. *Bottom:* Jaccard indices of TAD boundaries between samples.

Figure 3.19 The strength of TAD boundaries that are overlapped (light blue) or not overlapped (pink; unique boundaries) with those in the other sample from the same patient. Mann-Whitney U test is performed for each sample (***: p-value < 0.001).

3.3.6 FANT-C uncovers *chr16* inversion and *chr22* trisomy in one AML patient

Apart from the compartment and TAD analysis, we sought to extract information about genomic abnormalities from the FANT-C data (**Figure 3.12, 3.13**). It is known that the AML cells with FAB subtype M4Eo often harbor a pericentric chromosome inversion $\text{inv}(16)(\text{p13q22})^{283-285}$. This inversion results in *CBFB-MYH11* gene fusion, which is a factor that contributes to leukemogenesis²⁸⁶. Consistent with this characteristic feature, the heatmaps of both AD598 and AD598d cells (subtype M4Eo) show high contact probabilities between p13.1 and q22.1 regions of chromosome 16 (**Figure 3.20**). Next, using genomic DNA reads from FANT-C datasets, genome-wide read depth (RD) signal was computed for all patient samples. CNAta output shows that all chromosomes in the AD519, AD519d and AD598 samples have an estimated copy number of ~2 (**Figure 3.21**). Surprisingly, we found that the RD signal of chromosome 22 in AD598d increases to copy number ~2.8 (**Figure 3.21**). Compared to the same chromosome in AD598, the signal in AD598d is increased throughout the chromosome (**Figure 3.22**). These results suggest the existence of chromosome 22 trisomy in the post-therapy sample AD598d, which might be an effect of induction therapy. Interestingly, previous studies have shown the association between trisomy 22 and $\text{inv}(16)(\text{p13q22})$ in AML²⁸⁷⁻²⁸⁹. In those studies, it was debated whether trisomy 22 is a primary abnormality or a secondary chromosomal change following chromosome 16 inversion. Since both AD598 and AD598d carried the inversion and AD598d is the only sample with trisomy 22, we speculate that the inversion might be the earlier event in this particular patient. These

results validate the ability of FANT-C to provide sufficient information for the detection of genetic abnormalities.

Figure 3.20 The contact probability heatmap between chr16 p13.3-p12.1 and q12.1-q24.3 in AD598 (*top*) and AD598d (*bottom*). Arrows denote the region with highest contact probabilities (chr16 p13.3 and q12.1) in the heatmap. Color bar shows the scale of contact probability in the heatmap.

Figure 3.21 Genome-wide RD signal (bin size = 20 kb) in AD519, AD519d, AD598 and AD598d datasets. Y-axes represent the estimated copy numbers of the corresponding genomic segments (iso-copy number blocks; IBs).

Figure 3.22 Zoomed-in comparison of RD signal (bin size = 1 kb) in chromosome 22 of AD598 (*left*) and AD598d (*right*).

3.4 Discussion

The development of high-throughput 3C variants, Hi-C and 3C-seq, has helped to further our understanding of three-dimensional chromatin conformation and its role in phenotype change during cellular state transition. However, both Hi-C and 3C-seq protocols consist of multiple complex steps that result in lengthy processing time and a general requirement of higher input cell number, largely limiting their application in clinical settings. Here, we developed the FANT-C method for rapid preparation of 3C library with small cell input by introducing a series of modifications in conventional 3C protocol (**Table 3.1**). These modifications include 1) shortening of processing time, 2) removal/changing of steps or reagents to minimize sample loss, thus reducing the starting cell number, and 3) downsizing reaction volume to allow single-tube reaction. With this simplified experimental protocol, FANT-C is compatible with a small input sample in the clinical timescale. Notably, FANT-C protocol presented here employs standard DNA modifying enzymes. This leaves space for future improvements by introducing high-fidelity restriction endonuclease and DNA ligase, which can further reduce library preparation time and increase the reaction efficiency.

We generated FANT-C library from 50,000 K562 cells and performed a systematic comparison which revealed similar chromatin organization between FANT-C and publicly-available Hi-C datasets (**Figure 3.1-3.4**). We also found that FANT-C captures more *cis* and short-range (< 1 Mb) interactions (**Figure 3.5-3.7**). We think that it is possibly due to the absence of biotin-based enrichment procedures in FANT-C protocol, which permits it to detect chromatin

interactions in an unbiased manner. The lower number of long-range (≥ 1 Mb) *cis* interactions in the FANT-C dataset will lead to data sparsity at off-diagonal (distant) regions in the contact matrix. However, the main usage of these long-range reads is the prediction of A/B compartments. The similarities of PC1 eigenvectors as well as the identified A/B compartments between FANT-C and Hi-C datasets (**Figure 3.3**) suggest that these long-range interactions in FANT-C, though less in number, are sufficient for compartment calling. On the other hand, the abundant short-range interactions in FANT-C could contribute to a more detailed characterization of TAD structure, which might explain the higher number of called TADs in the FANT-C dataset compared to Hi-C datasets. To demonstrate the applicability of FANT-C in clinical settings, we performed library preparation using AML patient samples. Alterations in chromatin organization are observed between pre- and post-therapy samples, especially in the regions with low PC1 magnitudes or weak TAD boundaries (**Figure 3.14-3.19**).

In addition to chromatin interactions, FANT-C can also provide information about genomic abnormalities by taking advantage of the large number of genomic DNA reads (**Figure 3.12, 3.13**). By using these reads, trisomy 22 is detected in one of the AML patient samples (**Figure 3.21, 3.22**). We also found an intra-chromosomal inversion in chromosome 16 in the same patient before and after therapy (**Figure 3.20**). Although the inversion is revealed by the heatmaps constructed from valid pairs in the current study, the genomic DNA reads in the FANT-C datasets could theoretically be used for a more precise breakpoint detection of such structural variations. In a different context, similar to Hi-C, FANT-C method can be readily applied to chromosome-level scaffolding of genome assemblies^{290,291} and genome-wide haplotype reconstruction²⁹². However, the lower percentage of valid pairs in FANT-C compared to Hi-C also raises our concern. As a consequence, FANT-C requires ~1.5 folds higher sequencing coverage than Hi-C to achieve a similar number of valid pairs. Fortunately, this issue is gradually becoming less important with the increasing output and decreasing cost of NGS.

In summary, our results have established FANT-C as a rapid and low-input method to capture genome-wide chromatin interactions and genetic abnormalities simultaneously. We therefore envisage that FANT-C can potentially serve as a powerful protocol in NGS-based routine cancer test and eventually contribute to a fundamental step towards personal genomics and personalized therapy.

Work contribution

The AML patient samples were provided by Prof. (Dr.) Chng Wee Joo at the Department of Hematology and Oncology, National University Hospital. The FAB subtype and other information about the patient samples were provided by Prof. (Dr.) Chng Wee Joo. Under the guidance of Assistant Prof. Amartya Sanyal, Siti Rawaidah Binte Mohammad Muzaki developed the FANT-C protocol and prepared FANT-C libraries for K562 and AML patient samples. I have performed all the data analysis presented in this study.

Chapter 4 Scaffolding and characterizing the genome of *Candida tropicalis* using genome-wide chromatin interaction data

4.1 Introduction

The yeast genus *Candida* colonizes nearly all mucosal surfaces in the human body, especially on the gastrointestinal tract and the skin²⁹³. Many *Candida* species are generally commensals or endosymbionts of the hosts. However, abrupt disruption of epithelial cell barriers allows these opportunistic pathogens to invade host cells and proliferate causing a disease known as candidiasis²⁹⁴. In fact, *Candida* species are the most common cause of fungal infections in human, frequently leading to mortality in trauma patients²⁹⁵. Among the *Candida* species that can cause human candidiasis, *Candida albicans* is the most frequently isolated pathogen, especially in temperate climate¹⁴³. It has thus served as a model organism for studying human fungal pathogens. On the other hand, the second most virulent *Candida* species, *Candida tropicalis*, has emerged as the major cause of systemic ICU-acquired candidiasis in tropical countries^{144,145}. The increasing worldwide cases of drug resistance, especially multidrug resistance, have further constrained the treatment of candidiasis caused by *C. tropicalis*¹⁴⁴⁻¹⁴⁶. It is thus essential to dissect the genetic basis of pathogenicity and drug resistance in *C. tropicalis*.

The two *Candida* species, *C. albicans* and *C. tropicalis*, are closely related, both of which belong to the CUG-Ser1 clade of the fungal phylum of Ascomycota²⁹⁶. Previous studies in *C. albicans* have revealed its ability to dramatically evolve through various genetic mechanisms. Unlike other obligate fungal pathogens that generate genetic diversity through meiosis^{297,298}, *C. albicans* lacks conventional sexual stages and meiotic recombination²⁹⁹. Instead, the genomic heterogeneity is achieved by large-scale genetic changes. For instance, during adaptive evolution inside gut flora of the host, *C. albicans* has been shown to undergo loss of heterozygosity, genomic rearrangements and mutations³⁰⁰⁻³⁰². Furthermore, segmental and chromosomal aneuploidies have been associated with drug resistance in *C. albicans*¹⁴⁷⁻¹⁴⁹. These findings indicate the plasticity of the genome of *C. albicans* and its significant role in phenotypic changes under intense selection pressure inside the host. On the other hand, meiotic recombination has also been found to be absent in *C. tropicalis*³⁰³. However, in contrast to well-studied *C. albicans*, little is known about the genome of *C. tropicalis* due to the lack of a complete genome assembly. Although the current genome assembly, consisting of 23

contigs³⁰⁴, has enabled the identification of centromeres in *C. tropicalis*³⁰⁵, the genomic features remain largely unexplored³⁰⁶.

Meanwhile, comparison of *C. albicans* and *C. tropicalis* from the perspective of species evolution could also contribute to a more comprehensive understanding of the *Candida* genus and their pathogenic behaviors. The divergence between these two closely-related species from their last common ancestor (LCA) is estimated to occur around 50 million years ago³⁰⁷. Although the genomes of *C. albicans* and *C. tropicalis* are found to be highly similar³⁰⁴, they have evolved into distinct karyotypes comprising eight and seven chromosomes respectively. The unusual genomic plasticity of *Candida* sp. allows them to tolerate a wide range of chromosomal aneuploidy, segmental duplication or deletions. Thus, the study of karyotype evolution in these closely-related *Candida* species could possibly indicate the nature and key driving force of fungi speciation in the absence of meiosis.

Therefore, to facilitate the future genome-wide studies and to investigate the karyotype differences between closely related *Candida* species, we sought to construct a complete genome assembly of *C. tropicalis*. The objectives of this study include: 1) To generate a complete chromosome-level genome assembly of *C. tropicalis* strain MYA-3404 by combining the previously available contig information and data from single-molecule real-time (SMRT) sequencing and 3C-seq experiments. 2) To study the three-dimensional chromatin conformation in *C. tropicalis* using genome-wide contact map. 3) To detect large-scale structural variations, such as genomic rearrangements and segmental duplications, in the diploid genome of *C. tropicalis*. 4) To study the karyotype evolution in *C. tropicalis* and *C. albicans* by performing genome-wide syntenic analysis.

This chapter will describe the scaffolding and organizational characterization of *C. tropicalis* genome using genome-wide chromatin interaction data. This work was carried out in collaboration with the laboratory of Prof. Kaustuv Sanyal at Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), India. The work presented in this chapter has been included in a published manuscript³⁰⁸.

4.2 Materials and Methods

4.2.1 Cells and cell culture

C. tropicalis strain MYA-3404 was grown in non-selective YPDU (1% yeast extract, 2% peptone, 2% dextrose and 0.01% Uracil) media with incubation at 30 °C at 180 rpm.

4.2.2 The 3C library preparation and NGS

The general procedures in the 3C-seq experiment of *C. tropicalis* were same as in 2.2.2, including formaldehyde crosslinking, cell membrane lysis and chromatin solubilization, restriction enzyme (*DpnII*) digestion, sticky-end ligation, reverse crosslinking, and DNA purification and desalting. The library was then sonicated, size-selected, ligated with NGS adapters and submitted for paired-end sequencing.

(i) **Formaldehyde crosslinking** We have received the crosslinked cell pellet of *C. tropicalis* from our collaborator. Around 3×10^8 cells were used to generate the 3C-seq library. Crosslinking was performed as in 2.2.2, except that 1) the final concentration of formaldehyde was 1.5%, and 2) the final concentration of glycine was 400 mM.

(ii) **Cell membrane lysis and chromatin solubilization** Same as in 2.2.2, except that 1) the cross-linked cell pellet was resuspended in *DpnII* buffer, and 2) cells were lysed by liquid nitrogen grinding in a chilled mortar using a pestle. The cell lysate was distributed in 4 tubes for digestion reaction.

(iii) **Restriction enzyme digestion** Same as in 2.2.2, using 750 units of *DpnII* enzyme per tube.

(iv) **Sticky-end ligation** Same as in 2.2.2, except incubated at 16 °C for 6 hours.

(v) **Reverse crosslinking** Same as in 2.2.2.

(vi) **DNA purification and desalting** Same as in 2.2.2.

(vii) **Sonication and library size selection** Same as in 2.2.2.

(viii) **NGS library preparation** Sequencing library was prepared using NEBNext Ultra II DNA library prep kit for Illumina (catalog # E7645S) according to the manufacturer's protocol.

(ix) **Paired-end sequencing** Same as in 2.2.2.

4.2.3 The 3C-seq data analysis

(i) **Public data downloading** Publicly-available *C. albicans* Hi-C dataset was obtained from NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject>) with accession number PRJNA308106.

(ii) **Data processing** The 3C-seq/Hi-C raw reads were processed through the same pipeline as described in 2.2.3, including iterative mapping, fragment assignment and fragment filter, binning and bin filter, iterative bias correction and heatmap generation. The *C. albicans* Hi-C reads were aligned to *C. albicans* genome assembly 21 (Ca21), which was obtained from *Candida* Genome Database (<http://www.candidagenome.org>). The *C. tropicalis* 3C-seq reads were aligned to different versions of genome assembly containing contigs or chromosomes. Depending on the sequencing depth and resolution requirement, the binning was performed at 2-10 kb resolution. Since the interactions between anchor bin and its neighboring loci were particularly examined in this study, the high diagonal values could possibly affect the results. Therefore, after the bin filter, all the diagonal values were removed from the genome-wide matrix. The iterative correction was then performed on the matrix without diagonal values.

(iii) **Plotting of 3C profile** The connectivity between contigs can be inferred from the 3C profiles of genomic bins located at the end of a contig. The 3C profiles anchored on centromeres were also used to study the respective pericentric domains in *C. albicans*. Interactions between the anchor bin and other bins were plotted using values in a single row from the contact probability matrix.

(iv) **Histogram of trans interactions** To plot the histogram of *trans* interactions, all the *trans* contact probabilities, excluding zero-values, were taken from the genome-wide matrix with a bin size of 2 kb. The mean *trans* contact probabilities were computed respectively for centromere-centromere interactions, telomere-telomere interactions and all the *trans*

interactions in the bulk chromatin. The first and last bins in each chromosome were taken as telomeric bins. Mann-Whitney U test was performed to compare the interactions.

(v) Distance-dependent contact probability curves All the *cis* contact probabilities (bin size = 2 kb), excluding zero-values, were taken from pericentric regions, non-centromeric control regions and the entire chromosome-wide matrices. For each chromosome, 1) the pericentric region was defined as centromeric bins plus 10-kb upstream region and 10-kb downstream region, and 2) a non-centromeric region with the same size as the pericentric region was selected as control. For loci separated at each genomic distance, the mean contact probabilities in the pericentric region, control region and bulk chromatin were computed, and then Mann-Whitney U test was performed to compare the interactions between regions. Similarly, the box plot of interactions in pericentric regions was plotted as a function of distance, and Mann-Whitney U test was performed to compare the interactions between adjacent genomic distances.

(vi) Read depth signal Similar to 3.2.3, to calculate read depth, two sides of paired-end reads were separated and aligned to *C. tropicalis* genome assembly using bowtie2²²⁸, excluding the reads that contain the sequence of *DpnII* ligation junction. The reads aligned to each 1-kb genomic bin was counted and plotted as the read depth signal.

4.2.4 Synteny analysis

Genome-wide synteny analysis was performed using a few tools, including SyMAP³⁰⁹, Satsuma³¹⁰ and Synchro³¹¹. The single copy orthologs present in the genomes of *C. albicans* and *C. tropicalis* were identified by Orthofinder³¹² and their connections were plotted using Circos software³¹³.

4.2.5 Data availability

All the raw sequencing data used in this study are available at NCBI BioProject accession number PRJNA596050. The sequences of the seven chromosomes of *C. tropicalis* are available at GenBank accession numbers CP047869-CP047875.

4.3 Results

4.3.1 Constructing genome assembly of *C. tropicalis*

Several approaches were employed to assemble the sequences of seven chromosomes in *C. tropicalis* strain MYA-3404. The laboratory of Prof. Geraldine Butler (University College Dublin, Ireland) generated 16 contigs based on the 23 contigs from their earlier study³⁰⁴ and kindly shared the new contig information with our collaborator. SMRT sequencing using *C. tropicalis* was then performed by our collaborator on the PacBio platform to produce approximately 1 million reads with an average length of 5.8 kb. By filling gaps in previously available contigs using SMIS pipeline²⁴⁶, 12 contigs were successfully generated (**Table 4.1**). Seven out of the eight long contigs contain centromere information. The remaining four small contigs with lengths of 5-13 kb are then referred to as orphan contigs.

Contig	Contig length (bp)	Centromere
Contig 2	2,797,331	y
Contig 1	2,474,493	y
Contig 10	2,342,205	y
Contig 11	2,098,823	y
Contig 5	1,379,441	y
Contig 6	1,255,791	
Contig 8	1,225,766	y
Contig 7	921,083	y
Contig 12	12,816	
Contig 14	8,448	
Contig 15	6,389	
Contig 16	5,408	

Table 4.1 Information about 12 contigs, including eight numbers of Mb-sized long contigs and four numbers of kb-sized small contigs. The presence of centromere in a particular contig is denoted by “y”.

Unlike the other seven long contigs, one of the contig with 1.3 Mb in size (contig 6) contains no centromere sequence (**Table 4.1**). To obtain the information on contig 6 location with respect to other 7 long contigs, we used 3C-seq data of *C. tropicalis* which generated contig-wide contact maps from 75.5 million read pairs. As discussed in previous chapters, *cis* interactions, especially between nearby regions, are generally stronger than *trans* interactions. Within the same chromosome, the contact probability is inversely proportional to the linear

genomic distance between two *cis* loci. The scaffolding of two contigs could thus be achieved from the contact probability information between their end regions. Therefore, we used the genomic bins at 5' and 3'-end of contig 6 as anchors and investigated their 3C profiles throughout the genome. Strikingly, when setting anchor at the 3'-end of contig 6, the contact probability dramatically increases while approaching the 3'-end of contig 5 (**Figure 4.1**). Reversely, setting anchor at 3'-end of contig 5 results in a similar increase of contact probability towards 3'-end of contig 6 (**Figure 4.2**). The intensities of contacts between their respective end bins reach to a level comparable to the contacts between two adjacent *cis* loci, indicating that contig 5 and contig 6 belong to the same chromosome and they should be joined through their 3'-ends. We thus joined the genomic sequence of contig 5 and the reverse complementary sequence of contig 6 to generate a new long contig with a size of 2.6 kb (**Figure 4.3**). After aligning 3C-seq reads to the new assembly, we checked the 3C profile of three bins on the newly-generated contig: 1) a bin that belongs to 3'-end of the previous contig 5, 2) a bin that covers the junction of previous contig 5 and contig 6, and 3) a bin that belongs to 3'-end (now 5'-end) of the previous contig 6. All the three bins show expected distance-dependent decay of contact probability (data not shown), proving the continuity of the new contig.

Figure 4.1 The 3C profile anchored on the genomic bin (contig6:1,250,000-1,259,999; bin size = 10 kb) located at 3'-end of contig 6. *Top*: *cis* contact probabilities between the anchor bin and other bins on contig 6. *Bottom*: intra-contig contact probabilities between the anchor bin and bins on contig 5.

Figure 4.2 The 3C profile anchored on the genomic bin (contig5:1,370,000-1,379,999; bin size = 10 kb) located at 3'-end of contig 5. *Top*: *cis* contact probabilities between the anchor bin and other bins on contig 5. *Bottom*: intra-contig contact probabilities between the anchor bin and bins on contig 6.

Figure 4.3 Diagram of the joining of contig 5 and contig 6 through 3'-ends to form a new contig.

Similarly, we set anchors at orphan contigs and observed their behaviors of contact probability across all the large contigs. Although we were unable to assess the orientation of orphan contigs due to their short length, their possible location in large contigs could be predicted from the increase of contact probability (**Figure 4.4**). Despite the sparse data points at small bin sizes, we observed a distance-dependent decay in the 3C profile of each orphan contig at the predicted

location. On the other hand, the SMRT reads show continuous coverage in large contigs containing these predicted regions. These results hint that the orphan contigs are possibly present as heterozygous loci in the diploid genome. Meanwhile, using Canu³¹⁴ and FALCON³¹⁵, our collaborator has successfully built a *de novo* genome assembly from SMRT reads. Using sequences of orphan contigs as queries, BLAST³¹⁶ analysis successfully identified the corresponding homologous sequence in the *de novo* genome assembly. These heterozygous loci in the large contigs were then confirmed by our collaborators using Southern blot analysis.

Figure 4.4 The 3C profile anchored on genomic bins located in orphan contigs. *Top left:* The contact probabilities between a genomic bin in contig 12 (contig12:0-9,999; bin size = 10 kb) and all the bins on contig 8. *Top right:* The contact probabilities between a genomic bin in contig 14 (contig14:2,000-3,999; bin size = 2 kb) and all the bins on contig 8. *Bottom left:* The contact probabilities between a genomic bin in contig 15 (contig15:0-4,999; bin size = 5 kb) and all the bins on contig 7. *Bottom right:* The contact probabilities between a genomic bin in contig 16 (contig16:0-4,999; bin size = 5 kb) and all the bins on contig 8.

Following merging of contig 5 and contig 6 and identifying the nature of small contigs as heterozygous loci, the previous contig-level genome assembly was upgraded to chromosome-level assembly (**Table 4.2**). The long contigs were thus renamed as chromosomes, and the 3C-seq reads were realigned to the new chromosome-level assembly. The contact probability information was then used for 1) validation of previous joins and insertions of contigs and 2) further confirmation of orphan contigs as heterozygous loci in respective chromosomes. Furthermore, constructed based on previously available contig information, the genome assembly contains a total of 104 gaps with unknown sequence (represented as “N” in genome assembly) encompassing around 54 kb of the genome. Using 500 bases upstream and downstream regions of the gaps as queries and *de novo* genome assembly as the database, our collaborator performed BLAST analysis to successfully fill 78 out of 104 gaps. To fill the remaining 26 gaps, we developed an algorithm to iteratively obtain the sequence from BLAST database, which consists of SMRT reads, Illumina sequencing reads and the genomic DNA reads from the 3C-seq dataset. However, we found that many of these gaps contain repetitive sequences with unknown copy number, resulting in multiple hits and therefore difficulty in closing gaps using our approach. Our collaborators then avoided this complication by using 1 kb sequences from 10 kb upstream and downstream of the gap as queries. This allowed the filling of the remaining 26 gaps. All the filled gaps were verified by investigating the coverage of reads aligned to gap sequences and their surrounding regions in the genome assembly. The assembly was then polished by Pilon³¹⁷ using 3C-seq reads. Taken together, after a series of approaches, we built a chromosome-level gap-less genome assembly of *C. tropicalis*, which contains 14,609,527 bases in the seven chromosomes (**Table 4.2**).

Contig	Chromosome	Chromosome length (bp)
2	1	2,835,428
5	2	2,658,250
6		
1	3	2,504,192
10	4	2,347,188
8	5	1,239,682
7	6	927,416
11	R	2,097,371

Table 4.2 Information about the upgradation of genome assembly from eight long contigs to seven chromosomes. Chromosome lengths in the new genome assembly of *C. tropicalis* after gap filling and further polishing are shown.

4.3.2 Genome-wide contact maps reveal conserved principle of chromatin conformation

Next, we sought to study the three-dimensional chromatin organization in *C. tropicalis*. One of the most well-known features of yeast chromatin is Rabl configuration, where centromeres are clustered and located at one side of the nuclear envelope and the telomeres are dragging behind^{318,319}. The clustering of centromeres has been found to be evolutionally conserved in many yeast species including *C. albicans*^{137,320,321}. We first examined the spatial relationship between centromeres and between telomeres in *C. albicans* by analyzing a publicly-available Hi-C dataset from a previous study³²⁰ (**Table 4.3**). Consistent with the previous findings³²⁰, the genome-wide heatmap shows conspicuous puncta in the *trans* regions between centromeres (**Figure 4.5**), suggesting clustering of centromeres in *C. albicans*. In addition, telomere-telomere interactions are also observed in this heatmap (**Figure 4.5**). To study the features of the chromatin conformation in *C. tropicalis*, we realigned the 3C-seq reads to the complete genome assembly and processed the data through the same pipeline, which generated 3.5 million unique valid pairs (**Table 4.3**). Similar to *C. albicans*, the genome-wide heatmap of *C. tropicalis* also shows strong centromere-centromere interactions and telomere-telomere interactions (**Figure 4.6**). Future studies on the relative positions of centromeres and telomeres inside nucleus could examine the possible presence of Rabl configuration in *C. albicans* and *C. tropicalis*.

Species	<i>C. tropicalis</i>	<i>C. albicans</i>
Source	In-house	PRJNA308106 (SRR3381672)
Protocol	3C-seq	Hi-C
Total read pairs	75,540,633	34,602,688
Bases per read pair	101 X 2	80 X 2
Aligned pairs	63,196,040	26,897,784
<i>Cis</i> aligned pairs	61,727,841	17,086,632
Self-circles	444,308	53,338
Dangling ends	36,906,595	2,228,548
Valid pairs	3,500,320	21,910,659
Duplicates	43,501	2,000,471
Unique valid pairs	3,456,819	19,910,188
<i>Cis</i> unique valid pairs	2,007,031	10,975,844

Table 4.3 Mapping statistics of in-house *C. tropicalis* 3C-seq and publicly-available *C. albicans* Hi-C datasets.

Figure 4.5 Genome-wide heatmap (bin size = 5 kb) of *C. albicans* from a publicly-available Hi-C dataset. The start and end of each chromosome are demarcated by black lines. Chromosomes are labeled at the corresponding positions in the heatmap. Arrows point to examples of centromere-centromere interactions (black) and telomere-telomere interactions (blue) in the *trans* regions. Color bar denotes the scale of contact probability.

Figure 4.6 Genome-wide heatmap (bin size = 5 kb) of *C. tropicalis* using the newly-generated genome assembly. The start and end of each chromosome are demarcated by black lines. Chromosomes are labeled at the corresponding positions in the heatmap. Arrows point to 1) examples of centromere-centromere interactions (black) and telomere-telomere interactions (blue) in *trans* regions as well as 2) a butterfly-like pattern between chromosome 1 and 4 (green). Color bar denotes the scale of contact probability.

In another study using the same Hi-C dataset of *C. albicans*, we undertook a comprehensive strategy to examine the centromere-centromere interactions as well as the *cis* contacts within centromere-proximal regions³²² (**Figure 4.7, 4.8**). In *C. albicans*, upon native centromere deletion, the hotspots of neocentromere activation were found to be close to the location of native centromere³²³, suggesting the important role of centromere-proximal regions in performing centromere functions. Our collaborators integrated the transgene *URA3* into centromere-proximal regions on chromosome 7 of *C. albicans* and selected the colonies with reversible silencing of *URA3* by 5-fluoroorotic acid (5-FOA). Interestingly, the frequency of the reversible silencing decreases from the CENPA-bound core centromere to its periphery, suggesting a confined pericentric domain surrounding centromere. We then investigated the

properties of centromeres and the pericentric domains by analyzing Hi-C data of *C. albicans*. Firstly, the histogram of all non-zero *trans* interactions shows that the contact probabilities between centromeres are significantly higher than the bulk chromatin (Mann Whitney U test $p\text{-value} < 10^{-89}$) (**Figure 4.7A**). Secondly, we examined the *cis* interactions in different regions by generating distance-dependent contact probability curves. For each chromosome, we took 1) interactions within a roughly-defined pericentric region which comprised the core centromere, 10 kb upstream and 10 kb downstream of the centromere, 2) interactions within a non-centromeric control region which has the same size as the pericentric region, and 3) all the *cis* interactions in the chromosome. Interestingly, at any given distance, the contact probabilities in pericentric regions are significantly higher than the bulk chromatin (**Figure 4.7B**). In contrast, non-centromeric regions show a similar level of interactions as the bulk chromatin. Together with observations from subtraction matrices (**Figure 4.8**), these results suggest a compact chromatin environment in the pericentric domain. To explore the boundaries of the pericentric domain, we generated a box plot for contact probabilities in pericentric regions with increasing genomic distance (**Figure 4.7C**). By statistically comparing the values in adjacent distances, we observed a significant decay of contact probabilities at each increment in the genomic distance until 10 kb. Beyond 10-kb distance, the contact probabilities remain at a low level with almost no significant difference. These results suggest that, given a locus in the pericentric region (such as the core centromere), it could strongly interact with other pericentric loci in *cis* within ~10-kb distance. Since the centromeres in *C. albicans* are 3-5 kb in size³²⁴, we thus expected the pericentric domains to be ~25-kb regions centering the core centromeres. Consistently, the 3C profiles of the centromeres in chr7 and chr2 show ~25-kb centromere-flanking regions with high contact probabilities with the centromeres (**Figure 4.7D, E**). Taken together, we defined ~25-kb pericentric domains centering CENPA-bound core centromeres in *C. albicans*, which form centromere-flanking compact chromatin (CFCC) that is segregated from the bulk chromatin both in *cis* and in *trans*. Further ChIP-qPCR experiments performed by our collaborator have shown that the hotspots of neocentromere activation on chromosome 7 are located within the specific pericentric domain.

Figure 4.7 The compact chromatin environment of centromere-flanking pericentric domains in *trans* and in *cis* (bin size = 2 kb). **A**: Histogram of non-zero *trans* contact probabilities showing the mean values of bulk chromatin (blue vertical line) and between centromeres (red vertical line). **B**: The distance-dependent contact probability curves of pericentric (red), non-centromeric control (blue) and bulk chromatin (green) regions. (*Mann-Whitney U test p-value < 0.05; n.s.: not significant). The diagram (*inset*) demonstrates the source of data points in the distance-dependent contact probability curves, including the interactions within pericentric regions (red bins) surrounding the centromeric bin (black bin) and the interactions within non-centromeric control regions (blue bins). **C**: Box plot of contact probabilities in pericentric regions at each genomic distance. The red line in each box represents the mean contact

probability at that distance. (*Mann-Whitney U test p-value < 0.05; n.s.: not significant). **D**: the diagram (*top*) demonstrates a Hi-C matrix of a chromosome 7 region surrounding centromere 7 (CEN7; black bar), where the 3C profile of CEN7 is generated from the values (red box) in the contact probabilities matrix. The 3C profile (*bottom*) anchored on CEN7 bin (Chr7: 426000-428000; black bar) showing its *cis* contact probabilities (red dots) with the neighboring bins (Chr7:410000-450000) on chromosome 7. **E**: The 3C profile anchored on centromere 2 (CEN2) bin (Chr2:1926000-1928000; black bar) showing its *cis* contact probabilities (red dots) with the neighboring bins (Chr2:1910000-1950000) on chromosome 2. This figure was taken from the published article³²². The permission to reuse the figure was granted by Genetics Society of America.

Figure 4.8 The contact probability heatmap of regions in chromosome 7 of *C. albicans*. **A**: The chromosome-wide heatmap (bin size = 5 kb) showing three 26-kb regions (yellow boxes) including the pericentric region centering centromere 7 (CEN7 region) and two randomly-selected non-centromeric control regions (control 1 and control 2). Color bar denotes the scale of contact probability in the heatmap. **B-D**: Heatmaps (bin size = 2 kb) generated from direct subtraction of matrices between regions. The regions involved in subtraction are mentioned on

top of each heatmap. Color bars denote the differences of contact probabilities. This figure was taken from the published article³²². The permission to reuse the figure was granted by Genetics Society of America.

To test if such features are conserved in *C. tropicalis*, we performed the same analysis using our 3C-seq dataset. Consistent with the results in *C. albicans* (**Figure 4.7A**), centromere-centromere *trans* interactions are significantly stronger than the bulk chromatin (Mann Whitney U test p-value $< 10^{-11}$) (**Figure 4.9**). Additionally, in agreement with the observations in the genome-wide heatmap (**Figure 4.6**), telomeres of different chromosomes in *C. tropicalis* also interact with each other at high contact probabilities as compared to the bulk chromatin (Mann Whitney U test p-value $< 10^{-10}$) (**Figure 4.9**). We next sought to examine the *cis* interactions in centromere-proximal regions in *C. tropicalis*. However, no significant difference is found between the *cis* interactions in pericentric regions and bulk chromatin or non-centromeric control regions (data not shown). It is known that, unlike the unique centromeres in *C. albicans*, the centromeres of *C. tropicalis* are associated with homogenized inverted repeats (HIRs)³⁰⁵ and they have larger sizes (~10 kb) as compared to 2-5-kb sized centromeres in *C. albicans*. Therefore, we also directly examined the interactions in centromeric bins without taking pericentric regions. However, the *cis* interactions in centromeric regions also show no significant difference compared to bulk chromatin or the control regions. These results indicate that the CFCC, observed in *C. albicans*, may not be present in *C. tropicalis*. The divergence in pericentric chromatin organization in the two species may possibly be a result of distinct centromere types in their genomes.

Figure 4.9 Histogram of non-zero *trans* contact probabilities in *C. tropicalis* 3C-seq dataset. The mean values in the bulk chromatin (blue), between centromeres (red) and between telomeres (black) are marked as vertical lines.

4.3.3 Structural variations in diploid *C. tropicalis* genome

Apart from the interactions between centromeres and between telomeres, the genome-wide heatmap also reveals unexpected strong *trans* contacts showing a butterfly-like pattern between a region near 3'-end of chromosome 1 and another region near 5'-end of chromosome 4 (**Figure 4.6, 4.10**). Meanwhile, analysis of the *de novo* assembly by our collaborator has revealed two chimeric contigs encompassing sequences from both regions, suggesting a possible exchange of sequences between the two chromosomes. However, the continuous read coverage in both regions as well as the other contigs in the *de novo* assembly suggest that it is not due to an error in genome assembly. Therefore, we hypothesized that it could be a balanced translocation that is only present in one of the homologs. To test our hypothesis, we interchanged the sequences of respective regions in the genome assembly and realigned 3C-seq data to the alternative genome assembly. As expected, the contact maps using alternative genome assembly still shows the butterfly-like pattern in the same *trans* region (**Figure 4.11, 4.12**), proving the presence of this translocation as a heterozygous variation in the diploid *C. tropicalis* genome. This chromosomal exchange was then further verified by our collaborator using contour-

clamped homogeneous electric field (CHEF) gel electrophoresis followed by Southern blotting (CHEF-Southern).

Figure 4.10 Heatmap of contact probabilities (bin size = 5 kb) between chromosome 1 and 4. The 3C-seq reads are aligned to the newly-generated genome assembly of *C. tropicalis*. The green arrow points to the butterfly-like pattern in the *trans* region. Color bar denotes the scale of contact probability.

Figure 4.11 Genome-wide heatmap (bin size = 5 kb) of *C. tropicalis* using the alternative genome assembly. The start and end of each chromosome are demarcated by black lines. Chromosomes are labeled at the corresponding positions in the heatmap. The green arrow points to the butterfly-like pattern between chromosome 1 and 4. Color bar denotes the scale of contact probability.

Figure 4.12 Heatmap of contact probabilities (bin size = 5 kb) between chromosome 1 and 4 using the alternative genome assembly of *C. tropicalis*. The green arrow points to the butterfly-like pattern in the *trans* region. Color bar denotes the scale of contact probability.

The breakpoint of the balanced translocation is located at 440 kb from the 3'-end of chromosome 1 and 392 kb from the 5'-end of chromosome 4 (**Figure 4.13**). Thus, the intra-chromosomal exchange could result in ~50-kb size difference between two chromosome 1 homologs and between two chromosome 4 homologs. However, CHEF-Southern experiment performed by our collaborator has shown that the two homologs of chromosome 4 are approximately 2.3 and 2.6 Mb in size. This ~300 kb size difference could not be merely explained by this translocation, suggesting the possible existence of other genetic variations in chromosome 4. To explore these variations, we generated the read depth signal using genomic DNA reads from the 3C-seq dataset. Strikingly, we observed a 235-kb region near 3'-end of chromosome 4 where the read depth signal had increased to around 1.5 folds compared to the rest of the region in chromosome 4 (**Figure 4.14**). Interestingly, the 3C-seq heatmaps show strong interactions between the start and end bins of this amplified region (**Figure 4.15**), indicating the nature of this region as a result of head-to-tail tandem duplication. The 1.5-fold

read depth signal difference thus suggests the existence of a tandem duplication in one of the homologs of chromosome 4, which could then explain the size difference between two homologs observed in CHEF-Southern experiment.

Figure 4.13 Diagram of the balanced translocation between chromosome 1 and 4.

Figure 4.14 Read depth signal (read count per bin; bin size = 1 kb) of the amplified region in chromosome 4. *Left*: The chromosome-wide read depth signal. Arrow points to the 235-kb amplified region near 3'-end of chromosome 4. *Right*: Violin plots of the read counts in the amplified region and the remaining region on chromosome 4.

Figure 4.15 Contact probability heatmaps (bin size = 10 kb) showing the amplified region in chromosome 4. *Left*: heatmap of chromosome 4, where the high head-to-tail interactions of the amplified region are marked by yellow circles. *Right*: heatmap of the amplified region in chromosome 4. Arrow points to the interaction between the start and end bins of the amplified region. Color bars denote the scale of contact probability.

Inspired by the success of identifying the duplicated region in chromosome 4, we searched for other segmental copy number variations throughout the genome. Interestingly, we found another 85-kb region near the ribosomal DNA (rDNA) locus on chromosome R with increased read depth signal (**Figure 4.16**). With the exclusion of the rDNA region, the average read depth signal in this amplified region is around two folds compared to the rest of chromosome R (**Figure 4.16**). However, it is difficult to infer the nature of this amplification from the complex patterns of interactions between bins of this region in the contact map (**Figure 4.17**). Nevertheless, the mRNA sequencing results from our collaborator have shown concomitantly increased gene expression in these two amplified regions in chromosomes 4 and R (data not shown). Interestingly, by investigating the genes in these amplified regions, our collaborators identified the orthologs of a few genes that are implicated in pathobiology and antifungal drug resistance in *C. albicans*.

Figure 4.16 Read depth signal (read count per bin; bin size = 1 kb) of the amplified region in chromosome R. *Left*: The chromosome-wide read depth signal. Arrows point to the rDNA locus and the 85-kb amplified region on chromosome R. *Right*: Violin plots of the read counts in the amplified region and the remaining region on chromosome R, with the exclusion of rDNA region (ChrR:415,000-450,000).

Figure 4.17 Contact probability heatmaps showing the amplified region in chromosome R. *Left*: heatmap of chromosome R (bin size = 10 kb), where the complex interactions between start and end bins of the amplified region are marked in green circles. *Right*: heatmap of the amplified region (bin size = 5 kb) in chromosome R. Color bars denote the scale of contact probability.

4.3.4 Centromeres as hotspots for chromosomal recombination during speciation

Syntenic relationships in multiple *Candida* species and comparative evaluation of curated sets of orthologous *Candida* genes have been widely used for evolutionary analyses of this clade. The completion of *C. tropicalis* genome assembly enables a genome-wide synteny comparison between *C. albicans* and *C. tropicalis*. Following various approaches using SyMAP³⁰⁹, Satsuma³¹⁰ and Synchro³¹¹, our collaborator performed systemic synteny analysis for the newly-assembled *C. tropicalis* genome and publicly-available genome assembly of *C. albicans* strain SC5314. Total 4,974 single copy orthologs were detected by Orthofinder³¹² (**Figure 4.18**). Though the two species have evolved into distinct karyotypes with 8 and 7 chromosomes respectively, this analysis suggests that the orthologous protein-coding genes are arranged in syntenic clusters between these species (**Figure 4.18**). Nevertheless, notable chromosomal rearrangements are observed in the genomes of the two species, suggesting their significant role in karyotype evolution and speciation (**Figure 4.18**). Interestingly, we found an enrichment of inter-chromosomal synteny breakpoints (ICSBs) near the centromeres of the two species. Seven of the eight centromeres in *C. albicans* and four of the seven centromeres in *C. tropicalis* are located at those ICSBs, suggesting centromeres might be the hotspots for ancient chromosomal recombination during karyotype evolution in the common ancestor of the two species. A previous study using clinical isolates of *C. albicans* showed that the large-scale genomic rearrangements are mediated by 8 copies of major repeat sequence (MRS)³²⁵. While MRS elements are found to be absent in *C. tropicalis*, their centromeres are associated with HIRs³⁰⁵. Our analysis in several other members of the CUG-Ser1 clade has shown the conservation of HIR-associated centromeres in their genomes, which might have evolved from the centromeres of ancestral species³⁰⁸. Therefore, we put forward a model in which recombination-mediated rearrangements of chromosomes encompassing the HIR-associated centromeres might have occurred in the common ancestor of *C. albicans* and *C. tropicalis*³⁰⁸. In addition, the clustering of centromeres, as observed in many yeasts including *C. albicans* (**Figure 4.5**) and *C. tropicalis* (**Figure 4.6**), brings these centromeres in close spatial proximity, which may have facilitated the chromosomal recombinations in centromere-proximal regions. These inter-centromeric translocation events, followed by loss of HIR-associated centromeres and acquisition of evolutionary new (unique) centromeres, might have led to the karyotype evolution and speciation of *C. albicans*³⁰⁸.

Figure 4.18 The color-coded connections between the conserved single copy orthologs in the genomes of *C. albicans* (left; Ca1 to CaR) and *C. tropicalis* (right; Ct1 to CtR) as revealed by genome-wide synteny analysis. Red dots represent the genomic location of the centromere in each chromosome. Note that the arms of each chromosome in *C. tropicalis* genome are oriented in this figure to keep the short arm at the 5' side. This figure was provided by our collaborator.

4.4 Discussion

C. tropicalis is a common human fungal pathogen that can cause candidiasis, especially in tropical countries. However, unlike the well-studied organism *C. albicans*, the genome of *C. tropicalis* was not assembled completely. This largely limited the genome-wide studies in *C. tropicalis*, such as the genomic changes during the development of drug resistance and karyotype evolution in *C. tropicalis* and the closely-related species *C. albicans*. Here, we successfully constructed a complete genome assembly of *C. tropicalis* strain MYA-3404 (Table 4.2). We showed that the chromatin interaction information from 3C-seq data could be used for scaffolding and ultimately for generating the chromosomal-level genome assembly (Figure 4.1-4.4). Next, we generated the genome-wide chromatin contact map of *C. tropicalis* by aligning 3C-seq data to the newly-assembled complete genome (Figure 4.6). Similar to *C. albicans* (Figure 4.5, 4.7), the centromere-centromere interactions, as well as the telomere-telomere interactions, are detected in *C. tropicalis* (Figure 4.6, 4.9). Furthermore, we observed

a balanced translocation between chromosome 1 and chromosome 4 that exist in one of the homologs of the diploid genome (**Figure 4.10-4.13**). Using the genomic DNA reads from the 3C-seq dataset, we generated read depth signal throughout the genome, which reveals two amplified regions located on chromosome 4 and chromosome R respectively (**Figure 4.14-4.17**). Orthologs of a few genes in these two amplified regions have been found to be implicated in pathobiology and drug resistance in *C. albicans*. Finally, the completion of *C. tropicalis* genome assembly has enabled genome-wide synteny analysis in closely-related species. Synteny analysis performed by our collaborator has shown massive chromosomal recombination events between 7 chromosomes of *C. tropicalis* and 8 chromosomes of *C. albicans*, especially near the locations of the centromeres (**Figure 4.18**). The presence of several centromeres at chromosomal recombination breakpoints hints that centromeres could possibly be the hotspots of karyotype changes during speciation in the common ancestor of the two species. While our characterizations of *C. tropicalis* genome are preliminary, we believe that the complete genome assembly could facilitate future genome-wide research in this species.

Work contribution

As mentioned in **4.1 Introduction**, this study was carried out in collaboration with the laboratory of Prof. Kaustuv Sanyal at Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), India. The initial genome assembly from 23 to 16 contigs was performed by Caoimhe O'Brien in the laboratory of Prof. Geraldine Butler at School of Biomolecular and Biomedical Science, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Ireland. Under the guidance of Prof. Kaustuv Sanyal, Bhagya C. Thimmappa performed SMIS analysis that generated 12 contigs (**Table 4.1**) and built *de novo* genome assembly using Canu. Next, Krishnendu Guin generated *de novo* genome assembly using FALCON, performed BLAST analysis for gap filling, polished the genome assembly using Pilon, investigated the orthologs of genes in amplified regions and carried out genome-wide synteny analysis (**Figure 4.18**). Krishnendu Guin also performed chromoblot analysis and mRNA sequencing experiments in the work included in this chapter. Krishnendu Guin performed the formaldehyde crosslinking step in 3C-seq library preparation. SMRT-seq library preparation and sequencing were carried out using PacBio sequencing facility at Duke University with the aid of Dr. Joseph Heitman and Dr. Sheng Sun. In our laboratory, under the guidance of Assistant Prof. Amartya Sanyal, Siti Rawaidah Binte Mohammad Muzaki prepared the 3C-seq library using crosslinked *C. tropicalis* cells. I analyzed 3C-seq data for the

scaffolding of contigs (**Figure 4.1-4.4**), verification of genome assembly, generation of chromatin contact maps (**Table 4.3; Figure 4.6**), examination the *trans* and *cis* contacts in centromeric and telomeric regions (**Figure 4.9**) as well as identification of the balanced translocation (**Figure 4.10-4.13**) and amplified regions (**Figure 4.14-4.17**). I also analyzed publicly-available *C. albicans* Hi-C dataset for the generation of genome-wide contact map (**Figure 4.5**). This chapter also includes results from another published article where we studied pericentric domain in *C. albicans*³²² (**Figure 4.7, 4.8**). The work contribution in this article has been described in the Authorship Attribution Statement of this thesis. The permission to reuse the published figures was granted by Genetics Society of America.

Chapter 5 Conclusions and future work

The cell state transition is associated with transcriptional reprogramming. As discussed in Chapter 1, transcriptionally active (euchromatin) and repressed (heterochromatin) chromatin have distinct localization preferences in the interphase nucleus, which are corresponding to A and B compartments respectively. At the level of TADs and sub-TADs, the activity of genes can be modulated by distal regulatory sequences like enhancers through long-range chromatin loops, suggesting the significance of local chromatin conformation in gene regulation. In this thesis, by employing 3C-seq or in-house developed FANT-C protocol, we examined alterations in 3D chromatin topology during phenotypic changes of cancer cells.

Firstly, the contact heatmaps of SCLC chemoresistance cell line model show extensive differences in inter-chromosomal translocations, A/B compartments and TAD structures between drug-sensitive and resistant cells. These findings support our initial hypothesis that chromatin organization changes as the cells develop resistance to the drug challenge. These alterations in chromatin structure are often observed to be associated with concomitant changes in gene expression as well as H3K27ac ChIP-seq signal, suggesting chromatin organization as an epigenetic regulator during changes of drug-response states of the cancer cells. Next, to investigate the occurrence of drug-induced changes over time, we generated in-house drug-tolerant cell lines by culturing NCI-H69 cells in increasing concentration of doxorubicin (**Figure 2.20**). This experiment could provide insights into an interesting proposition: given the situation that we are exposing the same parental cell line NCI-H69 to nearly identical drug challenge as used to create H69AR previously, do cancer cells tread the same path to develop drug resistance? During somatic cell reprogramming, it was earlier observed that large-scale changes in compartments and TADs erased tissue-specific chromatin structures, leading to an embryonic stem-cell-like 3D chromatin organization in the induced pluripotent stem cells (iPSCs)¹²⁶. However, early-passage iPSCs displayed subtle cell-of-origin-dependent topological hallmarks, suggesting the presence of ‘residual’ molecular imprints of the original cell type during transitions between cell states. To assess whether the cancer cells will undergo definitive cell-of-origin-dependent transitions during the development of drug resistance, we investigated the time-resolved changes in the in-house derived cell lines from drug-sensitive to drug-tolerant states. Events of genomic rearrangements, compartment switching and TAD boundary changing are detected with the increasing concentration of the drug, suggesting that karyotype evolution and alterations in chromatin topology may be recurrent features in cancer

cells under drug-selection pressure. However, we found that different arrays of genes are regulated by the cells when exposed to the same drug challenge. For instance, *ABCC1* is the most differentially regulated gene between NCI-H69 and H69AR, where *ABCB1* is nearly off and not regulated in both cell lines. In contrast, *ABCB1* and not *ABCC1* becomes the most differentially regulated gene between H69AMC150 and H69DT150. This indicates that the overexpression of ABC transporters may be a common strategy by cancer cells to develop drug resistance, yet the member of ABC transporters to be overexpressed might be determined based on random selection.

Overall, our results contribute to a more sophisticated understanding of the role of chromatin organization during the development of drug resistance in SCLC, which can possibly regulate gene expression through 1) relocation of enhancers and/or genes between active and repressed chromatin (compartment switching) and 2) rewiring of enhancer-gene interactions (TAD changing). Compared to NCI-H69 and H69AR, changes in translocations, A/B compartments and TAD structures are also observed in H69PR. These changes may cause transcriptional reprogramming through above-mentioned mechanisms and eventually revert to the drug sensitive state. Another possible explanation is that the parental NCI-H69 cells comprise of multiple subclones with distinct genetic and epigenetic features. Under intense environmental pressure upon drug treatment, subclones that are more tolerant to the drug may be positively selected. After drug removal, in the absence of selection pressure, the cells no longer require the drug-induced genetic and epigenetic changes and may gradually regain the drug sensitivity. The alterations in chromatin organization, especially the switching between A and B compartments, are associated with changes in enhancer activity and gene expression during the development of drug resistance. However, their causal connections remain unresolved. For example, active transcription of a gene or enhancer could be a consequent event followed by $B \rightarrow A$ switch in compartments. On the other hand, processes such as recruitment of transcription factors and chromatin remodeling may possibly lead to a change in gene/enhancer location from inactive to active chromatin. At TAD level, although many variant H3K27ac ChIP-seq peaks and differentially regulated genes have been found in TAD changing regions, we did not observe a global association between TAD boundary alterations and gene/enhancer activity changes. A possible explanation is that gain or loss of TAD boundaries may not necessarily affect the contacts between enhancers and genes within the TADs. It was shown by a recent study that the fusion of adjacent TADs by removal of all CTCF sites had little impact

on gene expression³²⁶. On the other hand, significant changes in gene expression were achieved by inversions or repositioning of TAD boundaries that resulted in redirection of enhancer-gene interactions³²⁶. The enhancer-mediated gene activation through looping interactions also leads to a question that which process occurs first, enhancer activation or formation of chromatin loops. A study on β -globin locus showed that GATA1-mediated loop formation is critical for tethering the locus control region (LCR) and β -globin promoter and thus activating β -globin transcription³²⁷. On the contrary, TNF- α -responsive enhancers were found to be in contact with target promoters prior to TNF- α signalling³²⁸. In our observations at *ABCB1* locus in in-house derived cell lines, it is highly possible that the *ABCB1* activation is achieved by changing of TAD/sub-TAD boundary that brings active enhancer loci and *ABCB1* in close contact (**Figure 2.28**). Apart from our study, the epigenetic regulation in chemoresistance has also been investigated by other groups in different types of cancers. For examples, in response to kinase inhibitors, glioblastoma stem cells (GSCs) were found to reversibly transition into a slow-cycling drug-resistant state through widespread redistribution of repressive chromatin markers¹⁹⁹. In ovarian cancer, the transcription modulation by distal enhancers for acquiring and maintaining cisplatin resistance was characterized³²⁹. These studies may provide potential targets for the design of more specific therapeutic approaches against cancer chemoresistance.

Compared to the drug-resistant cell line H69AR, the levels of drug resistance in our in-house derived drug-tolerant cells are much lower (**Figure 2.1, 2.21**). As a result, only a small proportion of genomic regions and genes are found to be involved in drug-induced changes in these cells. This is largely due to time constraint, as the development of drug-tolerant cell lines can take months to years to reach a certain level of drug resistance. Therefore, further development of drug-tolerant cells would provide a more established model for the study of the role of chromatin organization with increasing concentration of doxorubicin. Moreover, the divergence between drug-sensitive and resistant cells be validated by other technologies. For instance, FISH experiments labelling genes in compartment switching regions would possibly suggest the localization changes of such genes in the nucleus. At a finer scale, ChIP-seq experiments targeting CTCF or cohesin could provide insights into the mechanism of enhancer-mediated gene activation during the development of drug resistance. In addition, CRISPR/dCas9-based methods could be applied to modulate the epigenetic status of enhancers and genes³³⁰ or manipulate the chromatin loops in enhancer-gene interactions³³¹, which can possibly revert the drug-response state of the cells.

Secondly, we present FANT-C, the rapid (10-hour) and low-input (as low as 50,000 cells) simplified protocol for 3C library preparation using clinical samples (**Table 3.1**). In a proof-of-concept study, we found that, with low starting cell number (50,000), FANT-C can recapitulate the hierarchical chromatin structures of K562 cells as shown previously using conventional as well as *in situ* Hi-C (**Figure 3.1**). Compared to publicly-available Hi-C and *in situ* Hi-C data, the FANT-C dataset comprises more *cis* and short-range (< 1 Mb) chromatin interactions (**Figure 3.7, 3.8**), which is possibly due to the absence of biotin-based enrichment procedures that allows FANT-C to capture ligation products in a more unbiased manner. Similar to results of above-mentioned SCLC cell lines, the application of FANT-C in AML patient samples shows alterations in compartments and TADs from pre- to post-therapy samples, indicating that changing of chromatin conformation is a common mechanism associated with phenotypic transitions in human cancers. Apart from the information about chromatin interactions, the majority of reads in a FANT-C dataset comprises genomic DNA sequences (**Figure 3.12, 3.13**), which can be used for the detection of genomic abnormalities like DNA copy number variations. For example, the FANT-C read coverage data has revealed trisomy 22 in one of the post-therapy patient samples (**Figure 3.21, 3.22**). Compared to the current application of WGS technique in clinical diagnosis for the discovery of genetic aberrations, FANT-C can provide additional information about genome-wide chromatin interactions without any additional sample collection from the patient. Together with the simplified protocol that allows library preparation using clinical samples within diagnostic time-scale, we envisage that FANT-C can be adapted to NGS-based routine cancer test in the future.

The FANT-C protocol can be further improved to achieve shorter library preparation time and lower starting cell number. As discussed in Chapter 3, the reaction efficiency can be largely enhanced and experimental time can be further reduced by replacing the standard restriction endonuclease in current FANT-C protocol with rapid digestion variety such as FastDigest (Thermo) or High-Fidelity restriction enzymes (NEB). Similarly, Quick ligase (NEB) variety can be replaced for standard T4 DNA ligase. In this thesis, as the first step to establish the implementation of FANT-C in clinical settings, we used around one million bone marrow cells from each of the AML patient samples. The number of cells used for the preparation of FANT-C library can be greatly reduced in future studies and the rest of the cells can be utilized to perform enhancer and gene expression profiling for a more comprehensive study of the underlying epigenetic regulatory system behind the phenotypic change upon therapy. In

addition, in this study, the pericentric inversion in AML patient samples is revealed by the abnormal interacting pattern in chromatin contact heatmaps (**Figure 3.20**). Apart from that, the genomic DNA reads in FANT-C datasets can also be employed to identify such genomic rearrangements. Furthermore, although FANT-C is designated to satisfy the criteria for clinical diagnosis, its application is not limited to patient samples. With the simple and straight-forward protocol, FANT-C can provide unbiased data of chromatin conformation and genomic abnormalities in research work at relatively low cost. In fact, FANT-C was used for our in-house derived drug-tolerant cell lines in the above-mentioned study. If SCLC patient samples are available, further experiments using FANT-C in drug sensitive and resistant samples would provide insights into chemoresistance in the real-world scenario. The development of genetic aberrations during cell state transitions could also be detected using FANT-C. On the other hand, enrichment of *cis* and short-range interactions in FANT-C datasets would provide more information for the research of TADs and looping interactions as well as in other applications such as scaffolding of genome assembly.

Apart from the study of chromatin organization in cancers, we also explored the employment of 3C-seq in constructing and characterizing the genome of fungal pathogen *C. tropicalis*. The *cis* interaction frequency shows a distance-dependent decay which can thus be used to infer the connectivity and/or orientation between two contigs in a genome assembly^{290,291}. With the aid of 3C-seq and SMRT data, we successfully assembled the genome of *C. tropicalis* type strain MYA-3404 in seven chromosomes (**Table 4.2**). Similar to the closely-related species *C. albicans*, the genome-wide chromatin contact heatmap of *C. tropicalis* shows enhanced centromere-centromere interactions as well as telomere-telomere interactions (**Figure 4.5, 4.6**). However, unlike *C. albicans* where the centromeres-proximal regions form a compact chromatin domain (**Figure 4.7, 4.8**), such feature may not be conserved in *C. tropicalis*. In fact, there are known differences between the centromeres of the two species, such as the HIR-associated centromeres of *C. tropicalis* as opposed to unique centromeres of *C. albicans*. Next, the chromosome-level genome assembly of *C. tropicalis* has enabled genome-wide synteny analysis between *C. albicans* and *C. tropicalis*, which reveals conserved arrangements of orthologous protein-coding genes in synteny clusters of the two species (**Figure 4.18**). Nevertheless, we observed enriched ICSBs near centromeres of both species. The possible presence and clustering of HIR-associated centromeres in their common ancestor hint that karyotype evolution in *C. albicans* and *C. tropicalis* might have occurred as a result of recombination-mediated rearrangement of chromosomes encompassing the centromeres. On

the other hand, we also sought to decipher the genetic alterations in *C. tropicalis*. In *C. albicans*, the genomic diversity is achieved by large-scale genetic changes, which play central roles in the adaptive evolution and are associated with drug resistance. Using 3C-seq data, we identified various types of genetic variations in the diploid genome of *C. tropicalis*, such as heterozygous loci, balanced translocation and segmental amplifications. We found the orthologs of a few genes in these amplified regions are known to be involved in pathobiology and drug resistance in *C. albicans*. Further investigation of genes involved in these genetic alterations may provide more information on the role of genome plasticity in the pathogenic behaviors and drug resistance of *C. tropicalis*.

The construction of the chromosome-level genome assembly is the first step towards understanding the genome-wide research in *C. tropicalis*. The successful application of 3C-seq data in contig scaffolding suggests the possibility that 3C-seq or FANT-C alone can be used for *de novo* genome assembly. While the information about chromatin interactions can imply the connectivity between contigs, the extensive genomic DNA reads in a 3C-seq or FANT-C dataset could be utilized for initial contig building. We also showed that the 3C-seq data can be applied to characterize the 3D chromatin organization as well as genetic alterations of *C. tropicalis*. In addition to current results that suggest clustering of centromeres as well as telomeres in *C. albicans* and *C. tropicalis*, further FISH experiments labelling centromeres and telomeres could be used to detect the relative locations of centromeres and telomeres. This would examine the possible presence of Rabl configuration in these organisms. In the future, by employing 3C-seq or FANT-C on different strains or clinical isolates of *C. tropicalis*, the study can be extended to investigate the changes of genomic features during phenotypic transitions in this organism, such as the process of adaptive evolution, the formation of biofilms as well as the development of drug resistance. The 3C-seq or FANT-C data could also be used to examine the role of chromatin organization in gene regulation during these transition events.

Taking all together, this thesis presents studies on the role of chromatin organization in establishing cell identity using various models including SCLC cell lines, AML patient samples and fungal pathogen *C. tropicalis*. We believe these studies will be stepping stones to future research in these directions.

References

- 1 Flemming, W. *Zellsubstanz, kern und zelltheilung*. (Vogel, 1882).
- 2 De Vries, H. *Intracellulare pangensis*. (G. Fischer, 1889).
- 3 H., D. V. Sur la loi de disjonction des hybrides. *C. R. Hebd. Séances Acad. Sci.* **130**, 845–847 (1900).
- 4 Johannsen, W. Elements of an exact theory of heredity. *Gustav Fischer, Jena.*[In German] (1909).
- 5 Morgan, T. H. *The mechanism of Mendelian heredity*. (Holt, 1915).
- 6 Morgan, T. H. The theory of the gene. *The American Naturalist* **51**, 513-544 (1917).
- 7 Winkler, H. Verbreitung und ursache der parthenogenesis im pflanzen-und tierreiche. (1920).
- 8 Miescher-Rüsch, F. *Ueber die chemische Zusammensetzung der Eiterzellen*. (1871).
- 9 Kossel, A. *Ueber die chemische Beschaffenheit des Zellkerns*. (PA Norstedt, 1911).
- 10 Griffith, F. The Significance of Pneumococcal Types. *J Hyg (Lond)* **27**, 113-159, doi:10.1017/s0022172400031879 (1928).
- 11 Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**, 137-158, doi:10.1084/jem.79.2.137 (1944).
- 12 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738, doi:10.1038/171737a0 (1953).
- 13 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563, doi:10.1038/227561a0 (1970).
- 14 Johns, E. W. in *Ciba Foundation Symposium-Homeostatic Regulators*. 128-143 (Wiley Online Library).
- 15 Johns, E. W., Phillips, D. M., Simson, P. & Butler, J. A. Improved fractionations of arginine-rich histones from calf thymus. *Biochem J* **77**, 631-636, doi:10.1042/bj0770631 (1960).
- 16 Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proc Natl Acad Sci U S A* **51**, 786-794, doi:10.1073/pnas.51.5.786 (1964).
- 17 Heitz, E. & Bauer, H. Beweise für die Chromosomennatur der Kernschleifen in den Knäuelkernen von *Bibio hortulanus* L. *Cell and Tissue Research* **17**, 67-82 (1933).

- 18 Painter, T. S. A New Method for the Study of Chromosome Rearrangements and the Plotting of Chromosome Maps. *Science* **78**, 585-586, doi:10.1126/science.78.2034.585 (1933).
- 19 Bridges, C. B. Salivary chromosome maps: with a key to the banding of the chromosomes of *Drosophila melanogaster*. *Journal of Heredity* **26**, 60-64 (1935).
- 20 Pardon, J. F., Richards, B. M. & Cotter, R. I. X-ray diffraction studies on oriented nucleohistone gels. *Cold Spring Harb Symp Quant Biol* **38**, 75-81 (1974).
- 21 Olins, A. L. & Olins, D. E. Spheroid chromatin units (v bodies). *Science* **183**, 330-332, doi:10.1126/science.183.4122.330 (1974).
- 22 Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-871, doi:10.1126/science.184.4139.868 (1974).
- 23 Woodcock, C. L., Safer, J. P. & Stanchfield, J. E. Structural repeating units in chromatin. I. Evidence for their general occurrence. *Exp Cell Res* **97**, 101-110, doi:10.1016/0014-4827(76)90659-5 (1976).
- 24 Oudet, P., Gross-Bellard, M. & Chambon, P. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* **4**, 281-300, doi:10.1016/0092-8674(75)90149-x (1975).
- 25 Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nat Rev Mol Cell Biol* **4**, 809-814, doi:10.1038/nrm1225 (2003).
- 26 Wolffe, A. *Chromatin: structure and function*. (Academic press, 1998).
- 27 Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-294, doi:10.1016/s0092-8674(00)81958-3 (1999).
- 28 Laybourn, P. J. & Kadonaga, J. T. Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II. *Science* **254**, 238-245, doi:10.1126/science.1718039 (1991).
- 29 Gilbert, N. *et al.* DNA methylation affects nuclear organization, histone modifications, and linker histone binding but not chromatin compaction. *J Cell Biol* **177**, 401-411, doi:10.1083/jcb.200607133 (2007).
- 30 Fischle, W., Wang, Y. & Allis, C. D. Histone and chromatin cross-talk. *Curr Opin Cell Biol* **15**, 172-183 (2003).
- 31 Smith, C. L. & Peterson, C. L. ATP-dependent chromatin remodeling. *Curr Top Dev Biol* **65**, 115-148, doi:10.1016/S0070-2153(04)65004-6 (2005).

- 32 Henikoff, S. & Smith, M. M. Histone variants and epigenetics. *Cold Spring Harbor perspectives in biology* **7**, a019364, doi:10.1101/cshperspect.a019364 (2015).
- 33 Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci U S A* **73**, 1897-1901, doi:10.1073/pnas.73.6.1897 (1976).
- 34 Woodcock, C. L., Frado, L. L. & Rattner, J. B. The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J Cell Biol* **99**, 42-52, doi:10.1083/jcb.99.1.42 (1984).
- 35 Langmore, J. P. & Paulson, J. R. Low angle x-ray diffraction studies of chromatin structure in vivo and in isolated nuclei and metaphase chromosomes. *J Cell Biol* **96**, 1120-1131, doi:10.1083/jcb.96.4.1120 (1983).
- 36 Widom, J. & Klug, A. Structure of the 300Å chromatin filament: X-ray diffraction from oriented samples. *Cell* **43**, 207-213, doi:10.1016/0092-8674(85)90025-x (1985).
- 37 Bordas, J., Perez-Grau, L., Koch, M. H., Vega, M. C. & Nave, C. The superstructure of chromatin and its condensation mechanism. I. Synchrotron radiation X-ray scattering results. *Eur Biophys J* **13**, 157-173 (1986).
- 38 Dorigo, B. *et al.* Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* **306**, 1571-1573, doi:10.1126/science.1103124 (2004).
- 39 Gilbert, N. *et al.* Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555-566, doi:10.1016/j.cell.2004.08.011 (2004).
- 40 Bystricky, K., Heun, P., Gehlen, L., Langowski, J. & Gasser, S. M. Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proc Natl Acad Sci U S A* **101**, 16495-16500, doi:10.1073/pnas.0402766101 (2004).
- 41 Schalch, T., Duda, S., Sargent, D. F. & Richmond, T. J. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **436**, 138-141, doi:10.1038/nature03686 (2005).
- 42 Robinson, P. J., Fairall, L., Huynh, V. A. & Rhodes, D. EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proc Natl Acad Sci U S A* **103**, 6506-6511, doi:10.1073/pnas.0601212103 (2006).
- 43 Kruithof, M. *et al.* Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. *Nat Struct Mol Biol* **16**, 534-540, doi:10.1038/nsmb.1590 (2009).

- 44 van Holde, K. & Zlatanova, J. Chromatin fiber structure: Where is the problem now? *Semin Cell Dev Biol* **18**, 651-658, doi:10.1016/j.semcdb.2007.08.005 (2007).
- 45 Sedat, J. & Manuelidis, L. A direct approach to the structure of eukaryotic chromosomes. *Cold Spring Harb Symp Quant Biol* **42 Pt 1**, 331-350 (1978).
- 46 Belmont, A. S., Braunfeld, M. B., Sedat, J. W. & Agard, D. A. Large-scale chromatin structural domains within mitotic and interphase chromosomes in vivo and in vitro. *Chromosoma* **98**, 129-143 (1989).
- 47 Belmont, A. S. & Bruce, K. Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. *J Cell Biol* **127**, 287-302, doi:10.1083/jcb.127.2.287 (1994).
- 48 Paulson, J. R. & Laemmli, U. K. The structure of histone-depleted metaphase chromosomes. *Cell* **12**, 817-828, doi:10.1016/0092-8674(77)90280-x (1977).
- 49 Laemmli, U. K. *et al.* Metaphase chromosome structure: the role of nonhistone proteins. *Cold Spring Harb Symp Quant Biol* **42 Pt 1**, 351-360 (1978).
- 50 Marsden, M. P. & Laemmli, U. K. Metaphase chromosome structure: evidence for a radial loop model. *Cell* **17**, 849-858, doi:10.1016/0092-8674(79)90325-8 (1979).
- 51 Dubochet, J., Adrian, M., Schultz, P. & Oudet, P. Cryo-electron microscopy of vitrified SV40 minichromosomes: the liquid drop model. *EMBO J* **5**, 519-528 (1986).
- 52 Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Q Rev Biophys* **21**, 129-228 (1988).
- 53 Dubochet, J. & Sartori Blanc, N. The cell in absence of aggregation artifacts. *Micron* **32**, 91-99 (2001).
- 54 Bouchet-Marquis, C., Dubochet, J. & Fakan, S. Cryoelectron microscopy of vitrified sections: a new challenge for the analysis of functional nuclear architecture. *Histochem Cell Biol* **125**, 43-51, doi:10.1007/s00418-005-0093-x (2006).
- 55 Fakan, S. & van Driel, R. The perichromatin region: a functional compartment in the nucleus that determines large-scale chromatin folding. *Semin Cell Dev Biol* **18**, 676-681, doi:10.1016/j.semcdb.2007.08.010 (2007).
- 56 Eltsov, M., Maclellan, K. M., Maeshima, K., Frangakis, A. S. & Dubochet, J. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proc Natl Acad Sci U S A* **105**, 19732-19737, doi:10.1073/pnas.0810057105 (2008).
- 57 Maeshima, K., Imai, R., Tamura, S. & Nozaki, T. Chromatin as dynamic 10-nm fibers. *Chromosoma* **123**, 225-237, doi:10.1007/s00412-014-0460-2 (2014).

- 58 Maeshima, K., Hihara, S. & Takata, H. New insight into the mitotic chromosome structure: irregular folding of nucleosome fibers without 30-nm chromatin structure. *Cold Spring Harb Symp Quant Biol* **75**, 439-444, doi:10.1101/sqb.2010.75.034 (2010).
- 59 Nishino, Y., Takahashi, Y., Imamoto, N., Ishikawa, T. & Maeshima, K. Three-dimensional visualization of a human chromosome using coherent X-ray diffraction. *Phys Rev Lett* **102**, 018101, doi:10.1103/PhysRevLett.102.018101 (2009).
- 60 Nishino, Y. *et al.* Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *EMBO J* **31**, 1644-1653, doi:10.1038/emboj.2012.35 (2012).
- 61 Joti, Y. *et al.* Chromosomes without a 30-nm chromatin fiber. *Nucleus* **3**, 404-410, doi:10.4161/nucl.21222 (2012).
- 62 Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357**, doi:10.1126/science.aag0025 (2017).
- 63 Maeshima, K. & Eltsov, M. Packaging the genome: the structure of mitotic chromosomes. *J Biochem* **143**, 145-153, doi:10.1093/jb/mvm214 (2008).
- 64 Rabl, C. Uber zelltheilung. *Morphol. Jahrb.* **10**, 214-330 (1885).
- 65 Boveri, T. Die Blastomerenkerne von *Ascaris megalocephala* und die Theorie der Chromosomenindividualitat. *Archiv fur Zellforschung* **3**, 181-268 (1909).
- 66 Wischnitzer, S. The submicroscopic morphology of the interphase nucleus. *Int Rev Cytol* **34**, 1-48 (1973).
- 67 Stack, S. M., Brown, D. B. & Dewey, W. C. Visualization of interphase chromosomes. *J Cell Sci* **26**, 281-299 (1977).
- 68 Zorn, C., Cremer, T., Cremer, C. & Zimmer, J. Laser UV microirradiation of interphase nuclei and post-treatment with caffeine. A new approach to establish the arrangement of interphase chromosomes. *Hum Genet* **35**, 83-89, doi:10.1007/bf00295622 (1976).
- 69 Zorn, C., Cremer, C., Cremer, T. & Zimmer, J. Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus. Distribution in interphase and metaphase. *Exp Cell Res* **124**, 111-119, doi:10.1016/0014-4827(79)90261-1 (1979).
- 70 Cremer, T. *et al.* Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Hum Genet* **60**, 46-56, doi:10.1007/bf00281263 (1982).

- 71 Cremer, T. *et al.* Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Hum Genet* **62**, 201-209, doi:10.1007/bf00333519 (1982).
- 72 Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc Natl Acad Sci U S A* **79**, 4381-4385, doi:10.1073/pnas.79.14.4381 (1982).
- 73 Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301, doi:10.1038/35066075 (2001).
- 74 Cremer, T. & Cremer, C. Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories. *Eur J Histochem* **50**, 161-176 (2006).
- 75 Edelmann, P., Bornfleth, H., Zink, D., Cremer, T. & Cremer, C. Morphology and dynamics of chromosome territories in living cells. *Biochim Biophys Acta* **1551**, M29-39, doi:10.1016/s0304-419x(01)00023-3 (2001).
- 76 Manuelidis, L. Individual interphase chromosome domains revealed by in situ hybridization. *Hum Genet* **71**, 288-293, doi:10.1007/bf00388453 (1985).
- 77 Schardin, M., Cremer, T., Hager, H. D. & Lang, M. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum Genet* **71**, 281-287, doi:10.1007/bf00388452 (1985).
- 78 Croft, J. A. *et al.* Differences in the localization and morphology of chromosomes in the human nucleus. *J Cell Biol* **145**, 1119-1131, doi:10.1083/jcb.145.6.1119 (1999).
- 79 Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum Mol Genet* **10**, 211-219, doi:10.1093/hmg/10.3.211 (2001).
- 80 Kurz, A. *et al.* Active and inactive genes localize preferentially in the periphery of chromosome territories. *J Cell Biol* **135**, 1195-1205, doi:10.1083/jcb.135.5.1195 (1996).
- 81 Dietzel, S. *et al.* The 3D positioning of ANT2 and ANT3 genes within female X chromosome territories correlates with gene activity. *Exp Cell Res* **252**, 363-375, doi:10.1006/excr.1999.4635 (1999).
- 82 Park, P. C. & De Boni, U. A specific conformation of the territory of chromosome 17 locates ERBB-2 sequences to a DNase-hypersensitive domain at the nuclear periphery. *Chromosoma* **107**, 87-95 (1998).

- 83 Mahy, N. L., Perry, P. E. & Bickmore, W. A. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J Cell Biol* **159**, 753-763, doi:10.1083/jcb.200207115 (2002).
- 84 Chambeyron, S. & Bickmore, W. A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev* **18**, 1119-1130, doi:10.1101/gad.292104 (2004).
- 85 Shopland, L. S. *et al.* Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *J Cell Biol* **174**, 27-38, doi:10.1083/jcb.200603083 (2006).
- 86 Brown, J. M. *et al.* Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *J Cell Biol* **182**, 1083-1097, doi:10.1083/jcb.200803174 (2008).
- 87 Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695, doi:10.1038/265687a0 (1977).
- 88 Sanger, F. *et al.* The nucleotide sequence of bacteriophage phiX174. *J Mol Biol* **125**, 225-246, doi:10.1016/0022-2836(78)90346-7 (1978).
- 89 Fiers, W. *et al.* Complete nucleotide sequence of SV40 DNA. *Nature* **273**, 113-120, doi:10.1038/273113a0 (1978).
- 90 Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162**, 729-773, doi:10.1016/0022-2836(82)90546-0 (1982).
- 91 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467, doi:10.1073/pnas.74.12.5463 (1977).
- 92 Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512, doi:10.1126/science.7542800 (1995).
- 93 Fraser, C. M. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science* **270**, 397-403, doi:10.1126/science.270.5235.397 (1995).
- 94 Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563-547, doi:10.1126/science.274.5287.546 (1996).
- 95 Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429**, 365-368, doi:10.1038/nature02390 (2004).

- 96 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
- 97 Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet* **24**, 344-352, doi:10.1016/j.tig.2008.04.005 (2008).
- 98 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).
- 99 Hsieh, A. C. *et al.* Cell type-specific abundance of 4EBP1 primes prostate cancer sensitivity or resistance to PI3K pathway inhibitors. *Sci Signal* **8**, ra116, doi:10.1126/scisignal.aad5111 (2015).
- 100 Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet* **46**, 1-19, doi:10.1146/annurev-genet-110711-155459 (2012).
- 101 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
- 102 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-825, doi:10.1038/nbt.1662 (2010).
- 103 Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502, doi:10.1126/science.1141319 (2007).
- 104 Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-1048, doi:10.1038/nbt1010-1045 (2010).
- 105 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318, doi:10.1038/ng1966 (2007).
- 106 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).
- 107 Giorgetti, L. & Heard, E. Closing the loop: 3C versus DNA FISH. *Genome Biol* **17**, 215, doi:10.1186/s13059-016-1081-2 (2016).
- 108 Dostie, J. & Bickmore, W. A. Chromosome organization in the nucleus - charting new territory across the Hi-Cs. *Curr Opin Genet Dev* **22**, 125-131, doi:10.1016/j.gde.2011.12.006 (2012).
- 109 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 (2002).

- 110 Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**, 1341-1347, doi:10.1038/ng1891 (2006).
- 111 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-1354, doi:10.1038/ng1896 (2006).
- 112 Splinter, E. *et al.* The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev* **25**, 1371-1383, doi:10.1101/gad.633311 (2011).
- 113 Tolhuis, B. *et al.* Interactions among Polycomb domains are guided by chromosome architecture. *PLoS Genet* **7**, e1001343, doi:10.1371/journal.pgen.1001343 (2011).
- 114 Hakim, O. *et al.* Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. *Genome Res* **21**, 697-706, doi:10.1101/gr.111153.110 (2011).
- 115 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309, doi:10.1101/gr.5571506 (2006).
- 116 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).
- 117 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:10.1038/nature08497 (2009).
- 118 Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472, doi:10.1016/j.cell.2012.01.010 (2012).
- 119 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593 (2013).
- 120 Zheng, M. *et al.* Multiplex chromatin interactions with single-molecule precision. *Nature* **566**, 558-562, doi:10.1038/s41586-019-0949-1 (2019).
- 121 Nagai, L. A. E., Park, S. J. & Nakai, K. Analyzing the 3D chromatin organization coordinating with gene expression regulation in B-cell lymphoma. *BMC Med Genomics* **11**, 127, doi:10.1186/s12920-018-0437-8 (2019).
- 122 McCord, R. P. *et al.* Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Res* **23**, 260-269, doi:10.1101/gr.138032.112 (2013).

- 123 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).
- 124 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:10.1038/nature11049 (2012).
- 125 Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res* **24**, 390-400, doi:10.1101/gr.163519.113 (2014).
- 126 Krijger, P. H. *et al.* Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. *Cell Stem Cell* **18**, 597-610, doi:10.1016/j.stem.2016.01.007 (2016).
- 127 Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110-114, doi:10.1038/nature16490 (2016).
- 128 Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269, doi:10.1038/nature19800 (2016).
- 129 Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458, doi:10.1126/science.aad9024 (2016).
- 130 Criscione, S. W. *et al.* Reorganization of chromosome architecture in replicative cellular senescence. *Sci Adv* **2**, e1500882, doi:10.1126/sciadv.1500882 (2016).
- 131 Lupianez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025, doi:10.1016/j.cell.2015.04.004 (2015).
- 132 Vakoc, C. R. *et al.* Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell* **17**, 453-462, doi:10.1016/j.molcel.2004.12.028 (2005).
- 133 Hadjur, S. *et al.* Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* **460**, 410-413, doi:10.1038/nature08079 (2009).
- 134 Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342**, 948-953, doi:10.1126/science.1236083 (2013).
- 135 Liu, C. *et al.* Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. *Genome Res* **26**, 1057-1068, doi:10.1101/gr.204032.116 (2016).
- 136 Umbarger, M. A. *et al.* The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol Cell* **44**, 252-264, doi:10.1016/j.molcel.2011.09.010 (2011).

- 137 Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363-367, doi:10.1038/nature08973 (2010).
- 138 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450 (2015).
- 139 Kaaij, L. J. T., van der Weide, R. H., Ketting, R. F. & de Wit, E. Systemic Loss and Gain of Chromatin Architecture throughout Zebrafish Development. *Cell Rep* **24**, 1-10 e14, doi:10.1016/j.celrep.2018.06.003 (2018).
- 140 Gottesman, M. M. Mechanisms of cancer drug resistance. *Annu Rev Med* **53**, 615-627, doi:10.1146/annurev.med.53.082901.103929 (2002).
- 141 Naumova, N., Smith, E. M., Zhan, Y. & Dekker, J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods* **58**, 192-203, doi:10.1016/j.ymeth.2012.07.022 (2012).
- 142 Kohler, J. R., Casadevall, A. & Perfect, J. The spectrum of fungi that infects humans. *Cold Spring Harb Perspect Med* **5**, a019273, doi:10.1101/cshperspect.a019273 (2014).
- 143 Tanure, M. A., Cohen, E. J., Sudesh, S., Rapuano, C. J. & Laibson, P. R. Spectrum of fungal keratitis at Wills Eye Hospital, Philadelphia, Pennsylvania. *Cornea* **19**, 307-312 (2000).
- 144 Chakrabarti, A. *et al.* **Incidence, characteristics and outcome of ICU-acquired candidemia in India.** *Intensive Care Med* **41**, 285-295, doi:10.1007/s00134-014-3603-2 (2015).
- 145 Goncalves, S. S., Souza, A. C. R., Chowdhary, A., Meis, J. F. & Colombo, A. L. Epidemiology and molecular mechanisms of antifungal resistance in *Candida* and *Aspergillus*. *Mycoses* **59**, 198-219, doi:10.1111/myc.12469 (2016).
- 146 Xiao, M. *et al.* Antifungal susceptibilities of *Candida glabrata* species complex, *Candida krusei*, *Candida parapsilosis* species complex and *Candida tropicalis* causing invasive candidiasis in China: 3 year national surveillance. *J Antimicrob Chemother* **70**, 802-810, doi:10.1093/jac/dku460 (2015).
- 147 Selmecki, A., Forche, A. & Berman, J. Aneuploidy and isochromosome formation in drug-resistant *Candida albicans*. *Science* **313**, 367-370, doi:10.1126/science.1128242 (2006).
- 148 Li, X. *et al.* Trisomy of chromosome R confers resistance to triazoles in *Candida albicans*. *Med Mycol* **53**, 302-309, doi:10.1093/mmy/myv002 (2015).

- 149 Anderson, M. Z., Saha, A., Haseeb, A. & Bennett, R. J. A chromosome 4 trisomy contributes to increased fluconazole resistance in a clinical isolate of *Candida albicans*. *Microbiology* **163**, 856-865, doi:10.1099/mic.0.000478 (2017).
- 150 www.who.int.
- 151 Faguet, G. B. A brief history of cancer: age-old milestones underlying our current knowledge database. *Int J Cancer* **136**, 2022-2036, doi:10.1002/ijc.29134 (2015).
- 152 Hassanpour, S. H. & Dehghani, M. Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice* **4**, 127-129 (2017).
- 153 Seto, M., Honma, K. & Nakagawa, M. Diversity of genome profiles in malignant lymphoma. *Cancer Sci* **101**, 573-578, doi:10.1111/j.1349-7006.2009.01452.x (2010).
- 154 Cigudosa, J. C. *et al.* Cytogenetic analysis of 363 consecutively ascertained diffuse large B-cell lymphomas. *Genes Chromosomes Cancer* **25**, 123-133 (1999).
- 155 Ezzati, M., Henley, S. J., Lopez, A. D. & Thun, M. J. Role of smoking in global and regional cancer epidemiology: current patterns and data needs. *Int J Cancer* **116**, 963-971, doi:10.1002/ijc.21100 (2005).
- 156 Aizawa, K. *et al.* Tobacco carcinogen induces both lung cancer and non-alcoholic steatohepatitis and hepatocellular carcinomas in ferrets which can be attenuated by lycopene supplementation. *Int J Cancer* **139**, 1171-1181, doi:10.1002/ijc.30161 (2016).
- 157 Poon, S. L., McPherson, J. R., Tan, P., Teh, B. T. & Rozen, S. G. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome Med* **6**, 24, doi:10.1186/gm541 (2014).
- 158 Trafialek, J. & Kolanowski, W. Dietary exposure to meat-related carcinogenic substances: is there a way to estimate the risk? *Int J Food Sci Nutr* **65**, 774-780, doi:10.3109/09637486.2014.917146 (2014).
- 159 Antwi, S. O. *et al.* Exposure to environmental chemicals and heavy metals, and risk of pancreatic cancer. *Cancer Causes Control* **26**, 1583-1591, doi:10.1007/s10552-015-0652-y (2015).
- 160 Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer* **118**, 3030-3044, doi:10.1002/ijc.21731 (2006).
- 161 Ricks-Santi, L. J. *et al.* BRCA1 polymorphisms and breast cancer epidemiology in the Western New York exposures and breast cancer (WEB) study. *Genet Epidemiol* **37**, 504-511, doi:10.1002/gepi.21730 (2013).

- 162 Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330-1334, doi:10.1126/science.aaf9011 (2017).
- 163 Shtivelman, E., Lifshitz, B., Gale, R. P. & Canaani, E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature* **315**, 550-554, doi:10.1038/315550a0 (1985).
- 164 Matlashewski, G. *et al.* Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene. *EMBO J* **3**, 3257-3262 (1984).
- 165 Wood, R. D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291**, 1284-1289, doi:10.1126/science.1056154 (2001).
- 166 Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150**, 12-27, doi:10.1016/j.cell.2012.06.013 (2012).
- 167 Geutjes, E. J., Bajpe, P. K. & Bernards, R. Targeting the epigenome for treatment of cancer. *Oncogene* **31**, 3827-3844, doi:10.1038/onc.2011.552 (2012).
- 168 Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat Biotechnol* **28**, 1057-1068, doi:10.1038/nbt.1685 (2010).
- 169 Goelz, S. E., Vogelstein, B., Hamilton, S. R. & Feinberg, A. P. Hypomethylation of DNA from benign and malignant human colon neoplasms. *Science* **228**, 187-190, doi:10.1126/science.2579435 (1985).
- 170 Fraga, M. F. *et al.* Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat Genet* **37**, 391-400, doi:10.1038/ng1531 (2005).
- 171 Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-950, doi:10.1093/nar/gkq929 (2011).
- 172 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).
- 173 Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-1113, doi:10.1126/science.1145720 (2007).
- 174 Knudson, A. G. Two genetic hits (more or less) to cancer. *Nat Rev Cancer* **1**, 157-162, doi:10.1038/35101031 (2001).
- 175 Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767, doi:10.1016/0092-8674(90)90186-i (1990).

- 176 Belikov, A. V. The number of key carcinogenic events can be predicted from cancer incidence. *Sci Rep* **7**, 12170, doi:10.1038/s41598-017-12448-7 (2017).
- 177 Maheswaran, S. & Haber, D. A. Circulating tumor cells: a window into cancer biology and metastasis. *Curr Opin Genet Dev* **20**, 96-99, doi:10.1016/j.gde.2009.12.002 (2010).
- 178 Souhami, R. & Tobias, J. S. *Cancer and its management*. (John Wiley & Sons, 2008).
- 179 Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. & Baradaran, B. The Different Mechanisms of Cancer Drug Resistance: A Brief Review. *Adv Pharm Bull* **7**, 339-348, doi:10.15171/apb.2017.041 (2017).
- 180 Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nat Rev Cancer* **13**, 714-726, doi:10.1038/nrc3599 (2013).
- 181 Cree, I. A. & Charlton, P. Molecular chess? Hallmarks of anti-cancer drug resistance. *BMC Cancer* **17**, 10, doi:10.1186/s12885-016-2999-1 (2017).
- 182 Gillet, J. P. & Gottesman, M. M. Mechanisms of multidrug resistance in cancer. *Methods Mol Biol* **596**, 47-76, doi:10.1007/978-1-60761-416-6_4 (2010).
- 183 Longo-Sorbello, G. S. & Bertino, J. R. Current understanding of methotrexate pharmacology and efficacy in acute leukemias. Use of newer antifolates in clinical trials. *Haematologica* **86**, 121-127 (2001).
- 184 Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. *Nature* **575**, 299-309, doi:10.1038/s41586-019-1730-1 (2019).
- 185 Borst, P. & Elferink, R. O. Mammalian ABC transporters in health and disease. *Annu Rev Biochem* **71**, 537-592, doi:10.1146/annurev.biochem.71.102301.093055 (2002).
- 186 Gottesman, M. M., Fojo, T. & Bates, S. E. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat Rev Cancer* **2**, 48-58, doi:10.1038/nrc706 (2002).
- 187 Liu, Y. Y., Han, T. Y., Giuliano, A. E. & Cabot, M. C. Ceramide glycosylation potentiates cellular multidrug resistance. *FASEB J* **15**, 719-730, doi:10.1096/fj.00-0223com (2001).
- 188 Lowe, S. W., Ruley, H. E., Jacks, T. & Housman, D. E. p53-dependent apoptosis modulates the cytotoxicity of anticancer agents. *Cell* **74**, 957-967 (1993).
- 189 Synold, T. W., Dussault, I. & Forman, B. M. The orphan nuclear receptor SXR coordinately regulates drug metabolism and efflux. *Nat Med* **7**, 584-590, doi:10.1038/87912 (2001).

- 190 Perez-Tomas, R. Multidrug resistance: retrospect and prospects in anti-cancer drug treatment. *Curr Med Chem* **13**, 1859-1876 (2006).
- 191 Juliano, R. L. & Ling, V. A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. *Biochim Biophys Acta* **455**, 152-162, doi:10.1016/0005-2736(76)90160-7 (1976).
- 192 Chen, C. J. *et al.* Internal duplication and homology with bacterial transport proteins in the *mdr1* (P-glycoprotein) gene from multidrug-resistant human cells. *Cell* **47**, 381-389 (1986).
- 193 Ramachandra, M. *et al.* Human P-glycoprotein exhibits reduced affinity for substrates during a catalytic transition state. *Biochemistry* **37**, 5010-5019, doi:10.1021/bi973045u (1998).
- 194 Cole, S. P. *et al.* Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line. *Science* **258**, 1650-1654, doi:10.1126/science.1360704 (1992).
- 195 Li, Q. & Shu, Y. Role of solute carriers in response to anticancer drugs. *Mol Cell Ther* **2**, 15, doi:10.1186/2052-8426-2-15 (2014).
- 196 Dalton, W. S. & Scheper, R. J. Lung resistance-related protein: determining its role in multidrug resistance. *J Natl Cancer Inst* **91**, 1604-1605, doi:10.1093/jnci/91.19.1604 (1999).
- 197 Mossink, M. H., van Zon, A., Scheper, R. J., Sonneveld, P. & Wiemer, E. A. Vaults: a ribonucleoprotein particle involved in drug resistance? *Oncogene* **22**, 7458-7467, doi:10.1038/sj.onc.1206947 (2003).
- 198 Bradshaw, D. M. & Arceci, R. J. Clinical relevance of transmembrane drug efflux as a mechanism of multidrug resistance. *J Clin Oncol* **16**, 3674-3690, doi:10.1200/JCO.1998.16.11.3674 (1998).
- 199 Liao, B. B. *et al.* Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell* **20**, 233-246 e237, doi:10.1016/j.stem.2016.11.003 (2017).
- 200 Toska, E. *et al.* PI3K pathway regulates ER-dependent transcription in breast cancer through the epigenetic regulator KMT2D. *Science* **355**, 1324-1330, doi:10.1126/science.aah6893 (2017).
- 201 Koren, S. & Bentires-Alj, M. Tackling Resistance to PI3K Inhibition by Targeting the Epigenome. *Cancer Cell* **31**, 616-618, doi:10.1016/j.ccell.2017.04.010 (2017).

- 202 Baker, E. K. & El-Osta, A. The rise of DNA methylation and the importance of chromatin on multidrug resistance in cancer. *Exp Cell Res* **290**, 177-194, doi:10.1016/s0014-4827(03)00342-2 (2003).
- 203 Jing, D. *et al.* Lymphocyte-Specific Chromatin Accessibility Pre-determines Glucocorticoid Resistance in Acute Lymphoblastic Leukemia. *Cancer Cell* **34**, 906-921 e908, doi:10.1016/j.ccell.2018.11.002 (2018).
- 204 Dubey, R. *et al.* Chromatin-Remodeling Complex SWI/SNF Controls Multidrug Resistance by Transcriptionally Regulating the Drug Efflux Pump ABCB1. *Cancer Res* **76**, 5810-5821, doi:10.1158/0008-5472.CAN-16-0716 (2016).
- 205 Chavez-Blanco, A. *et al.* Antineoplastic effects of the DNA methylation inhibitor hydralazine and the histone deacetylase inhibitor valproic acid in cancer cell lines. *Cancer Cell Int* **6**, 2, doi:10.1186/1475-2867-6-2 (2006).
- 206 Kim, M. S. *et al.* Inhibition of histone deacetylase increases cytotoxicity to anticancer drugs targeting DNA. *Cancer Res* **63**, 7291-7300 (2003).
- 207 Plumb, J. A., Strathdee, G., Sludden, J., Kaye, S. B. & Brown, R. Reversal of drug resistance in human tumor xenografts by 2'-deoxy-5-azacytidine-induced demethylation of the hMLH1 gene promoter. *Cancer Res* **60**, 6039-6044 (2000).
- 208 Qiu, Y. Y., Mirkin, B. L. & Dwivedi, R. S. Inhibition of DNA methyltransferase reverses cisplatin induced drug resistance in murine neuroblastoma cells. *Cancer Detect Prev* **29**, 456-463, doi:10.1016/j.cdp.2005.05.004 (2005).
- 209 Li, G.-D. Flexible Cancer-Associated Chromatin Configuration (CACC) Might Be the Fundamental Reason Why Cancer Is So Difficult to Cure. *arXiv preprint arXiv:1906.02710* (2019).
- 210 www.wcrf.org.
- 211 www.nrdo.gov.sg.
- 212 Furrukh, M. Tobacco Smoking and Lung Cancer: Perception-changing facts. *Sultan Qaboos Univ Med J* **13**, 345-358 (2013).
- 213 www.moh.gov.sg.
- 214 www.cancerresearchuk.org.
- 215 www.cancer.org.
- 216 Kellar, A., Egan, C. & Morris, D. Preclinical Murine Models for Lung Cancer: Clinical Trial Applications. *Biomed Res Int* **2015**, 621324, doi:10.1155/2015/621324 (2015).

- 217 Gazdar, A. F. *et al.* Establishment of continuous, clonable cultures of small-cell carcinoma of lung which have amine precursor uptake and decarboxylation cell properties. *Cancer Res* **40**, 3502-3507 (1980).
- 218 Mirski, S. E., Gerlach, J. H. & Cole, S. P. Multidrug resistance in a human small cell lung cancer cell line selected in adriamycin. *Cancer Res* **47**, 2594-2598 (1987).
- 219 Cole, S. P., Pinkoski, M. J., Bhardwaj, G. & Deeley, R. G. Elevated expression of annexin II (lipocortin II, p36) in a multidrug resistant small cell lung cancer cell line. *Br J Cancer* **65**, 498-502, doi:10.1038/bjc.1992.103 (1992).
- 220 Hassan, W. A. *et al.* Notch1 controls cell invasion and metastasis in small cell lung carcinoma cell lines. *Lung Cancer* **86**, 304-310, doi:10.1016/j.lungcan.2014.10.007 (2014).
- 221 Fujino, K. *et al.* Insulinoma-Associated Protein 1 Is a Crucial Regulator of Neuroendocrine Differentiation in Lung Cancer. *Am J Pathol* **185**, 3164-3177, doi:10.1016/j.ajpath.2015.08.018 (2015).
- 222 Li, X. *et al.* Fibroblast growth factor-inducible 14 regulates cell growth and multidrug resistance of small-cell lung cancer through the nuclear factor-kappaB pathway. *Anticancer Drugs* **25**, 1152-1164, doi:10.1097/CAD.0000000000000153 (2014).
- 223 Eriksson, H. *et al.* Quantitative membrane proteomics applying narrow range peptide isoelectric focusing for studies of small cell lung cancer resistance mechanisms. *Proteomics* **8**, 3008-3018, doi:10.1002/pmic.200800174 (2008).
- 224 Fang, S. *et al.* Long noncoding RNA-HOTAIR affects chemoresistance by regulating HOXA1 methylation in small cell lung cancer cells. *Lab Invest* **96**, 60-68, doi:10.1038/labinvest.2015.123 (2016).
- 225 Guo, L. *et al.* Gene expression profiling of drug-resistant small cell lung cancer cells by combining microRNA and cDNA expression analysis. *Eur J Cancer* **46**, 1692-1702, doi:10.1016/j.ejca.2010.02.043 (2010).
- 226 Creighton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).
- 227 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).
- 228 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

- 229 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 230 Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **9**, 90 (2007).
- 231 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).
- 232 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
- 233 v011014, M. L. C.-s. P. https://www.encodeproject.org/documents/6ecd8240-a351-479b-9de6-f09ca3702ac3/@@download/attachment/ChIP-seq_Protocol_v011014.pdf, 2014).
- 234 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
- 235 Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-1740, doi:10.1038/nprot.2012.101 (2012).
- 236 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).
- 237 Salido, M. *et al.* Cytogenetic characterization of NCI-H69 and NCI-H69AR small cell lung cancer cell lines by spectral karyotyping. *Cancer Genet Cytogenet* **191**, 97-101, doi:10.1016/j.cancergencyto.2009.01.016 (2009).
- 238 Slovak, M. L. *et al.* Localization of a novel multidrug resistance-associated gene in the HT1080/DR4 and H69AR human tumor cell lines. *Cancer Res* **53**, 3221-3225 (1993).
- 239 Whang-Peng, J. *et al.* A nonrandom chromosomal abnormality, del 3p(14-23), in human small cell lung cancer (SCLC). *Cancer Genet Cytogenet* **6**, 119-134 (1982).
- 240 Cuellar, K., Chuong, H., Hubbell, S. M. & Hinsdale, M. E. Biosynthesis of chondroitin and heparan sulfate in chinese hamster ovary cells depends on xylosyltransferase II. *J Biol Chem* **282**, 5195-5200, doi:10.1074/jbc.M611048200 (2007).
- 241 Lanzi, C., Zaffaroni, N. & Cassinelli, G. Targeting Heparan Sulfate Proteoglycans and their Modifying Enzymes to Enhance Anticancer Chemotherapy Efficacy and

- Overcome Drug Resistance. *Curr Med Chem* **24**, 2860-2886, doi:10.2174/0929867324666170216114248 (2017).
- 242 Haffner, C., Dettmer, U., Weiler, T. & Haass, C. The Nicastrin-like protein Nicalin regulates assembly and stability of the Nicalin-nodal modulator (NOMO) membrane protein complex. *J Biol Chem* **282**, 10632-10638, doi:10.1074/jbc.M611033200 (2007).
- 243 Schier, A. F. Nodal signaling in vertebrate development. *Annu Rev Cell Dev Biol* **19**, 589-621, doi:10.1146/annurev.cellbio.19.041603.094522 (2003).
- 244 Franke, W. W., Schmid, E., Osborn, M. & Weber, K. Different intermediate-sized filaments distinguished by immunofluorescence microscopy. *Proc Natl Acad Sci U S A* **75**, 5034-5038, doi:10.1073/pnas.75.10.5034 (1978).
- 245 Gard, D. L., Bell, P. B. & Lazarides, E. Coexistence of desmin and the fibroblastic intermediate filament subunit in muscle and nonmuscle cells: identification and comparative peptide analysis. *Proc Natl Acad Sci U S A* **76**, 3894-3898, doi:10.1073/pnas.76.8.3894 (1979).
- 246 Granger, B. L. & Lazarides, E. Desmin and vimentin coexist at the periphery of the myofibril Z disc. *Cell* **18**, 1053-1063, doi:10.1016/0092-8674(79)90218-6 (1979).
- 247 Kageyama, R., Ohtsuka, T. & Kobayashi, T. The Hes gene family: repressors and oscillators that orchestrate embryogenesis. *Development* **134**, 1243-1251, doi:10.1242/dev.000786 (2007).
- 248 Ishibashi, M. *et al.* Persistent expression of helix-loop-helix factor HES-1 prevents mammalian neural differentiation in the central nervous system. *EMBO J* **13**, 1799-1805 (1994).
- 249 Sun, H., Ghaffari, S. & Taneja, R. bHLH-Orange Transcription Factors in Development and Cancer. *Transl Oncogenomics* **2**, 107-120 (2007).
- 250 Sang, L., Roberts, J. M. & Collier, H. A. Hijacking HES1: how tumors co-opt the anti-differentiation strategies of quiescent cells. *Trends Mol Med* **16**, 17-26, doi:10.1016/j.molmed.2009.11.001 (2010).
- 251 Tian, C. *et al.* Hes1 mediates the different responses of hematopoietic stem and progenitor cells to T cell leukemic environment. *Cell Cycle* **12**, 322-331, doi:10.4161/cc.23160 (2013).
- 252 Maniati, E. *et al.* Crosstalk between the canonical NF-kappaB and Notch signaling pathways inhibits Ppargamma expression and promotes pancreatic cancer progression in mice. *J Clin Invest* **121**, 4685-4699, doi:10.1172/JCI45797 (2011).

- 253 Wang, S. C. *et al.* Hes1 triggers epithelial-mesenchymal transition (EMT)-like cellular marker alterations and promotes invasion and metastasis of nasopharyngeal carcinoma by activating the PTEN/AKT pathway. *Oncotarget* **6**, 36713-36730, doi:10.18632/oncotarget.5457 (2015).
- 254 Marfels, C., Hoehn, M., Wagner, E. & Gunther, M. Characterization of in vivo chemoresistant human hepatocellular carcinoma cells with transendothelial differentiation capacities. *BMC Cancer* **13**, 176, doi:10.1186/1471-2407-13-176 (2013).
- 255 Hazan, R. B., Phillips, G. R., Qiao, R. F., Norton, L. & Aaronson, S. A. Exogenous expression of N-cadherin in breast cancer cells induces cell migration, invasion, and metastasis. *J Cell Biol* **148**, 779-790, doi:10.1083/jcb.148.4.779 (2000).
- 256 Hazan, R. B., Qiao, R., Keren, R., Badano, I. & Suyama, K. Cadherin switch in tumor progression. *Ann N Y Acad Sci* **1014**, 155-163, doi:10.1196/annals.1294.016 (2004).
- 257 Nieman, M. T., Prudoff, R. S., Johnson, K. R. & Wheelock, M. J. N-cadherin promotes motility in human breast cancer cells regardless of their E-cadherin expression. *J Cell Biol* **147**, 631-644, doi:10.1083/jcb.147.3.631 (1999).
- 258 Fogar, P. *et al.* Neural cell adhesion molecule (N-CAM) in gastrointestinal neoplasias. *Anticancer Res* **17**, 1227-1230 (1997).
- 259 Sasaki, H. *et al.* Expression of the neural cell adhesion molecule in astrocytic tumors: an inverse correlation with malignancy. *Cancer* **82**, 1921-1931, doi:10.1002/(sici)1097-0142(19980515)82:10<1921::aid-cnrcr16>3.0.co;2-v (1998).
- 260 Tezel, E., Kawase, Y., Takeda, S., Oshima, K. & Nakao, A. Expression of neural cell adhesion molecule in pancreatic cancer. *Pancreas* **22**, 122-125 (2001).
- 261 Huerta, S. *et al.* Alternative mRNA splicing in colon cancer causes loss of expression of neural cell adhesion molecule. *Surgery* **130**, 834-843, doi:10.1067/msy.2001.116415 (2001).
- 262 Forcato, M. *et al.* Comparison of computational methods for Hi-C data analysis. *Nat Methods* **14**, 679-685, doi:10.1038/nmeth.4325 (2017).
- 263 Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59-64, doi:10.1038/nature21429 (2017).
- 264 Oliphant, T. E. Python for scientific computing. *Computing in Science & Engineering* **9**, 10-20 (2007).

- 265 Khalil, A. I. S., Khyriem, C., Chattopadhyay, A. & Sanyal, A. Hierarchical discovery of large-scale and focal copy number alterations in low-coverage cancer genomes. *BMC Bioinformatics* **21**, 147, doi:10.1186/s12859-020-3480-3 (2020).
- 266 Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276, doi:10.1016/j.ymeth.2012.05.001 (2012).
- 267 Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56-65, doi:10.1016/j.ymeth.2017.04.004 (2017).
- 268 Nagano, T. *et al.* Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat Protoc* **10**, 1986-2003, doi:10.1038/nprot.2015.127 (2015).
- 269 Diaz, N. *et al.* Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat Commun* **9**, 4938, doi:10.1038/s41467-018-06961-0 (2018).
- 270 Du, Z. *et al.* Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235, doi:10.1038/nature23263 (2017).
- 271 Ke, Y. *et al.* 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367-381 e320, doi:10.1016/j.cell.2017.06.029 (2017).
- 272 Jackson, D. A. & Cook, P. R. A general method for preparing chromatin containing intact DNA. *EMBO J* **4**, 913-918 (1985).
- 273 Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* **3**, 17-21, doi:10.1038/nmeth823 (2006).
- 274 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-1065, doi:10.1038/ng.947 (2011).
- 275 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
- 276 Harewood, L. *et al.* Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol* **18**, 125, doi:10.1186/s13059-017-1253-8 (2017).
- 277 Chakraborty, A. & Ay, F. Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics*, doi:10.1093/bioinformatics/btx664 (2017).

- 278 Davies, J. O. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* **13**, 74-80, doi:10.1038/nmeth.3664 (2016).
- 279 Engreitz, J. M., Agarwala, V. & Mirny, L. A. Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS One* **7**, e44196, doi:10.1371/journal.pone.0044196 (2012).
- 280 Eser, S. *et al.* Selective requirement of PI3K/PDK1 signaling for Kras oncogene-driven pancreatic cell plasticity and cancer. *Cancer cell* **23**, 406-420 (2013).
- 281 Niemann, C. U., Kjeldsen, L., Ralfkiaer, E., Jensen, M. K. & Borregaard, N. Serglycin proteoglycan in hematologic malignancies: a marker of acute myeloid leukemia. *Leukemia* **21**, 2406-2410, doi:10.1038/sj.leu.2404975 (2007).
- 282 Handschuh, L. *et al.* Gene expression profiling of acute myeloid leukemia samples from adult patients with AML-M1 and -M2 through boutique microarrays, real-time PCR and droplet digital PCR. *Int J Oncol* **52**, 656-678, doi:10.3892/ijo.2017.4233 (2018).
- 283 Le Beau, M. M. *et al.* Association of an inversion of chromosome 16 with abnormal marrow eosinophils in acute myelomonocytic leukemia. A unique cytogenetic-clinicopathological association. *N Engl J Med* **309**, 630-636, doi:10.1056/NEJM198309153091103 (1983).
- 284 Mitelman, F. & Heim, S. Quantitative acute leukemia cytogenetics. *Genes Chromosomes Cancer* **5**, 57-66 (1992).
- 285 Vardiman, J., Catovsky, D., Flandrin, G. R. & Pierre, R. Fourth International Workshop on Chromosomes in Leukemia 1982: Correlation between morphology and karyotype. *Cancer Genet Cytogenet* **11**, 275-281 (1984).
- 286 Liu, P. *et al.* Fusion between transcription factor CBF beta/PEBP2 beta and a myosin heavy chain in acute myeloid leukemia. *Science* **261**, 1041-1044 (1993).
- 287 Najfeld, V. *et al.* Trisomy 22--a new abnormality found in acute leukemia characterized by eosinophilia and monocytoid blasts expressing immature differentiation antigens. *Cancer Genet Cytogenet* **23**, 105-114 (1986).
- 288 Grois, N. *et al.* Is trisomy 22 in acute myeloid leukemia a primary abnormality or only a secondary change associated with inversion 16? *Cancer Genet Cytogenet* **43**, 119-129 (1989).
- 289 Wong, K. F. & Kwong, Y. L. Trisomy 22 in acute myeloid leukemia: a marker for myeloid leukemia with monocytic features and cytogenetically cryptic inversion 16. *Cancer Genet Cytogenet* **109**, 131-133 (1999).

- 290 Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA
interaction frequency. *Nat Biotechnol* **31**, 1143-1147, doi:10.1038/nbt.2768 (2013).
- 291 Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies
based on chromatin interactions. *Nat Biotechnol* **31**, 1119-1125, doi:10.1038/nbt.2727
(2013).
- 292 Selvaraj, S., J. R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction
using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31**, 1111-1118,
doi:10.1038/nbt.2728 (2013).
- 293 Toya, S. P., Schraufnagel, D. E. & Tzelepis, G. E. Candiduria in intensive care units:
association with heavy colonization and candidaemia. *J Hosp Infect* **66**, 201-206,
doi:10.1016/j.jhin.2007.03.028 (2007).
- 294 Eggimann, P., Garbino, J. & Pittet, D. Epidemiology of Candida species infections in
critically ill non-immunosuppressed patients. *Lancet Infect Dis* **3**, 685-702 (2003).
- 295 Manolakaki, D. *et al.* Candida infection and colonization among trauma patients.
Virulence **1**, 367-375, doi:10.4161/viru.1.5.12796 (2010).
- 296 Mancera, E. *et al.* Genetic Modification of Closely Related Candida Species.
Frontiers in microbiology **10**, 357, doi:10.3389/fmicb.2019.00357 (2019).
- 297 Sexton, A. C. & Howlett, B. J. Parallels in fungal pathogenesis on plant and animal
hosts. *Eukaryot Cell* **5**, 1941-1949, doi:10.1128/EC.00277-06 (2006).
- 298 Clay, K. & Kover, P. X. The Red Queen Hypothesis and plant/pathogen interactions.
Annu Rev Phytopathol **34**, 29-50, doi:10.1146/annurev.phyto.34.1.29 (1996).
- 299 Wertheimer, N. B., Stone, N. & Berman, J. Ploidy dynamics and evolvability in fungi.
Philos Trans R Soc Lond B Biol Sci **371**, doi:10.1098/rstb.2015.0461 (2016).
- 300 Tso GH, R.-C. J., Tan AS, Sem X, Le GT, Tan TG, Lai GC, Srinivasan KG, Yurieva
M, Liao W, Poidinger M. . Experimental evolution of a fungal pathogen into a gut
symbiont. *Science* **362(6414)**, 589-595 (2018).
- 301 Forche A, M. P., Selmecki A, Berman J, May G. . Evolution in Candida albicans
populations during a single passage through a mouse host. *Genetics* **182(3)**, 799-811
(2009).
- 302 Ene, I. V. *et al.* Global analysis of mutations driving microevolution of a
heterozygous diploid fungal pathogen. *Proc Natl Acad Sci U S A* **115**, E8688-E8697,
doi:10.1073/pnas.1806002115 (2018).

- 303 Seervai, R. N., Jones, S. K., Jr., Hirakawa, M. P., Porman, A. M. & Bennett, R. J. Parasexuality and ploidy change in *Candida tropicalis*. *Eukaryot Cell* **12**, 1629-1640, doi:10.1128/EC.00128-13 (2013).
- 304 Butler, G. *et al.* Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* **459**, 657-662, doi:10.1038/nature08064 (2009).
- 305 Chatterjee, G. *et al.* Repeat-Associated Fission Yeast-Like Regional Centromeres in the Ascomycetous Budding Yeast *Candida tropicalis*. *PLoS Genet* **12**, e1005839, doi:10.1371/journal.pgen.1005839 (2016).
- 306 Mancera, E., Porman, A. M., Cuomo, C. A., Bennett, R. J. & Johnson, A. D. Finding a Missing Gene: EFG1 Regulates Morphogenesis in *Candida tropicalis*. *G3 (Bethesda)* **5**, 849-856, doi:10.1534/g3.115.017566 (2015).
- 307 Shen, X. X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**, 1533-1545 e1520, doi:10.1016/j.cell.2018.10.023 (2018).
- 308 Guin, K. *et al.* Spatial inter-centromeric interactions facilitated the emergence of evolutionary new centromeres. *Elife* **9**, doi:10.7554/eLife.58556 (2020).
- 309 Soderlund, C., Bomhoff, M. & Nelson, W. M. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* **39**, e68, doi:10.1093/nar/gkr123 (2011).
- 310 Grabherr MG, R. P., Meyer M, Mauceli E, Alföldi J, Di Palma F, Lindblad-Toh K. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*. **26**, 1145-1151. (2010).
- 311 Drillon, G., Carbone, A. & Fischer, G. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* **9**, e92621, doi:10.1371/journal.pone.0092621 (2014).
- 312 Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157, doi:10.1186/s13059-015-0721-2 (2015).
- 313 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645, doi:10.1101/gr.092759.109 (2009).
- 314 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).
- 315 Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050-1054, doi:10.1038/nmeth.4035 (2016).

- 316 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
- 317 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).
- 318 Jin, Q. W., Fuchs, J. & Loidl, J. Centromere clustering is a major determinant of yeast interphase nuclear organization. *J Cell Sci* **113** (Pt 11), 1903-1912 (2000).
- 319 Jin, Q., Trelles-Sticken, E., Scherthan, H. & Loidl, J. Yeast nuclei display prominent centromere clustering that is reduced in nondividing cells and in meiotic prophase. *J Cell Biol* **141**, 21-29, doi:10.1083/jcb.141.1.21 (1998).
- 320 Burrack, L. S. *et al.* Neocentromeres Provide Chromosome Segregation Accuracy and Centromere Clustering to Multiple Loci along a *Candida albicans* Chromosome. *PLoS Genet* **12**, e1006317, doi:10.1371/journal.pgen.1006317 (2016).
- 321 Descorps-Declere, S. *et al.* Genome-wide replication landscape of *Candida glabrata*. *BMC Biol* **13**, 69, doi:10.1186/s12915-015-0177-6 (2015).
- 322 Sreekumar, L. *et al.* Cis- and Trans-chromosomal Interactions Define Pericentric Boundaries in the Absence of Conventional Heterochromatin. *Genetics* **212**, 1121-1132, doi:10.1534/genetics.119.302179 (2019).
- 323 Thakur, J. & Sanyal, K. Efficient neocentromere formation is suppressed by gene conversion to maintain centromere function at native physical chromosomal loci in *Candida albicans*. *Genome Res* **23**, 638-652, doi:10.1101/gr.141614.112 (2013).
- 324 Sanyal, K., Baum, M. & Carbon, J. Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. *Proc Natl Acad Sci U S A* **101**, 11374-11379, doi:10.1073/pnas.0404318101 (2004).
- 325 Chibana, H., Beckerman, J. L. & Magee, P. T. Fine-resolution physical mapping of genomic diversity in *Candida albicans*. *Genome Res* **10**, 1865-1877 (2000).
- 326 Despagne, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat Genet* **51**, 1263-1271, doi:10.1038/s41588-019-0466-z (2019).
- 327 Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244, doi:10.1016/j.cell.2012.03.051 (2012).

- 328 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).
- 329 Shang, S. *et al.* Chemotherapy-Induced Distal Enhancers Drive Transcriptional Programs to Maintain the Chemoresistant State in Ovarian Cancer. *Cancer Res* **79**, 4599-4611, doi:10.1158/0008-5472.CAN-19-0215 (2019).
- 330 Pulecio, J., Verma, N., Mejia-Ramirez, E., Huangfu, D. & Raya, A. CRISPR/Cas9-Based Engineering of the Epigenome. *Cell Stem Cell* **21**, 431-447, doi:10.1016/j.stem.2017.09.006 (2017).
- 331 Morgan, S. L. *et al.* Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat Commun* **8**, 15993, doi:10.1038/ncomms15993 (2017).